Journal of Financial Economics 110 (2013) 712–729

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

Word power: A new approach for content analysis ☆

Narasimhan Jegadeesh [a,*], Di Wu [b]

[a] Finance at the Goizueta Business School, Emory University and NBER, United States
[b] Finance at The Wharton School, University of Pennsylvania, United States

ARTICLE INFO

Article history:
Received 30 October 2012
Received in revised form
29 April 2013
Accepted 27 May 2013
Available online 7 September 2013

Keywords:
Content analysis
Lexicons
Term weighting

JEL classification:
G14
G19

ABSTRACT

We present a new approach for content analysis to quantify document tone. We find a significant relation between our measure of the tone of 10-Ks and market reaction for both negative and positive words. We also 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. Furthermore, we show that our approach circumvents the need to subjectively partition words into positive and negative word lists. Our approach reliably quantifies the tone of IPO prospectuses as well, and we find that the document score is negatively related to IPO underpricing.

1. Introduction

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 quantitative information such as accounting data in financial statements. In addition to such readily quantifiable data, firms 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 of 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 filed with the Securities and Exchange Commission (SEC) and their association with stock returns.

The most commonly used approach for content analysis in finance has two components. The first component is a word

☆ We would like to thank Tim Loughran, Bill McDonald, Paul Tetlock, Sheridan Titman, Francis Diebold, Joel Peress, Lyle Ungar, Bill Schwert (the Editor), an anonymous referee, and seminar participants at the Wharton School of the University of Pennsylvania, and participants at the American Finance Association 2012 annual meeting, Atlanta Federal Reserve Conference, and the Nomura Global Quantitative Equity Conference for helpful comments and suggestions.

* Corresponding author.
E-mail addresses: Jegadeesh@emory.edu (N. Jegadeesh), wudi1@wharton.upenn.edu (D. Wu).

0304-405X/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jfineco.2013.08.018

list or the algorithm's lexicon, in which 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, Saar-Tsechansky, and Macskassy (2008) use the word classification in the Harvard IV-4 Psychosociological Dictionary to categorize words as either positive or negative. Loughran and McDonald (2011), however, point out that the Harvard list might not be suitable for finance and accounting applications because many words that it classifies as positive or negative might not have such connotations in a financial context. Therefore, Loughran and McDonald create a comprehensive list (hereafter referred to as the "LM list") of positive and negative words based from 10-K reports, and they find that the negative word list captures the tone of 10-K reports better than the Harvard list.[1]

The second component of a content analysis algorithm is how each word in the lexicon is 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, in which 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.

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 objectively determine term weights based on market's reactions to 10-K filings. We apply our approach to determine the relative weights for the words in the lexicon to quantify the tone of 10-Ks.

We find several new results with our approach. For much of our analysis, we compute term weights using the positive and negative word list compiled by Loughran and McDonald. Interestingly, the quantitative scores 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 has been successful in doing so. In addition, our measure of tone is significantly related to filing period returns after controlling for additional variables.

One of the advantages of our approach is that it objectively extracts term weights from market reactions, and it does not rely on subjective judgment. While the first part of our tests subjectively classifies words as ones with either positive or negative connotation, our next set of tests examines whether our approach allows us to remove such subjectivity.

We use two word lists in these tests. The first list merges the LM lexicons with positive and negative words. We then determine term weights for words in the combined lexicon based on market reactions to 10-K filings. This approach uses market reactions to determine not only the strength of each word but also the direction as to whether the word had a positive or negative impact.

The second list generalizes our approaches to all tonal words in the English language from a wide variety of sources. Specifically, we merge the following lists of tonal words to compile a global lexicon[2]: the Harvard IV-4 Psychosociological Dictionaries,[3] the LM list, and the top and bottom two hundred words from the word list developed by Bradley and Lang (1999), in which each word is rated according to the standardized emotional responses from subjects in an experimental setting.[4]

This global list includes words with positive and negative connotations as well as words with neutral connotations in our context. We find that the document tone based on the combined word list is as informative as that based on separate positive and negative word lists. This finding demonstrates that our approach does not need preclassified positive and negative lexicons, and we can eliminate the subjectivity associated with such classifications.

The tests with the global lexicon also allow us to evaluate the robustness of our approach to the inclusion of extraneous words. The global lexicon includes words such as *tax* and *liability* that are neutral in a financial context and are, therefore, not tonal words. Our result that the document scores that we obtain are informative with the global lexicon indicates that we can apply our approach even when the list includes extraneous words.

Our next set of tests examines the extent to which the completeness of the words in the lexicon is important for quantifying document tone. To do so, we compute our measure using incomplete word lists composed of 50% of the words in the LM list, and in the global list that we randomly select. The relation between tone and stock returns using the partial lists is not significantly different from that using the complete lists. These results indicate that we can reliably quantify document tone of 10-Ks even if the lexicon is incomplete.

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 for up to two weeks. Therefore, the market initially underreacts to the tone, but the underreaction is corrected within two weeks.

We also explore the generalizability of our approach to a different economic context. Specifically, we examine the relation between tone scores of IPO prospectuses computed using the term weights that we previously determined using 10-Ks and IPO underpricing. We find a negative relation between the tone scores and underpricing as predicted by a number of models in the literature. This finding illustrates

[1] The LM list is available at http://www3.nd.edu/~mcdonald/Word_Lists.html.

[2] We also considered using all root words in the English language, rather than just the tonal words. But since there are more root words in the language than the number of 10-Ks, we need some data reduction. Therefore, we restrict the global list to all tonal words.

[3] Available at http://www.wjh.harvard.edu/~inquirer/homecat.htm. The Harvard-IV-4 dictionaries merge Harvard IV Dictionaries and the Lasswell Value Dictionary.

[4] Available through http://csea.phhp.ufl.edu/media/anewmessage.html.

that the term weights we determine in one context can be reliably used in a different context.

The rest of the paper is organized as follows. Section 2 describes the methodology. Section 3 describes our sample and data sources, and Section 4 reports the results of our empirical tests. Section 5 examines the timeliness of market reaction to the tone of 10-Ks. Section 6 examines the relation between the tone of initial public offering (IPO) prospectuses and underpricing, and Section 7 concludes.

2. Methodology

Content analysis aims to objectively characterize the message conveyed by descriptive information. 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 naïve Bayes and vector distance, to extract investor sentiment from posts on Yahoo! Finance message boards.

Typically, this branch of literature classifies some words as positive or negative words, and hypothesizes that market reaction is a function of the relative number of positive, negative and total words in a document. For example, Tetlock (2007) and Feldman, Govindaraj, Livnat, and Segal (2010) hypothesize a linear relation between returns and proportion of positive and negative words. Li (2006) finds 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 provide more useful document scores. Manning and Schütze (1999) describe 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. The intuition behind this weighting scheme is that, ceteris paribus, 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 weighting scheme described by Manning and Schütze (1999) assigns the following weight for word j:

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

where N is the total number of documents in the sample and df_j is the number of documents in which word j occurs at least once. The superscript *idf* denotes weights that are 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 *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 that the frequency of occurrence of a word in documents should be related to market's perception of its impact.

We propose an approach that assigns weights for each word based on market reactions to documents containing those words. 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 satisfies the above properties:

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

where w_j is the weight for word j and $F_{i,j}$ is the number of occurrences of word j in document i. The term $1/a_i$ 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 ith document is released.

While we can directly compute $F_{i,j}$ and a_i, we have to estimate the weights associated with each word. To do so, we fit the regression

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

In this regression, we treat B_j's as regression coefficients and the estimated values of these coefficients provide unbiased estimates of bw_j. We cannot separately estimate b and w_j at this stage because 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,

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

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

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

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

We obtain the estimate of $\hat{w}_j$ that we use in Eq. (8) using all 10-Ks except those filed in the same year in which document i is filed. 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 $b > 0$.

Because $\hat{w}_j$ 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 because the regression does not use $\hat{w}_j$ individually but uses the score that aggregates $\hat{w}_j$'s across all the negative or positive words that appear in the document. Our documents contain a large number of words of interest and therefore negative and positive measurement errors in individual word weights would offset each other, thereby attenuating any bias due to these errors. In Section 4.2 we evaluate the stability of document scores computed using weights estimated over sample periods of different lengths. We find that measurement errors are smaller when we use longer sample periods for term weight estimation. In any event, to the extent that residual measurement errors exist, the estimate of b in Eq. (8) would be biased toward zero, favoring the null hypothesis.

We can also adapt our approach to use naïve Bayes classifier instead of multivariate regressions to determine the relative impact of different words and compute document scores. For example, we can classify news associated with 10-K filings as good and bad based on the sign of contemporaneous abnormal returns and apply the naïve Bayes classifier.[5] However, a binary classification would not take advantage of the information in the magnitude of abnormal returns. Our regression-based approach uses both the sign and the magnitude of market reaction, and we thus avoid losing information entailed by binary classification. Moreover, a key assumption of the naïve Bayes classifier is that the occurrence of each word within a document is independent of other words. Our multivariate regression approach takes into account co-occurrences of different words within a document as well as the magnitude of returns to determine the term weight for each word.

3. Data

We obtain all 10-Ks filed from January 1995 through December 2010 from the SEC's Electronic Data Gathering, Analysis, and Retrieval (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 Wharton Research Data Services (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. 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 might not have negative connotations for financial firms. Therefore, we exclude all financial firms (Standard Industrial Classification code from 6000 through 6999).

The final sample contains 45,860 filings between 1995 and 2010 and 7,606 unique firms. Table 1 presents a summary of the sample that we use in our analysis. The mean market value is $3.09 billion and the book-to-market ratio has a mean value of 0.65.

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 to determine whether it is a word.[6] Common stop words are not included in the dictionary and proper nouns are removed prior to processing. Because the documents are

[5] Li (2010) uses the naïve Bayes approach to examine the information content of forward looking statements in the Management Discussion and Analysis (MD&A) sections of 10-Ks and 10-Qs. Li manually categorizes the tone and content of 30 thousand sentences to construct his training sample. In contrast, our approach uses contemporaneous market returns to measure the information conveyed by 10-Ks in their entirety.

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

Table 1
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*.

		Size ($bln)		*BM*		*Turnover*	
Year	Number of firms	Mean	Median	Mean	Median	Mean	Median
1995	1,429	2.03	0.34	0.738	0.714	6.364	6.354
1996	2,330	1.82	0.22	0.754	0.717	6.630	6.625
1997	3,607	1.65	0.20	0.715	0.676	6.783	6.677
1998	3,619	2.05	0.23	0.698	0.664	6.820	6.838
1999	3,337	2.79	0.23	0.568	0.659	6.876	6.872
2000	3,533	3.21	0.32	0.677	0.689	7.089	7.118
2001	3,066	3.20	0.33	0.764	0.667	6.957	6.998
2002	2,850	3.07	0.36	0.746	0.710	6.894	6.965
2003	2,629	2.70	0.33	0.828	0.706	6.916	7.030
2004	3,013	3.22	0.45	0.827	0.763	7.096	7.185
2005	2,940	3.46	0.50	0.386	0.673	7.131	7.215
2006	2,904	3.87	0.59	0.648	0.661	7.207	7.317
2007	2,845	4.38	0.65	0.338	0.649	7.309	7.422
2008	2,687	4.34	0.58	0.650	0.644	7.560	7.762
2009	2,326	3.54	0.50	0.721	0.618	7.389	7.534
2010	2,745	4.04	0.62	0.401	0.449	7.337	7.447
1995–2010	45.680	3.09	0.40	0.654	0.666	7.022	7.091

often in HTML format, we remove all encoded images, tables, exhibits, HTML languages and other non-text items from the documents.[7] We also remove the standard cover page, often the first page of the document with the filer's name and address. We do not count positive or negative words that are accompanied by a negator within a distance of three words.[8]

For most of our tests, we use the negative and positive word lists constructed by Loughran and McDonald (2011). The LM list contains 353 positive words and 2,337 negative words. In this list, different inflections of a word are counted as separate words. For example, the word *falsify* and its inflections *falsifies, falsified, falsifying, falsification*, and *falsifications* are all considered as separate words. Since we expect all these inflections to have the same strength, we group them together.[9] When we consider only the root words, the list reduces to 123 positive words and 718 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 a global lexicon that merge the LM list, the Harvard IV-4 Psychosociological Dictionary, and the top and bottom two hundred words from the word list developed by Bradley and Lang (1999).

4. 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. Next, 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.

4.1. Term weight estimates

We first estimate the term weights for each word in the LM list using historical data. Specifically, we are interested in quantifying the importance that the market attaches to each root word in the list at the time the information is revealed to the market. Because 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 investors who pay a fee through a qualified provider within ten minutes after filing and for free to other investors within the next one or two days after the filing date (see Griffin, 2003). Loughran and McDonald (2011) 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 define filing period return as the return within this four-day event window. We compute filing period abnormal return r_i as

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

where $ret_{i,t}$ and $ret_{vwi,t}$ are the returns on stock i and on the CRSP value-weighted index on date t. Fig. 1 plots the abnormal returns over the entire sample period from 1995 to 2010. The mean and median abnormal returns are −0.30% and −0.24%. The abnormal returns range from about −40% to over 40% during this period. Griffin (2003) finds that the volatility of returns in the filing period is significantly greater than the volatility during a ten-day

[7] Because some forms incorporate all text within tables, tables in which less than 10% of the characters are numeric are not removed.
[8] We use *not*, *no*, and *never* as negators.
[9] The word groupings we use are available at https://fnce.wharton.upenn.edu/profile/1661/.

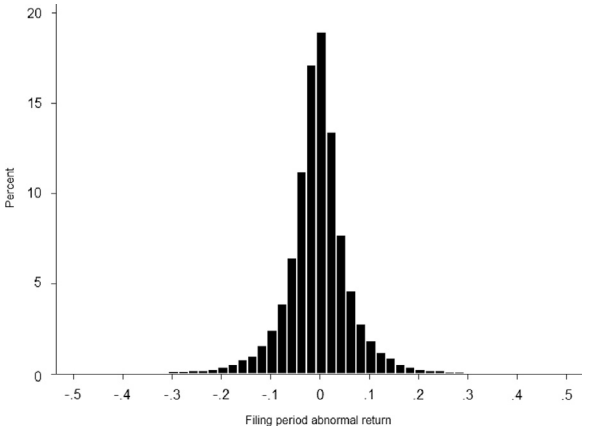

Fig. 1. Distribution of filing period abnormal returns. This figure plots that distribution of filing period abnormal return, defined as a firm's buy-and-hold return minus the CRSP value-weighted index return over the four-day window of (filing date, filing date + 3). Our sample contains 45,860 unique 10-Ks from 1995 to 2010.

pre- and post-announcement window and concludes that the market receives valuable information during this event window.

Fig. 2 presents the distribution of standardized weights for positive and negative words estimated using the entire 1995–2010 sample period. Because these are standardized weights, they are centered at zero. Sixteen negative and seven positive words have an absolute magnitude of weights greater than 2, and the figure presents their combined frequencies at the extreme ends. These words rarely appear in the sample of 10-Ks and our results are virtually identical when we winsorize their weights at +2 and −2.

4.2. Stability of document tone scores

Because we compute term weights using regression estimates, these weights contain measurement errors. Therefore, the document tone score that we compute (which we denote as $\widehat{Score}_i$) is also measured with error. Formally,

$$\widehat{Score}_i = Score_i + \text{Measurement Error}_i. \quad (10)$$

Because the measurement error is induced by the estimation errors in Eq. (6), it is a function of residual filing window returns during the estimation periods. Therefore, measurement errors in weights (and hence document scores) estimated in one period are uncorrelated with measurement errors in weights estimated over a nonoverlapping period. The true document scores, however, are uncorrelated with the measurement errors. Therefore, if a large fraction of the variation in $\widehat{Score}_i$ across documents arises because of variations in the true document score, then document scores based on weights computed in different nonoverlapping periods should be significantly correlated. But if variations in $\widehat{Score}_i$ across documents are largely due to measurement errors, then the correlation of scores based on weights computed in computed nonoverlapping periods would be small.

To examine document score stability, we partition the sample period into three five-year subperiods from 1996 to 2000, from 2001 to 2005, and from 2006 to 2010, and

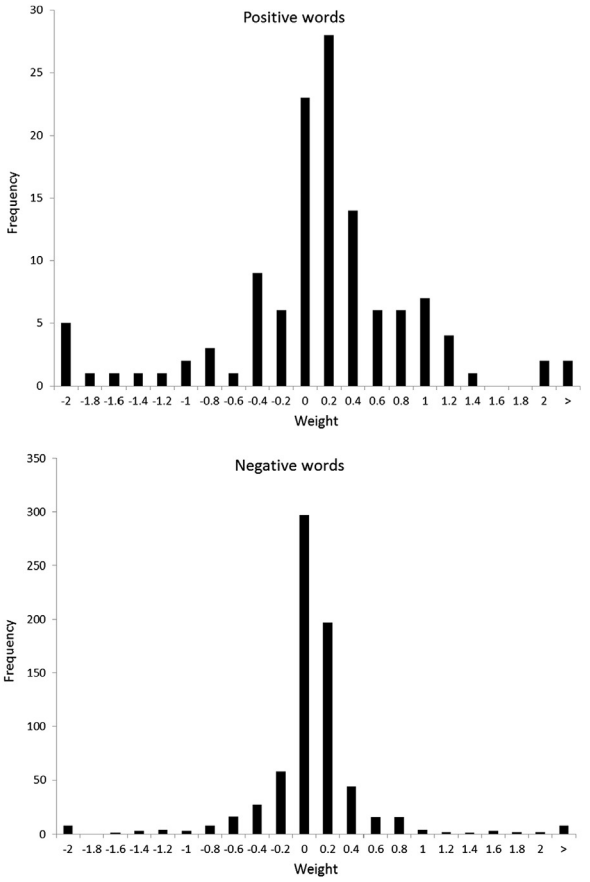

Fig. 2. Distribution of word weights. This figure presents the distribution of word power weights for positive and negative words. This figure plots the frequency distribution of weights based on Eqs. (6) and (7) fitted over the sample period of 1995–2010. Weights for negative and positive words are computed according to Eq. (7) for the entire sample period of 1995–2010. For ease of comparison, the weights are demeaned and divided by the standard deviation across the respective cross sections.

we estimate the weights during each subperiod.[10] We then compute three different tone scores for each 10-K using weights computed in each of these subperiods using negative words, and we repeat the same exercise for positive words as well.

Panel A of Table 2 presents the correlation between document scores computed using weights in one period with the corresponding score computed during a different sample period. For 10-Ks filed during the entire sample period, the rank correlation between negative word scores computed during this period and the scores computed during the next two subperiods are 0.461 and 0.466, respectively. The corresponding correlations for positive words are 0.560 and 0.604 respectively. The correlation between the document scores computed using weights estimated during 2001–2005 and 2006–2010 are 0.570 for negative words and 0.658 for positive words. All these correlations are statistically significant.

[10] We do not use 1995 data in Panel A of Table 2 so that the length of each subperiod is five years.

Table 2
Document score correlation and cross-tabulation of word weights and frequencies.

Panels A and B of this table present the correlation of document scores computed using weights estimated from Eqs. (6) and (7) fitted over five different sample subperiods. Panels C and D present 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 5 contains the quintile of words with the highest frequency. The word power (WP) weights in Panels C to E are computed using Eqs. (6) and (7) fitted over the sample period of 1995–2010. 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 E reports the rank correlation between WP and *idf* in both term weights and document scores.

Panel A: Document score correlation using weights calculated from three subperiods

Period	1996–2000	2001–2005	2006–2010
Negative words			
1996–2000	1.000		
2001–2005	0.461	1.000	
2006–2010	0.466	0.570	1.000
Positive words			
1996–2000	1.000		
2001–2005	0.560	1.000	
2006–2010	0.604	0.658	1.000

Panel B: Document score correlation using weights calculated from two subperiods

Period	1995–2002	2003–2010
Negative words		
1995–2002	1.000	
2003–2010	0.657	1.000
Positive words		
1995–2002	1.000	
2003–2010	0.734	1.000

Panel C: Cross-tabulation of word weights and frequencies, positive words

	Frequency Quintile (percent)					
Weight Quintile	1	2	3	4	5	Row total
1	48.00	28.00	16.00	8.00	0.00	25
2	12.00	4.00	4.00	40.00	40.00	25
3	0.00	0.00	8.33	37.50	54.17	24
4	4.00	24.00	52.00	16.00	4.00	25
5	37.50	45.83	16.67	0.00	0.00	24

Panel D: Cross-tabulation of word weights and frequencies, negative words

	Frequency Quintile (percent)					
Weight Quintile	1	2	3	4	5	Row total
1	50.69	30.56	15.28	2.78	0.69	144
2	9.03	13.89	23.61	30.56	22.92	144
3	0.00	6.99	16.78	29.37	46.85	143
4	4.17	14.58	23.61	30.56	27.08	144
5	36.36	34.27	20.28	6.99	2.10	143

Panel E: Correlation between WP and idf weights and document scores

Words	Word list	10-K
Negative words	−0.052	−0.089
Positive words	0.138	−0.341

Because the correlations are smaller than one, it is apparent that the scores contain measurement errors. To compute the scores more precisely, we increase the sample period over which we compute term weights. Specifically, we partition the sample period into two eight-year subperiods from 1995 to 2002 and from 2003 to 2010 and we estimate the weights during each subperiod.

Panel B of Table 2 presents the correlation between document scores computed during the two subperiods. The rank correlations between negative word scores computed

Table 3
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 5 contains the quintile of words with the highest frequency. This table reports the word power weights computed using Eqs. (6) and (7) fitted over the sample period of 1995–2010.

Frequency Quintiles				
1	2	3	4	5
Panel A: Top five most positive words				
ingenuity	influential	exceptional	adequately	favorable
acclaimed	optimistic	proficient	highest	strong
revolutionize	enthusiastic	transparency	progress	gain
courteous	excited	versatile	desirable	efficiency
incredible	regain	compliment	encouraged	opportunity
Panel B: Top five most negative words				
imperil	turbulent	disapprove	unplanned	unresolved
disavow	overestimate	reluctant	illegal	unsuccessful
insubordination	underinsured	uncontrollable	wasteful	discourage
bailout	aggravate	setback	misuse	unauthorized
dismal	unfortunate	turmoil	strain	insufficient

during the first and second subperiods are 0.657 for negative words and 0.734 for positive words. The larger correlations compared with Panel A indicate that the weights are estimated more precisely with longer sample periods.

These results indicate that we should use as long a sample period as possible to estimate term weights. However, we do not want to use the document for which we compute document scores to estimate term weights. Therefore, for each year T, we use term weights estimated during the entire sample periods excluding year T.

As a robustness check, we repeat all our tests using term weights estimated using only historical information and excluding all forward-looking data. Specifically, for each calendar year T we estimate term weights by fitting Eq. (6) using all filing period returns and 10-Ks in the sample period from 1995 to year T. For example, we fit the regression using 1995 data to estimate term weights for 1996, using 1995 through 1996 data to estimate the term weights for 1997, and so on. Our untabulated results using this procedure are qualitatively similar to those we report here.

4.3. Word power weights versus 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 Eq. (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}. Panels C and D of Table 2 present 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 words in Quintile 5 have the least negative impact.

The results in Table 2 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 E of Table 2 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 0.138 for the positive word list and −0.052 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 assigned 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 Eq. (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 Eq. (3).

The rank correlation between the word power scores and the *tf.idf* scores for the 10-Ks is −0.089 for the negative word score and −0.341 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 could be used to measure the tone of a document, how different words are weighted critically affects the measured tone of various documents.

Table 3 presents the top five most impactful positive and negative words within each document frequency quintile.

Table 4
Comparison of word power weights and *idf* term weights.

This table presents the top and bottom ten positive and negative words based on word power weights and their *idf* term weights. This table reports the words with the largest word power weights based on Eqs. (6) and (7) fitted over the sample period of 1995–2010. The *idf* term weighting scheme assigns weights inversely proportional to document frequency, as described in Eq. (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.

	Most impactful words			Least impactful words	
	WP rank	*idf* rank		WP rank	*idf* Rank
Panel A: Positive words					
ingenuity	1	14	lucrative	123	13
acclaimed	2	7	tremendous	122	35
influential	3	26	worthy	121	22
revolutionize	4	19	happy	120	9
optimistic	5	42	spectacular	119	21
enthusiasm	6	29	beautiful	118	15
excited	7	48	smooth	117	60
courteous	8	20	conducive	116	27
regain	9	39	receptive	115	30
incredible	10	3	proactive	114	38
Panel B: Negative words					
imperil	1	18	dispossess	718	8
disavow	2	22	ridicule	717	2
insubordination	3	20	mischief	716	27
bailout	4	31	derogatory	715	4
dismal	5	10	disorderly	714	3
untruthful	6	39	disassociate	713	35
unwelcome	7	5	immoral	712	23
turbulent	8	140	irreconciliable	711	19
vitiate	9	38	disgrace	710	1
undocumented	10	55	extenuating	709	34

Table 4 presents the list of ten most and least impactful positive and negative words and their rank based on w_j^{idf}.[11] 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 market's perception of word impact is an empirical issue that we address below.

4.4. Determinants of tone

The tone of 10-Ks would likely be affected by a number of firm-specific characteristics as 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 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 the month before the 10-K filing date.
- *BM*: Ratio of the book value of equity as of the fiscal year end in the 10-K.
- *Volatility*: 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 reflect a more 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 the management would likely be more cautious in setting investor expectations.

The next set of variables reflects recent events:

- *EADRet*: 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*: Computed as in Sloan (1996), 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.

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

Table 5
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 Eqs. (6) and (7) over the entire sample period except for the year when the current 10-K is filed. We compute the positive and negative tone for each 10-K using Eq. (4). *Size* is the natural logarithm of the market capitalization of equity at the end of the 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). We fit the annual regressions each year in the entire sample period of 1995–2010. A constant is also included in each regression. The coefficients are based on 13 annual Fama-MacBeth regressions. The estimates use a sample of 45,860 10-Ks over 1995–2010. All independent variables are standardized to a mean of 0 and standard deviation of 1.

Independent variables	Negative tone $Score_i$	Positive tone $Score_i$
Size	0.053	0.033
	(7.65)	(1.54)
BM	0.073	0.114
	(3.55)	(3.40)
Volatility	−0.497	−0.789
	(−5.78)	(−5.28)
Turnover	−0.055	−0.102
	(−3.14)	(−6.25)
EADRet	0.045	0.023
	(0.36)	(0.35)
Accruals	−0.085	−0.134
	(−1.57)	(−1.80)

The variable *EADRet* is the stock price response around earnings announcements, and it provides a measure of whether the earnings contained good news or bad news. Large accruals are generally considered bad news because they indicate an increase in working capital that could be due to bad business conditions or due to earnings manipulation. Most firms do not report balance sheet information necessary to compute accruals during their preliminary earnings announcements, and for these firms the market first receives information about accruals through 10-Ks.[12]

We examine the determinants of document tone using regression

$$Score_i = a + b \times Size_i + c \times BM_i + d \times Volatility_i + e \times Turnover_i + f \times EADRet_i + g \times Accruals_i + h \times Score_{i-1} + \epsilon_i, \quad (11)$$

where $Score_i$ is the word power score for the ith 10-K. 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 5 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 their 10-Ks contain a more negative or less positive tone.[13] Only negative score is significantly related to firm size. The lack of a significant relation between firm size and positive score indicates that, although smaller firms are more explicit about explaining possible negatives in their 10-Ks, 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 recent news as captured by *EADRet* or *Accruals*.

4.5. 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 on positive and negative scores, calculated with Eq. (4). Fig. 3 presents the average filing period returns for firms in various deciles. Decile 1 consists of firms with the highest scores and Decile 10 consists of firms with the lowest scores. For both positive and negative scores, the tone becomes more negative (or less positive) moving from Decile 1 to Decile 10.

For positive scores, the filing period returns decline monotonically from 1.84% for Decile 1 to −1.40% for Decile

[12] 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.

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

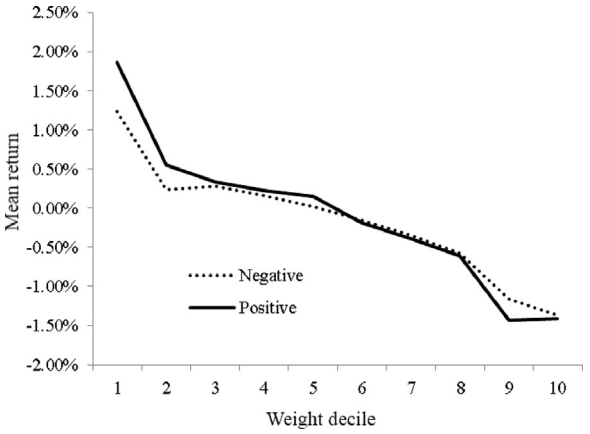

Fig. 3. Mean filing period abnormal return using word power weights. This figure presents the distribution of filing window abnormal returns, defined as a firm's buy-and-hold return minus the CRSP value-weighted index return over the four-day window of [filing date, filing date + 3] across various deciles of filings sorted based on the word power scores of the 10-Ks. We compute the word power weights for each year using Eqs. (6) and (7) over the sample period prior to the filing of 10-Ks. We compute the positive and negative tone for each 10-K using Eq. (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 covers 45,860 10-Ks over the 1995 to 2010 sample period.

10. The filing period returns progressively decline for negative scores as well, from 1.23% for Decile 1 to −1.37% for Decile 10. The negative returns for positive score Decile 10 firms and the positive returns for 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.

Fig. 3 also indicates that our methodology is able to quantify the tone of the document using only positive words. Previous studies do not find any relation between document scores that equally weight all words and stock returns. In contrast, when we use word power weights, we find a slightly 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 regression

$$r_i = a + b \times Score_i + \epsilon_i. \tag{12}$$

We fit the regression annually and estimate the coefficients and the standard errors using the Fama-MacBeth approach. Table 6 reports the regression coefficients. The slope coefficient is 0.402 for positive and 0.331 for negative words, which are reliably different from zero.

Table 6 also reports the coefficient estimates with tone measured using the *tf.idf* score as the independent variable. The coefficient is significantly negative when we use the positive word list. 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, the slope coefficient is significantly negative at the 10% level. 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 indicates that documents with more negative tone experience more negative returns during the filing period.

We next examine the relative power of our word power score and the *tf.idf* score in explaining filing period returns by fitting the regression[14]

$$r_i = a + b \times Score_i^{wp} + c \times Score_i^{tf.idf} + \epsilon_i. \tag{13}$$

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 Eq. (11), which is not surprising because the correlation between these scores is small.

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

$$r_i = a + \beta Score_i + b \times Size_i + c \times BM_i + d \times Volatility_i + e \times Turnover_i + f \times EADRet_i + g \times Accruals_i + \epsilon_i. \tag{14}$$

Table 6 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 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. 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 Eq. (14) than in Eq. (12). However, the differences are not statistically significant.[15] In addition, the coefficients in Eq. (14) are both statistically significant. When we use *tf.idf* score as the explanatory variable in Eq. (14) 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 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 lists, we compute the correlation between the scores based on each of these lists. The rank correlation between the two measures is 0.380, which is significantly positive.

[14] To avoid ambiguity, Eq. (13) adds the superscript *wp* to word power $Score_i$ defined in Eq. (4) and denotes this score as $Score_i^{wp}$.

[15] The average differences are −0.143 (−1.50) for negative list and −0.158 (−1.54) for positive list.

Table 6
Filing period abnormal return regressions.

This table reports the estimates of the regression of filing period abnormal return, defined as a firm's buy-and-hold return minus the CRSP value-weighted index return over the four-day window of [filing date, filing date+3] against document scores and various control variables. We compute the word power (WP) weights for each year using Eqs. (6) and (7) over the entire sample period except for the year when the current 10-K is filed, and we compute positive and negative WP scores for each 10-K using Eq. (4). The *idf* term weighting scheme assigns weights inversely proportional to document frequency, as described in Eq. (1). The document frequency of each word is the percentage of 10-Ks in which the word appears. See Table 5 for the definitions of the control variables. The coefficients are based on 15 annual Fama-MacBeth regressions. The estimates use a sample of 45,860 10-Ks over 1995–2010. 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	0.402		0.346	0.244		0.234
	(2.68)		(3.08)	(3.82)		(3.99)
.idf		−0.340	−0.202		−0.035	−0.086
		(−1.90)	(−1.33)		(−1.78)	(−1.09)
Control variables						
Size				−0.011	−0.020	−0.007
				(−0.13)	(−0.26)	(−0.09)
BM				2.262	2.766	2.215
				(1.26)	(1.00)	(1.22)
Volatility				−0.263	−0.322	−0.245
				(−1.57)	(−1.91)	(−1.56)
Turnover				−0.099	−0.118	−0.095
				(−1.34)	(−1.57)	(−1.33)
EADRet				0.570	0.543	0.568
				(5.67)	(6.36)	(5.65)
Accruals				−0.220	−0.315	−0.228
				(−1.49)	(−1.62)	(−1.56)

Panel B: Negative Words

	Models					
	(1)	(2)	(3)	(4)	(5)	(6)
Term weighting scheme						
WP	0.331		0.312	0.188		0.189
	(2.64)		(2.80)	(3.90)		(4.00)
idf		−0.314	−0.230		−0.099	−0.070
		(−1.92)	(−1.76)		(−1.47)	(−1.05)
Control variables						
Size				−0.018	−0.011	−0.011
				(−0.22)	(−0.09)	(−0.13)

Table 6 (*continued*)

Panel B: Negative Words

	Models					
	(1)	(2)	(3)	(4)	(5)	(6)
Term weighting scheme						
BM				2.611	3.018	2.662
				(1.45)	(1.60)	(1.47)
Volatility				−0.324	−0.365	−0.314
				(−1.80)	(−2.01)	(−1.79)
Turnover				−0.116	−0.133	−0.116
				(−1.57)	(−1.69)	(−1.58)
EADRet				0.576	0..573	0.574
				(5.78)	(5.71)	(5.76)
Accruals				−0.229	−0.230	−0.234
				(−1.58)	(−1.58)	(−1.62)

Panel C: Both Positive and Negative Scores

(Rank correlation of positive and negative scores=0.380)

	Models	
	(7)	(8)
Term weighting scheme		
WP (positive)	0.300	0.191
	(2.45)	(2.74)
WP (negative)	0.219	0.132
	(2.64)	(3.84)
Control variables		
Size		−0.018
		(−0.21)
BM		2.330
		(1.37)
Volatility		−0.238
		(−1.68)
Turnover		−0.109
		(−1.48)
EADRet		0.572
		(5.75)
Accruals		−0.225
		(−1.53)

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.

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

$$r_i = a + \beta \times Score_i^{positive} + \gamma \times Score_i^{negative} + \epsilon_i, \quad (15)$$

and

$$r_i = a + \beta \times Score_i^{positive} + \gamma \times Score_i^{negative} + b \times Size_i + c \times BM_i + d \times Volatility_i + e \times Turnover_i + f \times EADRet_i + g \times Accruals_i + \epsilon_i, \quad (16)$$

where the superscripts *positive* and *negative* on the scores indicate the particular word lists used to compute the scores. Panel C of Table 6 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 0.191 (2.74) and 0.132 (3.84) for positive and negative word lists, respectively. Therefore, each of these measures conveys incremental information relative to the other.

4.6. Combined lexicons

So far our approach uses positive and negative lexicons separately, but it uses market reactions to objectively determine term weights for the words in each of these lexicons. The compilation of separate lexicons is still subjective because it calls for the researchers' judgment about the connotation of each word. Some of the content analysis techniques, however, attempt to at least partly circumvent the need for researcher's subjective input to determine the connotation of each word. For example, in the discriminant-based classifier algorithm or the naïve Bayesian approach that Das and Chen (2007) evaluate, the researcher picks a lexicon and a sample of documents that are used to train the algorithm. The researcher first classifies the training documents as positive, neutral, or negative. The researcher then uses a discriminant function to determine the appropriate weights for each word in the lexicon so that the algorithm optimally differentiates the documents in the training sample. This algorithm is then applied to classify all documents into various tone categories.

The discriminant-based and the naïve Bayes classifier do not require the researcher to classify the connotation of each word, but they require that the researcher classify the tone of each document. In comparison, our approach relies on market reaction to measure both the sign and magnitude of the tone of each document and does not rely on the researcher to do so. We can also potentially allow the market reaction to determine the tone of each word, without relying on prior partitioning of the word list into positive and negative lexicons.

To do so, we create two lists of tonal words. The first list combines all the words in the positive and negative lexicons into one lexicon. To further remove subjectivity in the creation of the list, we create a second list that is a global list of tonal words. In particular, we combine the words from the following lexicons: Harvard IV-4 Psychosociological Dictionaries (both positive and negative), the LM list, and the top and bottom two hundred words from the word list compiled for the Affective Norms for English Words (ANEW) project by Bradley and Lang (1999), who score each word on the list by recording the standardized emotional responses to the word from subjects in an experimental setting. We perform a similar stemming procedure to collapse the list into the root word form. Because the Harvard lists and ANEW list do not include all inflections, we use a reverse version of the Porter stemming algorithm to mechanically inflect each term to get the complete set of inflections. The resulting list contains 2,068 unique root words.

We then fit Eq. (6) and compute the term weights for each word using Eq. (7). We now let market reactions determine both the magnitude and sign of the contribution of each word. We use these term weights and Eq. (4) to compute document scores. We repeat this procedure for each list.

Table 7 reports the estimates of Regressions (12) and (14) using Lougrhan and MacDonald combined-lexicon document scores in the first two columns. The slope coefficients for the score are significantly positive. Therefore, we can circumvent the need to subjectively assign positive or negative tonality to the words in the LM list.

The global tonal word list is more general. Moreover, the global list contains words that have neutral tonality in the financial context. For example, the global list contains words such as *tax*, *liability*, and *cost* from the Harvard negative words list but have a neutral connotation in a financial or accounting context as Loughran and McDonald (2011) point out. Therefore, the global list is a noisy lexicon that includes words that are extraneous in our context.

The last two columns of Table 7 report the regression estimates with the global list. Here again, we find that the slope coefficients on document score are positive and significant. Therefore, our approach removes much of the subjectivity inherent in compiling lexicons composed of words with positive or negative connotations. These results also illustrate that our methodology is robust to inclusion of extraneous words in the lexicon.

4.7. Completeness of word list

So far we have used a global list and the LM list to compute document scores. Although the underlying lists are carefully compiled, a possibility always exists that these lists do not include all English words with a negative or positive connotation. Because one cannot guarantee the completeness of any word list, it is important to examine whether the completeness of the word list critically influences the document tone score.

To examine the effect of omitting some of the relevant words, we construct partial lexicons. To do so, we separately sort the global list and the LM combined list, positive and negative word lists into quintiles based on document frequency. We then randomly remove 50% of the words from each quintile. We estimate Eq. (14) using these incomplete word lists.

The results in Table 8 show that the slope coefficients on word power scores based on all four incomplete lexicons are

Table 7
Filing period abnormal return regressions: global lexicons.

This table reports the estimates of the regression of filing period abnormal return, defined as a firm's buy-and-hold return minus the CRSP value-weighted index return over the four-day window of [filing date, filing date+3] against document scores computed using the combined Loughran and McDonald (LM) lexicon of both positive and negative words, scores computed with the global lexicon of all possible tonal words in the English language, and various control variables. We compute the word power (WP) weights for each year using Eqs. (6) and (7) over the entire sample period except for the year when the current 10-K is filed and compute positive and negative WP scores for each 10-K using Eq. (4). The document frequency of each word is the percentage of 10-Ks in which the word appears. See Table 5 for the definitions of the control variables. The coefficients are based on 15 annual Fama-MacBeth regressions. The estimates use a sample of 45,860 10-Ks over 1995–2010. All independent variables are standardized to a mean of 0 and standard deviation of 1.

	Models			
	Combined LM lexicon		Global lexicon	
	(1)	(2)	(3)	(4)
Term weighting scheme				
WP	0.343	0.192	0.294	0.190
	(2.67)	(3.81)	(2.44)	(3.58)
Control variables				
Size		−0.018		−0.019
		(−0.21)		(−0.14)
BM		2.631		2.461
		(1.45)		(1.00)
Volatility		−0.312		−0.334
		(−1.75)		(−1.84)
Turnover		−0.117		−0.123
		(−1.56)		(−1.62)
EADRet		0.575		0.546
		(5.80)		(6.38)
Accruals		−0.312		−0.312
		(−1.75)		(−1.62)

significant. The differences between the slope coefficients for the complete lexicons and the corresponding incomplete lexicons 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.

5. 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 regression

$$r_{i,(t+5,t+T)} = a + \beta \times Score_i^{wp} + \epsilon_i, \quad (17)$$

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 9 reports the regression estimates using the Fama-MacBeth approach. For positive words, the slope coefficient is 0.132 for the one-week window and 0.200 for the two-week window, which are significant at the 5% and 10% level, respectively. However, the slope coefficient is insignificant for the one-month window. We find similar results for negative words as well.[16]

To further examine the robustness of these results, we use size-adjusted returns as the dependent variable in place of market-adjusted returns in Eq. (17). 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 9 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 for positive words but insignificant for the one-month event window. For negative words, the point estimates of slope coefficients with size-adjusted returns are about the same as those with market-adjusted returns, although the two-week coefficient is now not statistically significant. The finding here indicates that the market does not fully respond to the tone of 10-Ks during the filing period. In contrast, in untabulated results, we did not find any evidence of market underreaction to document scores

[16] In comparison, Li (2006) finds that the market is much slower to respond to changes in risk sentiment expressed in 10-Ks with the use of words *risk* and *uncertainty* and their variations.

Table 8
Filing period abnormal return regressions on other word lists.

Positive_Omit is the Lougnran and Mcdonald (LM) positive word list with 50% of the words randomly removed from each frequency quintile. Negative_Omit is the LM negative word list with 50% of the words randomly removed from each frequency quintile. Combined_Omit is the LM combined word list with 50% of the words randomly removed from each frequency quintile. Global_Omit is the global tonal word list constructed in Section 5 with 50% of the words randomly removed from each frequency quintile. The dependent variable in each regression is the filing period abnormal return, defined as a firm's buy-and-hold return minus the CRSP value-weighted index return over the four-day window of [filing date, filing date+3]. We compute the word power (WP) weights for each year using Eqs. (6) and (7) over the entire sample period except for the year when the current 10-K is filed, and we compute positive and negative WP scores for each 10-K using Eq. (4). See Table 5 for the definitions of the control variables. The coefficients are based on 15 annual Fama-MacBeth regressions. The estimates use a sample of 45,860 10-Ks over 1995–2010. All independent variables are standardized to a mean of 0 and standard deviation of 1.

	Models			
	Positive_Omit	Negative_Omit	Combined_Omit	Global_Omit
Panel A: Additional word lists				
Term weighting scheme				
WP	0.224	0.230	0.265	0.276
	(2.51)	(2.95)	(2.98)	(2.52)
Control variables				
Size	−0.013	−0.010	−0.011	−0.086
	(−0.21)	(−0.23)	(−0.24)	(−0.48)
BM	2.598	2.111	2.246	2.300
	(0.38)	(0.76)	(0.74)	(0.53)
Volatility	−0.306	−0.290	−0.360	−0.329
	(−1.60)	(−1.91)	(−1.85)	(−1.95)
Turnover	−0.094	−0.113	−0.077	−0.092
	(−1.62)	(−1.46)	(−1.40)	(−1.33)
EADRet	0.573	0.565	0.526	0.574
	(6.35)	(5.98)	(5.74)	(6.32)
Accruals	−0.311	−0.280	−0.311	−0.295
	(−1.62)	(−1.71)	(−1.47)	(−1.70)
Panel B: Differences in slope coefficients				
ΔWP	0.018	−0.054	−0.044	−0.070
	(0.60)	(−0.45)	(−1.23)	(−0.36)

for 10-Ks computed using inverse document frequency weights. These results further reinforce the importance of accurately measuring the tone for fully understanding the timeliness of market's reaction to document tone.

6. Tone of IPO prospectus and underpricing

This section examines whether the term weights that we determine with the 10-Ks are generalizable to another context. Specifically, we examine the information content of IPO prospectuses as quantified by the tone scores that we compute using word power weights based on 10-Ks. Our tests examine whether the document scores of prospectuses predict IPO underpricing.

A number of papers in the literature offer explanations for the IPO underpricing phenomenon. For example, Tinic (1988) argues that IPO underpricing deters lawsuits against issuers and underwriters. Rock (1986) argues that some level of participation by uninformed investors is essential for the success of IPOs, but these investors are at a disadvantage relative to informed investors in differentiating between good and bad IPOs. Therefore, informed investors participate only in good IPOs and uninformed investors get a disproportionally large allocation of bad IPOs. In Rock's model underwriters underprice IPOs to offset the losses the uninformed investors suffer because of this winner's curse.

A common theme in these models and several others in the literature is that IPO underpricing is correlated with downside risk and the potential for negative outcomes. The tone of the IPO prospectuses can qualitatively convey the likelihood of negative outcomes. Therefore, under various IPO underpricing hypotheses we expect a negative relation between the WP score of the IPO prospectuses and underpricing.

We use the following regression model to examine the relation between IPO underpricing and document score:

$$r_i^{IPO} = a + b \times Score_i^{IPO} + c \times IPOSize_i + d \times Volatility_i + e \times Industry\ Dummy_i + f \times Year\ Dummy_i + u_i, \quad (18)$$

where

$$r_i^{IPO} = \frac{\text{First day closing price} - \text{IPO price}}{\text{IPO price}} - r^{MKT}, \quad (19)$$

Table 9
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 (WP) score calculated using lists of positive and negative words. We compute the word power weights for each year using Eqs. (6) and (7) over the sample period prior to the filing of 10-Ks, and we compute positive and negative WP scores for each 10-K using Eq. (4). The estimates use a sample of 45,860 10-Ks over 1995–2010. 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.

	Event windows		
Dependent variable	+5 to +9	+5 to +14	+5 to +26
Panel A: Positive words			
Market-adjusted returns	0.132	0.200	0.228
	(2.06)	(1.81)	(0.07)
Size-adjusted returns	0.093	0.123	0.130
	(1.98)	(1.80)	(0.25)
Panel B: Negative words			
Market-adjusted returns	0.101	0.132	0.191
	(1.93)	(1.51)	(0.83)
Size-adjusted returns	0.111	0.127	0.144
	(1.90)	(1.44)	(0.45)

r^{MKT} = CRSP value-weighted returns on the IPO date, (20)

and

$Score_i^{IPO}$ = WP score of the IPO prospectus. (21)

We compute the WP score of IPO prospectuses using the term weights that we computed with all available 10-Ks. We use *IPOSize*, *Volatility*, *IndustryDummy*, and *YearDummy* as control variables. We compute *IPOSize* as the number of shares offered on the IPO multiplied by the end-of-day price on the first day, and *Volatility* is the standard deviation of post-IPO returns from day 2 to day 30. We assign *IndustryDummy* based in the first two digits of the SIC code and *YearDummy* based on the calendar year of the IPO.

Our sample includes all IPOs during the 1995–2010 period for which we had an IPO prospectus on EDGAR, an IPO price from Thomson Reuters' Security Data Company (SDC) database, and first-day closing price and all necessary data to compute control variables on CRSP. To remain consistent with our weighting approach, we exclude financial firms from the sample.[17] There are a total of 1,475 IPOs in our sample. The standard deviation of first-day returns is 49.47%. We fit the regressions every year and compute the coefficients and standard errors using the Fama-MacBeth approach.

Table 10 reports the regression results. The slope coefficients are −2.834 for the negative lexicon, −4.898 for positive lexicon, and −3.305 for the combined lexicon, with all control variables. All these coefficients are significantly negative. We find similar results with only industry and year controls, and also when we include WP scores based on both positive and negative lexicons in the same regression.[18] Our results support the hypothesis that the potential for downside risk is positively related to IPO underpricing. Our results also indicate that the term weights that we determine using 10-Ks are useful in quantifying the tone of IPO prospectuses.

7. 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 is related to 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 factors such as earnings announcement date returns, accruals, and volatility.

We also apply our approach to determine term weights using a combined lexicon that merges the positive and negative word lexicons, and also using a global lexicon of tonal words. We extract term weights from market reactions when we use these combined lexicons and we circumvent the need for subjective judgment to partition selected words into positive or negative lexicons. We find that the document

[17] We identify 232 financial IPOs during this period. Including financial IPOs in the sample does not change our results.

[18] In a related study, Hanley and Hoberg (2010) compute standard and informative content scores for IPO prospectuses of each firm based on their similarities with the prospectuses of contemporaries from the same industry. They find that a 1 standard deviation increase in standard content increases IPO underpricing by 4% and a 1 standard deviation increase in the informative content decreases IPO underpricing by 8%.

Table 10
Prospectus tone and IPO underpricing.

This table reports the estimates of the regression of IPO first-day abnormal return, defined in Eq. (18) against document scores computed using the Loughran and Mcdonald (LM) positive and negative lists, as well as the combined LM lexicon of both positive and negative words. We compute the word power (WP) weights for each year using Eqs. (6) and (7) over the entire sample period and over all 10-Ks, and we compute positive and negative WP scores for each IPO prospectus using Eq. (4). *IPOSize* is defined as the number of shares offered on the IPO multiplied by the end-of-day price on the first day. *Volatility* is defined as the standard deviation of post-IPO returns from day 2 to day 30. Industry dummies based on firms' two-digit SIC codes and year dummies based on the IPO year are included in the regressions. The coefficients are based on 15 annual Fama-MacBeth regressions. The estimates use a sample of 1,475 IPOs over 1995–2010. All independent variables are standardized to a mean of 0 and standard deviation of 1.

	Models							
Independent variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
WP score negative	−4.391			−2.834			−3.356	−2.161
	(−3.62)			(−2.30)			(−2.74)	(−1.73)
WPscore positive		−5.984			−4.898		−5.231	−4.513
		(−4.69)			(−3.65)		(−4.01)	(−3.32)
WP score combined			−4.854			−3.305		
			(−3.96)			(−2.64)		
Volatility				2.317	1.543	2.231		1.433
				(1.63)	(1.09)	(1.57)		(1.05)
IPOSize				−4.811	−4.519	−4.732		−4.201
				(−2.72)	(−2.50)	(−2.66)		(−2.22)
Industry dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

tones based on the combined word lists are as informative as that based on separate positive and negative word lists. This finding indicates that our approach can be extended to minimize the level of subjectivity required for content analysis.

Content analysis should allow for the possibility that any underlying lexicon might not be entirely accurate or comprehensive. The results of our tests indicate that our approach reliably quantifies document tone even if half the words are excluded from the lexicon and if we use extraneous words in the lexicon. Therefore, our methodology extracts useful information even when 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.

We also explore the generalizability of our approach to a different context. Specifically, we examine the relation between scores of IPO prospectuses that we compute using the term weights previously determined using 10-Ks and IPO underpricing. We find a negative relation between tone scores and IPO underpricing as predicted by a number of models in the literature. This finding illustrates that the term weights we determine in one context can be reliably used in a different context.

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

References

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

Bradley, M., Lang, P., 1999. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. Technical Report C-1, Center for Research in Psychophysiology, University of Florida, Gainesville, FL.

Chen, S., Defond, M.L., Park, C.W., 2002. Voluntary disclosure of balance sheet information in quarterly earnings announcements. Journal of Accounting and Economics 33, 229–251.

Das, S., Chen, M., 2007. Yahoo! for Amazon: opinion extraction from small talk on the web. Management Science 53, 1375–1388.

Feldman, R., Govindaraj, S., Livnat, J., Segal, B., 2010. The incremental information content of tone change in management discussion and analysis. Review of Accounting Studies 15, 915–953.

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

Hanley, K.W., Hoberg, G., 2010. The information content of IPO prospectuses. Review of Financial Studies 23, 2821–2864.

Li, F., 2006. Do stock market investors understand the risk sentiment of corporate annual reports? Unpublished working paper. University of Michigan, Ann Arbor, MI.

Li, F., 2010. The information content of forward-looking statements in corporate filings—a naïve Bayesian machine learning approach. Journal of Accounting Research 48, 1049–1102.

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

Manning, C.D., Schütze, H., 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

Rock, K., 1986. Why new issues are underpriced. Journal of Financial Economics 15, 187–212.

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

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

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

Tinic, S.M., 1988. Anatomy of initial public offerings of common stock. Journal of Finance 43, 789–822.