The Use of Word Lists in Textual Analysis

Forthcoming in the *Journal of Behavioral Finance*

Tim Loughran
Mendoza College of Business
University of Notre Dame
Notre Dame, IN 46556-5646
574.631.8432 *voice*
Loughran.9@nd.edu

Bill McDonald
Mendoza College of Business
University of Notre Dame
Notre Dame, IN 46556-5646
574.631.5137 *voice*
mcdonald.1@nd.edu

February 1, 2015

Abstract
A commonly-used platform to assess the tone of business documents in the extant accounting and finance literature is Diction. We argue that Diction is inappropriate for gauging the tone of financial disclosures. About 83% of the Diction optimistic words and 70% of the Diction pessimistic words appearing in a large 10-K sample are likely misclassified. Frequently occurring Diction optimistic words like *respect*, *security*, *power*, and *authority* will not be considered positive by readers of business documents. Similarly, over 45% of the Diction pessimistic 10-K word-counts are *not* and *no*. The Loughran-McDonald (2011) dictionary appears better at capturing tone in business text than Diction.

Key words: Diction; word lists; sentiment analysis; Form 10-Ks; textual analysis.

JEL Classifications: C18, G14, and M41.

We thank Robert Battalio, Brad Badertscher, Peter Easton, Stephannie Larocque, and several anonymous referees for helpful comments. This paper won the 2014 Hillcrest Behavioral Finance Award.

Introduction

Textual analysis, the examination of document content, is becoming more common in accounting and finance research. In many applications, when researchers attempt to measure the tone of an annual report, earnings press release, or an IPO prospectus, the first step is selecting a dictionary.[1] The dictionary assigns words into positive, negative, or other sentiment categories (e.g., uncertainty, litigious, or weak modal). Once the total count of positive or negative words is tabulated, it is scaled, typically by the total number of words in the document. Documents with a relatively high frequency of positive words are considered optimistic and likewise those with a relatively high percentage of negative words are labeled pessimistic. In some cases, the scaled net count of positive minus negative is used collectively as a measure of tone.

One commonly used platform for assessing tone in the accounting and finance literature is Diction (www.dictionsoftware.com), which provides a program to parse documents into words, and then tabulates the words into functional categories. Researchers have used Diction to measure tone in earnings press releases (Rogers et al. (2011), Davis et al. (2012), and Demers and Vega (2014)), Form 10-Ks (Yuthas et al. (2002) and Li (2010)), restatement announcements (Durnev and Mangen (2011)), IPO prospectuses (Ferris et al. (2013)), and earnings conference calls (Davis et al. (2014)). It is important to note that Diction does not have specific positive and negative words lists. Instead, to quantify the optimism of a document, prior papers like Rogers et al. (2011) have tabulated the words in three Diction subgroups (praise, satisfaction, and inspiration), while for pessimistic tone, words in the three Diction subgroups of blame, hardship, and denial are targeted.

[1] In some fields, the distinction between a dictionary and a word list is important. We, however, will use the terms interchangeably.

Although widely-used in the literature, Diction's optimistic and pessimistic word lists were not specifically created to analyze financial documents. Indeed, many of the earliest users of the Diction word lists focused on political, not business, discourse (see Hart (1984a, 1984b) and Hart and Jarvis (1997)). The meaning of words a politician uses in a speech can be quite different from the meaning intended by a business executive. In this paper, we argue that Diction's classification dictionaries are inappropriate to gauge tone in important financial documents distributed to investors. Many of Diction's optimistic words proxy for specific industries and would not be considered optimistic by investors reading the business document.

For example, *respect, power,* and *trust* are all in the top ten of frequently used Diction optimism words in a sample of 77,158 annual reports (i.e., Form 10-Ks) during the 1994-2012 time period. *Respect* is not used in the classic sense of high opinion or deference to someone, but "with respect to loss contingencies" or "with respect to iOS devices" (examples from Apple's October 31, 2012 10-K filing). *Power* is commonly used by energy companies to describe a source of energy (i.e., nuclear power or electric power generation) while *trust* is frequently used by banks to note specific services or financial instruments (i.e., trust services, Trust Preferred Securities, and Executive Benefits Rabbi Trust).

Diction's pessimistic word list is also problematic. Over 45% of all the Diction pessimistic words used in a large 10-K sample are the words *not* and *no*. *Not* and *no* are common stop words in the context of business text; they do not generally indicate negative tone. Other frequently occurring Diction pessimistic words include *without, gross* (i.e., gross margin, gross interest), *none,* and *neither*. In a political speech, the Diction pessimistic word, *lynch,* has a very strong, negative connotation. However in business documents, the word merely denotes the name of a prominent investment banker, Merrill Lynch. We report that a large fraction of Diction's optimistic and pessimistic words are misclassified.

Unlike Diction, Loughran and McDonald (2011) created positive and negative word dictionaries with financial text specifically in mind. They examined all words appearing in at least 5% of the entire 10-K universe and placed words into a particular list if one could reasonably expect that the majority of the time the word would be used in a given context. There is a disappointingly low overlap between the Diction and Loughran and McDonald (2011) word lists. For optimistic words, 83% of the frequency counts of Diction words do not appear in the Loughran and McDonald (hereafter LM) positive word list. For pessimistic words, 70% of the Diction word frequencies do not overlap with the LM list.

The correlation between the Diction optimism and LM positive word percentages is actually slightly negative (-0.023). Thus, the higher the percentage of Diction optimism words in a 10-K, the lower is the percent of LM positive words. This is strong evidence that one of the dictionaries captures something besides optimistic tone by company managers. Partly due to frequently occurring negative words like *loss, losses, adverse,* and *failure,* which appear on both word lists, the correlation between the Diction pessimistic and LM negative words is fairly high (0.688).

To compare the two sets of dictionaries, we report regression results with stock return volatility in the year after the 10-K filing as the dependent variable. Included in the regressions are six control variables that all have significant explanatory power. One should expect that companies with a negative tone (i.e., a high proportion of pessimistic words) would have higher subsequent volatility following the 10-K filing.

In the regressions, the coefficients on both *% Diction optimism* and *% Diction pessimism* are positive and statistically significant. The sign on the *% Diction optimism* coefficient is somewhat surprising given the strong previous evidence that investors discount positive language (see Tetlock (2007) and Loughran and McDonald (2013)). The coefficient implies that the more frequent the occurrence of optimistic words, the higher is the firm's subsequent volatility. For the Loughran and McDonald word lists, the percent of positive words has an insignificant coefficient while the coefficient on percent negative is positive and statistically significant.

Instead of utilizing the optimism and pessimism word frequencies in isolation, researchers often examine the document's overall tone (i.e., optimism percentage minus pessimism percentage). When *Diction tone* is an explanatory variable, its coefficient is negative and significant only at the 7% level with subsequent stock return volatility as the dependent variable. In contrast, the coefficient on the *LM tone* variable is negative and significant at the 1% level. When the *Diction tone* and *LM tone* variables are included in the same regression, Diction's coefficient becomes insignificant while the coefficient on the *LM tone* variable has the expected negative sign and is statistically significant.

Arguably the current paper is simply a replication of Loughran and McDonald (2011), as applied to Diction. However, taken literally, their earlier paper simply shows that the Harvard General Inquirer word list is inappropriate in business applications. As we document in this paper, many researchers continue to use alternative sources, such as Diction, or factors built from a combination of sources. With this paper, we document that the Diction lists are inappropriate for business applications. As with the Harvard lists, many of the words driving the tabulated sentiment counts are either nonsensical or potentially industry flags (e.g., “death”, which occurs in both the Harvard and Diction negative lists is essentially a dummy variable for the pharmaceutical industry).

We do not attempt to replicate other studies in this paper for three reasons. First, we can show, definitively, by assessing the actual words driving the sentiment results, that the Diction dictionaries are miss-specified. Second, papers using factor combinations of the word lists, combined with unspecified parsing rules, are difficult to precisely replicate. Finally, because the word lists have the potential to increase both type I and type II errors, it is difficult to design a test that clearly shows the superiority of one list over the other. We will, however, attempt to provide a specific set of regression results that show the relative impact of using the alternative lists.

Our paper makes one primary contribution to the literature. Although other papers have expressed concern with the Diction optimism and pessimism word lists (see Li (2010) and Ferris et al. (2013)), we pinpoint the limitation of the Diction words. Using a sample of 77,158 10-Ks during 1994-2012, we report that about 83% of the Diction optimistic words and 70% of the Diction pessimistic words are incorrectly classified relative to the Loughran and McDonald (2011) lists. We explicitly itemize the dominant words driving the sentiment results, which reveals that many of the Diction words are clearly inappropriate as sentiment indicators in business writing. Thus, researchers should not use the widely-used Diction platform to gauge managerial tone in public documents, and, more generally, our paper underscores the importance of using dictionaries created specifically with business text in mind. The language managers use to describe their operations often has very different meaning than language used in a political speech, conversation, or a classic novel. Words like *gross*, *not*, *respect*, *power*, *lynch*, *needs*, *security*, and *pain* typically have neither a positive nor negative meaning in the context of business text and in some cases can simply serve as an unintended industry proxy.

Literature review

In the earlier stages of textual analysis research in accounting and finance, the decision of which dictionary to use was straightforward since only one predominate dictionary existed: Harvard University's General Inquirer (GI) IV-4 negative and positive word categories.[2] Like Diction, the Harvard's GI positive and negative word lists were not selected specifically for the language of business communication. Many of the early papers used Harvard's GI to gauge the tone of newspaper columns/articles or 10-Ks (see Tetlock (2007), Tetlock et al. (2008), Engelberg (2008), and Kothari et al. (2009)).

The first published paper of a custom finance-specific word list was Henry (2008). She created limited lists of positive and negative words to examine the tone of earnings press releases. For example, Henry's (2008) list had only 85 negative words (compared to 920 Diction pessimism and 2,329 LM negative words). Importantly, none of the most frequently occurring LM negative words in a 10-K sample (*loss*, *losses*, *claims*, *impairment*, *against*, *adverse*, *restated*, *adversely*, *restructuring*, and *litigation*) appear on Henry's list. Thus, the Henry (2008) word list is not an exclusive list of potentially negative words managers could use to describe current or future operations.

---

[2] Some early papers bypassed the challenges of dictionary selection by pursuing a machine learning approach (see Antweiler and Frank (2004), Das and Chen (2007), and Li (2010)).

Like Henry (2008), Loughran and McDonald (2011) created word lists specifically for business communications. Instead of having a relatively small list of positive and negative words, Loughran and McDonald created more comprehensive word lists: 354 positive and 2,329 negative words. In their paper, Loughran and McDonald (2011) document that about 75% of the Harvard Dictionary negative words are misclassified. Commonly appearing Harvard Dictionary negative words include *taxes, board, capital, liabilities,* and *mine*.[3]

Instead of using only one of the four dictionaries (Diction, Harvard, Henry, and LM), researchers often use several of the four or a combination of them to gauge tone in business communications. Rogers et al. (2011), for example, use the Diction, Henry, LM, and combination of all three in their analysis of the propensity of being sued. In his analysis of forward-looking statements in 10-K and 10-Q filings, Li (2010) used Diction and Harvard's GI lists to gauge text. Li (2010) concludes that either dictionary works well for corporate filings. To compare use of language in earnings press releases and Management's Discussion and Analysis (MD&A) in 10-K and 10-Q filings, Davis and Tama-Sweet (2012) utilize the Diction and LM word lists. Finally, Demers and Vega (2014), to gauge earnings announcements tone, use Diction, Harvard's GI, LM, and a factor model of all three measures.

Several papers have noticed particular problems in Diction's word lists. For example, Davis and Tama-Sweet (2012) note that the Diction optimism words *power* and *fair* have either a neutral meaning or might be correlated with specific industries. Ferris et al. (2013) mention that the Diction pessimism word *vice* (as in vice president) will not be viewed as a negative word by readers of the business communication. However, no paper

[3] Note that none of these Harvard negative words appear on Diction's pessimism word list.

has documented the very high misclassification problems existing with the Diction platform when it is applied to business communication.

Data

10-K sample creation

Table 1 shows how the available sample of 10-Ks on EDGAR is affected by our data filters. Initially, we download all 10-K, 10-K405, 10KSB, 10-KSB, and 10KSB40 filings (196,830 observations) from the SEC's EDGAR website (www.sec.gov) during 1994-2012. Amended filings are excluded from our analysis. The 10-K documents were parsed using the methods described in Loughran and McDonald (2014). The three data screens having the most impact on the sample are: requiring the firms to have a Center for Research in Security Prices (CRSP) PERMNO (dropping 103,325 observations); removing firms with a stock price less than or equal to $3 as of the day prior to the 10-K filing (dropping 14,381 observations); and requiring available daily stock returns prior to the 10-K filing date from CRSP (dropping 1,651 observations). Firms with low stock prices are removed to lessen the impact of micro-capitalization companies in our analysis. Our final sample is 77,158 firm-year observations during 1994-2012.

Following the literature, six of Diction's 35 various dictionaries are used to measure tone. In this paper, we use the Diction 7.0 word lists purchased from www.dictionsoftware.com. The variable *% Diction optimism* is the percentage of words in the 10-K that are in one of Diction's three optimism subgroups (praise, satisfaction, and inspiration). Of the 686 potential Diction optimism words, we drop 15 words because they contain hyphens or apostrophes. Examples of the dropped words include *first-rate, self-aware*, and *heaven's*. This leaves 671 words in the final Diction optimism list.

The variable *% Diction pessimism* is the percentage of words in the 10-K that are in Diction's pessimism subgroups of blame, hardship, and denial. Of the 920 possible Diction pessimism words, we remove 37 because they contain hyphens or apostrophes (i.e., *not-for-profit, won't,* and *small-minded*). Thus, our final count of Diction pessimism words is 883. The variables *% LM positive* and *% LM negative* are the percentage of words in the 10-K that are on Loughran and McDonald's positive and negative word lists, respectively. There are 354 positive words in the LM positive word list compared to 2,329 negative words.[4] The *LM tone* variable is defined as *% LM positive* minus *% LM negative*.

Summary statistics and time-series patterns for the word lists

Panel A of Table 2 reports the summary statistics of the paper's key variables. For the sample of 77,158 firm-year observations, the mean, 25th, 50th, 75th percentiles, and standard deviation are reported for each word list. The mean frequency of Diction optimism words (1.05%) is substantially higher than the average percentage of LM positive words (0.70%). The higher frequency of Diction optimism words is partly due to high occurrences of the Diction words *outstanding, respect, determined, necessary,* and *reasonable* which are not on the LM list.

For pessimistic words, the opposite is true. The mean percentage of Diction pessimism words is 1.33% versus 1.53% for the LM negative word list. Diction pessimism words have lower totals than the LM negative word frequencies due, in part, to different

[4] We use the August 2013 updated version of the Loughran and McDonald word list (available at ww.nd.edu/~mcdonald/Word_Lists.html).

dictionary sizes. Diction has 883 total pessimistic words while the LM negative word list is significantly larger (2,329).

Both Diction and LM tones for the 10-K sample are mostly negative. About 82% of the Diction tones are negative (i.e., % optimism < % pessimism) compared to 96% of the LM tones. The general negative 10-K tone likely reflects litigation concerns by managers. To potentially limit lawsuits, it is important to warn investors of all possible risks the company faces and to not oversell past/future operating performance.

Panel B of Table 2 reports the correlations between the word lists. The correlation of 0.156 between the Diction optimism and pessimism lists is somewhat surprising. One might have expected that optimistic and pessimistic word frequencies in a 10-K would have a slightly negative or close to zero correlation. For the LM positive and negative lists, the correlation is slightly less zero (-0.015).

The slightly negative correlation (-0.023) between Diction optimism and LM positive frequencies shows that the two lists are not capturing the same managerial sentiment. The negative correlation implies that the higher the percentage of Diction optimism words in a 10-K, the lower is the frequency of LM positive words. Although the two positive word lists have a negative relation, Diction pessimism and LM negative percentages have a strong positive correlation (0.688). This suggests that the Diction and LM negative words capture similar sentiment in business documents. Partly due to the strong link between the pessimistic word lists, the correlation between Diction tone and LM tone is positive and fairly strong (0.511).

Table 3 reports the median Diction and LM word frequencies and 10-K word counts by year. Diction optimism has a downward trend during the 1994 to 2012 time period. Form 10-Ks filed in 1994 have a median Diction optimism frequency of 1.11% compared to a median value of 0.71% in 2012. In contrast, the median LM positive word frequencies steadily rise during the sample period, 0.62% in 1994 to 0.89% in 2012. This dissimilar time-series pattern is further evidence that the two optimism lists are measuring different constructs.

The two pessimistic word lists both generally rise during 1994 to 2011. Interestingly, the LM negative percentages show a larger reaction to the 2008 financial crisis than Diction's pessimism words. Since the vast majority of 10-Ks are filed in February or March of each year, the 10-Ks filed in 2009 should reflect the deteriorating economic environment due to the financial meltdown of 2008's fourth quarter. Although both the Diction and LM pessimistic frequencies rise in 2009 relative to 2008, the median LM negative percentage rises more (6.1%) than the increase in the median Diction pessimism frequencies (2.9%).

As has been noted by others (see Li (2008)), there has been a sharp increase in the number of words in the typical 10-K since the 1990s. The median 10-K contains 19,886 words in 1996 compared to more than 43,000 by 2009.[5] In 2012, there was a massive jump to 92,242 in median 10-K word counts. Following a stated philosophy by the SEC in their Plain English initiative that document length is irrelevant, US firms are increasingly overloading investors and analysts with text. Three of the four word list frequencies experience a drop-off in 2012 partly due to the doubling of the median word counts. As firms increasingly place items like executive employment and operating lease contracts in their 10-Ks, specific word frequencies often decline as a result.

[5] The downward trend of words in a typical 10-K during the first three years of our sample is due to differing samples. Since electronic filings on EDGAR were not required until 1996, the early sample is tilted towards larger, more sophisticated companies. Larger firms tend to have longer 10-Ks.

Empirical findings

The most frequently occurring Diction optimism and pessimism words

Critical in assessing document word counts is an appreciation of Zipf's law, which is simply an observation that most non-pathological word list counts are well characterized by a power law distribution. That is, word counts are distributed much like the market capitalization of stocks, with a small minority accounting for the vast majority of total market size. In our specific application, words like *loss* will occur far more often in 10-Ks than words such as *unsavory*. Thus, it is critical in all applications of textual analysis to consider carefully a short list of top words driving the final counts.

In an attempt to see if misclassified Diction words account for the low correlations between the Diction and LM word lists, Table 4 reports the 25 most frequently occurring Diction optimistic words. Also reported in the table are the proportion, cumulative proportion, and whether the Diction word also appears in the Loughran-McDonald (2011) positive word list. Consistent with the power-law nature of these counts, the top 3.7% (25/671) of the words account for more than 70% of the counts

The obvious reason for the close to zero correlation between the Diction and LM positive word frequencies is the low overlap between the words driving the final counts. The two lists are quite different. For example, five words (*outstanding, respect, determined*, *necessary,* and *reasonable*) account for more than one-third of the total count of Diction optimistic words appearing in the 10-K sample. None of these five words appear on the LM positive word list. Of the 25 most commonly occurring Diction optimism words, only five (*greater, good, best, beneficial,* and *successful*) are also on the LM positive word list. This illustrates how dissimilar the two optimism word lists are.

In Table 4, the cumulative percentage of the 20 Diction words which are not on the LM positive word list is 63%. If we extend the tabulation to the entire Diction optimistic word list, we report that about 83% of the Diction optimistic frequencies are not on the Loughran and McDonald positive word list. Thus, since the LM positive word list was designed specifically for business text, the vast majority of the Diction optimism words appear to be misclassified.

The word *outstanding* accounts for 10.36% of the Diction optimism words occurring in the 10-K sample. In a large list of 671 Diction optimistic words, if one word accounts for more than 10% of the total optimism count, it is critical to ensure that the word is properly classified. At first glance, *outstanding* seems to be a well-specified optimistic word.[6] However, the word overwhelmingly refers to shares outstanding or options outstanding, not about outstanding firm performance.

As an example, Google's 10-K filed on February 11, 2011 uses the word *outstanding* 19 different times. None of Google's occurrences of *outstanding* would be interpreted as having optimistic tone by a reader. Google uses the word to merely note "Class A common stock outstanding;" "we had $3.0 billion of commercial paper outstanding;" "the actual number of RSUs outstanding;" or "the dilutive effect of outstanding stock options." The high frequency count of *outstanding* in the 10-K sample (more than 10% of all optimistic words) is suspicious since it is an extreme positive emotion word; managers will be hesitant to overuse the word in an optimistic context to

[6] The Loughran and McDonald (2011) paper classified *outstanding* as a positive word. In subsequent revisions of their word list, however, the authors removed it from their positive word dictionary.

lower litigation exposure. For example, Rogers et al. (2011) find that optimistic language in earnings announcements increases litigation risk. They also report that managerial sentences quoted by the plaintiffs in the class action complaint tend to be optimistic.

Other frequently occurring Diction optimism words like *respect, determined, necessary, reasonable, security, power,* and *trust* do not have positive sentiment when used in business text. *Respect* is not used to indicate deference; it is commonly used to imply "with respect to federal income taxes" or "with respect to its qualified defined benefit pension plans." *Respect*, like the word *outstanding*, has a substantial impact on the *% Diction optimism* value. The word accounts for more than 10% of all optimistic words appearing in the 10-K sample.

Necessary is mostly used as a requirement, not as a sign of optimism ("there can be no assurance that we will always obtain necessary renewals and approvals in the future"). In a political speech, *security* will certainly be seen as a positive word (President Reagan in a 1983 address: "The subject I want to discuss with you, peace and national security, is both timely and important"). However, in business text, the word commonly refers to a "financial security" or a "security deposit." *Power* and *trust* are commonly used by energy companies ("nuclear power" and "coal power") and banks ("trust services" and "trust preferred securities").

To provide insights into how well the Diction words capture the sentiment of business text, it is important to mention how infrequently some Diction words appear in 10-Ks. More than 100 of the 671 Diction optimism words occur less than 10 different times in the entire 10-K universe. About 28% of the words in the Diction optimism dictionary appear less than 100 times. As an example, the Diction words *chivalry, godliness,* and *tendernesses* appear a total of zero times in the 10-K sample universe. The sizable fraction of the Diction word list rarely appearing in business text highlights that the list was not designed with business communication in mind.

Many of Diction's pessimistic words are also misclassified. Table 5 reports the 25 most frequently occurring Diction pessimism words along with an indicator whether the word also appears on the LM negative word list. Unlike optimism words, there is more overlap between the Diction pessimism and LM negative word lists. Obviously negative words like *loss, losses, adverse, failure, bankruptcy,* and *fraud* appear in both the Diction and LM lists. The cumulative percentage of the 14 Diction words listed in Table 5 not on the LM negative word lists is 67%. If we extend the tabulation to the entire universe of Diction words, we report that 70% of the Diction pessimistic word frequencies are misclassified based on the LM categories.

The two most common Diction pessimism words, *not* and *no*, account for more than 45% of all the pessimistic words in the 10-K sample. These common stop words will not typically indicate negative tone by the manager's 10-K text. As an example, Google has three sentences in its mission statement (see its 10-K filed on March 16, 2006) that use *not*: "Our search results will be objective and we will not accept payment for inclusion or ranking in them;" "Advertisements should not be an annoying interruption;" and "We do not intend to compromise our user focus for short-term economic gain."

Besides *not* and *no*, other Diction words are clearly misclassified. The word *gross* accounts for 2.41% of the total Diction pessimism counts, yet it does not have a negative meaning in business communication (i.e., gross margin or gross interest). *None, nor,* and *cannot* are misclassified denial words which typically do not have pessimistic meaning in business text. Although not in the top 25, the Diction pessimism word *lynch* occurs in the 10-K universe a reasonable number of times. Used by a politician or a journalist, *lynch* is a powerfully negative word. However, when appearing in a 10-K, the word merely refers to one of the most prominent US investment bankers, Merrill Lynch.

Other Diction pessimism words merely proxy for specific industries like pharmaceutical, medical device, and insurance. The words, *death, disease,* and *pain* are all used fairly often in the 10-K universe, but the words' impact is limited to the word counts within the pharmaceutical industry. In our entire sample, the April 2, 2001 10-K filing by Pain Therapeutics has the third highest % Diction pessimism value (5.15%). Since Pain Therapeutics develops new generations of painkillers, their 10-K is full of discussions about alleviating patient pain (a positive, not negative, meaning).

In review, the evidence of Tables 4 and 5 underscores the weakness in using the Diction word lists for analyzing business text. About 83% of the Diction optimism and 70% of the Diction pessimism word frequencies are misclassified, based on the LM schemes, when applied to manager's text in 10-Ks. The poor overlap between the Diction and the LM dictionaries accounts for their surprisingly low correlations. This finding should discourage researchers from using Diction as a taxonomy for tone in business communications.

Explaining volatility with word lists

In the last two tables of the paper, we will analyze how well the dictionaries explain post 10-K filing stock return volatility, a measure used in previous papers (Tetlock (2007) and Loughran and McDonald (2011)) which has consistently been related to document tone. The Table 6 dependent variable, post-filing RMSE, is defined as the root mean square error (RMSE) from a market model estimated using trading days [1, 252] relative to the 10-K filing date. (A year is approximately 252 trading days.) To get a reasonable estimate, we require a minimum of 66 observations to be included in the analysis. This data requirement for the market model estimation drops an additional 1,227 firms from the sample.

In the regressions, six control variables are selected due to their ability to explain subsequent stock return volatility. The six control variables are: (1) natural logarithm of the firm's stock price as of one trading day prior to the 10-K filings date; (2) natural logarithm of the firm's market value (shares outstanding times stock price) as of one trading day prior to the 10-K filing date; (3) natural logarithm of one plus the two day cumulative abnormal stock returns around the 10-K filing date; (4) the alpha from a market model using trading days [-252, -1] which proxies for the firm's stock return performance in the year prior to the 10-K filing; (5) the root mean square error from a market model estimated using trading days [-252, -1] to control for the firm's stock return volatility in the year prior to the 10-K filing; and (6) dummy variable set to one if the firm is listed on Nasdaq at the time of the 10-K filing, else zero.

The variables are described in more detail in the Appendix. Besides the control variables, the regressions also include an intercept, Fama and French (1997) 48-industry dummies, and calendar year dummies. The *t*-statistics are in parentheses with the standard errors clustered by both year and industry. The sample size is 75,931 firm-year observations in all our regressions.

The first regression in Table 6 includes only the control variables. All of the independent variables are statistically significant at the 1% level. As should be expected, firms with lower stock prices, smaller market values, smaller filing date returns, higher stock return volatility in the prior year, worse pre-10-K filing performance, and a Nasdaq-listing have higher stock return volatility in the year after the 10-K filing. The R^2 value in the first regression is 56.37%.

When *% Diction optimism* is added in the second regression, the variable has a positive and statistically significant coefficient value (*t*-statistic 2.07). Thus, the higher is the frequency of optimism words in a firm's 10-K, the higher is the company's subsequent stock return volatility. This positive link between optimistic words and volatility is somewhat surprising given the prior empirical evidence that investors generally ignore or at least discount positive language in business text (see Tetlock (2007) and Engelberg (2008)). Investors might view overly optimistic 10-K language by managers as merely cheap talk. In the second regression, note that the coefficient signs and significant levels of the control variables remain the same when the *% Diction optimism* variable is added to the regression.

In column (3), *% Diction pessimism* has a positive and significant coefficient value (*t*-statistic of 2.95). This is exactly what should be expected; more pessimistic tone by managers foreshadows higher subsequent stock return volatility. That is, more discussion of losses, failure, and bankruptcy should be associated with increased subsequent volatility. When *% LM positive* is added as an independent variable in column (4), its coefficient value is insignificant. This is consistent with the notion that investors place little value in positive words used by managers to describe their operations.

In the last column of Table 6, *% LM negative* is added as a right-hand side variable. The coefficient value for *% LM negative* is positive and significant (*t*-statistic of 3.09). As with the Diction pessimism frequencies, a higher count of LM negative words is linked with significantly higher RMSE. Thus, although our prior results suggest substantial classification errors in the Diction word lists, the relatively high correlation between the Diction and LM negative percentages (0.688), seems to allow the Diction pessimism classification to provide a reasonable result.

Since a number of papers use overall tone (i.e., % optimistic - % pessimistic) instead of using the two components separately, we will follow the precedent. Table 7 reports the regression results of Diction and LM tone with subsequent stock return volatility as the dependent variable. In column (1), *Diction tone*, in the presence of the control variables, has a negative coefficient value which is only significant at the 7% level (*t*-statistic of -1.83). Hence, the more pessimistic is the overall tone (% optimism - % pessimism), the higher is the firm's subsequent volatility. Column (2) reports that in the presence of the six control variables, *LM tone* has a negative and statistically significant coefficient value (*t*-statistic of -4.06).

The last column of Table 7 runs a horserace between the two dictionaries. When both *Diction tone* and *LM tone* are included, along with the six control variables, in the same regression, only *LM tone* remains statistically significant. Although both tone measures have negative coefficients, the associated *t*-statistic for *Diction tone* is only -0.57 while the coefficient on *LM tone* has a *t*-statistic of -3.40. We interpret the Table 7 results as evidence that the Loughran and McDonald (2011) word lists are better than the Diction's dictionaries at measuring tone of business documents.

Conclusion

Diction is a commonly-used platform to gauge tone in the accounting and finance fields. We demonstrate that the Diction optimism and pessimism dictionaries are inappropriate for measuring tone of business communications. Analyzing text from a large sample of 77,158 10-Ks during 1994-2012, we pinpoint the limitation of the Diction word lists. The vast majority of the Diction optimistic and pessimistic words are not appropriate. We report that about 83% of the Diction optimism and 70% of the Diction pessimism word frequencies are misclassified, based on the LM classifications, when applied to manager's text in 10-Ks. Importantly, a substantial number of the targeted words occurring with the highest frequencies from the Diction word lists are patently misclassified.

Readers of 10-Ks will rarely consider *respect, necessary, security, power, trust, authority,* or *discretion* to be optimistic words when used by managers to describe present and future operations. Similarly, words like *not, gross, none, death, pain, vice,* and *lynch* will not be considered pessimistic when appearing in business communications written by firm managers. In contrast to Diction, Loughran and McDonald (2011) created positive and negative word lists with business communication specifically in mind.

We report that the correlation between Diction optimism and Loughran and McDonald (2011) positive frequencies is actually slightly negative (-0.023). This is strong evidence that the two word lists are not measuring the same managerial sentiment in the 10-Ks. The correlation between Diction pessimism and Loughran-McDonald negative word frequencies is reasonably high (0.688).

In regressions with subsequent volatility as the dependent variable, Diction tone is significant at the 7% level while the Loughran-McDonald tone coefficient is significant at the 1% level. When Diction and Loughran-McDonald tones are included in the same regression, only the LM tone is statistically significant in explaining subsequent firm stock return volatility. Even though our regression results suggest that in some cases the Diction dictionaries might provide noisy but useable proxies for the underlying sentiment, we also present evidence to suggest that the list could create false positives that are artifacts of industry related terminology (e.g., in optimism, *entertainment, health,* or *heal*; in pessimism, *abrasive, death, disease, pain,* or *sickness*.)

Our paper highlights a severe limitation in using Diction optimism and pessimism words: Diction's word lists are poorly suited for business communication. Given that the Diction words were not specifically designed for managerial communications, this finding should not be completely surprising. Emotional words used in a political speech (e.g., *lynch, security, power, pain,* and *respect*) will have very different meanings when used by managers describing their current/future operations. Researchers should only use dictionaries created specifically for business communication when gauging document tone. More generally, our results underscore the importance of carefully selecting a word list based on the context of its usage.

Appendix.

A.1 Definitions of the variables used in the paper

% Diction optimism	Percentage of words in the 10-K that are in one of Diction's three optimism subgroups (praise, satisfaction, and inspiration). Examples of Diction optimistic words include *outstanding, respect, determined,* and *necessary.* Of the 686 original Diction optimism words, we remove 15 words because they contain hyphens or apostrophes (i.e., *first-rate, self-aware,* and *heaven's*).
% Diction pessimism	Percentage of words in the 10-K that are in one of Diction's three pessimism subgroups (blame, hardship, and denial). Examples of Diction pessimistic words include *not, no, loss,* and *without.* Of the 920 original Diction pessimism words, we remove 37 words because they contain hyphens or apostrophes (i.e., *not-for-profit, won't,* and *small-minded*).
Diction tone	The % of Diction optimism words in the 10-K minus the % of Diction pessimism words in the 10-K.
% LM positive	Percentage of words in the 10-K that are on Loughran and McDonald's positive word list. Examples of positive words include *greater, effective, gain,* and *improvements* (see http://www.nd.edu/~mcdonald/word-lists).
% LM negative	Percentage of words in the 10-K that are on Loughran and McDonald's negative word list. Examples of negative words include *losses, impairment, adverse, bankruptcy,* and *closed.*
LM tone	The % of Loughran and McDonald positive words in the 10-K minus the % of Loughran and McDonald negative words in the 10-K.
Post-filing RMSE	The root mean square error from a market model estimated using trading days [1, 252] relative to the 10-K file date. We require a minimum of 66 observations to be included in the sample. This variable is multiplied by 100.
Words in 10-K	Total number of words contained in the 10-K.

Control Variables

Log(price)	The natural logarithm of the firm's stock price as of one day prior to the 10-K filing date.
Log(market value)	The natural logarithm of stock price multiplied by shares outstanding as of one day prior to the 10-K filing date.
Log(filing date abnormal returns)	The natural logarithm of one plus the buy-and-hold return starting on filing date (day 0) through day +1 minus the buy-and-hold return of the CRSP value-weighted index over the same two-day period.
Pre-filing alpha	The alpha from a market model using trading days [-252, -1]. We require a minimum of 66 observations of daily returns to be included in the sample. This variable is expressed as a percentage.
Pre-filing RMSE	The RMSE from a market model estimated using trading days [-252, -1]. We require a minimum of 66 daily returns to be included in the sample. This variable is multiplied by 100.
Nasdaq dummy	Dummy variable set to one if the firm is listed on Nasdaq at the time of the 10-K filing, else zero.

References

Antweiler, W. and Frank, M.Z. (2004). "Is all that talk just noise? The information content of internet stock message boards." *Journal of Finance 59*(3), pp. 1259-1294.

Das, S.R. and Chen, M.Y. (2007). "Yahoo! for Amazon: Sentiment extraction from small talk on the web." *Management Science 53*(9), pp. 1375-1388.

Davis, A. K., Ge, W., Matsumoto, D., and Zhang, J. L. (2014). "The effect of manager-specific optimism on the tone of earnings conference calls." Forthcoming *Review of Accounting Studies*.

Davis, A.K., Piger, J.M., and Sedor, L.M. (2012). "Beyond the numbers: Measuring the information content of earnings press release language." *Contemporary Accounting Research 29*(3), pp. 845-868.

Davis, A.K. and Tama-Sweet, I. (2012). "Managers' use of language across alternative disclosure outlets: Earnings press releases versus MD&A." *Contemporary Accounting Research 29*(3), pp. 804-837.

Demers, E. and Vega, C. (2014). "Understanding the role of managerial optimism and uncertainty in the price formation process: Evidence from the textual content of earnings announcements." Working paper, UVA Darden.

Durnev, A. and Mangen, C. (2011). "The real effects of disclosure tone: Evidence from restatements." Working paper, University of Iowa.

Engelberg, J. (2008). "Costly information processing: Evidence from earnings announcements." Working paper, University of North Carolina.

Fama, E.F. and French, K.R. (1997). "Industry costs of equity." *Journal of Financial Economics 43*(2), pp. 153-193.

Ferris, S.P., Hao, G., and Liao, M. (2013). "The effect of issuer conservatism on IPO pricing and performance." *Review of Finance 17*(3), pp. 993-1027.

Hart, R.P. (1984a). "Verbal style and the presidency: A computer-based analysis." New York: Academic Press.

Hart, R.P. (1984b). "Systematic analysis of political discourse: The development of diction." Political Communication Yearbook *1*, pp. 97-134.

Hart, R.P. and Jarvis, S.E. (1997). "Political debate forms, styles, and media." *American Behavioral Scientist 40*(8), pp. 1095-1122.

Henry, E. (2008). "Are investors influenced by how earnings press releases are written?" *Journal of Business Communication 45*(4), pp. 363-407.

Kothari, S.P., Li, X., and Short, J.E. (2009). “The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis.” *The Accounting Review 84*(5), pp. 1639-1670.

Li, F. (2008). “Annual report readability, current earnings, and earnings persistence.” *Journal of Accounting and Economics 45*(2), pp. 221-247.

Li, F. (2010). “The information content of forward-looking statements in corporate filings—A Naïve Bayesian machine learning approach.” *Journal of Accounting Research 48*(5), pp. 1049-1102.

Loughran, T. and McDonald, B. (2011). “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks.” *Journal of Finance 66*(1), pp. 35-65.

Loughran, T. and McDonald, B. (2013). “IPO first-day returns, offer price revisions, volatility, and form S-1 language.” *Journal of Financial Economics 109*(2), pp. 307-326.

Loughran, T. and McDonald, B. (2014). “Measuring Readability in Financial Disclosures.” *Journal of Finance 69*(4), pp. 1643-1671.

Rogers, J.L., Van Buskirk, A., and Zechman, S. (2011). “Disclosure tone and shareholder litigation.” *The Accounting Review 86*(6), pp. 2155-2183.

Tetlock, P.C. (2007). “Giving content to investor sentiment: The role of media in the stock market.” *Journal of Finance 62*(3), pp. 1139-1168.

Tetlock, P.C., Saar-Tsechansky, M., and Macskassy, S. (2008). “More than words: Quantifying language to measure firms' fundamentals.” *Journal of Finance 63*(3), pp. 1437-1467.

Yuthas, K., Rogers, R., and Dillard, J.F. (2002). “Communicative action and corporate annual reports.” *Journal of Business Ethics 41*(1-2), pp. 141-157.

Table 1

Sample creation

	Dropped	Sample Size
SEC 10-K files 1994–2012		196,830
Drop if missing CRSP PERMNO	103,325	93,505
Drop if number of 10-K words ≤ 2,000	81	93,424
Drop if less than 66 available returns before 10-K filing	1,651	91,773
Drop if missing 10-K filing date returns	188	91,585
Drop if missing CRSP market value	46	91,539
Drop if stock price ≤ $3	14,381	77,158
Final Sample		77,158

This table reports the effect of the data requirements on the initial 10-K sample available on the SEC's EDGAR website.

Table 2

Summary statistics

Panel A: Descriptive statistics

	Mean	25%	50%	75%	Std. Dev.
% Diction optimism	1.05	0.90	1.03	1.19	0.24
% Diction pessimism	1.33	1.15	1.33	1.51	0.29
% LM positive	0.70	0.57	0.67	0.80	0.19
% LM negative	1.53	1.23	1.51	1.81	0.46
Diction tone	-0.28	-0.49	-0.29	-0.08	0.34
LM tone	-0.83	-1.14	-0.83	-0.52	0.50

Panel B: Correlations

	% D-opt	*% D-pess*	*%LM-pos*	*%LM-neg*	*D-tone*
% Diction pessimism	0.156				
% LM positive	-0.023	-0.096			
% LM negative	0.065	0.688	-0.015		
Diction tone	0.567	-0.725	0.064	-0.528	
LM tone	-0.069	-0.670	0.387	-0.928	0.511

This table reports the summary statistics for the Diction optimism/pessimism and Loughran-McDonald (LM) positive/negative word lists. Diction tone is defined as % Diction optimism minus % Diction pessimism. LM tone is defined as % LM positive minus % LM negative. More detailed variable definitions are in the Appendix. The sample size for each variable is 77,158 firm-year observations.

Table 3
Median time series patterns of the Diction and Loughran-McDonald (LM) word lists

Year	*% Diction optimism*	*% Diction pessimism*	*% LM positive*	*% LM negative*	*Words in 10-K*
1994	1.11	1.16	0.62	1.27	23,746
1995	1.10	1.17	0.63	1.26	21,159
1996	1.08	1.21	0.62	1.30	19,886
1997	1.09	1.26	0.62	1.31	21,540
1998	1.08	1.29	0.63	1.34	22,605
1999	1.07	1.35	0.63	1.39	23,989
2000	1.07	1.33	0.66	1.40	23,468
2001	1.07	1.32	0.65	1.41	24,741
2002	1.05	1.35	0.65	1.50	27,389
2003	1.03	1.37	0.64	1.56	31,391
2004	1.04	1.37	0.66	1.60	32,905
2005	1.03	1.34	0.67	1.58	34,626
2006	1.02	1.36	0.68	1.63	35,677
2007	1.01	1.36	0.69	1.62	37,898
2008	1.00	1.38	0.69	1.64	39,293
2009	1.00	1.42	0.69	1.74	43,712
2010	1.00	1.41	0.68	1.78	43,171
2011	0.99	1.40	0.69	1.77	46,058
2012	0.71	1.16	0.89	1.37	92,242
Total	1.03	1.33	0.67	1.51	32,059

This table reports the time-series pattern of the median Diction/Loughran-McDonald word list percentages and median 10-K word count during 1994-2012.

Table 4
The top 25 most frequently occurring Diction optimistic words

Diction Word	Proportion	Cumulative proportion	LM Positive
OUTSTANDING	10.36%	10.36%	
RESPECT	10.24%	20.61%	
DETERMINED	4.93%	25.54%	
NECESSARY	4.75%	30.29%	
REASONABLE	4.30%	34.60%	
SECURITY	3.81%	38.40%	
POWER	3.17%	41.57%	
GROWTH	2.87%	44.44%	
KNOWLEDGE	2.61%	47.05%	
TRUST	2.26%	49.31%	
SECURED	1.98%	51.28%	
GREATER	1.91%	53.19%	✓
AUTHORITY	1.76%	54.95%	
DETERMINATION	1.65%	56.60%	
GOOD	1.58%	58.19%	✓
QUALITY	1.47%	59.66%	
BEST	1.41%	61.07%	✓
DISCRETION	1.40%	62.48%	
HEALTH	1.38%	63.86%	
BENEFICIAL	1.17%	65.03%	✓
CARE	1.16%	66.19%	
QUALIFIED	1.16%	67.36%	
RESPONSIBILITY	1.08%	68.43%	
COMMITMENT	1.02%	69.45%	
SUCCESSFUL	0.98%	70.43%	✓

This table reports the most frequently occurring Diction optimistic words in the 10-K sample during the 1994-2012 time period. The last column indicates if the Diction optimism word also appears on the Loughran and McDonald (2011) positive word list.

Table 5
The top 25 most frequently occurring Diction pessimistic words

Diction Word	Proportion	Cumulative Proportion	LM Negative
NOT	33.41%	33.41%	
NO	12.34%	45.75%	
LOSS	7.59%	53.34%	✓
LOSSES	5.07%	58.42%	✓
WITHOUT	4.85%	63.26%	
RISK	4.01%	67.28%	
GROSS	2.41%	69.68%	
RISKS	2.12%	71.81%	
ADVERSE	2.03%	73.84%	✓
FAILURE	1.78%	75.62%	✓
NONE	1.36%	76.98%	
NOR	1.30%	78.27%	
CANNOT	1.29%	79.56%	
NOTWITHSTANDING	0.93%	80.49%	
NEEDS	0.89%	81.37%	
BANKRUPTCY	0.72%	82.09%	✓
DEATH	0.70%	82.79%	
FRAUD	0.68%	83.47%	✓
NEGATIVE	0.64%	84.11%	✓
NEITHER	0.64%	84.74%	
WEAKNESSES	0.62%	85.36%	✓
UNTRUE	0.57%	85.93%	
COMPLAINT	0.55%	86.49%	✓
FAIL	0.50%	86.99%	✓
DEFICIT	0.49%	87.48%	✓

This table reports the 25 most frequently occurring Diction pessimistic words in the 10-K sample during the 1994-2012 time period. The last column indicates if the Diction pessimism word also appears on the Loughran and McDonald (2011) negative word list.

Table 6
Regressions for the Diction and Loughran-McDonald (2011) word lists with post-filing return volatility as the dependent variable

	(1)	(2)	(3)	(4)	(5)
% Diction optimism		0.107			
		(2.07)			
% Diction pessimism			0.254		
			(2.95)		
% LM positive				-0.040	
				(-0.49)	
% LM negative					0.136
					(3.09)
Log(price)	-0.238	-0.239	-0.235	-0.238	-0.225
	(-4.26)	(-4.27)	(-4.22)	(-4.26)	(-3.98)
Log(market value)	-0.066	-0.067	-0.064	-0.065	-0.073
	(-3.14)	(-3.19)	(-3.05)	(-3.09)	(-3.51)
Log(filing date abnormal returns)	-2.286	-2.287	-2.280	-2.286	-2.284
	(-6.03)	(-6.02)	(-6.07)	(-6.03)	(-6.05)
Pre-filing alpha	-0.987	-0.986	-0.980	-0.987	-0.979
	(-4.26)	(-4.25)	(-4.23)	(-4.25)	(-4.21)
Pre-filing RMSE	0.579	0.579	0.575	0.580	0.575
	(11.73)	(11.74)	(11.58)	(11.70)	(11.67)
Nasdaq dummy	0.296	0.296	0.290	0.296	0.293
	(3.42)	(3.42)	(3.43)	(3.47)	(3.44)
R-squared	56.37%	56.38%	56.45%	56.37%	56.43%

This table reports the coefficients and *t*-statistics for regressions of the Diction and Loughran-McDonald (2011) words lists with post-filing return volatility as the dependent variable. Included in the regressions but not tabulated are a constant, year dummies, and Fama-French (1997) industry dummies. Variable definitions are in the Appendix. The *t*-statistics are in parentheses with the standard errors clustered by both year and industry. Each regression includes 75,931 firm-year observations.

Table 7
Regressions for the Diction and Loughran-McDonald (2011) tone with post-filing return volatility as the dependent variable

	(1)	(2)	(3)
Diction tone	-0.106		-0.036
	(-1.83)		(-0.57)
LM tone		-0.112	-0.101
		(-4.06)	(-3.40)
Log(price)	-0.236	-0.227	-0.227
	(-4.23)	(-4.03)	(-4.03)
Log(market value)	-0.065	-0.070	-0.069
	(-3.07)	(-3.38)	(-3.38)
Log(filing date abnormal returns)	-2.283	-2.285	-2.284
	(-6.05)	(-6.05)	(-6.05)
Pre-filing alpha	-0.985	-0.981	-0.981
	(-4.26)	(-4.24)	(-4.24)
Pre-filing RMSE	0.578	0.576	0.576
	(11.64)	(11.76)	(11.72)
Nasdaq dummy	0.293	0.296	0.295
	(3.44)	(3.43)	(3.46)
R-squared	56.39%	56.43%	56.43%

This table reports the coefficients and *t*-statistics for regressions of the Diction tone and Loughran-McDonald tone with post-filing return volatility as the dependent variable. Included in the regressions but not tabulated are a constant, year dummies, and Fama-French (1997) industry dummies. Variable definitions are in the Appendix. The *t*-statistics are in parentheses with the standard errors clustered by both year and industry. Each regression includes 75,931 firm-year observations.