Internet Appendix for "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks"*

Tim Loughran and Bill McDonald

In the Internet Appendix we provide a detailed description of the parsing procedure for the 10-K sample outlined in Section II.B of the main article, links for the word lists used in the study, and additional analyses not tabulated in the main text.

I. Parsing Procedure for the 10-K Sample

We first download all 10-K and 10-K405 documents identified in the quarterly master index files appearing on the EDGAR website for the period 1994 to 2008. Each complete text filing is read into a single string variable and parsed using the following sequence:

- Remove graphics (ASCII encoded graphics) —embedded graphics increase by orders of magnitude the character count and file size of the 10-K documents. The inclusion of graphics in 10-K filings has increased each year. All encoded graphics must be purged, or the use of document size-related variables will be severely affected.
- Identify self-reported SIC code on the first page of the filing. If the SIC code does not appear in the 10-K, we programmatically go to the general web page for the firm on the EDGAR site to see if a SIC code is reported. If no SIC code is found, the industry is classified as "Other."
- Remove SEC header—we remove the standard first page of the filing appearing between the HTML <IMS-HEADER> or <SEC-HEADER> tags.
- Re-encode characters—translates "encoded" characters such as   (blank space) or &AMP (&) back to their original ACSII form.
- Remove exhibits—removes all text appearing within "<TYPE>EX" HTML tagged document segments.

---

* Citation format: Loughran, Tim, and Bill McDonald, 2011, Internet Appendix for "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *Journal of Finance* 66, 35-65, http://www.afajof.org/supplements.asp. Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the authors of the article.

- Remove tables—we remove all text appearing within <TABLE> HTML tags, where more than 25% of the nonblank characters are numbers. (Note that this filter is important since some filers embed all of their text within TABLE tags.)
- Remove HTML—the quantity of HTML increased substantially in the late 1990s. Some 10-Ks include more HTML than actual content.
- Parse into tokens—we use a regular expression (regex) to parse the remaining string variable into all collections of two or more alphabetic characters. (Hyphens are also allowed in the character collections.) We first replace all hyphens followed by a line-feed with a hyphen so that the word boundary regex works correctly.
- Create word counts—at this point we have a collection of alphabetic characters (tokens), which we then look up in our master dictionary. This parsing process also accounts for hyphenation. We keep a word count for all words in the master dictionary for each document. This allows us to subsequently go back and create word counts based on the various tonal word lists using the document dictionaries.

Our master dictionary is based on release 4.0 of the 2of12inf dictionary from http://wordlist.sourceforge.net/12dicts-readme.html, which includes word inflections but does not include abbreviations, acronyms, or names. Extensive dictionaries are generally available on the internet due to their usefulness in hacking, where they are used for lookups to crack passwords. We have added more than 800 words taken from a list of tokens from 10-Ks that did not appear in the 2of12inf dictionary. Our final master dictionary used to determine whether a token is classified as a word is available at http://www.nd.edu/~mcdonald/Word_Lists.html.

II. Word Lists

- Harvard IV Negative word list with inflections
- Loughran-McDonald lists:
  - Negative words
  - Positive words
  - Uncertainty words
  - Litigious words
  - Modal Strong
  - Modal Weak

III. Additional Analyses

In this section we present the results of two analyses not tabulated in the main text for brevity. Specifically, in Table IA.I we consider the regressions appearing in Table IV of the main text using normalized differences in the word measures instead of levels. The results show that the essential results remain the same whether we use levels or differences. In Table IA.II we show the results for a trading strategy based on negative word counts using a four-factor model discussed in the main text at the end of Section IV.C. The results for the trading strategy are not significant.

Table IA.I
Comparison of Negative Word Lists using Filing Period Excess Return Regressions: Normalized Differences

All variables are defined as in Table IV, except the negative word list variables are now normalized differences. The normalized difference is defined as: (current period negative word proportion – prior year average negative word proportion for the same Fama-French 48 industry) / standard deviation of the prior year Fama-French 48 industry proportion. The dependent variable in each regression is the event period excess return (defined as the firm's buy-and-hold stock return minus the CRSP value-weighted buy-and-hold market index return over the four-day event window, expressed as a percent). The word lists are available in the Internet Appendix or at http://www.nd.edu/~mcdonald/Word_Lists.html. See the Appendix of the main text for the other variable definitions. Fama-French (1997) industry dummies (based on 48 industries) and a constant are also included in each regression. The coefficients are based on 59 quarterly Fama-MacBeth (1973) regressions with Newey-West (1987) standard errors using one lag. The estimates use a sample of 44,829 (after differencing) 10-Ks over 1994 to 2008.

	(1)	(2)
Word Lists		
H4N-Inf : Normalized difference	-0.043	
	(-1.16)	
Fin-Neg : Normalized difference		-0.078
		(-2.16)
Control Variables		
Log(size)	0.132	0.134
	(3.01)	(3.02)
Log(book-to-market)	0.257	0.256
	(2.83)	(2.88)
Log(share turnover)	-0.286	-0.273
	(-2.35)	(-2.27)
Pre_FFAlpha	0.885	-0.470
	(0.02)	(-0.01)
Institutional ownership	0.260	0.241
	(0.82)	(0.75)
NASDAQ dummy	0.095	0.095
	(0.99)	(0.98)
Average R^2	2.29%	2.46%

Table IA.II
Trading Strategy Returns

This table shows the monthly four-factor adjusted returns (Alpha) from a trading strategy using the negative word counts contained in 10-Ks. The dependent variable is the monthly difference in the portfolio of returns between the quintiles with the lowest and highest measure of negative words for monthly periods during 199707 to 200706. Stock portfolios are formed in June of each year. The first two columns use proportional weights of negative words to categorize firms into quintiles while the last two columns use term weighting. The tf.idf weights are as defined in equation (1). The four factors are the three Fama-French (1993) factors (the contemporaneous market return (Market), size (SMB), and book-to-market (HML)) plus Carhart's (1997) momentum (MOM) factor. All reported coefficients are multiplied by 100.

	Proportional Weights		tf.idf Weights	
	H4N-Inf	Fin-Neg	H4N-Inf	Fin-Neg
Alpha	0.237	0.173	0.082	0.099
	(1.64)	(0.87)	(0.38)	(0.48)
Market	0.161	0.225	0.261	0.260
	(3.70)	(4.39)	(4.46)	(4.76)
SMB	0.079	0.302	0.405	0.311
	(2.00)	(4.76)	(4.54)	(4.18)
HML	-0.014	-0.272	-0.516	-0.458
	(-0.30)	(-4.85)	(-7.86)	(-7.39)
MOM	-0.057	-0.075	-0.086	-0.089
	(-1.86)	(-1.59)	(-1.80)	(-1.82)
R^2	32.06%	62.41%	71.71%	69.29%

REFERENCES

Carhart, Mark, 1997, On the persistence of mutual fund performance, *Journal of Finance* 52, 57-82.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns of stocks and bonds, *Journal of Financial Economics* 33, 3-56.

Fama, Eugene F., and Kenneth R. French, 1997, Industry costs of equity, *Journal of Financial Economics* 43, 153-193.

Fama, Eugene F., and James MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy* 81, 607-636.

Newey, Whitney K., and Kenneth D. West, 1987, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, 703–708.