Word Power: A New Approach for Content Analysis

Article *in* Journal of Financial Economics · April 2013

DOI: 10.2139/ssrn.1787273

Narasimhan Jegadeesh
Goizueta Business School, Emory University

Word Power: A New Approach for Content Analysis

Narasimham Jegadeesh

Di Wu*

November 2011

ABSTRACT

We present a new approach for content analysis to determine the impact of various words in conveying positive or negative tone. We apply our approach to quantify the tone of 10-K filings and find a significant relation between document tone and market reaction for both negative and positive words. Previous research has not been successful in using positive words to quantify tone. We find that our measures of positive and negative tone are significantly related to filing period returns after controlling for factors such as return around earnings announcements and accruals, while earlier approaches in the literature are not. In addition, we find that the appropriate choice of term weighting in content analysis is at least as important as, and perhaps more important than, a complete and accurate compilation of the word list. We find that the market underreacts to the tone of 10-Ks, and this underreaction is corrected over the next two weeks.

*Narasimhan Jegadeesh is the Deans Distinguished Professor at the Goizueta Business School, Emory University, and NBER, ph: 404-727-4821, email: Narasimhan_Jegadeesh@bus.emory.edu. Di (Andrew) Wu is a Ph.D. student of Finance at The Wharton School, University of Pennsylvania, ph: 215-898-2543, email: wudi1@wharton.upenn.edu. We would like to thank Tim Loughran, Bill McDonald, Paul Tetlock and Sheridan Titman for helpful comments and suggestions

Capital markets allocate resources efficiently when investors use all available information to determine the marginal returns and values of investments. Numerous papers in the finance and accounting literature examine the flow of information in capital markets and the timeliness of market reaction to such information. Most of these papers focus on examining the flow of quantifiable information such as accounting data in financial statements. In addition to such readily quantifiable data, firms also present detailed descriptive information in their annual reports. Sell side analysts and the financial press also provide extensive descriptive information about firms. While a voluminous literature examines market reactions to quantitative information such as earnings, revenues and analyst recommendations, relatively few papers explore in detail how investors interpret descriptive information, and whether investors efficiently incorporate that information into prices.

The paucity of research into how investors interpret descriptive information is primarily due to the difficulty in objectively quantifying such information. With recent advances in statistical natural language processing, a growing body of literature uses content analysis to quantify the tone and content descriptive information and learn how the market interprets such information. For example, Tetlock (2007) examines the tone of the Abreast of the Market column in the Wall Street Journal and finds that a pessimistic tone is associated with lower market returns. Tetlock, Saar-Tsechansky, and Macskassy (2008) examine market reactions to tone of news stories. Feldman, Govindaraj, Livnat, and Segal (2010) and Loughran and McDonald (2011) examine the tone of 10-K reports and their association with stock returns.

Broadly, content analysis uses algorithms that have two important components. The first component is a word list or the algorithms lexicon, where each word is categorized as positive or negative (or as bullish or bearish, etc). Many of the early papers in the literature including Tetlock (2007), and Tetlock et al. (2008) use the word classification in the Harvard Psychosociological Dictionary to categorize words as either positive or negative. Loughran and McDonald (2011), however, point out that the Harvard list may not be suitable for finance and accounting applications since many words that it classifies as positive or negative may not have such connotations in a financial context. As Loughran and McDonald note, *tax(es), cost(s), capital, expense(s), and liability(ies)* are categorized as words with a negative connotation in the Harvard list but they are generally neutral in a finance context. Loughran and McDonald create a fairly comprehensive list of positive and negative words based on the words in 10-K reports, and they show that their negative word list captures the tone of 10-K reports but the Harvard list is not as successful.

The second component of a content analysis algorithm is how each word in the lexicon should be weighted, which along with its lexicon enables the algorithm to map descriptive content of any document into a quantitative score. Many of the papers in this literature use a proportional weighting scheme, where the tone is measured by the ratio of negative or positive words to the total number of words in the document. This weighting scheme implicitly assumes that all words within a category are equally important.

Loughran and McDonald (2011) also propose a weighting scheme discussed in Manning and Schütze (1999), where the weights for words are inversely proportional to the number of documents in the sample that the word found. This weighting scheme assumes that words that are found in fewer documents are more impactful than words that occur in more documents. Manning and Schütze discuss this weighting scheme in the context of information retrieval in situations where one is interested in quantifying document similarity. As an example, Manning and Schtze note that the total number of occurrences of the words *try* and *insurance* were about the same in a corpus of New York Times articles, but *insurance* appears in fewer articles. Words such as *insurance* that occur in fewer documents typically refer to narrowly defined concepts compared with words such as *try* that occur in more documents because they are likely relevant in a wide range of contexts. Therefore, articles that contain words that occur in a smaller number of documents are more likely to be similar to one another than articles that contain words that are used in a larger number of documents. The inverse document frequency weighting scheme uses this logic to assign larger weights to words that occur less frequently across documents than words than occur more frequently. However, in the context of inferring tone, there is no reason why words that are found in fewer documents would be more impactful than words that found in more documents.

This paper presents a new approach to determine the strength of various words in conveying negative or positive tone that is particularly suitable for finance and accounting applications. We apply our approach to quantify the tone of 10-Ks and we find several new results. For much of our analysis, we compute term weights using the positive and negative word list compiled by Loughran and McDonald. Interestingly, the quantitative score that our approach assigns to various 10-Ks have very low correlation with the scores assigned by Loughran and McDonald using the same list of words.

More importantly, our approach provides a more reliable measure of document tone than the earlier approaches. In particular, we find a significant relation between document tone and market reaction for the positive words, while none of the other papers in the literature have been successful in doing so. Additionally, we find that the negative tone as measured by Loughran and McDonald does not convey incremental information during the filing period after controlling for certain firm characteristics and accruals. However, our measure of tone is significantly related to filing period returns after controlling for these variables, and also after including Loughran and McDonald measure of tone.

Our next set of tests examines the extent to which the accuracy and completeness of the words in the lexicon are important for quantifying document tone. Loughran and McDonald (2011) argue that their word lists more accurately characterize positive and negative words in a context of finance than the Harvard lists and present empirical evidence supporting their claim. To examine the importance of lexicon accuracy, we implement our term weighting scheme using the Harvard lists, which presumably is less accurate from a finance applications oriented perspective. When we use our approach, the relation between tone and stock returns using the Harvard list is not significantly different from that using Loughran and McDonald list.

To examine extent to which the comprehensiveness of the word list affects the accuracy of our measure of document tone, we compute our measure using a randomly selected word list that contains only 50% of the words in the Loughran and McDonald list. Here again, we do not find that the relation between tone and stock returns using the partial list is significantly different from that using the complete list. These results indicate that the appropriate choice of the term weighting scheme is more important than a complete and accurate compilation of the lexicon to which the weighting scheme is applied.

We also examine the timeliness of market reactions to the tone of 10-K filings. For positive and negative words, we find that tone is significantly related to returns up to the next two weeks. Therefore, the market initially underreacts to the tone, but the underreaction is corrected within two weeks.

The rest of the paper is organized as follows: Section I describes the term weighting methodology we propose. Section II describes our sample and data sources and Section III reports the results of our empirical tests and Section IV concludes.

I. Methodology

This section explains our methodology. Content analysis aims to objectively characterize the message conveyed by descriptive information in various documents. In many finance and accounting applications, content analysis examines how the market reacts to such qualitative information by quantifying document tone. For example, Tetlock (2007), Tetlock, Saar-

Tsechansky, and Macskassy (2008), Feldman, Govindaraj, Livnat, and Segal (2010) and Loughran and McDonald (2011) examine how the market reacts to the tone of newspaper articles and statutory filings. Das and Chen (2007) and Antweiler and Frank (2004) employ several alternative classifiers, such as Naive Bayes and Vector Distance, to extract investor sentiment from posts on Yahoo! Finance message boards. However, the most common approach is word characterization (bag of words).

Typically, this branch of literature classifies some words as positive or negative words, and hypothesizes that the market reaction is a function of the relative number of positive, negative and total words in a document. For example, Tetlock (2007) and Feldman, et al. (2010) hypothesize a linear relation between returns and proportion of positive and negative words. Li (2006) find similar relations in 10-K filings by focusing on two root words: risk and uncertainty. The approaches in these papers implicitly assume that all words in the negative word list are equally negative and all words in the positive list are equally positive.

It is quite likely that some words are more impactful than others and an approach that assigns document scores that take into account the relative word impact could potentially provide more useful document scores. Manning and Schütze (1999) propose a weighting scheme that is widely used in the document retrieval literature that weights each word inversely proportional to document frequency, or the frequency with which the word appears in the sample of documents. This weighting scheme assigns the following weight for word j:

$$w_j^{idf} = log\frac{N}{df_j} , \tag{1}$$

where where N is the total number documents in the sample, df_j is the number of documents where word j occurs at least once. The superscript idf denotes weights that inversely related to document frequency (idf). Although the idf weighting scheme does not have any theoretical justification, Manning and Schütze (1999) report that document retrieval applications find it useful in practice. Loughran and McDonald (2011) use this weighting scheme in addition to a weighting scheme that assigns equal weights to all negative or positive words. For each word j in the negative or positive word lists, the *idf* weighted value for document i is defined as:

$$w_{i,j}^{tf.idf} = \begin{cases} 1 + log(tf_{i,j}) w_j^{idf}, & \text{if } tf_{i,j} > 0 \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

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

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

where a_i is the total number of words in document i and J is the total number of positive or negative words in the lexicon.

Although *idf* weights have an appeal in other contexts, there is no particular reason why the frequency of occurrence of a word in documents should be related to markets perception of its impact. Therefore, we propose an approach that assigns weights for each word based on how the market reacted to those words in the past. We expect that our term weighting methodology would be particularly suitable for finance and accounting applications where we can observe market reactions based on stock returns around specific events.

We use a lexicon of positive and negative words and seek a mapping between the occurrence of these words in the document and a quantitative score that has the following intuitive properties:

1. The score is positively related to the number of occurrences of each positive or negative word.
2. The score is positively related to the strength of the negative or positive words.

3. The score is inversely related to the total number of words in the document.

We propose the following functional form for the score for document i that satisfy the above properties:

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

where where w_j is the weight for word j, $F_{i,j}$ is the number of occurrences of word j in document i. The term $\frac{1}{a_i}$reflects the fact that the score is negatively related to the total number of words in the document. To the extent that the tone of a document conveys information to the market, the document score should be correlated with the stock return that accompanies the release of the document to the public. We specify the following relation between the score and the contemporaneous stock return:

$$\begin{aligned} r_i &= a + b \left(\sum_{j=1}^{J} (w_j F_{i,j}) \frac{1}{a_i} \right) + \epsilon_i \\ &= a + \left(\sum_{j=1}^{J} (b w_j F_{i,j}) \frac{1}{a_i} \right) + \epsilon_i , \end{aligned} \tag{5}$$

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

While we can directly compute $F_{i,j}$ and a_i, we have to estimate the weights associated with each word. We use the sample period prior to the release of the document to estimate w_j 's. Specifically, we fit the following regression using a sample period prior to the release of the document under consideration:

$$r_i = a + \left(\sum_{j=1}^{J} (B_j F_{i,j}) \frac{1}{a_i} \right) + \epsilon_i . \tag{6}$$

In this regression, we treat B_j's as regression coefficients and the estimated values of these coefficients provide unbiased estimates of bw_j. Note that we cannot separately estimate b

and w_j at this stage since the weights measure the relative strength of each word in the lexicon, and the weights can be scaled arbitrarily. We standardize the estimates of B_j's to obtain an estimate of the weight for each word. Specifically:

$$\hat{w}_j = \frac{\hat{B}_j - \bar{B}}{\text{Standard Deviation}(\hat{B}_j)}, \tag{7}$$

where $\hat{w}_j$ is our estimate of w_j, $\hat{B}_j$ is the slope coefficient estimate in from Regression (6), and $\bar{B}$ is the mean of $\hat{B}_j$ across all words.

To examine whether our estimate of score is related to score, we fit the following regression:

$$r_i = a + b\left(\sum_{j=1}^{J}(w_j F_{i,j})\frac{1}{a_i}\right) + \epsilon_i\,. \tag{8}$$

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

Since $\hat{w}_j$ is only an estimate of word strength (and not true w_j), we measure Score with error. However, we do not expect this source of bias to severely affect the power of our test since the regression does not use $\hat{w}_j$ individually, but uses the score which aggregates $\hat{w}_j$'s across all the negative/positive words that appear in the document. Our documents contain a large number of words of interest, and hence negative and positive measurement errors in individual word weights would offset each other, and thereby attenuate any bias due to these errors. In any event, to the extent that there are any residual measured errors, the estimate of b in Regression (8) would be biased downwards, favoring the null hypothesis.

II. Data

We obtain all 10-Ks filed in the 1995 through 2008 sample period from SEC's EDGAR database using a customized web crawling algorithm. We use the following criteria to construct our sample of 10-Ks:

1. The 10-K should be the first filing for the year by the company. We exclude any subsequent filing because most of the information in the 10-Ks would be revealed in the first filing.
2. EDGAR identifies firms that file 10-Ks using Central Index Key (CIK). We use the WRDS CIK-PERMNO file to match CIK with PERMNO from the CRSP-COMPUSTAT Merged database. We exclude all firms for which we are not able to match CIK to PERMNOs.
3. Our tests use market capitalization, book-to-market ratio and turnover as control variables. We exclude all firms for which we do not have these data for the years when the data are not available.
4. In order to mitigate the effect of bid-ask bounces, the stock price should be at least $3.00 on the filing date.
5. A number of words such as risk and casualty that are perceived as negative words in the context of non-financial firms may not have negative connotations for financial firms. Therefore, we exclude all financial firms (SIC code from 6000 through 6999). The final sample contains 40,789 filings and 8,633 unique firms. Summary statistics for the sample is reported in Table I.

We process the downloaded 10-K documents into vectors of tokens consisting of two or more alphabetic characters. We exclude tables and exhibits in the 10-Ks in our analysis. We then compare each token with a comprehensive English dictionary[1] to determine whether it is

[1]We use the 2of12inf dictionary, available at wordlist.sourceforge.net/12dicts-readme.html.

a word. Common stop words are not included in the dictionary and proper nouns are removed prior to processing. Since the documents are often in HTML format, we remove all encoded images, tables, exhibits, HTML languages and other non-text items from the documents.[2] We also remove the standard cover page, often the first page of the document with filers name and address. We do not count positive or negative words they are accompanied by a "negator"[3] within a distance of three words.

For most of our tests, we use the negative and positive word lists constructed by Loughran and McDonald (2011) (LM).[4] The LM list considers contains 353 positive words and 2,337 negative words. In this list, different inflections of a root word are counted as different words. For example, the root word *falsify* and its inflections *falsifies, falsified, falsifying, falsification* and *falsifications* are all considered as separate words. Since we expect these inflections to have the same strength as the root word, we consider any inflection to be the same as the root words. When we consider only the root words, we reduce the list to 122 positive words and 716 negative words. We perform this process manually to ensure that inflections that have different meanings, such as *defend* and *defendant*, are treated as different root words.

In some of our tests, we use the word list from Harvard-IV-4- Psychosociological Dictionary[5] that has been previously used by Tetlock (2007) and others. The Harvard list contains 544 positive words and 1,412 negative words when we consider only the root words. LM find that their word list is more suitable for finance applications since the Harvard negative list includes words such as *tax* and *capital* although these words do not have a negative connotation in finance. The LM negative list includes words such as *misstatement* and *unanticipated* that are not in the Harvard list, but have a negative connotation in finance. Table I presents a summary of the sample that we use in our analysis. In total, we examine 40,789 10-K

[2]Since some forms incorporate all text within tables, tables where less than 10% of the characters are numeric are not removed.

[3]We use "not", "no", "never" as negators.

[4]Available at www.wjh.harvard.edu/ inquirer/homecat.htm

[5]Available at www.wjh.harvard.edu/ inquirer/homecat.htm

and 10-K405 filings between 1995 and 2008. The mean market value is $3.02 billion and the book-to-market ratio has a mean value of 0.664.

III. Empirical Tests and Results

This section presents the empirical tests and results. We first estimate term weights using the approach we propose. We also present a comparison of the weights we estimate with the inverse document frequency weights used in the document retrieval literature. We then examine the factors that affect document tone and we examine the relation between tone and filing date returns. Finally, this section examines the extent to which accuracy and completeness of the word list are important for quantifying document tone.

A. Term-weight Estimates

We use the root words that we construct from the LM positive and negative word lists as our basic lexicon. We refer to this lexicon as the LM list. We first estimate the term weights for each word in the list using historical data. Specifically, we are interested in quantifying the importance that the market attaches to each root word in the LM list at about the time the document was released to the market. Since we are interested in measuring the tone of 10-K reports, we focus on stock returns around the time of 10-K filings.

Companies are required to file their 10-Ks annually. These filings are available to the investors who pay fee through a qualified provider within ten minutes after filing and to other investors for free within the next one or two days after the filing date (see Griffin (2003)). LM use stock returns in a four-day window from event days 0 through +3 relative to the filing dates to measure the information conveyed in 10-K filings. To facilitate comparability, we also use measure filing period return within this 4-day event window.

We compute filing period abnormal return r_i as follows:

$$r_i = \prod_{t=0}^{3} ret_{i,t} - \prod_{t=0}^{3} ret_{vwi,t}, \tag{9}$$

where $ret_{i,t}$ and $ret_{vwi,t}$ are the returns on stock i and on the CRSP value-weighted index on date t. Figure 1 plots the abnormal returns over the entire sample period from 1995 to 2008. The mean and median abnormal returns are -0.14% and -0.13%. The abnormal returns range from about -30% to over 30% during this period. Griffin (2003) finds that the volatility of returns in the filing period is significantly greater than the volatility during a 10-day pre- and post-announcement window and concludes that the market receives valuable information during this event window.

We estimate the term weight for each word in the LM for each calendar year T by fitting Regression (6) using all filing period returns and 10-Ks in the sample period from 1995 to year T. Specifically, we fit the regression using 1995 to estimate the term weights for 1996, using data from 1995 through 1996 to estimate the term weights for 1997 and so on. To mitigate the effect of extreme returns, we use the percentile rank of returns for this stage. Using the coefficient estimates from Regression (6), we compute term weights by standardizing the coefficients using Equation (7). We fit the regressions separately for positive and negative words and compute their term weights. Since we need at least one-year of data to estimate the first set of term weights, the sample period for all our subsequent tests is from 1996 to 2008.

Figure 2 presents the distribution of standardized weights for positive and negative words estimated using the entire 1995 to 2008 sample period. Because these are standardized weights, they are centered at zero. The wide range of weights here shows that some words are more impactful than others.

B. Word Power Weights Vs. Inverse Document Frequency Weights

We first examine the relation between the term weights we estimate, which we refer to as word power weights, or WP weights, and the inverse document frequency (*idf*) weights given by Equation (1). For this analysis, we independently rank each word based on its word power weight computed over the entire sample period, and its document frequency, which is inversely related to w_j^{idf}. Table II presents the frequency distribution of words within the intersection of word power weight quintiles and document frequency quintiles. For the positive word list, words in word power quintile 5 are the ones that have the most positive impact and the words in quintile 1 are the ones that have the least positive impact. For the negative word list, words in quintile 1 are the ones that have the most negative impact and the words in document frequency quintile 1 have the most negative impact and words in quintile 5 have the least negative impact.

The results in Table II indicate that both for positive and negative words, the least frequent words are the most impactful, which is consistent with *idf* weights. However, the least frequent words are also the least impactful ones which is the opposite of that implied by *idf* weights. Words that occur most frequently across documents tend to be neither most impactful nor least impactful, although the *idf* weights would consider them least impactful. Panel C of Table II presents the rank correlation between the word power weights we compute and the *idf* weights. We find that the correlation between these two weights is .117 for the positive word list and -.045 for the negative word list. Therefore, the term weight assigned by our approach has a low correlation with the *idf* weights for the both positive and negative word lists.

Although the weights for individual words materially differ depending on the approach, what is perhaps more important is the relation between the quantitative scores that are assign to various documents based on these weights. To examine this issue, we compute positive and negative word power scores for each 10-K as specified by Equation (4) using the estimated term weights for the particular calendar year during which the document was filed. We compute the positive and negative *tf.idf* scores using Equation (3).

The rank correlation between the word power scores and the *tf.idf* scores for the 10-Ks is .021 for the negative word score and .124 for the positive word score. Here again, the document score assigned by our approach is virtually uncorrelated with the *tf.idf* scores for negative words, and it has a low correlation for the positive word score. These results indicate that although the same word list may be used to measure the tone of a document, how different words are weighted critically affect the measured tone of various documents.

Table III presents the top five most impactful positive and negative words within each document frequency quintile. Table IV presents the list of ten most and least impactful positive and negative words and their rank based on w_j^{idf}.[6] Among negative words, some of the words ranked as most impactful based on *idf* ranking are ranked as one of the least impactful words by WP rankings. The results here further highlight the stark differences that result from different approaches to term weighting. Which of the approaches more accurately capture markets perception of word impact is an empirical issue that we address below.

C. Determinants of Tone

The tone of 10-Ks would likely be affected by a number of firm specific characteristics since they are statutory filings and firms generally tend to use them to guard against potential future liabilities. For instance, risky firms are more likely to state potential negative consequences of the risks they face than relatively safe firms. Also, firms that had recent poor

[6]For the purposes of comparability, we compute w_j^{idf} rank using only the word root for all inflections.

performance are more likely to use the 10-Ks to offer some reasons for such performance. This subsection examines the factors that potentially affect the tone of 10-Ks.

The first set of factors that we consider are the following firm specific factors:

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

Size and volatility proxy for the risks that firms face and we expect that the 10-Ks of riskier firms would contain reflect a negative tone. Firms with smaller BM are growth firms that are valued more for their growth opportunities and hence are likely to be more cautious in their 10-Ks. High turnover firms are the ones that attract more investor interest, and perhaps the management would likely be more cautious in setting investor expectations.

The next set of variables reflects recent events. These variables are:

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

The variable EADRet is the stock price response on earnings announcement date and hence provides a measure of whether the earnings contained good news or bad news. Large accruals are generally considered bad news either because it indicates an increase in working capital that may be due to bad business conditions or due to earnings manipulation. Most firms do not report balance sheet information necessary to compute accruals at the time of their preliminary earnings announcements. For example, Chen, Defond, and Park (2002) report that only 38% of the firms in their sample report balance sheet information in their press releases accompanying preliminary earnings announcements. Balsam, Bartov, and Marquardt (2002) note that not all firms issue such press releases. Therefore, for a majority of firms, the market first receives information about accruals through 10-Ks.

Finally, we include the tone of the 10-K filed by the firm in the previous fiscal year. This measure reflects the extent to which a firm tends to be cautious in the tone of 10-Ks. We examine the determinants of document tone using the following regression:

$$Score_i = a + bSize_i + cBM_i + dVolatility_i + eTurnover_i + fEADRet_i + gAccruals_i + hScore_{i-1} + \epsilon_i , \tag{10}$$

where $Score_i$ is the word power score for the i^{th} 10-K, and $Score_{i-1}$ is the score for that company's 10-K filed the previous year. To facilitate interpretation, we standardize all independent variables by subtracting the mean and dividing by the cross-sectional standard deviation. We fit the regression separately for positive and negative word scores every year and compute the coefficients and standard errors using the Fama-MacBeth approach.

Table V presents the regression estimates. Both positive and negative tones are significantly related to BM and volatility, and the signs of these coefficients are consistent with our expectations that risky firms are more explicit about potential risks and hence their 10-Ks contain a more negative or less positive tone.[7] Only negative score is significantly related

[7]For both negative and positive scores, a larger score indicates a more positive or a less negative tone.

firm size, which indicates that although smaller firms are more explicit about explaining possible negatives in their 10-Ks, the lack of a significant relation between firm and positive score indicates that larger firms do not tend to present a more optimistic picture in their 10-Ks.

We do not find any significant relation between document tone and the nature of recent news as captured by EADRet or accruals. However, the score is significantly positively related to the score the previous year. Therefore, firms tend to be similarly optimistic or pessimistic in their annual reports from one year to the next.

Our findings that the tone of 10-Ks is significantly related to a number of commonly used firm characteristics have important implications for this literature. For instance, Feldman et al. (2010) report that document tone is related to post-announcement drift. However, our results indicate that firm size, volatility and book-to-market ratios significantly predict document tone and we also know that these characteristics are related to post-announcement drift. LM report that their measure of negative tone predicts return volatility during the one-year post filing period. However, we show that the 10-Ks of high volatility firms tend have a more negative tone, and since volatility is persistent, there may not be a causal relation between tone and future volatility. Therefore, any tests that aim to test for a causal relation between document tone and future events or future stock characteristics should control for the characteristics that are significant predictors of the tone itself.

D. Document Tone and Filing Date Returns

This subsection examines the relation between document tone and stock returns during the 10-K filing period. As a first cut, we examine the filing period returns for firms sorted based on positive and negative scores. Figure 3 presents the average filing period returns for firms in various deciles. Decile 1 is firms in the largest score decile and Decile 10 is firms with the smallest score decile. For both positive and negative scores, the tone becomes more negative (or less positive) as we move from Decile 1 to Decile 10.

For positive scores, the filing period returns decline monotonically from 1.97% for Decile 1 to -1.43% for Decile 10. The filing period returns progressively decline for negative scores as well, from 1.03% for Decile 1 to -1.17% for Decile 10. The negative returns for positive score Decile 10 firms and the positive returns negative score Decile 1 firms indicate that the market interprets a low score on positive tone as bad news and a high score on negative tone as good news.

Figure 3 also indicates that our methodology is able to quantify the tone of the document using only positive words. Previous attempts to quantify document tone using positive words in Tetlock (2007), LM and others were not successful. We are successful in doing so using the same list of positive words as in LM, and this result highlights the importance of using appropriate term weights. In fact, with our approach, we find a larger difference between filing period returns of extreme deciles with positive words than with negative words.

We also examine the relation between filing period returns and document tone using the following regression:

$$r_i = a + bScore_i + \epsilon_i \, . \tag{11}$$

We fit the regression annually and Table VI reports the regression coefficients. The slope coefficient is .429 both for positive and negative words, which is reliably different from zero.

Table VI also reports the coefficient estimates with tone measured using the *tf.idf* score as the independent variable. Interestingly, the coefficient is significantly negative when we use the positive word list, which implies that documents with more positive tone based on *tf.idf* score experience more negative returns during the filing period. Therefore, the positive word list cannot be used to reliably quantify the document tone using *tf.idf* scores.

When we use the *tf.idf* score based on negative words, we find the slope coefficient to be negative. In contrast with the WP scores, a bigger *tf.idf* score indicates a more negative tone. Therefore, the negative slope coefficient on *tf.idf* score word indicates that documents with more negative tone experience more negative returns during the filing period. Although the sign of the coefficient indicates that the *tf.idf* score is useful for quantifying negative tone, the coefficient is statistically significant only at the 10% level.

We next examine the relative power of our word power score and the *tf.idf* score in explaining filing period returns by fitting the following regression:[8]

$$r_i = a + bScore_i^{wp} + cScore_i^{tf.idf} + \epsilon_i \ . \tag{12}$$

For positive words, the coefficient on the word power score is significantly positive, and the coefficient of the *tf.idf* score is again negative, although it is not statistically significant. For negative words, both the coefficients on the word power and *tf.idf* scores are about the same as they were in Regression (11), which is not surprising since these scores are uncorrelated.

In our set of tests, we examine the incremental effect of document tone after accounting for the effect of the control variables. Specifically, we fit the following regression:

$$r_i = a + \beta Score_i + bSize_i + cBM_i + dVolatility_i + eTurnover_i + fEADRet_i + gAccruals_i + \epsilon_i \ . \tag{13}$$

Table VI reports the regression coefficients. For both positive and negative words, the coefficient on EADRet is significantly positive and the coefficient on volatility is negative. The positive coefficient on EADRet indicates that the market does not fully react to the news

[8]To avoid ambiguity, Equation (12) adds the superscript wp to word power $Score_i$ defined in Equation (4), and denotes this score as $Score_i^{wp}$.

at the time of earnings announcement that precedes the filing of 10-Ks and it is pleasantly surprised by their contents for good news firms and negatively surprised for bad news firms. The market also does not seem to fully anticipate the relatively negative contents of high volatility firms. The coefficient on accruals is marginally negative, which is consistent with the fact that the market finds high accruals to be bad news. We do not find any significant relation between filing period returns and Size, BM and Turnover.

The point estimates on the slope coefficients for both positive and negative words are smaller in Regression (13) than in Regression (11). However, the differences are not statistically significant.[9] In addition, the coefficients in Regression (13) are both statistically significant. When we use *tf.idf* score as the explanatory variable in Regression (13) in place of the word power score, however, we find that the slope coefficient is not significant both for positive words and negative words. We find similar results when we simultaneously use word power and *tf.idf* scores in the regression in addition to the control variables. The difference between the results using our score and *tf.idf* score illustrate the critical importance of using our term weighting approach to reliably measure document tone.

The results so far indicate that we can use both positive and negative word lists to quantify document tone. To examine the relation between the scores based on the two word list, we compute the correlation between the scores based on each of these lists. The rank correlation between the two measures is 0.3452, which is significantly positive. Therefore, firms that convey good news using more impactful positive words in their 10-Ks on average use negative words that are not as impactful.

[9] The average differences are -0.102% (-1.50) for negative list and -0.193% (-1.62) for positive list.

To examine whether the positive scores contain incremental information after controlling for negative word score and vice versa, we fit the following regressions:

$$r_i = a + \beta Score_i^{positive} + \gamma Score_i^{negative} + \epsilon_i, \tag{14a}$$

$$\begin{aligned} r_i = a + \beta Score_i^{positive} + \gamma Score_i^{negative} + bSize_i + cBM_i + dVolatility_i + eTurnover_i \\ + fEADRet_i + gAccruals_i + \epsilon_i \ , \end{aligned} \tag{14b}$$

where the superscripts positive and negative on the scores indicate the particular word lists used to compute the scores. Panel C of Table VI presents the regression results. We find that the slope coefficients on both positive and negative word list scores are significant in these regressions. For example, in the regression that includes the control variables, the slope coefficients (t-statistics) are 0.230 (2.96) and 0.211(2.09) for positive and negative word lists, respectively. Therefore, each of these measures conveys incremental information relative to the other.

E. Accuracy and Completeness of Word List

Our analyses use the LM list of positive and negative words to determine term weights. Although LM carefully select these lists, there is always a possibility that these lists or any other list of words subjectively classified as positive or negative does not include all possible negative or positive words that may be found in 10-Ks. In addition, it is also possible that any list would contain words that the market may not perceive the same way as the compiler of the list. This subsection examines the extent to which accuracy and completeness of the word list are important for quantifying document tone.

LM point out that the Harvard positive and negative list includes many words that do have positive or negative connotations in the financial context. We use the Harvard list instead of the LM list as our lexicon to examine the extent to which our conclusions are sensitive to inclusion of inaccurately classified words. Specifically, we estimate Regression (12) using the Harvard lists as our lexicon, and compare the coefficients with ones we obtained for the LM lists.

Table VII reports the coefficients as well as the differences. For both positive and negative Harvard lists, the coefficients for the word power score are statistically significant. In addition, the differences between coefficients for Harvard and LM lists are not statistically significant. Therefore, a correct and robust term weighting scheme is more important than the accuracy of the underlying lexicon in capturing document tones.

To examine the effect of omitting some of the relevant words, we construct a lexicon with only 50% of the words from the LM list. To do so, we sort the LM lists into quintiles based on document frequency and randomly remove 50% of the words from each quintile. The reduced lists contain 63 words for the positive list and 360 words for the negative list. We estimate Regression (13) using these reduced lists.

The results in Table VII show that the coefficients for both positive and negative word power scores based on the incomplete lexicon are significant. The differences in the coefficients between the full and reduced lists are also minimal and not statistically significant. As such, our term weighting measure reliably quantifies tone even when presented with an incomplete word list, which in turn shows that the choice of term weighting scheme is at least as important as the completeness of the lexicon.

F. Timeliness of Market Reaction to Tone

This section examines whether the market fully reacts to document tone around the 10-K filing or whether it underreacts or overreacts to the tone. To examine this issue, we test whether the tone of 10-Ks predicts future returns over various horizons. Specifically, we fit the following regression:

$$r_{i,(t+5,t+T)} = a + \beta Score_i^{wp} + \epsilon_i , \tag{15}$$

where we measure returns in the event window from five days to T days after the filing window. We consider event windows of one week (five trading days), two weeks (ten trading days) and one month (22 trading days).

Table VIII reports the regression estimates using the Fama-MacBeth approach. For positive words, the slope coefficient is .143 for the one-week window and .330 for the two-week window, which are both statistically significant. However, the slope coefficient is insignificant for the one-month window. We find similar results for negative words as well.

To further examine the robustness of these results, we use size-adjusted returns as the dependent variable in place of market-adjusted returns in Regression (15). To compute size-adjusted returns, we first identify the NYSE size decile of the firm at the end of the month prior to the filing date. We compute size-adjusted returns over various event windows as the buy-and-hold returns of the stock minus the contemporaneous buy-and-hold returns for the matching size decile portfolio.

Table VIII (Panel A) also reports the regression estimates using size-adjusted returns as the dependent variable. Here again, the slope coefficients are significant for one- and two-week event windows but insignificant for the one-month event window and the results are similar for both positive and negative words. Therefore, the results are not sensitive to the choice of benchmarks.

Table VIII (Panel B) also presents the regression estimates using *tf.idf* scores. The slope coefficients for one week market-adjusted returns are .034 and -.043 for positive and negative words, respectively, and these coefficients are not reliably different from zero. In contrast with the results in Panel A, none of the slope coefficients are different from zero for either positive words or negative words.

The finding in this section indicates that the market does not fully respond to the tone of 10-Ks during the filing period. However, we can detect the underreaction only when we use our approach to determining document score. The filing date underreaction is corrected fairly quickly and we do not find any delayed reaction beyond two weeks.

IV. Conclusion

This paper proposes a new return-based term weighting scheme for content analysis for finance and accounting applications. Our measure of document tone based on this term weighting scheme for 10-Ks is significantly related to market returns of filing firms around their 10-K filing dates. Furthermore, our measure of tone is reliably related to filing date returns for both positive and negative word lists, while none of the other measures in the literature are related market reaction when only the positive word list is used. In addition, we find that our measure of tone is significantly related to filing date returns after controlling for additional factors such as earnings announcement date returns, accruals and volatility. Third, we find that the accuracy and completeness of the underlying lexicon is secondary to the term weighting scheme. In other words, with our term weighting method, useful information can be extracted even if the underlying word lists contain extraneous words or when they are incomplete.

We also find that the market does not fully respond to the tone of 10-Ks during the filing period. The underreaction during the filing period, however, is corrected fairly quickly and we do not find any delayed reaction beyond two weeks. In contrast to these findings, in untabulated results, we did not find any evidence of market underreaction to document scores for 10-Ks computed using inverse document frequency weights as in Loughran and

McDonald (2011). This contrast further reinforces the importance of accurately measuring the tone for fully understanding the timeliness of markets reaction to document tone.

Our term weighting methodology can be extended beyond examining the tone of statutory financial filings. Since our approach is not sensitive to the underlying word list, we expect that it would be useful in other scenarios as well where quantifying tone is beneficial, such as the analysis of financial news reports, firms press releases, etc. We plan to explore these issues in future research.

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.

Balsam, Steven, Eli Bartov, and Carol Marquardt, 2002, Accruals management, investor sophistication, and equity valuation: Evidence from 10-Q filings, *Journal of Accounting Research* 40, 987–1012.

Chen, Shuping, Mark L. Defond, and Chul W. Park, 2002, Voluntary disclosure of balance sheet information in quarterly earnings announcements, *Journal of Accounting and Economics* 33, 229–251.

Das, Sanjiv, and Mike Chen, 2007, Yahoo! for Amazon: Opinion extraction from small talk on the web, *Management Science* 53, 1375–1388.

Feldman, Ronen, Suresh Govindaraj, Joshua Livnat, and Benjamin Segal, 2010, The incremental information content of tone change in management discussion and analysis, *Review of Accounting Studies* 15, 915953.

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

Li, Feng, 2006, Do stock market investors understand the risk sentiment of corporate annual reports?, *Working Paper, University of Michigan.*

Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *Journal of Finance* 66, 35–65.

Manning, Christopher D., and Hinrich Schütze, 1999, *Foundations of Statistical Natural Language Processing.* (MIT Press).

Sloan, Richard G., 1996, Do stock prices fully reflect information in accruals and cash flows about future earnings?, *Accounting Review* 71, 289–315.

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., Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More than words: Quantifying language to measure firms fundamentals, *Journal of Finance* 63, 1437–1467.

Figure 1. Distribution of Filing Period Abnormal Returns

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

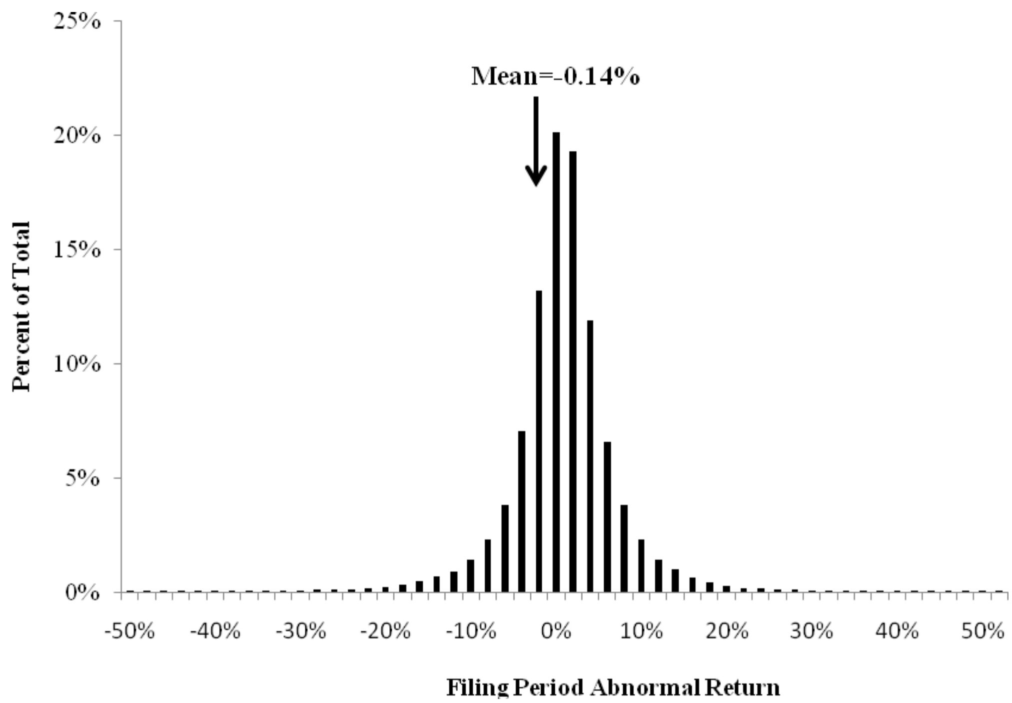

Figure 2. Distribution of Word Weights

This figure presents the distribution of word power weights for positive and negative words. This figure plots the frequency distribution of weights based on Regression (6) and Equation (7) in the text fitted over the sample period of 1995-2008. Weights for negative and positive words are computed according to Equation (7) of the text for the entire sample period of 1995-2008. For ease of comparison, the weights are demeaned and divided by the standard deviation across the respective cross-sections.

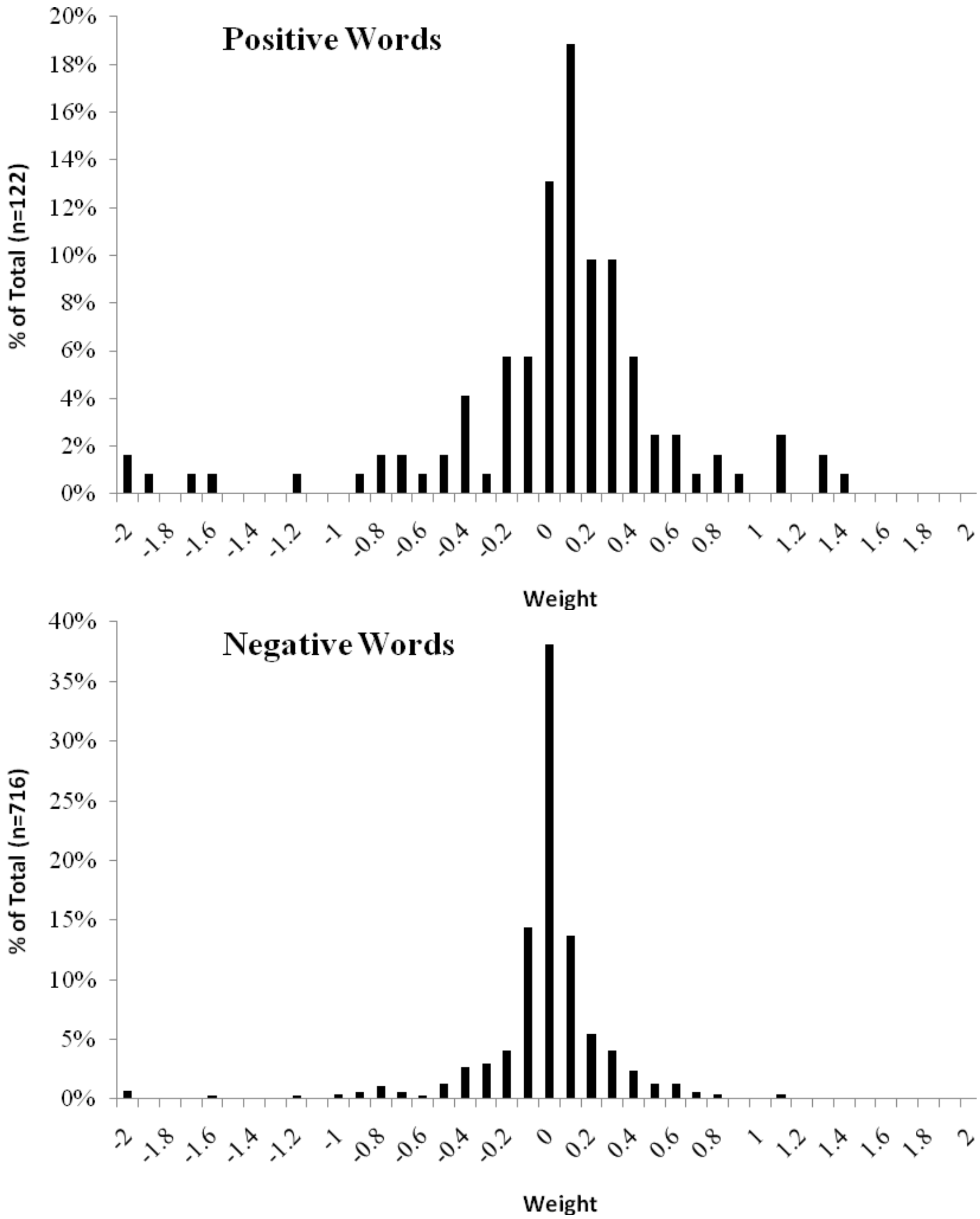

Figure 3. Mean Filing Period Abnormal Return Using WP Weights

This figure presents the distribution of filing window abnormal returns, defined as a firms buy-and-hold return minus the CRSP value-weighted index return over the four-day window of [filing date, filing date + 3] across various deciles of filings sorted based on the word power scores of the 10-Ks. We compute the word power weights for each year using Regression (6) and Equation (7) over the sample period prior to the filing of 10-Ks. We compute the positive and negative tone for each 10-K using Equation (4). Decile 1 is comprised of the decile of firms with the most positive (or least negative) document scores and decile 10 is comprised of the decile of firms with the least positive (or most negative) document scores. Mean return is the average filing period abnormal returns for all firms in that decile. The sample comprises 40,789 10-Ks over the 1995 to 2008 sample period.

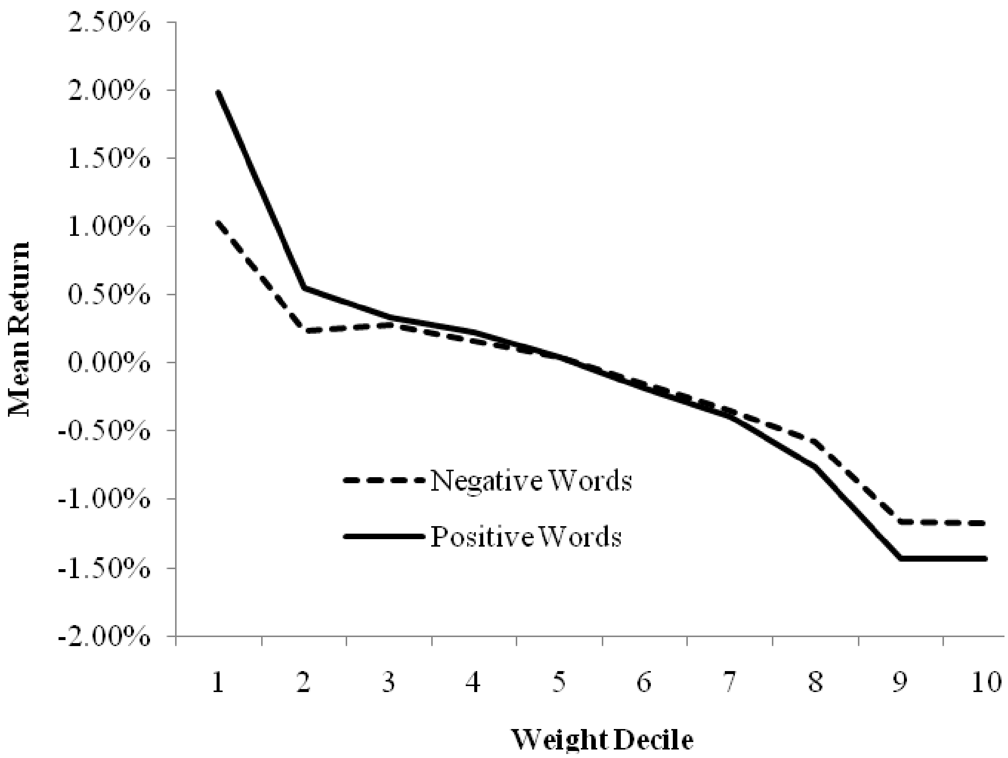

Table I
Summary Statistics

This table presents the number of firms in the sample and the mean and median of the market capitalization of equity (Size) at the beginning of each year, the book-to-market ratio (BM) and the annual turnover.

Year	# of Firms	Size ($bln)		BM		Turnover	
		Mean	Median	Mean	Median	Mean	Median
1995	1,429	$2.03	$0.34	0.738	0.714	0.918	0.575
1996	2,330	$1.82	$0.22	0.754	0.717	1.229	0.754
1997	3,607	$1.65	$0.20	0.715	0.676	1.333	0.877
1998	3,619	$2.05	$0.23	0.698	0.664	1.378	0.932
1999	3,337	$2.79	$0.23	0.568	0.659	1.558	0.965
2000	3,533	$3.21	$0.32	0.677	0.689	1.898	1.235
2001	3,066	$3.20	$0.33	0.764	0.667	1.709	1.094
2002	2,850	$3.07	$0.36	0.746	0.710	1.664	1.059
2003	2,629	$2.70	$0.33	0.828	0.706	1.638	1.130
2004	3,013	$3.22	$0.45	0.827	0.763	1.890	1.318
2005	2,940	$3.46	$0.50	0.386	0.673	1.961	1.360
2006	2,904	$3.87	$0.59	0.648	0.661	1.998	1.505
2007	2,845	$4.38	$0.65	0.338	0.649	2.184	1.672
2008	2,687	$4.34	$0.58	0.65	0.644	2.595	2.121
1995-2008	40,789	$3.02	$0.36	0.664	0.681	1.731	1.15

Table II
Cross-Tabulation of Word Weights and Frequencies

This table presents the distribution of term weights for words in various term frequency quintiles. Term frequency of each word is the percentage of 10-Ks in which the word appears. Frequency quintile 1 contains the quintile of words with the lowest frequency and frequency quintile 2 contains the quintile of words with the highest frequency. This table reports the word power weights computed using Regression (6) and Equation (7) in the text fitted over the sample period of 1995-2008. We independently sort the words based on word power weights, and Weight Quintile 1 contains the words with the smallest weights and Weight Quintile 5 contains the words with the largest weights. Panel A presents the frequency distribution for positive words and Panel B presents that for negative words. Panel C reports the rank correlation between term frequency and term weights.

Panel A: Positive Words

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

Panel B: Negative Words

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

Panel C: Correlation Between WP and idf weights and document scores

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

Table III
Top Five Most Positive and Negative Words within Frequency Quintiles

This table presents the five positive and words with the largest word power weights within each term frequency quintile. Term frequency of each word is the percentage of 10-Ks in which the word appears. Frequency quintile 1 contains the quintile of words with the lowest frequency and frequency quintile 2 contains the quintile of words with the highest frequency. This table reports the word power weights computed using Regression (6) and Equation (7) in the text fitted over the sample period of 1995-2008.

Panel A: Top 5 Most Positive Words

		Frequency Quintiles		
1	2	3	4	5
ingenuity	influential	exceptional	attractive	favorable
acclaimed	regain	versatile	desirable	opportunity
revolutionize	enthusiasm	compliment	progress	efficiency
courteous	optimistic	empower	highest	leading
incredible	excited	solves	stable	gained

Panel B: Top 5 Most Negative Words

		Frequency Quintiles		
1	2	3	4	5
imperil	turbulent	overestimate	misuse	unsuccessful
insubordination	underinsured	uncorrected	wasted	discourage
vitiate	unwarranted	frustrated	illegal	prosecute
bailout	aggravate	reluctant	worsen	insufficient
unwelcome	inexperience	unconscionable	strain	unauthorized

Table IV

Comparison of Word Power Weights and idf Term Weights

This table presents the top and bottom ten positive and negative words based in word power weights and their *idf* term weights. This table reports the words with the largest word power weights based on Regression (6) and Equation (7) in the text fitted over the sample period of 1995-2008. The *idf* term weighting scheme assigns weights inversely proportional to document frequency, as described in Equation (1). The document frequency of each word is the percentage of 10-Ks in which the word appears. Panels A and B present the top and bottom positive and negative words, respectively.

Panel A: Positive Words

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

Panel B: Negative Words

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

Table V
Determinants of Negative and Positive Tone

This table reports the relation between document tone computed using word power weights and firm characteristics. We compute the word power weights for each year using Regression (6) and Equation (7) over the sample period prior to the filing of 10-Ks. We compute the positive and negative tone for each 10-K using Equation (4). Size is the natural logarithm of the market capitalization of equity at the end of month before the 10-K filing date, BM is the ratio of the book value of equity as of the fiscal year end in the 10-K, Volatility is the standard deviation of the firm-specific component of returns estimated using up to 60 months of data as of the end of the month before the filing date and Turnover is the natural logarithm of the number of shares traded during the period from six to 252 trading days before the filing date divided by the number of shares outstanding on the filing date. EADRet is the buy-and-hold returns within the three-day earnings announcement window (earnings announcement date to earnings announcement date plus 2) minus the CRSP value-weighted index return and Accruals is computed as in Sloan (1996). is the Scorei-1 is the document score for the previous year. We fit the annual regressions each year in the entire sample period of 1995-2008. A constant is also included in each regression. The coefficients are based on 13 annual Fama-MacBeth regressions. The estimates use a sample of 40,789 10-Ks over 1995 to 2008. All independent variables are standardized to a mean of 0 and standard deviation of 1.

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

Table VI
Filing Period Abnormal Return Regressions

This table reports the estimates of the regression of filing period abnormal return, defined as a firms buy-and-hold return minus the CRSP value-weighted index return over the four-day window of [filing date, filing date+3] against document scores and various control variables. We compute the word power weights for each year using Regression (6) and Equation (7) over the sample period prior to the filing of 10-Ks, and compute positive and negative WP scores for each 10-K using Equation (4). The idf term weighting scheme assigns weights inversely proportional to document frequency, as described in Equation (1). The document frequency of each word is the percentage of 10-Ks in which the word appears. See Table V for the definitions of the control variables. The coefficients are based on 13 annual Fama-MacBeth regressions. The estimates use a sample of 40,789 10-Ks over 1995 to 2008. All independent variables are standardized to a mean of 0 and standard deviation of 1.

Panel A: Positive Words

	Models					
	(1)	(2)	(3)	(4)	(5)	(6)
Term Weighting Scheme						
WP	0.429		0.387	0.254		0.247
	(2.54)		(2.60)	(2.95)		(2.96)
tf.idf		-0.092	-0.224		-0.035	-0.09
		(-1.86)	(-1.53)		(-1.78)	(-1.22)
Control Variables						
Size				-0.057	-0.025	-0.047
				(-0.68)	(-0.71)	(-0.54)
BM				-0.457	0.004	0.234
				(-0.04)	(0.07)	(0.02)
Volatility				-0.363	-0.772	-0.335
				(-1.97)	(-2.63)	(-2.01)
Turnover				-0.109	-0.078	-0.106
				(-1.23)	(-1.04)	(-1.22)
EADRet				0.640	0.616	0.638
				(5.51)	(5.52)	(5.47)
Accruals				-0.282	-0.877	-0.287
				(-1.58)	(-1.79)	(-1.61)

Panel B: Negative Words

	Models					
	(1)	(2)	(3)	(4)	(5)	(6)
Term Weighting Scheme						
WP	0.429		0.396	0.268		0.263
	(2.47)		(2.59)	(2.50)		(2.58)
tf.idf		-0.321	-0.313		-0.106	-0.12
		(-1.82)	(-1.77)		(-1.39)	(-1.62)
Control Variables						
Size				-0.069	-0.058	-0.052
				(-0.84)	(-0.69)	(-0.62)
BM				1.67	8.074	2.256
				(0.14)	(0.64)	(0.19)
Volatility				-0.434	-0.472	-0.41
				(-2.27)	(-2.42)	(-2.27)
Turnover				-0.118	-0.136	-0.114
				(-1.34)	(-1.50)	(-1.33)
EADRet				0.641	0.607	0.638
				(5.54)	(5.86)	(5.50)
Accruals				-0.28	-0.283	-0.286
				(-1.61)	(-2.07)	(-1.64)

Panel C: Both Positive and Negative Scores
(Rank Correlation of Positive and Negative Scores=0.3452)

	Models	
	(7)	(8)
Term Weighting Scheme		
$\text{WP}_{Positive}$	0.371	0.23
	(2.54)	(2.96)
$\text{WP}_{Negative}$	0.274	0.211
	(2.33)	(2.09)
Control Variables		
Size		-0.066
		(-0.81)
BM		-1.727
		(-0.16)
Volatility		-0.345
		(-1.91)
Turnover		-0.1
		(-1.19)
EADRet		0.641
		(5.54)
Accruals		-0.278
		(-1.59)

Table VII
Filing Period Abnormal Return Regressions on Other Word Lists

Hvd-Neg is the Harvard IV-4 Negative dictionary expanded for inflections by Loughran & McDonald (2011). Hvd-Pos is the Harvard IV-4 Positive dictionary expanded for inflections by the authors. Neg_Omit is the negative word list with 50% of the words randomly removed from each frequency quintile. Pos_Omit is the negative word list with 50% of the words randomly removed from each frequency quintile. LM_Pos and LM_Neg are WP coefficients reported in Column (4) of Table VI. The dependent variable in each regression is the filing period abnormal return, defined as a firms buy-and-hold return minus the CRSP value-weighted index return over the four-day window of [filing date, filing date+3]. We compute the word power weights for each year using Regression (6) and Equation (7) over the sample period prior to the filing of 10-Ks, and compute positive and negative WP scores for each 10-K using Equation (4). The idf term weighting scheme assigns weights inversely proportional to document frequency, as described in Equation (1). The document frequency of each word is the percentage of 10-Ks in which the word appears. See Table V for the definitions of the control variables. The coefficients are based on 13 annual Fama-MacBeth regressions. The estimates use a sample of 40,789 10-Ks over 1995 to 2008. All independent variables are standardized to a mean of 0 and standard deviation of 1.

Panel A: Additional Word Lists

	Models			
	Hvd-Pos	Hvd-Neg	Pos_Omit	Neg_Omit
Term Weighting Scheme				
WP	0.196	0.409	0.296	0.253
	(2.13)	(2.51)	(3.75)	(4.00)
Control Variables				
Size	-0.062	-0.064	-0.044	-0.062
	(-0.74)	(-0.77)	(-0.52)	(-0.74)
BM	2.249	2.861	-0.207	1.464
	(0.19)	(0.23)	(-0.02)	(0.12)
Volatility	-0.43	-0.449	-0.37	-0.435
	(-2.26)	(-2.31)	(-1.98)	(-2.22)
Turnover	-0.120	-0.123	-0.096	-0.114
	(-1.33)	(-1.36)	(-1.15)	(-1.27)
EADRet	0.639	0.640	0.640	0.639
	(5.48)	(5.51)	(5.50)	(5.47)
Accruals	-0.284	-0.288	-0.298	-0.276
	(-1.59)	(-1.63)	(-1.71)	(-1.55)

Panel B: Differences in Slope Coefficients

	Models			
	Hvd-Pos	Hvd-Neg	Pos_Omit	Neg_Omit
WP	0.058	-0.155	-0.040	-0.030
	(0.81)	(-0.49)	(-1.32)	(-0.62)

Table VIII
Document Tone and Future Returns

This table reports the slope coefficient of the regression of future stock returns against document score. Market-adjusted returns is stock return minus contemporaneous CRSP value-weighted index return, and Size-adjusted return is stock return minus the contemporaneous return on matched size decile portfolio (available at http://mba.tuck.dartmouth.edu/pages/ faculty/ken.french/data_library.html). The dependent variable is the abnormal returns computed within the Event Windows specified at the top of the respective columns. The independent variables in all regressions are the Word Power score calculated using lists of positive and negative words. We compute the word power weights for each year using Regression (6) and Equation (7) over the sample period prior to the filing of 10-Ks, and compute positive and negative WP scores for each 10-K using Equation (4). The estimates use a sample of 40,789 10-Ks over 1995 to 2008. The independent variables are standardized to a mean of 0 and standard deviation of 1. The table reports the coefficients and t-statistics computed using the Fama-MacBeth approach with annual regressions.

Panel A: Positive Words

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

Panel B: Negative Words

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