THE JOURNAL OF FINANCE • VOL. LXVI, NO. 1 • FEBRUARY 2011

When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks

TIM LOUGHRAN and BILL MCDONALD*

ABSTRACT

Previous research uses negative word counts to measure the tone of a text. We show that word lists developed for other disciplines misclassify common words in financial text. In a large sample of 10-Ks during 1994 to 2008, almost three-fourths of the words identified as negative by the widely used Harvard Dictionary are words typically not considered negative in financial contexts. We develop an alternative negative word list, along with five other word lists, that better reflect tone in financial text. We link the word lists to 10-K filing returns, trading volume, return volatility, fraud, material weakness, and unexpected earnings.

A GROWING BODY of finance and accounting research uses textual analysis to examine the tone and sentiment of corporate 10-K reports, newspaper articles, press releases, and investor message boards. Examples are Antweiler and Frank (2004), Tetlock (2007), Engelberg (2008), Li (2008), and Tetlock, Saar-Tsechansky, and Macskassy (2008). The results to date indicate that negative word classifications can be effective in measuring tone, as reflected by significant correlations with other financial variables.

A commonly used source for word classifications is the Harvard Psychosociological Dictionary, specifically, the Harvard-IV-4 TagNeg (H4N) file. One positive feature of this list for research is that its composition is beyond the control of the researcher. That is, the researcher cannot pick and choose which words have negative implications. Yet English words have many meanings, and a word categorization scheme derived for one discipline might not translate effectively into a discipline with its own dialect.

In a survey of textual analysis, Berelson (1952) notes that: "Content analysis stands or falls by its categories. Particular studies have been productive to the extent that the categories were clearly formulated and well adapted to the problem" (p. 92). In some contexts, the H4N list of negative words may effectively capture the tone of a text. The question we address in this paper is whether a word list developed for psychology and sociology translates well into the realm of business.

*Loughran and McDonald are with University of Notre Dame. We are indebted to Paul Tetlock for comments on a previous draft. We also thank Robert Battalio, Peter Easton, James Fuehrmeyer, Paul Gao, Campbell Harvey (Editor), Nicholas Hirschey, Jennifer Marietta-Westberg, Paul Schultz, an anonymous referee, an anonymous associate editor, and seminar participants at the 2009 FMA meeting, University of Notre Dame, and York University for helpful comments. We thank Hang Li for research assistance.

While measuring document tone using any word classification scheme is inherently imprecise, we provide evidence based on 50,115 firm-year 10-Ks between 1994 and 2008 that the H4N list substantially misclassifies words when gauging tone in financial applications. Misclassified words that are not likely correlated with the variables under consideration—for example, *taxes* or *liabilities*—simply add noise to the measurement of tone and thus attenuate the estimated regression coefficients. However, we also find evidence that some high frequency misclassifications in the Harvard list, such as *mine* or *cancer,* could introduce type I errors into the analysis to the extent that they proxy for industry segments or firm attributes.

We make several contributions to the literature on textual analysis. Most notably, we find that almost three-fourths (73.8%) of the negative word counts according to the Harvard list are attributable to words that are typically not negative in a financial context. Words such as *tax*, *cost*, *capital*, *board*, *liability*, *foreign*, and *vice* are on the Harvard list. These words also appear with great frequency in the vast majority of 10-Ks, yet often do no more than name a *board* of directors or a company's *vice*-presidents. Other words on the Harvard list, such as *mine, cancer, crude* (oil), *tire,* or *capital*, are more likely to identify a specific industry segment than reveal a negative financial event.

We create a list of 2,337 words that typically have negative implications in a financial sense. The prevalence of polysemes in English—words that have multiple meanings—makes an absolute mapping of specific words into financial sentiment impossible. We can, however, develop lists based on actual usage frequency that are most likely associated with a target construct. We use the term Fin-Neg to describe our list of negative financial words. Some of these words also appear on the H4N list, but others, such as *felony*, *litigation*, *restated*, *misstatement*, and *unanticipated* do not.

When testing the 10-K sample, whether tone should be gauged by the entire document or just the Management Discussion and Analysis (MD&A) section is an empirical question. We show that the MD&A section does not produce tone measures that have a more discernable impact on 10-K file date excess returns. Thus, the MD&A section does not allow us to assess tone through a clearer lens.

In our results, we find that dividing firms into quintiles according to the proportion of H4N words (with inflections) in their 10-Ks produces no discernable pattern. That is, the proportion of H4N words does not systematically increase as 10-K filing returns decrease. However, when we use our financial negative list to sort firms, we observe a strong pattern. Regressions with multiple control variables confirm the univariate findings of no effect for the proportional counts from the Harvard list versus a significant impact for the Fin-Neg list.

We also show that the attenuation bias introduced by misclassifications, especially by high frequency words (which may be overweighted based on simple proportional measures), can be substantially mitigated by using term weighting. Most textual analysis uses a "bag of words" method where a document is summarized in a vector of word counts, and then combined across documents into a term-document matrix. In other disciplines, term weighting is typically used in any vector space representation of documents.[1] With term weighting, where the enormous differences in frequencies are dampened through a log transformation and common words are weighted less, both the Harvard list and our Fin-Neg list generally produce similar results.

To expand the word classification categories, we create five additional word lists. Specifically, in addition to the negative word lists, we consider positive, uncertainty, litigious, strong modal, and weak modal word categories.[2] When we assess whether these word lists actually gauge tone, we find significant relations between our word lists and file date returns, trading volume, subsequent return volatility, standardized unexpected earnings, and two separate samples of fraud and material weakness. We also examine whether negative tone classifications are related to future returns in terms of a trading strategy, and find no evidence of return predictability based on the competing measures.

The nature of word usage in firm-related news is not identical across media. Whether our results hold for samples beyond 10-Ks is an important question. We provide preliminary evidence in alternative contexts showing that in comparison with the Harvard list, the Fin-Neg list has larger correlations with returns in samples of seasoned equity offerings and news articles.

The remainder of the paper is organized as follows. Section I discusses related research on textual analysis. Section II introduces the data sources, variables, and term weighting method used in our analysis. Section III describes the various word lists and Section IV reports the empirical results. Finally, Section V concludes.

I. Research on Textual Analysis

Textual analysis is a subset of a broader literature in finance on qualitative information. This literature is confronted by the difficult process of accurately converting qualitative information into quantitative measures. Examples of qualitative studies not based on textual analysis include Coval and Shumway (2001), who examine the relation between trading volume in futures contracts and noise levels in the trading pits, and Mayew and Venkatachalam (2009), who analyze conference call audio files for positive or negative vocal cues revealed by managers' vocal signatures.

Although we focus on the more common word categorization (bag of words) method for measuring tone, other papers consider alternative approaches based on vector distance, Naïve Bayes classifications, likelihood ratios, or other classification algorithms. (See, for example, Das and Chen (2001), Antweiler and Frank (2004), or Li (2009)). Li discusses the benefits of using a statistical

[1] See Manning and Schütze (2003), Jurafsky and Martin (2009), or Singhal (2009).

[2] Modal verbs are used to express possibility (weak) and necessity (strong). We extend this categorization to create our more general classification of modal words.

approach over a word categorization one, arguing that categorization might have low power for corporate filings because "there is no readily available dictionary that is built for the setting of corporate filings" (p. 12). Tetlock (2007, p. 1440) discusses the drawbacks of using methods that require the estimation of likelihood ratios based on difficult to replicate and subjective classification of texts' tone.[3]

Authors commonly use external word lists, like Harvard's General Inquirer, to evaluate the tone of a text. The General Inquirer has 182 tag categories. Examples include positive, negative, strong, weak, active, pleasure, and even pain categories. Finance and accounting researchers generally focus on the Harvard IV-4 negative and positive word categories, although none seems to find much incremental value in the positive word lists.

The limitations of positive words in prior tests, as noted by others, is likely attributable to their frequent negation. It is common to see the framing of negative news using positive words ("did not benefit"), whereas corporate communications rarely convey positive news using negated negative words ("not downgraded").

While not every prior work uses the Harvard negative word list to gauge text tone, it is a typical example of word classification schemes. We choose to use the Harvard list for our tests because, unlike many other word lists, the Harvard list is nonproprietary. This allows us to assess exactly which words contribute most to the aggregate counts.

Perhaps the best known study in this area is Tetlock (2007), who links the *Wall Street Journal*'s popular "Abreast of the Market" column with subsequent stock returns and trading volume. Tetlock finds that high levels of pessimistic words in the column precede lower returns the next day. Pessimism is initially determined by word counts using a factor derived from 77 General Inquirer categories in the Harvard dictionary. However, later in his paper, Tetlock focuses on both negative words and weak words, as these are most highly correlated with pessimism. Tetlock notes that "negative word counts are noisy measures of qualitative information" and that the noisy measures attenuate estimated regression coefficients. In a subsequent study, Tetlock, Saar-Tsechansky, and Macskassy (2008) focus exclusively on the Harvard negative word list using firm-specific news stories. Our study shows that the noise of misclassification (nontonal words classified as negative) in the Harvard list is substantial when analyzing 10-Ks and that some of these misclassified words might unintentionally capture other effects.

[3] Other researchers link the tone of newspaper articles (Kothari, Li, and Short (2008)) or company press releases (Demers and Vega (2008), Engelberg (2008), and Henry (2008)) with lower firm earnings, earnings drift, or stock returns. Also considered are a firm's 10-K or IPO prospectus (Li (2008, 2009), Hanley and Hoberg (2010), and Feldman et al. (2008)). The main point of these papers is that the linguistic content of a document is useful in explaining stock returns, stock volatility, or trading volume.

Table I
10-K Sample Creation

This table reports the impact of various data filters on initial 10-K sample size.

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

II. Data, Variables, and Term Weights

A. The 10-K Sample

We download all 10-Ks and 10-K405s, excluding amended documents, from the EDGAR website (www.sec.gov) over 1994 to 2008.[4] Table I shows how the original sample of 10-Ks is impacted by our data filters and data requirements. Most notably, the requirement of a CRSP PERMNO match reduces the original sample of 121,217 10-Ks by 44,822 firms.[5] This is not surprising as many of the

[4] A 10-K405 is a 10-K where a box on the first page is checked indicating that a "disclosure of delinquent filers pursuant to Item 405" was not included in the current filing. Until this distinction was eliminated in 2003, a substantial portion of 10-Ks were categorized as 10-K405. The SEC eliminated the 405 classification due to confusion and inconsistency in its application. The choice does not impact our study, so we include both form types in our sample and simply refer to their aggregation as 10-Ks.

[5] We use the Wharton Data Services CIK file to link SEC CIK numbers to the CRSP PERMNOs.

firms with missing PERMNOs are real estate, nonoperating, or asset-backed partnerships/trusts that are required to file with the SEC.

We require data to be available for regression variables (e.g., returns, size, book-to-market, institutional ownership), and firms to be listed on the NYSE, Amex, or NASDAQ with a reported stock price immediately before the file date of at least $3. Eliminating low-priced firms reduces the role of bid-ask bounce when we examine the market reaction to the 10-K filing. We require the firm to have at least 60 days of trading in the year before and the year after the filing date. We also require the 10-K document to include more than 2,000 words (some filings simply reference other filings). We include only one filing per firm in each calendar year, with at least 180 days between filings. These sample selection criteria yield a sample of 50,115 observations consisting of 8,341 unique firms.

Part of our analysis focuses on the Management Discussion and Analysis subsection of the 10-K. It can be argued that the MD&A section is where management is most likely to reveal information through the tone that they use. We require at least 250 words to appear in the MD&A section, because in many cases this information is "incorporated by reference" (typically deferring to the shareholders annual report). Using the 250-word filter, it was much more common for firms to incorporate the MD&A section by reference earlier in the sample period (55% in 1994 vs. 9% in 2008).[6] We have 37,287 firm-year observations for the MD&A results.

B. Parsing the 10-Ks

As in most studies using textual analysis in finance, we use a bag of words method that requires us to parse the 10-K documents into vectors of words and word counts. A detailed description of the parsing process is provided in the Internet Appendix.[7] As we describe in the Internet Appendix, we exclude 10-K tables and exhibits from the analysis as these items are more likely to contain template language that is less meaningful in measuring tone. For example, post–Sarbanes-Oxley, most 10-Ks contain Exhibit 31.1, pertaining to the certification of the 10-K by the CEO. The standard exhibit includes a number of negative words, for example, *untrue*, *omit*, *weaknesses*, and *fraud*. Thus, when we refer to the "full 10-K" in the text, we are referring to the document excluding the tables and exhibits.[8]

[6] When the MD&A section is incorporated by reference to the annual report, it usually appears in an exhibit that is part of the filing (often with other material). Within the exhibit, the beginning and especially the ending point for the MD&A material typically is not demarcated in a manner that facilitates accurate parsing. Thus, we only include MD&A material that appears in the body of the primary document.

[7] The Internet Appendix is available on the *Journal of Finance* website at http://www.afajof.org/supplements.asp.

[8] The essential conclusions of the paper remain the same if we include the exhibits.

C. 10-K Subsamples

To evaluate the economic relevance of our word lists, we also consider two samples documenting negative financial events in other studies. First, we consider a sample of 10-Ks filed by firms subject to shareholder litigation under Rule 10b-5.[9] All of the firms in this limited sample have been accused of accounting fraud. The sample was created by a keyword search for "GAAP" and "Restatement" in 10b-5 class action suits. The Rule 10b-5 allegations argue that material omissions by managers led to inflated stock prices. Firms in the 10b-5 sample include Enron, Boston Chicken, and Cardinal Health.

Our second sample considers Doyle, Ge, and McVay's (2007) firms disclosing at least one material weakness in internal control between August 2002 and November 2005.[10] These disclosures are an artifact of Sections 302 and 404 of the Sarbanes-Oxley Act. A material weakness in internal control is described by the Public Company Accounting Oversight Board (2004) as "more than a remote likelihood that a material misstatement of the annual or interim financial statements will not be prevented or detected."

D. Variables

Our primary tests examine stock returns relative to the 10-K filing date. The file date return is measured as the 4-day holding period excess return over days 0 through 3. In all cases, the excess return refers to the firm's buy-and-hold stock return minus the CRSP value-weighted buy-and-hold market index return over the 4-day event window. The event period for the file date return is based on Griffin (2003, Table II), who documents that 10-Ks' "elevated response extends to day 3" (p. 447).[11]

In our regressions we include as control variables firm size, book-to-market, share turnover, prefile date Fama–French alpha (*Pre_FFAlpha*), institutional ownership, and a dummy variable for NASDAQ listing.[12] The first four of these variables are used in Tetlock, Saar-Tsechansky, and Macskassy (2008). To adjust for microstructure effects and different stock trading behaviors, a NASDAQ dummy and the proportion of the firm owned by institutions are also added. Postevent returns are considered using a long-short portfolio based on the proportion of negative words in the 10-K.

We also examine whether the 10-K tone measures are related to subsequent standardized unexpected earnings (SUE). In the SUE regressions, we extend

[9] We thank Peter Easton, Greg Sommers, Mark Zmijewski, and Chicago Partners LLC for providing us with the 10b-5 data.

[10] The data were downloaded from http://faculty.washington.edu/geweili/ICdata.html in August 2008.

[11] The exact point at which the 10-K becomes public is confounded by changes in the SEC's dissemination process during the sample period and the assumption that the public has completely read and assessed the document. Under the dissemination system established in November 1998, a filing is checked for simple perfunctory errors and then a private contractor makes the document available to the public.

[12] See the Appendix at the end of the main text for detailed definitions of the variables described in this section.

the control variables to include analyst dispersion and analyst revisions, as in Tetlock, Saar-Tsechansky, and Macskassy (2008).

Finally, we include industry dummy variables to control for cross-sectional effects in the data. We use the 48-industry classification scheme of Fama and French (1997), except for the logit regressions based on the 10b-5 and material weakness data, where we use a five-industry classification due to the sample size.

E. Term Weighting

In the information retrieval literature, a critical first step in the vector space (bag of words) model is the selection of a term weighting scheme. In the context of information retrieval, Jurafsky and Martin (2009, p. 771) note that term weighting "has an enormous impact on the effectiveness of a retrieval system." Essentially, term weighting acknowledges that raw word counts are not the best measure of a word's information content. Weighting schemes address three components: the importance of a term within a document (often measured by proportional occurrence or the log of frequency); some form of normalization for document length; and the importance of a term within the entire corpus (typically measured by inverse document frequency).

Weighting schemes are generically labeled tf.idf, where tf (term frequency) represents the method used to account for the word frequency and normalization, and idf (inverse document frequency) denotes the method used to adjust for impact across the entire collection. We use one of the most common term weighting schemes with a modification that adjusts for document length.[13] If N represents the total number of 10-Ks in the sample, df_i the number of documents containing at least one occurrence of the i^{th} word, $tf_{i,j}$ the raw count of the i^{th} word in the j^{th} document, and a_j the average word count in the document, then we define the weighted measure as

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

The first term attenuates the impact of high frequency words with a log transformation. For example, the word *loss* appears 1.79 million times in our sample while the word *aggravates* appears only 10 times. It is unlikely that the collective impact of the word *loss* is 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

[13] In information retrieval, a document is frequently compared with a search query, in which case it is less important to adjust for document length. Since we are comparing different documents, length matters. For a more general discussion of term weighting see Manning and Schütze (2003) or Chisholm and Kolda (1999).

by more than 90%. Alternatively, because *aggravates* appears in relatively few documents, the second term now increases the first by a factor of approximately eight. In our empirical results, we examine both the simple proportion of words for a given tonal classification and the tf.idf weighted measures.

III. Textual Analysis and Word Lists

There are other labels for textual analysis in different disciplines; terms such as content analysis, natural language processing, information retrieval, or computational linguistics describe a similar set of text-based methods. Many different disciplines use textual analysis, including psychology, anthropology, linguistics, political science, journalism, and computer science.

Dovring (1954) documents the use of textual analysis as far back as the early 1600s, when hymns were examined for word choices that threatened a particular religious group. As computers became more accessible, textual analysis attracted more attention beginning in the 1950s. The arrival of the internet and the development of search engines prompted new interest in the topic and generated more sophisticated techniques, primarily appearing under the rubric of information retrieval.[14]

The initial General Inquirer (GI) group that produced the early versions of the Harvard word lists became most active in the 1960s. The current version of the *Harvard Psychosociological Dictionary*, is available through the GI website (see http://www.wjh.harvard.edu/~inquirer/). We focus on the TAGNeg file used in Tetlock, Saar-Tsechansky, and Macskassy (2008) because it is a nonproprietary list that has been used frequently in prior studies and should be representative of negative word lists developed in disciplines other than finance and accounting.

The classification of words in a document should account for inflections, or different forms of the same word. For example, if we consider *accident* a negative word, we would probably also want to include words such as *accidental, accidentally,* and *accidents*. We expand the H4N list by inflecting each word to forms that retain the original meaning of the root word.[15] The original H4N list includes 2,005 words; our inflected version increases the number to 4,187. Our tests focus on the infected H4N list, which we label H4N-Inf.

Tetlock (2007), Engelberg (2008), and Kothari, Li, and Short (2008) find little incremental information in the Harvard positive word list, and thus we do not include this GI category in the analysis. Engelberg (2008) argues that the Harvard positive word list may fail to correlate with financial disclosures because of erroneous classification (he cites *company*, *shares*, and *outstanding* from the Harvard list). For completeness, however, when we create additional

[14] For a discussion of the early research in textual analysis, see Stone et al. (1966).

[15] We expand the word list by explicitly identifying appropriate inflections to avoid errors associated with stemming (i.e., assigning morphological variants to common root words). The problem with stemming is that often a word's meaning changes when common prefixes or suffixes are added. From the H4N list, for example, *odd* and *bitter* take on different meanings when made plural: *odds* and *bitters*.

word lists we include a positive category even though our primary focus is comparing the H4N-Inf word list with our Fin-Neg list.

We propose five other word lists: positive (Fin-Pos); uncertainty (Fin-Unc); litigious (Fin-Lit); strong modal words (MW-Strong); and weak modal words (MW-Weak). All these lists are available in the Internet Appendix or at http://www.nd.edu/∼mcdonald/Word_Lists.html.

To create the above word lists, one strategy would be to let the data empirically determine the most impactful words. This approach would allow us to develop a relatively short list of tonal words. The limitation of this approach is the endogeneity problem that would arise going forward. If, for example, managers know there is a list of words that have a significant negative impact on returns, then they will systematically avoid those words. A second strategy, which we follow, is to create a relatively exhaustive list of words that makes avoidance much more challenging.

To create the Fin-Neg, Fin-Pos, Fin-Unc, and Fin-Lit word lists, we first develop a dictionary of words and word counts from all 10-Ks filed during 1994 to 2008. We carefully examine all words occurring in at least 5% of the documents, to consider their most likely usage in financial documents (including inflections). Words that we include beyond the 5% level are typically inflections of root words that made the original cut.

We account for simple negation only for Fin-Pos words. Simple negation is taken to be observations of one of six words (*no, not, none, neither, never, nobody*) occurring within three words preceding a positive word. We would not expect to see phrases such as “not terrible earnings” in a report, so we do not consider negation for the negative word lists.

Unlike the H4N-Inf list, the Fin-Neg list is specific to business terminology. In the language of business, words like *increase* or *decrease* are tonally ambiguous. In this case, what these words imply depends on whether they precede words such as revenues or costs. Words from the Harvard lists, such as *liability* or *tax,* are expected to appear in both positive and negative contexts simply as a structural artifact of accounting language. The critical empirical question is thus whether such words appear often enough to impact the statistical power of word lists derived from other disciplines. Any nontrivial word list applied to as many documents as in our 10-K sample will misclassify—the issue is to what extent. A discipline-specific word list can reduce measurement error, thus increasing power and reducing the associated attenuation bias in parameter estimates. Also, as we will see below, some misclassified words can unintentionally proxy for other effects.

Of the 2,337 words in our Fin-Neg list, about half (1,121) overlap with the H4N-Inf list. Frequently occurring words in our list that are not on the H4N-Inf list include: *restated, litigation, termination, discontinued, penalties, unpaid, investigation, misstatement, misconduct, forfeiture, serious, allegedly, noncompliance, deterioration,* and *felony*.

Our Fin-Pos word list consists of 353 words including inflections, substantially fewer words than in the negative word list. In attempting to select positive words, one quickly realizes that there are few positive words that are not easily compromised. Knowing that readers are using a document to evaluate the value of a firm, writers are likely to be circumspect and avoid negative language, instead qualifying positive words, often in ways not easily detected by a parsing program. The tone of negative words has a much more pervasive effect. For example, *felony*, even if appearing in the phrase "whose *felony* conviction was overturned," is still negative. Words in our Fin-Pos list, such as *achieve, attain, efficient, improve, profitable,* or *upturn* are more unilateral in potential tone. We include a positive word list more in the interest of symmetry than in an expectation of discerning an impact on tone identification.

The Fin-Unc list includes words denoting uncertainty, with emphasis on the general notion of imprecision rather than exclusively focusing on risk. The list includes 285 words, such as *approximate, contingency, depend, fluctuate, indefinite, uncertain,* and *variability*.

The Fin-Lit list categorizes words reflecting a propensity for legal contest or, per our label, litigiousness. The list includes 731 words such as *claimant, deposition, interlocutory, testimony,* and *tort*. We also include words like *legislation* and *regulation,* which do not necessarily imply a legal contest but may reflect a more litigious environment. Note that many words from the Fin-Neg, Fin-Unc, and Fin-Lit lists overlap.

We extend Jordan's (1999) categories of strong and weak modal words to include other terms expressing levels of confidence. Examples of strong modal words (MW-Strong) are words such as *always, highest, must,* and *will*. Examples of weak modal words (MW-Weak) are *could, depending, might,* and *possibly*. There are 19 MW-Strong words in our list and 27 MW-Weak words.

How generalizable are our word lists to other documents, such as newspaper articles or press releases? Since our list is generated by examination of a large collection of words used in 10-Ks, we believe that the Fin-Neg list could be applied successfully to other financial documents. Although certain negative words might be used less often in some media releases, there is no reason to believe that misclassification would be more likely. We provide some preliminary evidence on this question in Section IV.G of the paper.

IV. Results

A. Sample Description

Summary statistics for the full sample of 50,115 10-Ks and subsample of 37,287 MD&As are reported in Table II. In total, we examine 2.5 billion words in the 10-Ks. For the seven word list variables, a comparison of the mean and median values in both the 10-K and MD&A samples suggests that none of the frequencies exhibit substantial skewness that might be caused by outliers. As the Fin-Neg list has only about half as many words as the H4N-Inf list, it is not surprising that, on average, a lower percentage of 10-K words are in the Fin-Neg word list (3.79% vs. 1.39%).

We also examine the means and medians for H4N-Inf and Fin-Neg by year—both show a gradual yet steady upward trend over the sample period, with

Table II

Summary Statistics for the 1994 to 2008 10-K Sample

The first seven variables represent the proportion of occurrences for a given word list relative to the total number of words. The word lists are available in the Internet Appendix or at http://www.nd.edu/~mcdonald/Word_Lists.html. See the Appendix for the other variable definitions. The sample sizes for the last three earnings-related variables are 28,679 for the full 10-K sample and 21,240 for the MD&A subsample.

	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%
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 preevent 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%

the mean for H4N-Inf rising from about 3.5% to 4.3% and the mean for Fin-Neg rising from about 1.1% to 1.7%. For the 10-K sample, the correlation between the H4N-Inf and Fin-Neg lists is a positive 0.701. Yet the correlations of the two negative word lists with the event period excess returns are notably different, with H4N-Inf and Fin-Neg having correlations of −0.004 and −0.021, respectively.

It is interesting to note the substantial rise in the proportion of negative Harvard words in the MD&A section. The median proportion of H4N-Inf negative words is 3.84% for the full 10-K, compared to 4.79% in the MD&A section. Fin-Neg reports a much smaller percentage increase for the median value (1.36%

vs. 1.43%). Also note the sharp drop in litigious words contained in the MD&A section compared to the full 10-K.

The mean market value is about $3.1 billion for the full 10-K sample, while for the MD&A sample it is only $2.1 billion. Recall that in the early years of the sample, more than half of all firms incorporate the MD&A section by reference. This large difference in mean market values between the two samples indicates that larger firms were more likely to have MD&A sections incorporated by reference and that focusing on the MD&A section produces a nonrandom sample of the 10-Ks.

B. Examining the Composition of Negative Tone

Which words have a stronger weight in determining the tone of a text? Does the frequency of a limited number of common negative words dominate the likelihood that a text will be classified as pessimistic? Are the most common Harvard words truly negative in a financial sense? Are some negative words specific to an industry?

For the 10-K sample of 50,115 firms, Table III reports, for both the full 10-K document and the MD&A subsection, the 30 most frequently occurring words from the H4N-Inf (Panel A) and Fin-Neg (Panel B) lists. The check mark in Panel A indicates whether the word is also on the Fin-Neg list and, similarly, the check mark in Panel B indicates that the word is also on the H4N-Inf list.

The words that do not appear in both the full 10-K sample and the MD&A subsample are indicated by italics. In Panel A, note that the two samples differ by only two words, with *vice* and *matters* in the 10-K sample replacing *decreased* and *decline* in the MD&A sample. Clearly, both the full 10-K and MD&A section use very similar negative words. For this reason, we initially focus our comments on the full 10-K sample results.

The first column following each list of words reports, for each word, the fractional percentage of the total negative word count. For example, the word *costs* accounts for 4.61% of the total count of all the H4N-Inf negative words.

Panel A of Table III demonstrates the considerable misclassification of negative words in 10-K documents according to the Harvard word list. The first seven words (*tax, costs, loss, capital, cost, expense,* and *expenses*) account for more than one-fourth of the total count of "negative" words. Yet in the financial world, firm *costs*, sources of *capital,* or the amount of *tax* paid are neutral in nature; managers using this language are merely describing their operations.

In some nonbusiness situations, *foreign* or *vice* might appear as negative words. In 10-K text, however, it is far more likely that *foreign* is used in the context of international operations or *vice* is used to refer to *vice*-presidents of the firm.

In textual analysis research, a higher negative word frequency indicates a more pessimistic or negative tone for the text. When we eliminate the five words that also appear on the Fin-Neg list (*loss, losses, impairment, against,* and *adverse*), Panel A reveals that from just the remaining 25 words, almost 50% of the Harvard negative word count is attributable to words that are not

Table III
Thirty Most Frequent Words Occurring in 10-Ks from the H4N-Inf and Fin-Neg Word Lists

The H4N-Inf word list is based on the Harvard-IV-4 Psychosociological Dictionary TagNeg file. We extend the original word list to include appropriate inflections. The Fin-Neg word list includes negative words from a list of all words occurring in the full sample of 10-Ks filed over 1994 to 2008. The word lists are available in the Internet Appendix or at http://www.nd.edu/∼mcdonald/Word_Lists.html. There are 4,187 H4N-Inf words, based on 2,005 root words in the original H4N list. Fin-Neg consists of 2,337 words, including inflections. The results in this table are based on the sample of 50,115 10-Ks with complete data for our regression variables downloaded from EDGAR for the period 1994 to 2008. The MD&A subsample comprises 37,287 observations. Results are presented for the full 10-K documents and the corresponding MD&A portion of the 10-Ks. Words not appearing in both the full 10-K and MD&A subsample lists are italicized.

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%

(*continued*)

Table III—*Continued*

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%

typically negative in the context of financial reporting. If more than the top 25 words are examined, we find that almost three-fourths (73.8%) of the Harvard negative word count typically does not have negative meaning in financial documents based on our classification.

Words such as *costs* from the Harvard list simply add noise to the measure. Misclassification, however, can also bias the measure of tone for specific industries. For example, for both the precious metals and coal industries, *mine* is the most common Harvard negative word (it is the second most common for nonmetallic mining). In this case the word is used in the 10-Ks merely to describe operations (e.g., Rochester *Mine*, gold *mine*, or coal *mine*). In an extreme example, the word *mine* in the 1999 10-K of Coeur d'Alene Mines Corporation accounts for over 25% of all the H4N-Inf negative word counts.

Another example is the word *cancer*, the tenth most common H4N-Inf negative word in the Fama–French pharmaceutical products industry. For both the banking and trading industries, by far the most common negative word is *capital*. For autos, *tire* appears in the top 20 most frequent H4N-Inf negative words; for oil, *crude* is ranked seventh in frequency. The tendency for some of the misclassified words to potentially proxy for other effects provides part of the basis for why our negative word list might be a better choice for financial researchers. Some of the association of the HN4-Inf list with financial variables might be attributable to the Harvard list unintentionally capturing industry effects.

Panel B of Table III reports the 30 most common words according to the Fin-Neg list. Of these 30 words, 21 of them also appear on the H4N-Inf list, whereas 9 words do not (*claims, restated, restructuring, litigation, discontinued, termination, unable, critical,* and *terminated*). Unlike the Harvard list, the most common Fin-Neg words are ones that are more likely to be negative in a financial sense.

Both panels of Table III highlight a well-known phenomenon in natural language processing popularly referred to as Zipf's law.[16] Essentially, this law tells us that there are typically a small number of very high–frequency words and a large number of low-frequency words. This reveals why term weighting could be important in financial applications where the impact of frequency should probably be muted—for example, a word occurring 10 times more frequently is most likely not 10 times more informative—and a word's impact is likely diminished by its commonality. The tf.idf weighting scheme we employ in subsequent regressions attempts to address this phenomenon.

In sum, for any nontrivial list of words, we expect a small number of the words to dominate the overall count. For the H4N-Inf word list, the majority of high-frequency words happen to also be words that would typically be considered misclassified in the context of financial documents, introducing substantial noise in measures based on the Harvard list. Further examination of word rankings within industries and for individual firms indicates that some of this misclassification error, beyond simply adding noise to the measure, is likely to introduce spurious correlations.

C. 10-K Filing Period Returns and Negative Word Lists

One way to test word lists is to examine the market's reaction at the time of a 10-K filing. If tone matters, firms filing 10-Ks with a high measure of negative words should, on average, experience negative excess returns around the filing date. Figure 1 reports the median filing period excess returns by quintiles according to the H4N-Inf and Fin-Neg word lists. Median returns

[16] Zipf's law is based on the observation that word frequency is approximately a fixed proportion of the inverse of frequency rank. More accurately, the distribution of words is similar to a power law distribution such as the distribution of firm size. See Baayen (2001) for a discussion of word frequency distributions.

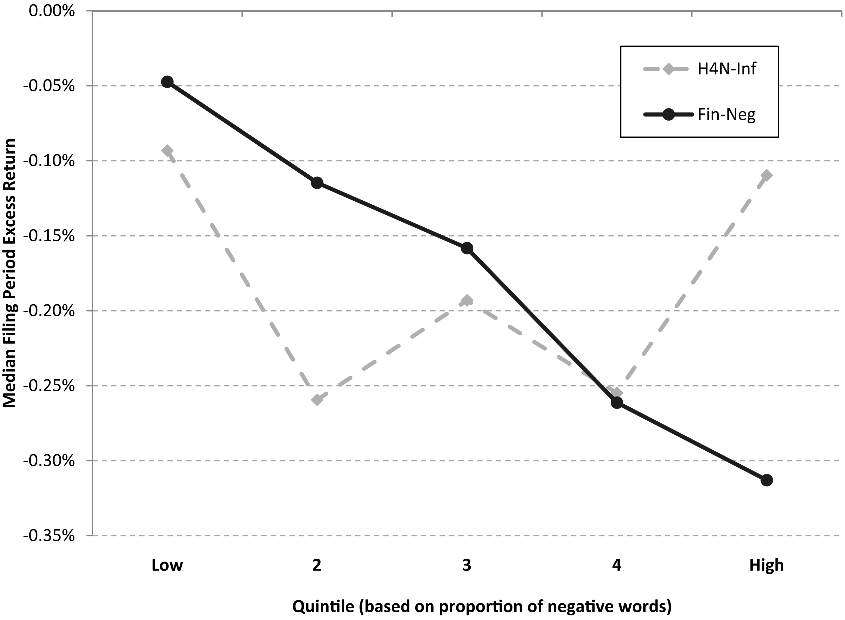

Figure 1. Median filing period excess return by quintile for the Harvard-IV-4 Psychosociological Dictionary TagNeg word list with inflections (H4N-Inf) and the Financial-Negative (Fin-Neg) word list. For each of the word lists, the sample of 50,115 10-Ks is divided into five portfolios based on the proportion of negative words. The filing period return is the holding period excess return for the 10-K file date through the subsequent 3 days, where the excess return is a firm's common stock buy-and-hold return minus the CRSP value-weighted market index buy-and-hold return.

for the H4N-Inf list do not reflect a consistent relation with the proportion of negative words. Firms with a high proportion of Harvard negative words have only a slightly lower filing period return in comparison with firms having relatively few negative words on those lists.

The pattern produced by the Fin-Neg list is what we would expect if the word lists capture useful information. Firms including a lower percentage of pessimistic words have slightly negative returns on the 4 days around the 10-K filing date compared to sharply negative median returns for the quintiles including a high percentage of negative words. The return pattern for Fin-Neg across the quintiles is monotonic.

We next examine the relation between the negative word lists and filing period returns for the 10-K sample in a multivariate context using various control variables. Table IV reports regression results defining the dependent variable as the day [0,3] filing period buy-and-hold excess return expressed as a percent. In columns (1) and (2) of Table IV, the first two independent variables are the proportions of words classified as negative using the Harvard and Fin-Neg word lists, while in columns (3) and (4) the measures for each list are based on the tf.idf weights. The control variables are size, book-to-market,

Table IV
Comparison of Negative Word Lists Using Filing Period Excess Return Regressions

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 4-day event window, expressed as a percent). The proportional weights are the word list counts relative to the total number of words appearing in a firm's 10-K. The tf.idf weighted values are defined in equation (1) of the text. See the Appendix 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 60 quarterly Fama-MacBeth (1973) regressions with Newey–West standard (1987) errors using one lag. The estimates use a sample of 50,115 10-Ks over 1994 to 2008.

	Proportional Weights		tf.idf Weights	
	(1)	(2)	(3)	(4)
Word Lists				
H4N-Inf (Harvard-IV-4-Neg	−7.422		−0.003	
with inflections)	(−1.35)		(−3.16)	
Fin-Neg (negative)		−19.538		−0.003
		(−2.64)		(−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^2	2.44%	2.52%	2.64%	2.63%

share turnover, prefile date Fama–French alpha, institutional ownership, and a NASDAQ dummy. Included in each regression are Fama and French (1997) 48-industry dummies and a constant. Because of the panel structure of the data, we use a Fama-MacBeth (1973) methodology where the firms are grouped by quarter, with Newey-West (1987) standard errors using one lag. The estimates for each period are weighted by frequency, since the calendar distribution of file dates is clustered around specific dates (see Griffin (2003)).

In columns (1) and (2), we consider the simple proportional measures for H4N-Inf and Fin-Neg. Consistent with the bivariate correlations, H4N-Inf is not significantly related to the file date excess returns while Fin-Neg has a significantly negative coefficient (t-statistic of −2.64). Thus, higher proportions of negative words, as measured by the Fin-Neg list, are associated with lower excess returns.

We know, however, from Table III that the word counts from these lists are dominated by a relatively small number of words. For the H4N-Inf list, many of these words are simply adding noise to the measure. In columns (3) and (4) we run the same regressions with the term-weighted measures of H4N-Inf and Fin-Neg. In this case, both the word lists are negative in sign, significant, and essentially identical in their impact.

This result captures the essence of subsequent results. That is, as an unadjusted measure Fin-Neg appears superior, which is not surprising since it does not contain some of the common words that H4N-Inf misclassifies. The term weighting method, however, mitigates the noise in both measures—especially for the H4N-Inf measure—to an extent that the Fin-Neg list does not dominate.

We should not overstate the regression results. Even for the regressions using the weighted measures and all of the control variables, the adjusted R^2 is still low (around 2.6% for both regressions). Only a small amount of the variation in filing period returns is explained by the independent variables. Textual analysis is not the ultimate key to the returns cipher.[17]

While some studies use the level of word counts as we have, others standardize the measure by looking at changes in proportional occurrence relative to a historic benchmark (see, for example, Tetlock, Saar-Tsechansky, and Macskassy (2008) or Feldman, et al. (2008)). Under some conditions, differencing or some form of standardization might have the advantage of reducing the impact of words contextually misclassified. Given the results from Table III, it is likely that much of the variation in differences will be driven by random variation in the frequency of common words. The differencing method also assumes that a reader can remember the frequency of negative words in previous news articles, columns, or 10-Ks—for example, that today's column or 10-K has fewer negative words than previous editions, so it may convey a bullish signal.

We report in the Internet Appendix regressions paralleling those in columns (1) and (2) of Table IV, where the proportional measures are normalized differences. The essential conclusions in terms of signs and significance remain identical. The results do not suggest, therefore, that differencing mitigates the noise problem.

In Table V we return to the empirical question of whether the MD&A section of the 10-K is a more appropriate measure of tone in a 10-K. For both the H4N-Inf and Fin-Neg lists we consider regressions that analyze word counts from only the MD&A section of the 10-K. These regressions have the same control variables as the previous table, yet have a smaller sample size since firms must have an identifiable MD&A section with more than 250 words to be

[17] To assess the appropriateness of the event window, we consider a simple t-test comparing the absolute excess return for days [0,5] relative to the file date with the average absolute excess return for the 5 days [−5, −1]. The returns remain elevated through day 4, with t-statistics for days 0 through 5 of 12.8, 16.8, 12.9, 6.9, 3.6, and 1.7. In the regressions of Table IV, only the weighted measures remain significant if we shrink the event window down to days [0,2] and none of the measures are significant if we use only days [0,1]. In addition to the potential lag between the file date and the release date, the median 10-K contains 20,000 words. The document length would require the average investor some period of time to absorb the information.

Table V
Comparison of Negative Word Lists Using Filing Period Excess Return Regressions: MD&A Section

The sample is now based on the MD&A section of 10-Ks over 1994 to 2008, where the MD&A section contains at least 250 words (N = 37,287). 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 4-day event window, expressed as a percent). The proportional weights are the word list counts relative to the total number of words appearing in a firm's 10-K. The tf.idf weighted values are defined in equation (1) of the text. See the Appendix 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 60 quarterly Fama-MacBeth (1973) regressions with Newey-West (1987) standard errors using one lag.

	Proportional Weights		tf.idf Weights	
	(1)	(2)	(3)	(4)
Word Lists				
H4N-Inf (only MD&A)	1.892		−0.005	
	(0.35)		(−1.96)	
Fin-Neg (only MD&A)		−5.344		−0.006
		(−0.68)		(−1.96)
Control Variables				
Log(size)	0.157	0.162	0.172	0.172
	(3.06)	(3.10)	(3.32)	(3.28)
Log(book-to-market)	0.324	0.330	0.334	0.335
	(3.57)	(3.59)	(3.64)	(3.65)
Log(share turnover)	−0.364	−0.362	−0.345	−0.341
	(−2.84)	(−2.82)	(−2.79)	(−2.76)
Pre_FFAlpha	−21.977	−22.279	−23.050	−23.168
	(−0.45)	(−0.45)	(−0.47)	(−0.47)
Institutional ownership	0.271	0.264	0.244	0.245
	(0.82)	(0.79)	(0.74)	(0.74)
NASDAQ dummy	0.146	0.144	0.141	0.139
	(1.45)	(1.39)	(1.40)	(1.38)
Average R^2	2.45%	2.70%	2.65%	2.76%

included. The results for both word lists using proportional measures are not significant and are only marginally significant for the weighted measures. The results of this table indicate that the MD&A section does not contain richer tonal content.

Also, as previously noted, the sample for which the MD&A section is available varies systematically over time. More firms incorporate the MD&A section by reference in the early period, locating the discussion in an exhibit. We do not include these cases in the sample because it becomes difficult to accurately parse the MD&A section (typically because the termination point is not identified). The tendency to incorporate by reference is also correlated with firm size. Thus, the sample changes in a nonrandom way through time. Because the empirical results do not indicate that the MD&A section provides more discriminating content and because of the systematic shifts in the sample, we only consider the full 10-Ks in subsequent tests.

How successful would a trading strategy using the proportional or term weighted negative word counts be? We calculate the Fama and French (1993) four-factor portfolio returns generated by taking a long position in stocks with a low negative word count and a short position in stocks with a high negative count. More precisely, in June of each year starting in 1997, we sort all available firms into quintiles based on the prior year's 10-K Fin-Neg or H4N-Inf word counts.[18] Over the next 12 months, the return differences between the long/short portfolios are regressed against the four factors. Although the alphas across the four regressions are positive, none of the values are statistically significant. Hence, after controlling for various factors, the relation between 1-year returns and negative word counts is not enough to warrant active trading by investors.

D. Additional Word Lists, Filing Date Returns, Volume, and Postevent Volatility

We have some evidence that Fin-Neg is related to short-term returns when the 10-K is filed. Is there also a relation between Fin-Neg and abnormal trading volume or subsequent stock return volatility? In addition, is there a relation between the other word lists and these effects around or after the firm's filing date?

Table VI reports regression results for three different dependent variables: event period excess returns, event period abnormal volume, and postevent return volatility. Panel A reports the regression results using proportional weights while Panel B uses the term weights (tf.idf). Each entry in the table is based on separate Fama-MacBeth (1973) regressions (42 different Fama–MacBeth regressions in all) with the specified word list along with the control variables and Fama–French industry dummy variables appearing in Table IV. Only the coefficients associated with the word lists are reported in the table. The word lists are H4N-Inf and our six word lists (negative, positive, uncertainty, litigious, modal strong, and modal weak).

In Table VI, Panel A, which is based on the proportional measure, when filing period returns are the dependent variable, the coefficient is negative and insignificant if H4N-Inf is the only word list included as an independent variable (besides the control variables). This is the identical regression as in column (1) of Table IV. When event period excess returns are the dependent variable, we find that only the Fin-Neg, uncertainty, modal strong, and modal weak word lists are statistically significant. All these coefficients are negatively signed. Firms using fewer negative, uncertain, modal strong, and modal weak words realize a more positive reaction from the market in the filing date event window.

[18] We begin in 1997 because this is the first year all firms are required to file digital forms. The estimated alphas and factor coefficients are reported in the Internet Appendix.

Table VI
Additional Word Lists, Filing Period Returns, Filing Period Abnormal Volume, and Postevent Return Volatility

The table reports the coefficients for the tone-related variables. Each entry in the table is based on a separate Fama-MacBeth (1973) regression (i.e., 42 different Fama-MacBeth regressions). The event period excess return is defined as the firm's buy-and-hold stock return minus the CRSP value-weighted buy-and-hold market index return over the 4-day event window, expressed as a percent. Event period abnormal volume is the sum of volumes for the 4-day filing event period, where volume is standardized based on the 60 days prior to the file date. Postevent return volatility is the root-mean square error from a Fama-French (1993) regression on daily returns data for the 252 days following the file date, with the first 5 days following the file date excluded. The coefficients are based on 60 (59 for Postevent return volatility) quarterly Fama-MacBeth (1973) regressions with Newey–West (1987) standard errors using one lag. Each coefficient reported is from a regression also containing the control variables appearing in Table IV (see the Appendix for definitions), Fama-French (1997) industry dummies (based on 48 industries), and a constant. The estimates are based on the sample of 50,115 (49,179 for Postevent return volatility) 10-Ks over 1994 to 2008.

		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)

In the second row of Panel A, with abnormal trading volume in the 4-day filing date period as the dependent variable, only the Harvard and Fin-Neg word lists are significant in separate regressions after controlling for other variables. Since both coefficients have positive signs, the more the negative words (as measured by the Harvard or Fin-Neg word lists) that appear in the 10-K, the higher is the abnormal trading volume during the event window.

The last row of Panel A reports the results when the subsequent return volatility is the left-hand-side variable. This variable is calculated outside of the 4-day 10-K filing period. In these seven separate Panel A regressions, all of the different word lists are highly significant (with the exception of litigious). Since all of the word lists have positive coefficients, a higher proportion of positive, negative, or modal words is linked with larger stock return volatility in the year after the filing. Thus, while event period trading volume is more difficult to explain, the word lists do a better job at explaining postevent return volatility. This volatility-word tone linkage is consistent with the internet stock message board evidence presented in Antweiler and Frank (2004).

Panel B reports the separate regression results using the term weighting procedure (tf.idf). As noted before, this weighting procedure raises the significance of the word lists by improving the signal-to-noise in the lists. For excess filing period returns, all of the word lists are significant except for litigious. When event period abnormal trading volume is the dependent variable, all the word lists, except for positive, modal strong, and modal weak, are significant. The last row of Panel B reports that all of the word lists are positively signed and significantly related to subsequent stock return volatility.[19]

E. 10-b5 Filings and Material Weakness in Internal Controls

We next examine two different firm samples to see whether the various word lists capture language usage differences. Do companies accused of accounting improprieties or firms that self-report material weaknesses in internal controls use different language from other firms in their 10-Ks? Table VII reports the logit regression results; Panel A reports the results using proportional weights while Panel B uses the term weighting procedure (tf.idf). Each entry in the table is based on a separate logit regression, for 28 different regressions in all. The independent variables are the control variables from Table IV and a separate word list in each regression.

In the first row of each panel, the binary dependent variable is a dummy variable equal to one if a 10b-5 suit was filed against the firm alleging accounting improprieties in the year after the 10-K filing date or if the 10-K was filed during the alleged violation period. Between January 1, 1994 and September 23, 2004, there were 585 firms in a potential universe of 35,992 firm-year

[19] We do not include all of our word lists in one regression due to their high degree of collinearity. An alternative is to create an omnibus measure where the proportions are aggregated (with appropriate signs). We conducted the same regressions with an omnibus measure, but in no case did it dominate simply using the Fin-Neg measure.

Table VII

Logit Regressions for Shareholder Class Action Suits for Accounting Issues and Self-reported Material Weakness in Internal Controls

The table reports the coefficients for the tone-related variables. Each entry in the table is based on a separate logit regression (i.e., 28 separate regressions). The binary dependent variable for the logit regression when the dependent variable is Fraud is equal to one if a firm had a 10b-5 class action lawsuit filed in the year after the 10-K filing date or if the 10-K file date falls within the purported violation period reported in the 10b-5 ($N = 585$). The 10b-5 sample includes only firms accused of accounting fraud. To parallel the dates available in the 10b-5 sample, all 10-Ks from 1/1/1994 through 9/23/2004 are included in the sample ($N = 35{,}992$). For the dependent variable labeled Material weakness, the binary dependent variable is set equal to one if within 18 months of the 10-K file date a disclosure of material weakness is reported in a subsequent 10-K, 10-Q, or 8-K. The material weakness sample is taken from Doyle, Ge, and McVay (2007). To parallel the dates of their data set, the observation period for this second sample is all 10-Ks from January 1, 2001 through October 30, 2005. The total material weakness sample is 17,143 with 708 cases indicating a material weakness event. For all regressions, the standard errors used to calculate the z-statistics, in parentheses, are corrected for rare-event bias using the method of Tomz, King, and Zeng (2003). The control variables from Table IV, Fama–French five industry dummies, year dummies, and a constant are also included in the logit regressions. See the Appendix for the control variable definitions.

		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)

observations. For the fraud results in Panel A, using simple proportional measures, none of the seven word lists have a significant coefficient after controlling for other variables.

In the last row of each panel in Table VII, the binary dependent variable is equal to one if within 18 months of the 10-K file date a firm disclosed a material weakness in a subsequent 10-K, 10-Q, or 8-K ($N = 708$). The material weakness sample is taken from Doyle, Ge, and McVay (2007). The coefficients on the Fin-Neg and modal strong word lists are positive and statistically significant (respective z-statistics of 3.95 and 3.50). Thus, firms with a higher proportion of negative financial words or strong modal words are more likely to report material weaknesses in their internal accounting controls.

As noted before, the term weighting procedure (tf.idf) improves the explanatory power of the various word lists. For the fraud regressions, Panel B reports that the Fin-Neg, uncertainty, and litigious word lists are all significantly linked to the 10b-5 fraud lawsuits. For the material weakness category in the second row in Panel B, all of the separate word lists (excluding modal weak) have a positive coefficient and are significant. Thus, firms using stronger language (i.e., more positive, more negative, more modal strong words) are more likely to disclose a material weakness in internal controls.

For both the 10b-5 and material weakness regressions, it is not clear what we should expect about the word list coefficients. That is, we might expect a higher proportion of negative words for firms acknowledging underlying problems, or we could expect a lower proportion if managers were trying to disguise underlying problems. Our logit regressions suggest, nevertheless, that word lists can play a role in identifying firms experiencing unusual events.

F. Negative Word Lists and Standardized Unexpected Earnings (SUE)

Tetlock, Saar-Tsechansky, and Macskassy (2008) find that the proportion of negative words in a news article can be used to predict quarterly earnings. They find that the more negative words that are used in a firm-specific news story, the lower are the firm's subsequent standardized unexpected earnings.

Table VIII provides the relation between standardized unexpected earnings and the negative word lists. In the four regressions, the dependent variable is the earnings surprise based on analyst estimates, standardized by price, for quarterly information reported within 3 months after the 10-K filing. As before, the reported coefficients are based on 60 quarterly Fama-MacBeth (1973) regressions with Newey-West (1987) standard errors using one lag. The first two columns use proportional weights for the negative word lists while the last two columns use tf.idf weighting. In the regressions, we also add analyst dispersion and analyst revisions as additional control variables.

A number of the control variables are statistically significant. As in Tetlock, Saar-Tsechansky, and Macskassy (2008), the prior period Fama–French alpha, analyst dispersion, and analyst revisions have the expected coefficient sign and significance levels. That is, the higher the prior performance, the higher the analyst revisions, and the lower the analyst dispersion, the larger is the firm's subsequent earnings surprise.

In all four columns of Table VIII, the coefficients on both the Harvard and Fin-Neg word lists are positive and statistically significant. This is the opposite of what Tetlock, Saar-Tsechansky, and Macskassy (2008) find for news stories in the days prior to the quarterly earnings announcement. Clearly, firm-specific news articles before quarterly earnings announcements appear to be an accurate reflection of the direction of subsequent earnings surprises. More negative words used by independent journalists indicate pessimism (i.e., lower

Table VIII
Standardized Unexpected Earnings Regressions

The dependent variable is the earnings surprise, based on analysts' estimates (standardized by price and expressed as a percentage), for the quarterly number reported within 3 months of the 10-K file date. The coefficients are based on 60 quarterly Fama-MacBeth (1973) regressions with Newey-West (1987) standard errors using one lag. Each coefficient reported is from a regression also containing Fama-French (1997) industry dummies (based on 48 industries) and a constant. In addition, this regression contains Analyst dispersion, which is the standard deviation of analysts' forecasts in the most recent period prior to the announcement scaled by stock price in the month prior to the announcement. Analyst revisions are calculated using the mean forecast from the 4 months prior to the earnings announcement to compute the forecast revision. See the Appendix for the other variable definitions. The estimates are based on a sample of 28,679 10-Ks over 1994 to 2008. The coefficients for the word lists in columns (3) and (4) are multiplied by 100.

	Proportional Weights		tf.idf Weights	
	(1)	(2)	(3)	(4)
Word Lists				
H4N-Inf	1.937		0.035	
	(2.58)		(4.03)	
Fin-Neg		2.683		0.030
		(2.41)		(2.87)
Control Variables				
Log(size)	0.012	0.011	0.011	0.011
	(3.22)	(2.93)	(2.94)	(2.90)
Log(book-to-market)	0.021	0.021	0.022	0.022
	(2.23)	(2.20)	(2.27)	(2.27)
Log(share turnover)	0.001	0.000	−0.002	−0.002
	(0.18)	(−0.02)	(−0.23)	(−0.21)
Pre_FFAlpha	16.913	16.646	16.520	16.630
	(2.65)	(2.61)	(2.59)	(2.60)
Institutional ownership	−0.035	−0.031	−0.028	−0.028
	(−1.65)	(−1.48)	(−1.33)	(−1.36)
NASDAQ dummy	−0.005	−0.006	−0.004	−0.003
	(−0.35)	(−0.43)	(−0.34)	(−0.27)
Analyst dispersion	−45.218	−45.515	−45.647	−45.643
	(−5.93)	(−5.99)	(−6.01)	(−6.02)
Analyst revisions	11.403	11.460	11.434	11.447
	(4.45)	(4.47)	(4.48)	(4.49)
Average R^2	18.38%	18.54%	18.56%	18.53%

subsequent earnings surprises). For 10-Ks, managers might be attempting to lower expectations with a higher proportion of negative words. When insiders are the document's authors, more negative words as measured by either the Harvard or Fin-Neg word lists point to more positive subsequent earnings surprises for the firm.[20]

[20] As we would expect, the correlation between the H4N-Inf and Fin-Neg lists with the SUE from the quarter prior to the 10-K filing date is negative (−0.02 and −0.04, respectively).

G. Negative Words in Alternative Contexts—Some Preliminary Evidence

Our study focuses on the comparative power of word lists in assessing document tone using the specific context of 10-K filings. Although developing complete empirical models for other media and other events is beyond the scope of this paper, we do offer preliminary correlation evidence of the relative performance of the Harvard and Fin-Neg dictionaries in two other applications.

First, we consider another context based on SEC filings, but a context much different from the 10-Ks. Specifically, we use a sample of seasoned equity offerings (SEOs) obtained from Thomson Financial and compare the negative tone measures with the discount of the offering price relative to the market price. We are able to match Form 424 filings for 3,623 firms over the May 1996 to June 2007 period. The Form 424 is typically filed within 1 or 2 days of the offering and consists of the prospectus, issuance expense, indemnification of directors and officers, and other material contracts. The average discount in the SEO sample is 4.0%.

The simple correlation between the SEO discount and the Harvard dictionary is 0.042. For the Fin-Neg dictionary the correlation is 0.064. If we modify the tone measures using term weighting, the correlations for the Harvard and Fin-Neg list drop to 0.026 and 0.029, respectively. As with the 10-K sample, Fin-Neg appears superior when using simple proportional measures, while the term-weighted measures are more similar in impact.

Second, the reviewer of this paper calculated the negative tone measures for all news stories relating to U.S. firms in the Dow Jones archive from 1979 to 2007, and correlated the tone measures with each firm's stock returns on the day of and days following the news release.[21] The correlation between H4N-Inf and day 0 returns was −0.039. For the Fin-Neg list, the day 0 correlation was −0.044. The correlations for both lists drop to a range between −0.008 and −0.009 for the day 1 and days [2,10] periods, with Fin-Neg being slightly more negative. The day 0 results again suggest that Fin-Neg is potentially an improved measure, although the advantage is not substantial.

V. Conclusions

We find that almost three-fourths of negative word counts in 10-K filings based on the Harvard dictionary are typically not negative in a financial context. Common words like *depreciation, liability, foreign,* and *board* are not tonal when occurring in a 10-K.

Our paper proposes two solutions to this measurement problem. First, by examining all words that occur in at least 5% of the SEC's 10-K universe, we create a list of words that we believe typically have a negative meaning in financial reports. In tests on the 10-K filing date, our negative word list is significantly related to announcement returns. Second, we create a term

[21] The reviewer used a prior version of the Fin-Neg word list that is very similar to the current version.

weighting scheme that attenuates the impact of high frequency words and allows less frequently used words to have greater impact. We find that such a scheme can lower the noise introduced by word misclassifications. In particular, with term weighting, both the Harvard and our negative word lists have improved explanatory power and the impact of misclassification appears to be mitigated.

However, while many of the misclassified words simply add noise to the tonal measure, some of the misclassified Harvard words, such as *cancer, capital,* or *mine,* are strongly linked to the language of specific industry segments. Thus, some of the power of the Harvard lists in relating tone to other financial variables might be attributable to misclassifications that unintentionally proxy for other variables. This tendency for some of the misclassified words to proxy for industry effects provides additional support for the use of our word list in financial research.

In additional analysis, we also create five other word classifications (positive, uncertainty, litigious, strong modal, and weak modal words). The paper finds evidence that some word lists are related to market reactions around the 10-K filing date, trading volume, unexpected earnings, and subsequent stock return volatility. Some of our word lists are also linked to firms accused of accounting fraud and to firms reporting material weaknesses in their accounting controls.

Given our results, we recommend the use of term weighting when creating word counts. Even though the apparent power (with term weighting) of the two negative word lists is similar, we suggest the use of our list to avoid those words in the H4N list that might proxy for industry or other unintended effects. The other word lists that we created should be used primarily to address specific topics of interest, especially since in some categories many of the words overlap with the negative word lists.

Our results do not suggest that textual analysis will resolve, to paraphrase Roll (1988), our profession's modest ability to explain stock returns. Additionally, the existing literature on financial text does not actually determine the causal link between tone and returns. Tone could simply proxy for other contemporaneous information—such as the accounting numbers revealed in the 10-K—that drives returns. Our results and others', however, suggest that textual analysis can contribute to our ability to understand the impact of information on stock returns, and even if tone does not directly cause returns it might be an efficient way for analysts to capture other sources of information.

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. All textual analysis ultimately stands or falls by the categorization procedures.

Appendix: Variable Definitions

This appendix provides definitions for the variables used in the paper.

Size

The number of shares outstanding times the price of the stock as reported by CRSP on the day before the file date.

Book-to-market

Derived from the Compustat and CRSP data items as specified in Fama and French (2001). The variable is based on the most recent Compustat data no more than 1 year before the file date. After eliminating observations with negative book-to-market, we winsorize the book-to-market variable at the 1% level.

Share turnover

The volume of shares traded in days [−252, −6] prior to the file date divided by shares outstanding on the file date. At least 60 observations of daily volume must be available to be included in the sample.

Pre_FFAlpha

The prefile date Fama–French alpha based on a regression of their three-factor model using days [−252, −6]. At least 60 observations of daily returns must be available to be included in the sample.

Institutional ownership

The percent of institutional ownership reported in the CDA/Spectrum database for the most recent quarter before the file date. The variable is considered missing for negative values and winsorized to 100% on the positive side.

Abnormal volume

The average volume of the 4-day event window [0, 3], where volume is standardized based on its mean and standard deviation from days [−65, −6].

Postevent return volatility

The root-mean square error from a Fama–French three-factor model for days [6, 252], with a minimum of 60 daily observations.

SUE

Standardized unexpected earnings for the quarterly earnings announced within 90 days after the 10-K file date. The actual earnings and the analyst forecast consensus (mean) are from I/B/E/S unadjusted files, which are used to avoid the rounding issue. The unexpected earnings are standardized with stock price.

Analyst dispersion

The standard deviation of analysts' forecasts in the most recent period prior to the earnings announcement used to calculate SUE, scaled by the stock price at the end of the quarter.

Analyst revisions

The monthly change in the mean of analysts' forecasts, scaled by the stock price in the prior month.

NASDAQ dummy

A dummy variable set equal to one for firms whose shares are listed on the NASDAQ stock exchange, else zero.

REFERENCES

Antweiler, Werner, and Murray Z. Frank, 2004, Is all that talk just noise? The information content of Internet stock message boards, *Journal of Finance* 59, 1259–1293.

Baayen, R. Harald, 2001, *Word Frequency Distributions* (Kluwer Academic Publishers, The Netherlands).

Berelson, Bernard R., 1952, *Content Analysis in Communication Research* (The Free Press, Glencoe, IL).

Chisholm, Erica, and Tamara G. Kolda, 1999, New term weighting formulas for the vector space method in information retrieval, Technical Report Number ORNL-TM-13756, Oak Ridge National Laboratory, Oak Ridge, TN.

Coval, Joshua D., and Tyler Shumway, 2001, Is sound just noise? *Journal of Finance* 56, 1887–1910.

Das, Sanjiv, and Mike Chen, 2001, Yahoo! for Amazon: Opinion extraction from small talk on the web, Working paper, Santa Clara University.

Demers, Elizabeth, and Clara Vega, 2008, Soft information in earnings announcements: News or noise? Working paper, INSEAD.

Dovring, Karin, 1954, Quantitative semantics in 18th century Sweden, *Public Opinion Quarterly* 18, 389–394.

Doyle, Jeffrey, Weili Ge, and Sarah McVay, 2007, Accruals quality and internal control over financial reporting, *The Accounting Review* 82, 1141–1170.

Engelberg, Joseph, 2008, Costly information processing: Evidence from earnings announcements, Working paper, Northwestern University.

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 Kenneth R. French, 2001, Disappearing dividends: Changing firm characteristics or lower propensity to pay? *Journal of Financial Economics* 60, 3–43.

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

Feldman, Ronen, Suresh Govindaraj, Joshua Livnat, and Benjamin Segal, 2008, The incremental information content of tone change in management discussion and analysis, Working paper, INSEAD.

Griffin, Paul, 2003, Got information? Investor response to Form 10-K and Form 10-Q EDGAR filings, *Review of Accounting Studies* 8, 433–460.

Hanley, Kathleen Weiss, and Gerard Hoberg, 2010, The information content of IPO prospectuses, *Review of Financial Studies* 23, 2821–2864.

Henry, Elaine, 2008, Are investors influenced by the way earnings press releases are written? *The Journal of Business Communication 45*, 363–407.

Jordan, R.R., 1999, *Academic Writing Course* (Longman, London).

Jurafsky, Daniel, and James H. Martin, 2009, *Speech and Language Processing* (Prentice Hall, Upper Saddle River, NJ).

Kothari, S.P., Xu Li, and James Short, 2008, The effect of disclosures by management, analysts, and financial press on cost of capital, return volatility, and analyst forecasts: A study using content analysis, Working paper, MIT.

Li, Feng, 2008, Annual report readability, current earnings, and earnings persistence, *Journal of Accounting and Economics* 45, 221–247.

Li, Feng, 2009, The determinants and information content of the forward-looking statements in corporate filings—a Naïve Bayesian machine learning approach, Working paper, University of Michigan.

Manning, Christopher D., and Hinrich Schütze, 2003, *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA).

Mayew, William J., and Mohan Venkatachalam, 2009, The power of voice: Managerial affective states and future firm performance, Working paper, Duke University.

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

Public Company Accounting Oversight Board (PCAOB), 2004, An audit of internal control over financial reporting performed in conjunction with an audit of financial statements. Auditing Standard No. 2, (Washington, DC.)

Roll, Richard, 1988, R^2, *Journal of Finance* 43, 541–566.

Singhal, Amit, 2009, Modern information retrieval: A brief overview, Working paper, Google, Inc.

Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie, 1966, *The General Inquirer: A Computer Approach to Content Analysis* (MIT Press, Cambridge, MA).

Tetlock, Paul C., 2007, Giving content to investor sentiment: The role of media in the stock market, *Journal of Finance* 62, 1139–1168.

Tetlock, Paul C., M. Saar-Tsechansky, and S. Macskassy, 2008, More than words: Quantifying language to measure firms' fundamentals, *Journal of Finance* 63, 1437–1467.

Tomz, Michael, Gary King, and Langche Zeng, 2003, ReLogit: Rare events logistic regression, *Journal of Statistical Software* 8, 1–27.