When is a Liability not a Liability?

Textual Analysis, Dictionaries, and 10-Ks

Tim Loughran
and
Bill McDonald

University of Notre Dame

UNIVERSITY OF NOTRE DAME
Mendoza College of Business

- Researchers are increasingly using textual analysis to examine the tone and sentiment of 10-Ks, newspaper articles, and company press releases.

- A commonly used source for word classifications is the Harvard Psychosociological Dictionary, specifically the Harvard-IV-4-TagNeg (H4N).

- To measure tone, typically the proportional count of negative words is used.

- More negative words scaled by total document words is gauged as being more pessimistic.

- Why are researchers using the Harvard list?

- Two positive features of the H4N list:
  - One, the list's composition is beyond the control of the researcher.
  - Two, the list is free.

- The main question of our paper is whether a word list developed for psychology and sociology translates well into the language of business.

- We examine over 1.2 billion words contained in 50,115 firm-year 10-Ks during 1994-2008.

- We find that almost 75% of the Harvard negative word counts are typically not negative in a financial context according to our classification.

- Here are some example of high frequency Harvard negative words: *tax*, *costs*, *board*, *liabilities*, *foreign*, and *vice*.

- Clearly, these are not negative financial words. The firm is merely naming their *board* of directors or company *vice*-presidents.

- Some of the misclassification simply adds noise, thus attenuating coefficient estimates.

- There is another concern when using the Harvard negative word list.

- Other H4N words (such as *mine*, *cancer*, *crude*, and *capital*) are likely to identify a specific industry rather than reveal negative financial information about the text of the 10-K.

- So what are we going to do about this apparent misclassification problem with the Harvard list?

- We created a list of 2,337 words (Fin-Neg) that typically have negative implications in a financial sense.
- All of our word lists are available on Bill McDonald's Notre Dame website.
- Some of our negative words (like *loss* and *losses*) also appear on the Harvard list.
- Others, like *felony*, *litigation*, *restated*, *misstatement*, and *unanticipated* do not.

- Another contribution is to show that a common term weighting scheme reduces the noise introduced by misclassifications.

- With term weighting, common words are weighted less. Infrequent negative words are given more impact.

- We also create five other word lists (positive, uncertain, litigious, strong modal, and weak modal).

- When we look at whether our word lists actually gauge tone, we find significant relations between these word lists and file date returns, trading volume, subsequent return volatility, standardized unexpected earnings, and two separate samples of fraud and material weakness.

The 10-K Sample

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 $\geq$ $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 preceding 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 $\geq$ 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 & Analysis (MDA) Subsection		
Subset of 10-K sample where MDA section could be identified	49,179	936
MDA section $\geq$ 250 words	37,287	11,892

- The fraud sample:
  - 10b-5 class action lawsuits (like Enron, Boston Chicken, and Cardinal Health)
  - Dates: 1/1/1994 - 9/23/2004
  - 586 cases (of 35,992) where either:
    - The 10-K is filed during the period of alleged fraud
    - A 10b-5 filing occurs within one year of the 10-K filing

- The Doyle, Ge and McVay (2007) sample:
  - Disclosure of a material weakness
  - Dates: 2/1/2001 – 10/30/2005
  - 708 cases (of 17,143) where, within 18 months of a 10-K filing, a disclosure of material weakness is reported in a subsequent 10-K, 10-Q, or 8-K.

- Summary of technical literature:

> *"Natural languages are messy and difficult to parse with computers."*

Current Issues in Parsing Technology, Masaru Tomita, p. 1

- Remove all ASCII-encoded graphics
- Remove exhibits
- Remove Tables
- Remove all HTML
- Parse into tokens
- Create word counts

- Term Weighting (tf.idf) addresses three issues in aggregating the word counts:
  - *The importance of a term within a document*
  - *Normalization for document length*
  - *The importance of a term within the entire corpus*

- We use a relatively common weighting scheme:

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

  - *N = total number of 10-Ks*
  - *df_i = # of docs with at least one occurrence of the i^{th} word.*
  - *$tf_{i,j}$ = the raw count of the i^{th} word in the j^{th} doc*
  - *a = the average word count in the 10-K*

- The first term of equation (1) attenuates the impact of large counts (Zipf's law). For example:

  The word *loss* appears 1.79 million times in our sample while the word *aggravates* appears only 10 times. Is the likely impact of the word *loss* more than 179,000 times that of *aggravates?*

- The second term of equation (1) modifies the impact of a word based on its commonality. For example:

  The word *loss* appears in more than 90% of the documents which implies that the second term will decrease the first term by more than 90%.

- We control for inflection
  - depreciate=>depreciated/depreciates/depreciating/depreciation
  - "Selected" inflection avoids morphologies like:
    - blind / blinds
    - odd / odds
    - bitter / bitters

- H4N-Inf
  - General Inquirer Harvard-IV-4 TAGNeg file extended to include appropriate inflections. N = 4,187
  - http://www.wjh.harvard.edu/~inquirer/

- Our word lists
  - Create a dictionary of all words occurring in 10-Ks from 1994-2008.
  - Classify words occurring in 5% or more of the documents, plus inflections.

- Our word lists

  – Fin-Neg – negative words (e.g., *restated*, *termination*, *felony*, *misstated*, *discontinued*, *misconduct*, *unable)*. N=2,337 (1,121 overlap with H4N-Inf)

  – Fin-Pos – positive words (e.g., *delighted*, *excellent*, *innovative*), excluding cases of simple negation (e.g., *not good*). N = 353

  Notice that in financial reporting it is unlikely that negative words will be negated (e.g., *not terrible earnings*), whereas positive words are easily qualified or compromised.

- Our word lists

  – Fin-Unc – uncertainty words. Note here the emphasis is more so on uncertainty than risk (e.g., *ambiguity*, *approximate*, *fluctuate*). N = 285

  – Fin-Lit – litigious words (e.g., *admission*, *breach*, *defendant*, *plaintiff*, *remand*, *testimony*). N = 731

- Our word lists

  – Modal Word-Strong – e.g., *always*, *best*, *definitely*, *highest*, *lowest*, *will*. N = 19

  – Modal Word-Weak – e.g., *could*, *depending*, *may*, *possibly*, *sometimes*. N = 27

Adapted from: Jordon, R. R. (1990) *Academic Writing Course*. Edinburgh (1992)

- Event period excess return – day[0,3] return. It is the firm's buy-and-hold stock return minus the CRSP value-weighted buy-and-hold market index return over the four day period.

- Size – CRSP market capitalization on the day before the file date. [Log]

- Book-to-market – Fama-French (2001), missing for negative values then winsorized at 1% level. [Log]

- Share Turnover – volume of shares trade in days [-252,-6] divided by shares outstanding on file date. [Log]

- Institutional ownership – CDA/Spectrum. Missing for negative values and winsorized to 100% for cases > 100%
- Nasdaq dummy.
- Fama-French (1997) industry dummies (based on 48 industries)
- One year Pre-Fama-French Alpha
- SUE, analyst dispersion, analyst revisions. SUE is the standardized unexpected earnings for the quarterly earnings announced within 90 days after the 10-K filing date. Actual earnings minus mean analyst forecast scaled by stock price.
- Other dependent variables:
  - Abnormal volume
  - Post-Event Return Volatility

Summary Statistics for the 1994-2008 10-K Samples

	Full 10-K Document (N=50,115)			MD&A Subsection (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%

Summary Statistics (continued)

	Full 10-K Document (N=50,115)			MD&A Subsection (N=37,287)		
Variable	Mean	Median	Standard Deviation	Mean	Median	Standard Deviation
Other Variables						
Event-period(0,3) excess return	-0.12%	-0.19%	6.82%	-0.23%	-0.28%	7.26%
Size ($billions)	$3.09	$0.33	$14.94	$2.12	$0.30	$9.62
Book-to-market	0.613	0.512	0.459	0.611	0.501	0.477
Turnover	1.519	0.947	2.295	1.695	1.104	2.508
One-year pre-event FF alpha	0.07%	0.04%	0.20%	0.07%	0.05%	0.21%
Institutional ownership	48.34%	48.07%	28.66%	49.23%	48.52%	29.33%
Nasdaq dummy	56.15%	100.00%	49.62%	60.12%	100.00%	48.97%
Standardized unexpected earnings	-0.02%	0.03%	0.76%	-0.03%	0.03%	0.82%
Analysts' earnings forecast dispersion	0.17%	0.07%	0.33%	0.19%	0.08%	0.36%
Analysts' earnings revisions	-0.21%	-0.04%	0.69%	-0.24%	-0.05%	0.74%

Panel A: H4N-Inf

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%

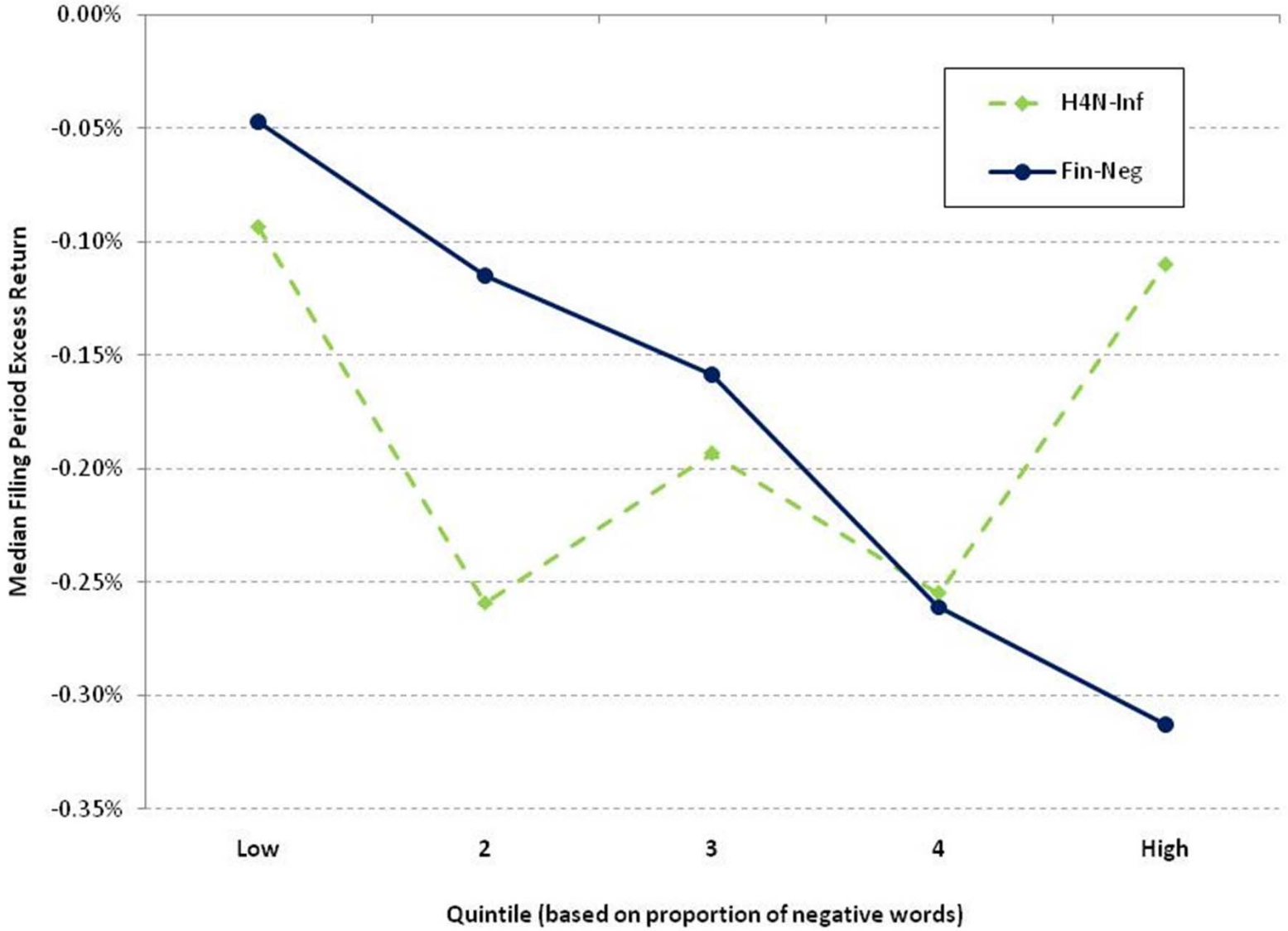
Panel B: Fin-Neg

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%

WorldCom Example

- WorldCom's 2002 10-K just prior to filing for bankruptcy protection
  - Most frequent Fin-Neg words: *loss, impairment, losses, termination, complaint, liquidation, litigation, restated*
  - Most frequent H4N-Inf words: *costs, expense, service, tax, loss, liabilities, expenses*

Median Filing Period Excess Returns by H4N-Inf and Fin-Neg



Comparison of Negative Word Lists using Filing Period Excess Returns as Dependent Variable

Variable	(1)	(2)	(3)	(4)
Word Lists				
H4N-Inf (Harvard-IV-4-Neg with inflections)	-7.422			
	(-1.35)			
Fin-Neg (negative)		-19.538		
		(-2.64)		
$H4N\text{-}Inf_w$ (tf.idf weighted H4N-Inf)				
$Fin\text{-}Neg_w$ (tf.idf weighted Fin-Neg)				
Control Variables				
Log(size)	0.123	0.127		
	(2.87)	(2.93)		
Log(book-to-market)	0.279	0.280		
	(3.35)	(3.45)		
Log(share turnover)	-0.284	-0.269		
	(-2.46)	(-2.36)		
Pre_FFAlpha	-2.500	-3.861		
	(-0.06)	(-0.09)		
Institutional Ownership	0.278	0.261		
	(0.93)	(0.86)		
Nasdaq Dummy	0.073	0.073		
	(0.86)	(0.87)		
Average R-square	2.44%	2.52%		

Comparison of Negative Word Lists using Filing Period Excess Returns as Dependent Variable

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

MD&A Regressions

Panel A: H4N-Inf Word Lists				
	Proportional Weights		tf.idf Weights	
	(1)	(2)	(1)	(2)
H4N-Inf (MD&A section)	1.892	0.846	-0.005	-0.002
	(0.35)	(0.16)	(-1.96)	(-0.42)
H4N-Inf (Non-MD&A section)		1.772		-0.002
		(0.18)		(-0.94)

MD&A Regressions

Panel A: H4N-Inf Word Lists				
	Proportional Weights		tf.idf Weights	
	(1)	(2)	(1)	(2)
H4N-Inf (MD&A section)	1.892	0.846	-0.005	-0.002
	(0.35)	(0.16)	(-1.96)	(-0.42)
H4N-Inf (Non-MD&A section)		1.772		-0.002
		(0.18)		(-0.94)

Panel B: Fin-Neg Word Lists				
	Proportional Weights		tf.idf Weights	
	(1)	(3)	(1)	(3)
Fin-Neg (MD&A section)	-5.344	-4.573	-0.006	-0.006
	(-0.68)	(-0.51)	(-1.96)	(-1.98)
Fin-Neg (Non-MD&A section)		-8.720		-0.001
		(-0.84)		(-1.07)

Tone Regressions

Dependent Variable	H4N-Inf	Finance Dictionaries					
		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)
Post_FFRMSE	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)

Tone Regressions

Dependent Variable	H4N-Inf	Finance Dictionaries					
		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)
Post_FFRMSE	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)
Post_FFRMSE	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)

Fraud and Material Weakness Logits

		Finance Dictionaries					
Dependent Variable	H4N-Inf	Negative	Positive	Uncertainty	Litigious	Modal Strong	Modal Weak
		Panel A: Proportional Weights					
Fraud	3.109	9.207	-6.031	19.425	-0.003	1.066	-45.369
	(0.52)	(1.20)	(-0.34)	(1.42)	(-0.00)	(0.03)	(-1.94)
Material Weakness	9.082	31.342	-10.396	-9.738	3.421	152.445	8.844
	(1.43)	(3.95)	(-0.51)	(-0.61)	(0.36)	(3.50)	(0.40)

Fraud and Material Weakness Logits

		Finance Dictionaries					
Dependent Variable	H4N-Inf	Negative	Positive	Uncertainty	Litigious	Modal Strong	Modal Weak
Panel A: Proportional Weights							
Fraud	3.109	9.207	-6.031	19.425	-0.003	1.066	-45.369
	(0.52)	(1.20)	(-0.34)	(1.42)	(-0.00)	(0.03)	(-1.94)
Material Weakness	9.082	31.342	-10.396	-9.738	3.421	152.445	8.844
	(1.43)	(3.95)	(-0.51)	(-0.61)	(0.36)	(3.50)	(0.40)
Panel B: TF.IDF Weights							
Fraud	0.001	0.003	0.006	0.012	0.005	0.057	0.010
	(1.56)	(2.85)	(1.69)	(2.43)	(3.34)	(1.11)	(0.39)
Material Weakness	0.004	0.004	0.012	0.014	0.006	0.153	0.041
	(4.45)	(5.10)	(3.94)	(2.97)	(3.56)	(3.63)	(1.65)

SUE Regressions

	(1)	(2)	(3)	(4)
Word Lists				
H4N-Inf	1.937			
	(2.58)			
Fin-Neg		2.683		
		(2.41)		

SUE Regressions

	(1)	(2)	(3)	(4)
Word Lists				
H4N-Inf	1.937			
	(2.58)			
Fin-Neg		2.683		
		(2.41)		
$H4N\text{-}Inf_w$ (tf.idf weighted)			0.035	
(coef x 100)			(4.03)	
$Fin\text{-}Neg_w$ (tf.idf weighted)				0.030
(coef x 100)				(2.87)

- Generic word lists
  - *Substantial misclassification*
  - *Potentially misleading*

- Term weighting is useful.

- MD&A section is not more informative.

- Most important, we show that financial researchers should be cautious when relying on word classification schemes derived outside the domain of business usage.

- Applying non-business word lists to accounting and finance topics can lead to a high misclassification rate and spurious correlations.

- Our word lists work – subjective but transparent.
  - *They are related to file date returns.*
  - *They are related to other financial/accounting variables.*
  - *They are less likely to produce spurious correlations.*