



Management Science

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Disclosure Sentiment: Machine Learning vs. Dictionary Methods

Richard Frankel, Jared Jennings, Joshua Lee

To cite this article:

Richard Frankel, Jared Jennings, Joshua Lee (2022) Disclosure Sentiment: Machine Learning vs. Dictionary Methods. Management Science 68(7):5514-5532. <https://doi.org/10.1287/mnsc.2021.4156>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2021, INFORMS

Please scroll down for article—it is on subsequent pages





With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Disclosure Sentiment: Machine Learning vs. Dictionary Methods

Richard Frankel,^a Jared Jennings,^a Joshua Lee^b

^aOlin Business School, Washington University in St. Louis, St. Louis, Missouri 63130; ^bMarriott School of Business, Brigham Young University, Provo, Utah 84602

Contact: frankel@wustl.edu,  <https://orcid.org/0000-0001-6736-0738> (RF); jaredjennings@wustl.edu,  <https://orcid.org/0000-0002-3658-4779> (JJ); joshlee84@byu.edu,  <https://orcid.org/0000-0001-7303-6029> (JL)

Received: January 15, 2020

Revised: February 25, 2021; May 5, 2021

Accepted: May 13, 2021

Published Online in Articles in Advance:
November 11, 2021

<https://doi.org/10.1287/mnsc.2021.4156>

Copyright: © 2021 INFORMS

Abstract. We compare the ability of dictionary-based and machine-learning methods to capture disclosure sentiment at 10-K filing and conference-call dates. Like Loughran and McDonald [Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Finance* 66(1):35–65.], we use returns to assess sentiment. We find that measures based on machine learning offer a significant improvement in explanatory power over dictionary-based measures. Specifically, machine-learning measures explain returns at 10-K filing dates, whereas measures based on the Loughran and McDonald dictionary only explain returns at 10-K filing dates during the time period of their study. Moreover, at conference-call dates, machine-learning methods offer an improvement over the Loughran and McDonald dictionary method of a greater magnitude than the improvement of the Loughran and McDonald dictionary over the Harvard Psychosociological Dictionary. We further find that the random-forest-regression-tree method better captures disclosure sentiment than alternative algorithms, simplifying the application of the machine-learning approach. Overall, our results suggest that machine-learning methods offer an easily implementable, more powerful, and reliable measure of disclosure sentiment than dictionary-based methods.

History: Accepted by Brian Bushee, accounting.

Funding: The authors are thankful to Olin Business School and the Marriott School of Business for financial support.

Supplemental Material: Data and the online appendix are available at <https://doi.org/10.1287/mnsc.2021.4156>.

Keywords: textual analysis • machine learning • disclosure • conference calls

1. Introduction

We compare the ability of dictionary methods and machine-learning methods to capture the sentiment of 10-K filings and earnings conference calls. Our research follows Loughran and McDonald (2011) who develop a measure of disclosure sentiment based on a business-context dictionary (LM dictionary) and provide evidence of a stronger association with 10-K filing date returns than a measure based on the Harvard Psychosociological Dictionary (Harvard Dictionary).¹ Although dictionary methods are easy to implement, they have flaws that machine-learning methods can alleviate. For example, dictionaries do not reflect language change over time or across industries; do not allow for variation in word importance; can be costly to update; and are subject to researcher subjectivity, which may lead to bias, incompleteness, or overfitting.

We provide evidence that machine-learning methods yield an improvement over the LM sentiment measure in capturing returns at both 10-K filing dates and conference-call dates—an improvement similar to that of the LM dictionary over the Harvard Dictionary. Moreover, the ability of dictionary-based methods to

capture sentiment at 10-K filing dates is time-period specific. Using the Loughran and McDonald (2011) sample period, we confirm their results; but we find the LM and Harvard sentiment measures are unassociated with 10-K filing returns when the sample period is extended to 2019. In contrast, machine-learning methods consistently explain 10-K filing and conference call returns in both samples. We also compare multiple machine-learning models and find that the random-forest regression-tree method explains 10-K and conference call returns better than the support-vector regression and supervised-latent-Dirichlet allocation methods. Overall, our results suggest that machine-learning methods offer a more powerful and reliable measure of disclosure sentiment than dictionary-based methods.

Since Ball and Brown (1968), researchers have sought to understand the usefulness of disclosures to investors. Although early research studies the information content of numerical information, more recent work studies qualitative information content. One hurdle in this line of research is quantifying and summarizing qualitative disclosure content. Studies use

different techniques to extract disclosure signals such as readability, similarity, and sentiment. For example, Loughran and McDonald (2011) use dictionary-based approaches to measure disclosure sentiment. We use the LM dictionary-based sentiment measures as a benchmark of comparison because of their wide use.² The rise in computing technology and resources has given researchers the ability to extract disclosure content using automated methods, and researchers have promoted the use of these techniques (Core 2001, Beyrer et al. 2010). Research has begun to examine whether machine-learning methods can effectively extract information from qualitative disclosures (e.g., Frankel et al. 2016, Huang et al. 2018, Donovan et al. 2021). Our objective is to merge these two research lines by examining whether machine-learning methods can improve upon dictionary-based measures of disclosure sentiment. Our research shares the objective of Loughran and McDonald (2011), whose goal was to identify a better method to capture disclosure sentiment.

Like Loughran and McDonald (2011), we use contemporaneous returns to validate the disclosure sentiment measures. We first replicate the Loughran and McDonald (2011) results. Using a sample of 40,922 10-K firm-year observations between 1996 and 2008, we find evidence consistent with Loughran and McDonald (2011) that negative and net tone measures based on the LM dictionary explain 10-K filing date returns. However, when we include all firm-year observations between 1996 and 2019 (75,363 firm-year observations), we find that dictionary-based sentiment measures no longer explain 10-K filing returns. Using the Harvard Dictionary to measure 10-K sentiment, we find that tone measures are insignificant even during the LM sample period and yield coefficients inconsistent with expectations in the full sample. These results might suggest that the language in the 10-K has lost information content over time or that dictionary-based measures lack the power to capture information content.

Next, we examine whether machine-learning methods consistently capture 10-K sentiment over time. We use three machine-learning methods: random forest regression trees (RF), support vector regression (SVR), and supervised latent Dirichlet allocation (sLDA). We use a rolling training sample approach including observations from previous years to train each model to identify one- and two-word phrases in the 10-K that explain 10-K filing date returns. We then apply these models to out-of-sample observations in the current year to calculate machine-learning sentiment measures based on the language in the 10-K. We use factor analysis to create a fourth measure that combines all three machine-learning sentiment measures into a single combined measure. All four machine-learning measures explain

10-K filing returns during the LM sample period. The RF, sLDA, and combined measures explain 10-K filing returns in the full sample. The RF measure yields an adjusted- R^2 that is higher than that of the combined measure. In summary, the machine-learning methods more consistently capture disclosure sentiment than dictionary-based methods in the 10-K setting.

Prior research suggests that 10-Ks, given the muted market reactions to these filings, are less informative than other disclosures (Li and Ramesh 2009). Conference calls typically elicit stronger price reactions than 10-K releases, perhaps because they are timelier and permit interaction between the analysts and management. Thus, machine learning's incremental benefit compared with dictionary-based methods may be different when extracting information content from conference calls. Using 106,151 firm-quarter observations between 2004 and 2019, we calculate conference call sentiment measures using the dictionary and machine-learning approaches, employing similar techniques when examining the 10-K. We model the relation between each sentiment measure and conference call returns. All coefficients on the dictionary-based sentiment measures have the expected signs, and their explanatory power ranges between 4.7% and 6.4%. The coefficients on all machine-learning measures are significant with the expected signs, and their adjusted- R^2 range is between 7.1% and 12.7%. Once again, the RF measure yields the highest adjusted- R^2 of all four machine-learning measures, including the combined measure.

The conference call results yield two notable implications. First, the adjusted- R^2 of the regression including the RF measure is 99% larger than the model including the LM sentiment measure, which is the dictionary-based measure that yields the highest adjusted- R^2 . The RF measure produces a significant improvement over dictionary-based sentiment measures even when disclosures contain more information content. Second, an approach using multiple machine-learning models does not necessarily yield a better sentiment measure than using the RF method alone. Thus, future researchers can simply use the RF method to capture disclosure sentiment rather than estimate multiple machine-learning models.

Although our primary objective is to assess the ability of dictionary-based and machine-learning measures to capture 10-K and conference call sentiment, we also examine whether these methods can be used to capture information that investors overlook. We retrain the machine-learning measures to identify language that correlates with both future earnings surprises and future stock market reactions. We find that the RF measure predicts future earnings surprises and yields positive hedge portfolio returns using the conference call and 10-K. These results suggest that

machine-learning methods are useful for capturing disclosure sentiment that correlates with investors' immediate response and information overlooked by investors that is impounded into prices with a delay. We also find evidence that the sentiment measure based on the LM dictionary yields positive future hedge portfolio returns but only in the conference call setting; however, the hedge portfolio return using the RF measure is 3.9 times larger than the hedge portfolio return using the LM dictionary-based method. We find no evidence that the LM dictionary measure yields positive hedge portfolio returns in the 10-K setting, and that it negatively predicts future earnings surprises, contrary to expectations. The Harvard Dictionary measure also negatively predicts future earnings surprises in the 10-K setting and is unrelated to future earnings surprises in the conference call setting. It also yields negative and significant future hedge portfolio returns in both the 10-K and conference call settings, which is contrary to expectations. Thus, only the machine-learning approach consistently identifies information overlooked by investors.

As an additional test, we examine the dictionary-based and machine-learning measures' relative ability to capture earnings-announcement press-release sentiment. Using a sample of 177,900 earnings announcement 8-K press release disclosures, we provide evidence that the RF sentiment measure best explains returns around the earnings announcement date (adjusted- R^2 equal to 4.5%), but the relative improvement in explanatory power over the LM sentiment measure is only about 6%. These results reinforce two key takeaways. First, the RF measure captures investor reactions better than dictionary-based sentiment measures. Second, machine-learning measures only provide modest improvements over the dictionary-based measures when disclosures are less spontaneous and dynamic. Like the 10-K, managers carefully craft the earnings announcement with legal personnel.

Our study contributes to the literature in three ways. First, we provide a more powerful and robust measure of disclosure sentiment using three prominent disclosures: 10-Ks, conference calls, and earnings announcement press releases. These results suggest that machine-learning methods reduce measurement error and improve construct validity. Given that uncorrelated measurement error can lead to both type I and type II errors (Jennings et al. 2021) and the abundance of research examining the sentiment of disclosures, reducing measurement error is an important consideration, especially in a relatively new area of research (i.e., capturing the information content of qualitative disclosure).

Second, dictionary-based measures do not consistently capture disclosure sentiment across time. Dictionaries are static and do not reflect changes in disclosure

language. Managers might also adjust language they believe investors perceive to be negative. Cao et al. (2021) show that firms with a higher level of machine downloads are less likely to include negative LM dictionary words. Thus, the efficacy of dictionaries in capturing the intended construct can vary over time. In contrast, machine-learning models that use rolling training windows adapt to changing disclosures over time.

Third, rather than estimating many different machine-learning models to capture investor reactions, our results suggest that researchers can simplify their approach and use only the random forest regression tree method to measure disclosure sentiment. The RF method yields the highest adjusted- R^2 relative to the other machine learning and combined measures when explaining investors' reactions to the 10-K filing and conference call. Lastly, we provide evidence that machine-learning measures not only capture disclosure sentiment but can also be used to identify information that may not be readily apparent to investors. Our evidence that the language in the 10-K and conference call predicts future earnings surprises and future earnings announcement returns reinforces our conclusions about the robustness of machine-learning methods relative to dictionary-based methods.

Overall, our results suggest that researchers can use machine-learning techniques to capture information that leads investors to revise valuations at the time of the conference call, 10-K, or earnings announcement date as well as information overlooked by investors. We caution future research that the results in this paper may not be generalizable to the measurement of other nondisclosure-sentiment proxies. Different methods may be better suited to capture other textual constructs. For example, dictionary methods are likely more appropriate when testing whether a specific attribute of narrative (e.g., euphemisms) drives an association or causation (e.g., Bochkay et al. 2020, Suslava 2021).

2. Prior Literature

Textual analysis techniques have substantially progressed over the past 20 years (Gentzkow et al. 2019). The first attempts to analyze the text of disclosures relied upon manual classification.³ For example, Botosan (1997) manually identifies items discussed in annual reports to measure disclosure quality, and Bryan (1997) manually classifies the favorability of items included in the Management's Discussion and Analysis. Manual classification is beneficial because it can be more precise (Li 2010). However, it also has its disadvantages. First, manual classification is time consuming, leading to studies with small sample sizes, limited scope, and reduced statistical power. Second,

manual coding involves subjectivity, reducing its replicability and limiting follow-up studies.

The accounting literature adopted alternative methods to classify and extract information in disclosures to improve replicability and applicability. These include counting words based on dictionaries (e.g., tone, competition)⁴ and using alternative computational linguistic algorithms to measure specific text characteristics (e.g., readability, cosine similarity).⁵ Once researchers establish a dictionary or linguistic algorithm, measures become standardized, thereby improving replicability, applicability to larger samples, and increasing the likelihood of follow-up studies. Dictionary-based approaches also present drawbacks. First, dictionaries depend on context. Tone dictionaries developed for 10-K reports may not be applicable to conference call transcripts (e.g., the word “question” has negative implications in a 10-K but is commonly used in conference calls during the question-and-answer portion of the call). In addition, readability measures might more readily apply to written 10-K reports by firms in low-tech industries but might not be applicable to transcripts of spoken, spontaneous language or situations using specialized, multisyllable words. Second, researcher judgment in the creation of these methods can reduce their reliability. For example, researchers must justify why specific words are included or excluded from a dictionary or why measures of readability include their various dimensions (e.g., sentence length). Third, these methods are narrowly applied to specific constructs (e.g., tone or readability), requiring researchers to create new dictionaries or new methods to examine alternative constructs.

Prior literature has primarily used two dictionaries to create sentiment measures. The first is the Harvard Psychosociological Dictionary, which identifies positive and negative words that are used to calculate the sentiment of a document or disclosures. The Harvard Dictionary was originally developed to capture sentiment in psychological and social contexts. As suggested by Loughran and McDonald (2011, p. 35), “English words have many meanings, and a word categorization scheme derived for one discipline might not translate effectively into a discipline with its own dialect.” The English language is complex and words can have different meanings based on the context in which they are used. One drawback to the Harvard Dictionary is that it was not developed for financial contexts.

Because words can have many different meanings depending on the context, Loughran and McDonald developed a sentiment dictionary specific to financial documents. Although Loughran and McDonald (2011) provide evidence that sentiment measures based on

the Harvard Dictionary explain returns, their evidence suggests that sentiment measures based on the Loughran and McDonald (2011) dictionary more consistently capture returns around the 10-K. Therefore, much of the subsequent research in accounting and finance uses sentiment measures based on the Loughran and McDonald (2011) dictionary (e.g., Rogers et al. 2011, Law and Mills 2015).

To advance the literature, Core (2001, p. 452) encouraged researchers to adopt “techniques in natural language processing from fields like computer science, linguistics, and artificial intelligence,” and Beyer et al. (2010) echoes this advice. Many papers followed that use various automated machine-learning techniques to identify information contained in disclosures.

The first applications of automated machine learning in the accounting literature rely upon manual text classification to aid in the learning process (e.g., Li 2010, Huang et al. 2014). For example, Li (2010) randomly selects a subset of forward-looking sentences from the MD&A and manually categorizes their favorability. He then applies a Naïve Bayes Classifier to these manually classified sentences. The Naïve Bayes Classifier assigns values to word patterns to maximize the likelihood that the total values will produce categorizations that match the manually-assigned favorability categories. Once values of word patterns are established, the classifier can be used to produce favorability scores for unclassified sentences. This crossover between manual text classification and automated learning yields several benefits relative to manual classification alone. First, the researcher can train the classifier on a subset of categorized observations and apply the categorization to a holdout sample of uncategorized observations, thereby allowing the researcher to expand the sample size. Second, the manual machine-learning method allows the researcher to precisely tailor the content analysis to a specific setting. However, manual machine learning has some similar drawbacks to those based on manual classification alone. For example, manual classification is time-intensive and subject to researcher judgment, which limits the applicability of these methods to new constructs or new disclosures, reduces replicability, and limits the possibility of follow-up studies.

A natural progression in this literature is the implementation of automated machine-learning methods, which rely less on researcher judgment and instead allow an algorithm to categorize a disclosure. Rather than focusing on ill-defined constructs, such as “tone” or “readability,” automated methods identify topics or words contained in a disclosure that can be used to predict. The most widely used automated machine-learning method in the accounting and finance literature is unsupervised latent Dirichlet allocation (LDA) (Blei et al. 2003).⁶ Supervised latent Dirichlet allocation

is an alternative topic modeling approach that creates topics that predict a dependent variable (Blei and McAuliffe 2007). Alternative supervised methods (i.e., methods that are trained to predict a dependent variable) include support vector regression and random forest regression trees. For example, Frankel et al. (2016) use accruals to train a support vector regression model and then apply the model to out-of-sample documents to generate an estimate of accruals based on MD&A text. Manela and Moreira (2017) use support vector regressions to estimate the VIX using the front page of the *Wall Street Journal*. Donovan et al. (2021) use supervised machine-learning methods to measure credit risk. Machine-learning methods have several benefits. First, these methods reduce researcher subjectivity from the classification process. Second, automation can identify information that might not immediately be apparent to the researcher upon manual inspection. Third, the cost of applying machine-learning methods to new documents, constructs, or languages is lower relative to methods that rely on manual classification (e.g., Li 2010). As a result, researchers can increase sample sizes, expand scope, increase statistical power, and improve replicability. In contrast, these methods are subject to the disadvantage that they may identify words and topics that have little economic intuition.

Prior research examines the power of machine-learning methods to explain firm fundamentals (e.g., Frankel et al. 2016, Donovan et al. 2021) but has not assessed whether machine-learning measures are superior to dictionary-based measures for capturing sentiment. Because automated machine-learning methods are based on data rather than on intuition, they may fail to capture information that investors glean from disclosures. Our paper begins to fill these gaps in the literature by employing computer scientists' methods to extract information from text and compare their explanatory power to existing dictionary-based measures.

3. Measures of Narrative Content

We compare the ability of dictionary-based and machine-learning measures to capture disclosure sentiment of both 10-K reports and earnings conference call transcripts. Although these two disclosures discuss financial results, they are different in two fundamental ways. First, the conference call is timelier than the 10-K. Conference calls are typically released soon after the earnings announcement press release and are associated with a large stock market response relative to that generated by the release of the subsequent 10-K report. The muted market response to 10-K reports suggests that much of the information in the 10-K is redundant to previously released information

and is thus less informative to investors (Li and Ramesh 2009). Second, conference calls are typically more spontaneous and dynamic than the 10-K. Conference call participants can ask questions about value-relevant information released in the earnings announcement. The firm's legal and management team carefully craft the 10-K, which possibly decreases its transparency and reduces its information content.

3.1. Dictionary-Based Sentiment Measures

Following prior research, we use two dictionaries to measure disclosure sentiment or tone: the Harvard Psychosociological Dictionary and the Loughran and McDonald (2011) business-context dictionary. We calculate positive, negative, and net tone using the words included in each dictionary. The positive (negative) tone measure is equal to the sum of the positive (negative) words divided by the total words in the disclosures. $HARV\ 10-K\ POS\ TONE_{i,t}$ ($HARV\ 10-K\ NEG\ TONE_{i,t}$) is the positive (negative) tone variable using the Harvard Dictionary for the 10-K sample. $LM\ 10-K\ POS\ TONE_{i,t}$ and $LM\ 10-K\ NEG\ TONE_{i,t}$ are the equivalent measures using the Loughran and McDonald (2011) dictionary. $HARV\ CC\ POS\ TONE_{i,q}$ ($HARV\ CC\ NEG\ TONE_{i,q}$) is the positive (negative) tone variable using the Harvard Dictionary for the conference call sample. $LM\ CC\ POS\ TONE_{i,q}$ and $LM\ CC\ NEG\ TONE_{i,q}$ are the equivalent measures using the Loughran and McDonald (2011) dictionary. The net tone measures are equal to the sum of the positive words less the sum of the negative words and then divided by the sum of the positive and negative words in each disclosure. $HARV\ 10-K\ TONE_{i,t}$ ($LM\ 10-K\ TONE_{i,t}$) is the net tone measure based on the Harvard Dictionary (Loughran and McDonald dictionary) in the 10-K sample. $HARV\ CC\ TONE_{i,q}$ and $LM\ CC\ TONE_{i,q}$ are the equivalent measures in the conference call sample.

3.2. Supervised Machine-Learning Sentiment Measures

We create machine-learning measures of disclosure sentiment for both 10-K filings and conference calls using three machine-learning models: support vector regression, supervised latent Dirichlet allocation, and random forest regression trees. We use these models to map the counts of all one- and two-word phrases contained in each disclosure (10-Ks and conference calls) to the cumulative abnormal returns from day t to day $t+1$ surrounding the date of the release of each disclosure ($10-K\ CAR[0,1]_{i,t}$ and $CC\ CAR[0,1]_{i,q}$).

Similar to prior research (e.g., Frankel et al. 2016), we stem all words using the Porter Stemmer algorithm and remove any one- or two-word phrase that is used in fewer than 10 disclosures in each training

sample. We also remove highly frequent words (i.e., stop words) such as “and” and “the” and remove all words containing digits. We then estimate out-of-sample text-based predictions for 10-K $CAR[0,1]_{i,t}$ and CC $CAR[0,1]_{i,q}$ using rolling training samples for each calendar year in our sample. We use rolling training samples to estimate the parameters for each of our statistical methods because language that is useful in determining the sentiment of the disclosure could change over time (Frankel et al. 2016, Brown et al. 2020).⁷ Because the out-of-sample estimation method is slightly different for SVR, sLDA, and RF, we describe each method in more detail below. We also provide a detailed description of the data preparation process and the assumptions required for each machine-learning model in the online appendix.

SVR allows the estimation of a unique weight for each one- and two-word phrase count that is included in the disclosure (e.g., 10-K or conference call). Ordinary least squares cannot generate weights for each unique one- and two-word phrase because the number of phrases is greater than the number of observations in each training sample. For example, the training sample used for 2019 contains 496,841 (466,305) unique phrases in the 10-K (conference call) but only 6,270 (32,647) unique observations. SVR estimates weights on each one- and two-word phrase by simultaneously minimizing both the coefficient vector magnitude and the prediction error. The simultaneous minimization of the coefficient vector magnitude and the prediction error works to reduce overfitting. Frankel et al. (2016) and Manela and Moreira (2017) provide a more in-depth discussion of the SVR estimation procedure. Using our rolling training samples, we use SVR to estimate weights for each one- and two-word phrase count. We then apply the weights to the one- and two-word phrase counts of the 10-Ks (conference calls) in year t to estimate an out-of-sample prediction of 10-K (conference call) returns, which we label $SVR\ 10-K\ CAR[0,1]_{i,t}$ ($SVR\ CC\ CAR[0,1]_{i,q}$). SVR is likely to be effective if the individual weights on words and phrases are predictive of the dependent variable.

sLDA categorizes the words and phrases of a disclosure into a set of latent (i.e., unknown) topics that are predictive of a dependent variable (Blei and McAuliffe, 2007). The algorithm assumes that all disclosures share the same set of topics; however, the mix of each topic varies by disclosure. The algorithm sorts the words and phrases from the disclosures into topics based on the probability of words co-occurring within the disclosures while simultaneously creating topics that are predictive of a dependent variable. The prior accounting and finance literature has primarily used unsupervised LDA for text classification and prediction (e.g., Bao and Datta 2014, Campbell et al. 2014, Dyer et al. 2017). Because our goal is prediction

rather than simply text classification, we use the supervised version of the LDA model.

We fit the sLDA model to each rolling training sample using the counts of all one- and two-word phrases found in the 10-Ks (conference calls). We allow the algorithm to identify 200 topics and extract the topic weightings for each 10-K (conference call) that are predictive of 10-K (conference call) returns.⁸ We apply the topics and topic weightings from the training sample to all 10-Ks (conference calls) in year t to estimate an out-of-sample prediction for 10-K (conference call) returns, which we label $sLDA\ 10-K\ CAR[0,1]_{i,t}$ ($sLDA\ CC\ CAR[0,1]_{i,q}$). Because of the diversity of the English language, disclosures often discuss similar items or activities in different ways. sLDA groups words that discuss similar activities into topics. Topics likely include words with similar meanings (e.g., income versus earnings versus earnings per share) as well as words that are used to discuss a particular topic (e.g., covenant, debt, threshold).

We also use RF to predict 10-K and conference call returns. The standard regression tree method uses an iterative process called binary recursive partitioning, which creates a decision tree by recursively partitioning observations based on features (e.g., one- and two-word phrases) to predict a specific value or characteristic. Each partition made by the algorithm is identified with a node (or branch), which is a binary classification of the data using one of the data set's features. At each node, the algorithm examines each of the remaining binary splits of the data using the remaining features and chooses the feature that minimizes the sum of squared errors within each partition. The algorithm continues to partition the data using nodes until the number of observations within each partition falls below a prespecified number (e.g., two or five observations) or when the sum of the squared errors within the partition is equal to zero. When the process stops, the average value of the response value at each terminal (i.e., final) node represents the predicted value of the response given the preceding binary partitions of the data. We provide additional description of the regression tree method in the online appendix.

Random forest is an application of the regression tree method that works to reduce overfitting and improve the generalization of the model (Breiman 2001). The random-forest method has two characteristics that set it apart from the standard regression tree method. First, the random-forest method constructs a predetermined number of regression trees (e.g., 500 or 5,000) using different bootstrapped samples of the training data for each tree (i.e., random sampling with replacement). Second, the method constructs each tree using a randomly selected subset of features (e.g., the square root of the total features in the data). As a

result, each tree uses a different feature as a starting node, which allows the random-forest method to identify many possible nonlinearities in the data. The predicted value of the response is then equal to the average predicted value generated by all bootstrapped trees (i.e., the forest). Breiman (2001) argues that as the number of trees in the random forest becomes large, the error of the forest converges almost surely to a limit.⁹

For each of our rolling training samples, we use the RF method to create 5,000 regression trees using the one- and two-word phrase counts as features and then apply the model to the one- and two-word phrase counts in year t to predict the 10-K (conference call) returns in year $t+1$ (quarter $q+1$), which we label $RF\ 10\text{-}K\ CAR[0,1]_{i,t}$ ($RF\ CC\ CAR[0,1]_{i,q}$).¹⁰ The strength of the RF method is that its algorithm allows for interactions between words.

Ex ante, we expect to extract nonoverlapping signals from disclosures using the different machine-learning models because the models use different processes to extract information from text (e.g., placing weight on specific words/phrases, identifying topics, and identifying interactions among words/phrases). Thus, we calculate a combined measure of all machine-learning predictions of the 10-K (conference call) return using factor analysis estimated separately for each sample observation using all available observations in the sample over the previous 365 days. We label the combined measure for the 10-K (conference call) sample as $ML\ 10\text{-}K\ CAR[0,1]_{i,t}$ ($ML\ CC\ CAR[0,1]_{i,q}$).

4. Sample and Main Results

4.1. Sample Selection and Descriptive Statistics

We obtain a sample of 75,363 firm-year 10-Ks between 1996 and 2019. Panel A of Table 1 presents the descriptive statistics for the variables used in our primary analyses. As expected, the mean of each machine-learning measure is similar to the 10-K filing date return ($10\text{-}K\ CAR[0,1]_{i,t}$). As suggested by the mean of $LM\ 10\text{-}K\ NEG\ TONE_{i,t}$ ($LM\ 10\text{-}K\ POS\ TONE_{i,t}$), 1.1% (0.6%) of the words in the 10-K are negative (positive) words as defined by the LM dictionary. The mean of $HARV\ 10\text{-}K\ NEG\ TONE_{i,t}$ ($HARV\ 10\text{-}K\ POS\ TONE_{i,t}$) suggests that 3.4% (5.9%) of the words in the 10-K are negative (positive) words as defined by the Harvard Dictionary. Consistent with these statistics, the mean value of $LM\ 10\text{-}K\ TONE_{i,t}$ ($HARV\ 10\text{-}K\ TONE_{i,t}$) is negative (positive), which suggests that the average 10-K tone is negative (positive) based on these dictionaries. Control variable values are consistent with expectations. All variable definitions are included in the appendix.

We also obtain a sample of 106,151 firm-quarter conference call transcripts from Factiva's Fair Disclosure Wire between 2004 and 2019. Panel B of Table 1

presents the descriptive statistics for the variables used in our primary analyses. We include all variables that are also included in Panel A of Table 1. We note a few differences between the 10-K and conference call samples. First, using the LM dictionary, the average sentiment of the conference call ($LM\ CC\ TONE_{i,q}$ is equal to 0.227) is more positive than the average sentiment of the 10-K ($LM\ 10\text{-}K\ TONE_{i,q}$ is equal to -0.273). The average firm-quarter in the conference call sample is larger (MVE), has higher institutional ownership ($INSTOWN$), has higher turnover ($TURNOVER$), has higher growth Book to market (BTM), and is less likely to be included on the NASDAQ than the average firm-year in the 10-K sample.

Table 2 presents the Pearson (above diagonal) and Spearman (below diagonal) correlations for all dictionary-based and machine-learning sentiment measures. The correlations for the 10-K (conference call) sample are presented in Panel A (Panel B). Generally, the correlations between the dictionary-based sentiment measures and the machine-learning measures are as expected. The correlations between the dictionary-based measures and the machine-learning measures are relatively weaker than the correlations between the two dictionary-based measures in the 10-K setting. The Spearman correlation between $RF\ 10\text{-}K\ CAR[0,1]$ and $LM\ 10\text{-}K\ TONE$ is equal to 0.03, whereas the correlation between $LM\ 10\text{-}K\ TONE$ and $HARV\ 10\text{-}K\ TONE$ is equal to 0.42, which suggests that the machine-learning models identify a different disclosure signal than that obtained using the dictionary approach. In the conference call setting, the Spearman correlation between $RF\ CC\ CAR[0,1]$ and $LM\ CC\ TONE$ is equal to 0.47 and is similar to the correlation between $LM\ CC\ TONE$ and $HARV\ CC\ TONE$ (0.44), suggesting a greater but not complete overlap in these conference call measures.

4.2. 10-K Sample—Dictionary-Based Sentiment Measures

We next examine whether the sentiment measures based on the LM dictionary explain the two-day return on the 10-K filing date. Similar to Loughran and McDonald (2011), we use the following equation.

$$\begin{aligned} 10 - K\ CAR[0,1]_{i,t} = & \alpha_0 + \alpha_1 SENTIMENT_{i,t} \\ & + \alpha_2 \ln(MVE_{i,t}) + \alpha_3 BTM_{i,t} \\ & + \alpha_4 TURNOVER_{i,t} \\ & + \alpha_5 PRE_FFALPHA_{i,t} \\ & + \alpha_6 INSTOWN_{i,t} \\ & + \alpha_7 NASDAQ_{i,t} + \varepsilon_{i,t} \end{aligned} \quad (1)$$

We estimate Equation (1) separately for each of the following measures of $SENTIMENT_{i,t}$: $LM\ 10\text{-}K$

Table 1. Descriptive Statistics

Panel A: 10-K sample					
Variable	Mean	Std. dev.	Q1	Median	Q3
10-K CAR[0,1] _{i,t}	−0.001	0.049	−0.020	−0.001	0.018
FUT EA CAR[0,1] _{i,t}	0.001	0.085	−0.034	0.000	0.036
FUT EARN SURP _{i,t}	0.000	0.014	−0.001	0.000	0.002
LM 10-K TONE _{i,t}	−0.273	0.178	−0.399	−0.298	−0.175
LM 10-K POS TONE _{i,t}	0.006	0.001	0.005	0.006	0.007
LM 10-K NEG TONE _{i,t}	0.011	0.004	0.009	0.011	0.014
HARV 10-K TONE _{i,t}	0.273	0.106	0.204	0.283	0.351
HARV 10-K POS TONE _{i,t}	0.059	0.008	0.053	0.058	0.064
HARV 10-K NEG TONE _{i,t}	0.034	0.009	0.027	0.032	0.039
RF 10-K CAR[0,1] _{i,t}	−0.001	0.014	−0.006	0.000	0.005
SVR 10-K CAR[0,1] _{i,t}	0.000	0.057	−0.029	0.001	0.031
sLDA 10-K CAR[0,1] _{i,t}	−0.001	0.006	−0.004	−0.001	0.003
ML 10-K CAR[0,1] _{i,t}	0.000	1.317	−0.667	0.041	0.720
RF FUT EA CAR[0,1] _{i,t}	0.002	0.022	−0.007	0.001	0.010
RF FUT EARN SURP _{i,t}	0.000	0.003	−0.001	0.000	0.001
MVE _{i,t}	4,159.140	11,539.890	220.986	716.308	2,557.930
BTM _{i,t}	0.540	0.446	0.252	0.450	0.718
TURNOVER _{i,t}	1.966	1.754	0.793	1.462	2.512
PRE_FFALPHA _{i,t}	0.000	0.002	−0.001	0.000	0.001
INSTOWN _{i,t}	0.479	0.351	0.106	0.519	0.798
NASDAQ _i	0.532	0.499	0.000	1.000	1.000
DISPERSION _{i,t}	0.003	0.006	0.000	0.001	0.002
REVISIONS _{i,t}	−0.003	0.010	−0.004	−0.001	0.000
Panel B: Conference call sample					
Variable	Mean	Std. dev.	Q1	Median	Q3
CC CAR[0,1] _{i,q}	0.000	0.082	−0.042	0.000	0.043
FUT EA CAR[0,1] _{i,q}	0.000	0.091	−0.042	0.000	0.043
FUT EARN SURP _{i,q}	0.000	0.013	−0.001	0.001	0.002
LM CC TONE _{i,q}	0.227	0.199	0.095	0.239	0.371
LM CC POS TONE _{i,q}	0.015	0.004	0.012	0.015	0.018
LM CC NEG TONE _{i,q}	0.009	0.003	0.007	0.009	0.011
HARV CC TONE _{i,q}	0.324	0.086	0.266	0.326	0.384
HARV CC POS TONE _{i,q}	0.073	0.008	0.067	0.073	0.078
HARV CC NEG TONE _{i,q}	0.037	0.006	0.033	0.037	0.041
RF CC CAR[0,1] _{i,q}	0.000	0.007	−0.005	0.000	0.005
SVR CC CAR[0,1] _{i,q}	−0.001	0.071	−0.047	0.000	0.046
sLDA CC CAR[0,1] _{i,q}	0.002	0.010	−0.005	0.002	0.009
ML CC CAR[0,1] _{i,q}	0.000	1.369	−0.922	−0.018	0.914
RF FUT EA CAR[0,1] _{i,q}	0.000	0.005	−0.003	0.000	0.003
RF FUT EARN SURP _{i,q}	0.000	0.002	0.000	0.000	0.001
EARN SURP _{i,q}	0.000	0.012	−0.001	0.001	0.002
MVE _{i,q}	8,211.500	20,557.590	559.614	1,686.640	5,697.880
BTM _{i,q}	0.476	0.424	0.221	0.391	0.634
TURNOVER _{i,q}	2.596	2.067	1.276	2.014	3.223
PRE_FFALPHA _{i,q}	0.000	0.001	−0.001	0.000	0.001
INSTOWN _{i,q}	0.589	0.359	0.288	0.712	0.892
NASDAQ _i	0.456	0.498	0.000	0.000	1.000
DISPERSION _{i,q}	0.003	0.005	0.000	0.001	0.002
REVISIONS _{i,q}	−0.003	0.010	−0.003	−0.001	0.000

Notes. This table presents the descriptive statistics for the variables used in the empirical analyses. Panel A reports all variables included in the 10-K analyses and includes 75,363 observations from 1996–2019. Panel B reports all variables included in the conference call analyses and includes 106,151 observations from 2004–2019. All variables are defined in the appendix. All continuous variables are winsorized at the 1st and 99th percentiles. Std. dev., standard deviation.

TONE_{i,t}, LM 10-K POS TONE_{i,t}, and LM 10-K NEG TONE_{i,t}. We also estimate the model using two sample periods. First, we use a sample of firm-year observations between 1996 and 2008 to replicate the sample period in Loughran and McDonald (2011). We then extend the sample through 2019 to examine the stability of the results in more recent years. The results are reported in Panel A of Table 3. Consistent with Loughran

Table 2. Correlations

Panel A: 10-K sample											
		I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.
I.	LM 10-K TONE _{i,t}	1.00	0.39	−0.80	0.41	−0.21	−0.45	0.03	0.02	0.08	0.03
II.	LM 10-K POS TONE _{i,t}	0.40	1.00	0.16	−0.02	0.26	0.15	−0.04	−0.02	−0.03	−0.04
III.	LM 10-K NEG TONE _{i,t}	−0.79	0.18	1.00	−0.47	0.37	0.58	−0.05	−0.03	−0.09	−0.05
IV.	HARV 10-K TONE _{i,t}	0.42	−0.03	−0.48	1.00	0.07	−0.86	−0.02	0.00	0.00	−0.01
V.	HARV 10-K POS TONE _{i,t}	−0.19	0.27	0.38	0.07	1.00	0.43	0.02	−0.01	0.01	0.02
VI.	HARV 10-K NEG TONE _{i,t}	−0.48	0.18	0.64	−0.83	0.45	1.00	0.03	0.00	0.02	0.02
VII.	RF 10-K CAR[0,1] _{i,t}	0.03	−0.05	−0.07	−0.01	0.02	0.01	1.00	0.58	0.34	0.86
VIII.	SVR 10-K CAR[0,1] _{i,t}	0.02	−0.01	−0.03	0.00	−0.01	−0.01	0.52	1.00	0.15	0.79
IX.	sLDA 10-K CAR[0,1] _{i,t}	0.07	−0.03	−0.09	0.01	0.02	−0.01	0.34	0.13	1.00	0.51
X.	ML 10-K CAR[0,1] _{i,t}	0.02	−0.04	−0.06	−0.01	0.02	0.01	0.81	0.78	0.51	1.00
Panel B: Conference call sample											
		I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.
I.	LM CC TONE _{i,q}	1.00	0.73	−0.77	0.46	0.36	−0.32	0.47	0.18	0.41	0.47
II.	LM CC POS TONE _{i,q}	0.72	1.00	−0.15	0.30	0.44	−0.06	0.36	0.15	0.34	0.38
III.	LM CC NEG TONE _{i,q}	−0.76	−0.15	1.00	−0.40	−0.13	0.42	−0.36	−0.13	−0.30	−0.34
IV.	HARV CC TONE _{i,q}	0.44	0.29	−0.38	1.00	0.64	−0.82	0.24	0.12	0.23	0.25
V.	HARV CC POS TONE _{i,q}	0.36	0.43	−0.13	0.62	1.00	−0.10	0.11	0.08	0.14	0.15
VI.	HARV CC NEG TONE _{i,q}	−0.30	−0.05	0.40	−0.81	−0.09	1.00	−0.23	−0.09	−0.19	−0.21
VII.	RF CC CAR[0,1] _{i,q}	0.47	0.36	−0.36	0.24	0.12	−0.22	1.00	0.43	0.59	0.86
VIII.	SVR CC CAR[0,1] _{i,q}	0.18	0.14	−0.13	0.11	0.08	−0.09	0.41	1.00	0.28	0.67
IX.	sLDA CC CAR[0,1] _{i,q}	0.41	0.34	−0.28	0.22	0.14	−0.17	0.59	0.27	1.00	0.78
X.	ML CC CAR[0,1] _{i,q}	0.47	0.37	−0.34	0.24	0.16	−0.20	0.85	0.65	0.78	1.00

Notes. This table presents the Pearson and Spearman correlations for the tone and machine learning measures used in the 10-K (Panel A) and conference call (Panel B) analyses. Pearson (Spearman) correlations are above (below) the diagonal. Correlations significant at the 5% level or lower are bolded. All variables are defined in the appendix. All variables are winsorized at the 1st and 99th percentiles.

and McDonald (2011), we find a significantly negative (1% level) coefficient on LM 10-K NEG TONE_{i,t} equal to −0.239 using the early sample period (column (3)). We then extend the sample to include firm-year observations between 1996 and 2019 and fail to find a significant coefficient on LM 10-K NEG TONE_{i,t} (column (6)). These results suggest that negative tone measures based on the LM dictionary do not consistently relate to stock market reactions at the 10-K filing date.

Although Loughran and McDonald (2011) focus on negative tone measures to capture sentiment, we also examine whether positive and net tone measures explain 10-K filing returns. Most subsequent research uses a net sentiment measure, which combines negative and positive tone dictionaries. Consistent with expectations, we find a positive and significant (1% level) coefficient equal to 0.003 on LM 10-K TONE_{i,t} in column (1) using the early sample period. Similar to LM 10-K NEG TONE_{i,t}, we cease to find a significant coefficient on LM 10-K TONE_{i,t} when the sample is extended to 2019 (column (4)). These results reinforce the inference that the LM dictionary does not consistently capture investor responses to the 10-K. Contrary to expectations, we find a negative and significant (5% level) coefficient on LM 10-K POS TONE_{i,t} when estimating the regression using the early sample (column (2)) and an insignificant coefficient when extending the sample to 2019.

Given the insignificance of the sentiment measures based on the LM dictionary for the full sample, we next examine whether sentiment measures based on the Harvard Dictionary provide different implications. We re-estimate Equation (1), replacing SENTIMENT_{i,t} with the Harvard measures (HARV 10-K TONE_{i,t}, HARV 10-K POS TONE_{i,t}, HARV 10-K NEG TONE_{i,t}). We present the regression results using sentiment measures based on the Harvard Dictionary in Panel B of Table 3. We suppress the coefficients on the control variables for the sake of brevity. Each cell represents a separate estimation of Equation (1). We report redundant results from Panel A in Table 3 for measures based on the LM dictionary for ease in comparison.

In the early sample (1996–2008), none of the coefficients on the sentiment measures based on the Harvard Dictionary are significant at conventional levels. In the full sample (1996–2019), none of the coefficients on the Harvard measures have the expected sign at conventional significance levels. Contrary to expectations, the coefficient on HARV 10-K TONE_{i,t} is significantly negative (1% level) and the coefficient on HARV 10-K NEG TONE_{i,t} is significantly positive (1% level). Consistent with the general takeaway from Loughran and McDonald (2011), these results suggest that the sentiment measures based on the Harvard Dictionary are not well suited for measuring disclosure sentiment in a 10-K setting.

Table 3. Sentiment Measures and 10-K Filing Returns

Panel A: Loughran and McDonald tone measures						
	Early sample period (1996–2008)			Full sample period (1996–2019)		
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Intercept</i>	0.003* (1.773)	0.004** (2.478)	0.004** (2.550)	−0.001 (−1.076)	−0.000 (−0.318)	−0.001 (−0.875)
<i>LM 10-K TONE_{i,t}</i>	0.003*** (2.635)			0.001 (1.141)		
<i>LM 10-K POS TONE_{i,t}</i>		−0.369** (−2.101)			−0.160 (−1.078)	
<i>LM 10-K NEG TONE_{i,t}</i>			−0.239*** (−3.463)			−0.050 (−0.994)
<i>ln(MVE)_{i,t}</i>	−0.000 (−0.143)	−0.000 (−0.198)	0.000 (0.060)	0.000 (1.562)	0.000 (1.531)	0.000 (1.561)
<i>BTM_{i,t}</i>	0.000 (0.505)	0.000 (0.146)	0.000 (0.527)	0.000 (0.567)	0.000 (0.324)	0.000 (0.520)
<i>TURNOVER_{i,t}</i>	−0.002*** (−10.112)	−0.002*** (−10.316)	−0.002*** (−9.848)	−0.001*** (−9.993)	−0.001*** (−10.075)	−0.001*** (−9.959)
<i>PRE_FFALPHA_{i,t}</i>	0.569*** (3.224)	0.581*** (3.290)	0.561*** (3.173)	0.335** (2.364)	0.337** (2.377)	0.335** (2.360)
<i>INSTOWN_{i,t}</i>	0.004*** (5.348)	0.004*** (5.443)	0.004*** (5.499)	0.006*** (10.134)	0.006*** (10.173)	0.006*** (10.112)
<i>NASDAQ_i</i>	−0.004*** (−7.060)	−0.004*** (−6.996)	−0.004*** (−6.756)	−0.002*** (−5.135)	−0.002*** (−5.140)	−0.002*** (−5.036)
#OBS	40,922	40,922	40,922	75,363	75,363	75,363
Adjusted R ²	0.787%	0.782%	0.800%	0.457%	0.457%	0.457%
Panel B: Comparison of all tone and machine learning measures						
	Early sample period (1996–2008)		Full sample period (1996–2019)			
Sentiment measure	Coefficient (<i>t</i> -statistic)	Adj. R ²	Coefficient (<i>t</i> -statistic)	Adj. R ²		
<i>LM 10-K TONE_{i,t}</i>	0.003*** (2.635)	0.787%	0.001 (1.141)	0.457%		
<i>LM 10-K POS TONE_{i,t}</i>	−0.369** (−2.101)	0.782%	−0.160 (−1.078)	0.457%		
<i>LM 10-K NEG TONE_{i,t}</i>	−0.239*** (−3.463)	0.800%	−0.050 (−0.994)	0.457%		
<i>HARV 10-K TONE_{i,t}</i>	−0.001 (−0.275)	0.771%	−0.005*** (−3.042)	0.467%		
<i>HARV 10-K POS TONE_{i,t}</i>	−0.029 (−0.966)	0.773%	−0.010 (−0.458)	0.456%		
<i>HARV 10-K NEG TONE_{i,t}</i>	−0.010 (−0.259)	0.771%	0.051*** (2.606)	0.464%		
<i>RF 10-K CAR[0,1]_{i,t}</i>	0.107*** (4.378)	0.843%	0.070*** (3.955)	0.493%		
<i>SVR 10-K CAR[0,1]_{i,t}</i>	0.009* (1.756)	0.782%	0.006 (1.448)	0.460%		
<i>sLDA 10-K CAR[0,1]_{i,t}</i>	0.129*** (2.745)	0.795%	0.147*** (4.331)	0.486%		
<i>ML 10-K CAR[0,1]_{i,t}</i>	0.001*** (2.900)	0.798%	0.001*** (3.119)	0.476%		

Notes. This table reports results of regressions where the dependent variable is equal to the cumulative abnormal return during the [0, +1] trading window surrounding the 10-K filing date (*10-K CAR[0,1]_{i,t}*). The regressions are estimated for the early sample period from 1996–2008 in columns (1)–(3) and for the full sample period from 1996–2019 in columns (4)–(6). Panel A reports results for the Loughran and McDonald tone measures including control variables. Panel B includes a comparison of all tone and machine learning measures. Control variables are also included in the regressions in Panel B but are not reported for brevity. Standard errors are clustered by firm. All variables are defined in the appendix. All continuous variables are winsorized at the 1% and 99% levels.

*, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

In summary, the sentiment measures based on both the Loughran and McDonald (2011) and Harvard Dictionaries do not consistently capture information that is ultimately reflected in 10-K filing returns. We next examine whether machine-learning methods improve the researcher's ability to capture 10-K sentiment.

4.3. 10-K Sample—Machine-Learning Sentiment Measures

We re-estimate Equation (1) replacing $SENTIMENT_{i,t}$ with $RF\ 10\text{-K}\ CAR[0,1]_{i,t}$, $SVR\ 10\text{-K}\ CAR[0,1]_{i,t}$, $sLDA\ 10\text{-K}\ CAR[0,1]_{i,t}$, and $ML\ 10\text{-K}\ CAR[0,1]_{i,t}$. We present the results in Panel B of Table 3. We find that the coefficients on $RF\ 10\text{-K}\ CAR[0,1]_{i,t}$, $sLDA\ 10\text{-K}\ CAR[0,1]_{i,t}$, and $ML\ 10\text{-K}\ CAR[0,1]_{i,t}$ are positive and statistically significant at the 1% level in both the early (i.e., 1996–2008) and full sample (i.e., 1996–2019). The coefficient on $SVR\ 10\text{-K}\ CAR[0,1]_{i,t}$ is statistically significant at the 10% level in the early sample but insignificant in the full sample. When inspecting the adjusted- R^2 for the full sample period, we note that $RF\ 10\text{-K}\ CAR[0,1]_{i,t}$ yields the highest adjusted- R^2 (0.493%) across all dictionary-based and machine-learning measures included in Panel B of Table 3, including the combined machine-learning measure ($ML\ 10\text{-K}\ CAR[0,1]_{i,t}$).¹¹ The adjusted- R^2 for the model including $RF\ 10\text{-K}\ CAR[0,1]_{i,t}$ is statistically higher than the adjusted- R^2 for the model including $LM\ 10\text{-K}\ TONE_{i,t}$ using a Vuong likelihood ratio test (p -value equal to 0.059); however, the economic significance of the difference in explanatory power is not large. The relatively small adjusted- R^2 s in the 10-K setting are likely due to the weak stock market reaction to 10-K filings documented in Li and Ramesh (2009). Thus, any improvement in measurement error for sentiment measures is unlikely to yield economically large improvements in explanatory power in the 10-K market reaction tests.¹²

To put these results in perspective, in Figure 1 we graph the average two-day 10-K filing return ($10\text{-K}\ CAR[0,1]_{i,t}$) for quintiles based on $HARV\ 10\text{-K}\ TONE_{i,t}$, $LM\ 10\text{-K}$

$TONE_{i,t}$, and $RF\ 10\text{-K}\ CAR[0,1]_{i,t}$. These sentiment measures yield the highest adjusted- R^2 when using the LM dictionary, Harvard Dictionary, and machine-learning methods. We expect to observe a monotonic increase in average returns as the quintiles for each measure increase if these measures capture 10-K filing sentiment. A steeper line suggests that the sentiment measure better captures 10-K filing returns. Both $HARV\ 10\text{-K}\ TONE_{i,t}$ and $LM\ 10\text{-K}\ TONE_{i,t}$ yield relatively flat lines, whereas the line for $RF\ 10\text{-K}\ CAR[0,1]_{i,t}$ is steeper and monotonic, with the exception of the fifth quintile for which there is a slight decrease in the average return from the fourth to fifth quintiles. These results provide graphical evidence that $RF\ 10\text{-K}\ CAR[0,1]_{i,t}$ better captures 10-K filing returns than the dictionary-based measures and thus is a superior measure of 10-K sentiment, although we recognize the inherent limitation of market reaction tests in the 10-K setting.

4.4. Conference Call Sample

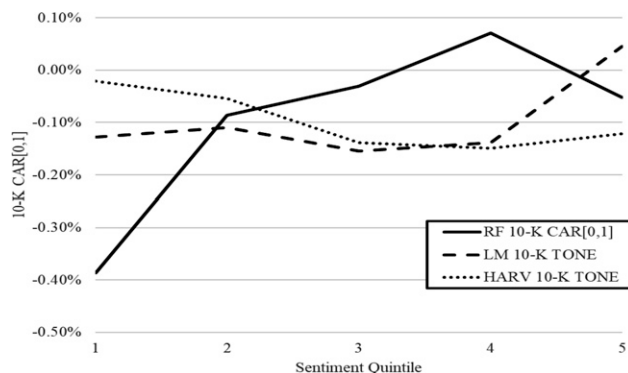
Our previous analysis uses 10-K reports to examine the incremental power of dictionary-based and machine-learning methods to extract signals of sentiment from disclosures based largely on an attempt to replicate seminal research in this area. However, more recent research extends results to capture sentiment in various other contexts. We therefore extend our analysis to the conference call setting to examine whether machine-learning methods have superior power for capturing sentiment in a context which elicits greater stock price movement relative to the 10-K setting (Li and Ramesh 2009). The incremental benefit of using more sophisticated machine-learning techniques to capture disclosure sentiment relative to dictionary-based methods may have different implications in the conference call setting relative to the 10-K setting. Prior research suggests that conference calls are more spontaneous and dynamic than other disclosure settings and that investors respond to information released during conference calls (Matsumoto et al. 2011).

We use Equation (2) to examine the extent to which dictionary-based and machine-learning sentiment measures relate to conference call returns. All variables are calculated at the firm-quarter level.

$$\begin{aligned} CC\ CAR[0,1]_{i,q} = & \alpha_0 + \alpha_1 SENTIMENT_{i,q} \\ & + \alpha_2 EARN\ SURP_{i,q} + \alpha_3 \ln(MVE_{i,q}) \\ & + \alpha_4 BTM_{i,q} + \alpha_5 TURNOVER_{i,q} \\ & + \alpha_6 PRE_FFALPHA_{i,q} \\ & + \alpha_7 INSTOWN_{i,q} + \alpha_8 NASDAQ_i + \varepsilon_{i,q} \end{aligned} \quad (2)$$

$CC\ CAR[0,1]_{i,q}$ is equal to the two-day cumulative abnormal return on day t and day $t+1$ relative to the date of the conference call. In Panel A of Table 4, we estimate Equation (2), replacing $SENTIMENT_{i,q}$ with

Figure 1. 10-K Filing Returns and Sentiment



Note. Figure 1 plots the average value of $10\text{-K}\ CAR[0,1]$ for quintiles based on $RF\ 10\text{-K}\ CAR[0,1]$, $LM\ 10\text{-K}\ TONE$, and $HARV\ 10\text{-K}\ TONE$.

Table 4. Sentiment Measures and Conference Call Returns

Panel A: Comparison of tone and random forest measures			
	(1)	(2)	(3)
Intercept	−0.014*** (−7.678)	−0.037*** (−17.372)	0.011*** (6.364)
LM CC $TONE_{i,q}$	0.063*** (40.709)		
HARV CC $TONE_{i,q}$		0.094*** (27.363)	
RF CC $CAR[0,1]_{i,q}$			3.433*** (76.888)
EARN $SURP_{i,q}$	1.314*** (33.994)	1.369*** (34.830)	1.138*** (31.596)
$\ln(MVE)_{i,q}$	−0.001*** (−2.930)	0.000 (0.977)	−0.002*** (−10.258)
$BTM_{i,q}$	0.008*** (9.404)	0.007*** (8.720)	0.008*** (10.119)
$TURNOVER_{i,q}$	−0.000*** (−2.671)	−0.001*** (−5.019)	−0.000** (−1.973)
$PRE_FFALPHA_{i,q}$	−0.508** (−2.281)	−0.072 (−0.327)	−3.597*** (−15.619)
$INSTOWN_{i,q}$	0.004*** (4.729)	0.008*** (9.419)	−0.001 (−0.684)
$NASDAQ_i$	−0.002*** (−3.386)	−0.003*** (−4.760)	0.000 (0.283)
#OBS	106,151	106,151	106,151
Adjusted R^2	6.382%	5.131%	12.676%
Panel B: Comparison of all tone and machine learning measures			
Sentiment measure	Coefficient (t -statistic)		Adj. R^2
LM CC $TONE_{i,q}$	0.063*** (40.709)		6.382%
LM CC POS $TONE_{i,q}$	2.401*** (33.889)		5.573%
LM CC NEG $TONE_{i,q}$	−3.168*** (−31.008)		5.325%
HARV CC $TONE_{i,q}$	0.094*** (27.363)		5.131%
HARV CC POS $TONE_{i,q}$	0.701*** (20.510)		4.679%
HARV CC NEG $TONE_{i,q}$	−1.093*** (−21.249)		4.726%
RF CC $CAR[0,1]_{i,q}$	3.433*** (76.888)		12.676%
SVR CC $CAR[0,1]_{i,q}$	0.211*** (50.672)		7.499%
sLDA CC $CAR[0,1]_{i,q}$	1.423*** (47.545)		7.058%
ML CC $CAR[0,1]_{i,q}$	0.018*** (74.688)		12.452%

Notes. This table includes observations from 2004–2019 and reports results of regressions where the dependent variable is equal to the cumulative abnormal return during the $[0, +1]$ trading window surrounding the conference call date ($CC\ CAR[0,1]_{i,t}$). Panel A reports results for the net tone and random forest measures including control variables. Panel B includes a comparison of all tone and machine learning measures. Control variables are also included in the regressions in Panel B but are not reported for brevity. Standard errors are clustered by firm. All variables are defined in the appendix. All continuous variables are winsorized at the 1% and 99% levels. Adj., adjusted.

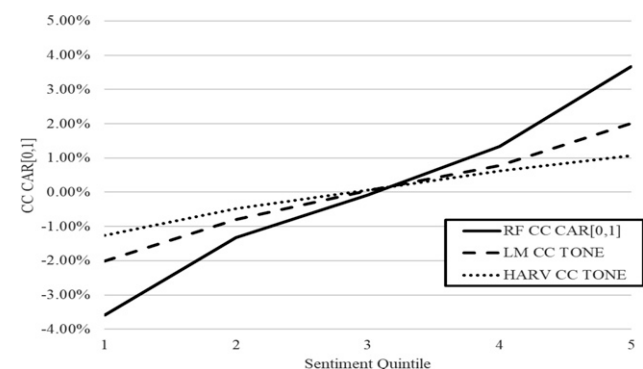
*, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

LM CC $TONE_{i,q}$, HARV CC $TONE_{i,q}$, and RF CC $CAR[0,1]_{i,q}$ in columns (1)–(3), respectively. We include similar control variables as those in Equation (1) to reduce the likelihood that the dictionary-based and machine-learning sentiment measures capture information related to other observable firm characteristics on the conference call date. In addition to the control variables included in Equation (1), we also include the earnings surprise ($EARN\ SURP_{i,q}$) for firm i in quarter q as a control variable, because the conference call tends to occur soon after the earnings announcement date.

As expected, the coefficients on all three sentiment measures are positive and significant at the 1% level. However, there are notable differences in the explanatory power of each model.¹³ HARV CC $TONE_{i,q}$ has the lowest adjusted- R^2 equal to 5.13% in column (2). The adjusted- R^2 for LM CC $TONE_{i,q}$ is equal to 6.38% and is approximately 24.1% larger than the adjusted- R^2 for HARV CC $TONE_{i,q}$. Using a Vuong test, we find that the difference in adjusted- R^2 is significant at the 1% level. The adjusted- R^2 for RF CC $CAR[0,1]_{i,q}$ is equal to 12.68% and is approximately 98.6% larger than the adjusted- R^2 for LM CC $TONE_{i,q}$. Using a Vuong test, we find that the difference in adjusted- R^2 is significant at the 1% level. These results suggest that the RF machine-learning measure yields a significant improvement in the ability to capture the sentiment of the conference call as reflected in the stock price movement surrounding the conference call date.

Similar to Figure 1, we graph the average two-day conference call return ($CC\ CAR[0,1]_{i,q}$) by quintile of LM CC $TONE_{i,q}$, HARV CC $TONE_{i,q}$, and RF CC $CAR[0,1]_{i,q}$ in Figure 2. A steeper line suggests that the measure better captures the information released in the conference call. The RF CC $CAR[0,1]_{i,q}$ measure yields the steepest line of the three measures, which provides graphical evidence that the RF method better captures conference call sentiment relative to the dictionary-based methods.

Figure 2. Conference Call Returns and Sentiment



Note. Figure 2 plots the average value of $CC\ CAR[0,1]$ for quintiles based on RF CC $CAR[0,1]$, LM CC $TONE$, and HARV CC $TONE$.

We also tabulate the results for all dictionary-based and machine-learning sentiment measures in Panel B of Table 4. Coefficients for all sentiment measures are significant at the 1% level with the expected sign. These results suggest that the dictionary-based and machine-learning measures more consistently capture disclosure sentiment in a conference call setting, which tends to elicit larger stock market reactions than in a 10-K setting. The positive and negative tone measures explain approximately the same amount of the variation in conference call returns as the net tone measure, which includes positive and negative words. Of the machine-learning measures, $RF\ CC\ CAR[0,1]_{i,q}$ explains the highest percentage of the variation in conference call returns, with an adjusted- R^2 equal to 12.68%. $SVR\ CC\ CAR[0,1]_{i,q}$ and $sLDA\ CC\ CAR[0,1]_{i,q}$ only explain approximately 7.50% and 7.06% of the variation in conference call returns.¹⁴ In fact, $ML\ CC\ CAR[0,1]_{i,q}$ explains slightly less of the variation in conference call returns (12.45%) than $RF\ CC\ CAR[0,1]_{i,q}$, which suggests that the RF method yields a superior measure of sentiment than a measure that extracts signals using several different machine-learning models.

In summary, machine-learning measures explain returns around the conference call and 10-K filing dates better than dictionary-based sentiment measures based on the LM and Harvard Dictionaries. Among the three machine-learning models, the RF measure best explains returns around the conference call and 10-K filing dates and therefore provides the most powerful measure of disclosure sentiment. In addition, neither the SVR measure nor the sLDA measure provide additional incremental information beyond that extracted by the RF measure. Thus, we recommend that future research use the RF method to capture disclosure sentiment in both the conference call and 10-K settings. In all subsequent analyses, we focus on the net tone and RF machine-learning measures to provide additional evidence on the relative usefulness of the dictionary and machine-learning approaches.

5. Additional Analyses

5.1. Future Earnings Surprises and Future Cumulative Abnormal Returns

Although our primary focus in this paper is to examine the ability of the dictionary-based and machine-learning measures to capture disclosure sentiment, an additional benefit of these measures is their potential to identify information that may not be readily apparent to investors. We therefore examine whether the dictionary-based sentiment measures and machine-learning measures predict earnings surprises and stock market returns at the next earnings announcement date. We define $FUT\ EARN\ SURP_{i,t}$ as the difference between actual earnings per share reported in

quarter $q + 1$ less the mean analyst estimate of earnings per share available immediately before the quarter $q + 1$ earnings announcement date scaled by stock price at the end of quarter q . We define $FUT\ EA\ CAR[0,1]_{i,t}$ as the cumulative market-adjusted return for firm i on day t and day $t + 1$ relative to the earnings announcement date in quarter $q + 1$. Using a similar process that we use to train the machine-learning models for concurrent returns in our previous analysis, we retrain the RF models to predict both $FUT\ EARN\ SURP_{i,t}$ and $FUT\ EA\ CAR[0,1]_{i,t}$. We create separate predictions for each of these variables using both the text of the 10-K filing and the text of the conference call transcript. We label the predictions of the future earnings surprise as $RF\ FUT\ EARN\ SURP_{i,t}$ and the predictions of the future earnings announcement cumulative abnormal return as $RF\ FUT\ EA\ CAR[0,1]_{i,t}$.

In Table 5, we report results of estimating models including $FUT\ EARN\ SURP_{i,t}$ as the dependent variable and each disclosure content measure ($LM\ TONE$, $HARV\ TONE$, and $RF\ FUT\ EARN\ SURP$) as the primary independent variable of interest. We include all prior control variables in the models with the addition of $DISPERSION$ and $REVISIONS$, following Loughran and McDonald (2011). The sample size for this analysis reduces to 53,749 observations with the inclusion of these additional control variables. In the first three columns, we present results using the 10-K sample. We find the coefficient on the $RF\ FUT\ EARN\ SURP$ measure is positive and significant, which suggests that this measure predicts future earnings surprises. The adjusted- R^2 of the model including $RF\ FUT\ EARN\ SURP$ is equal to 5.775%. Contrary to expectations, the coefficients on the $LM\ 10-K\ TONE$ and $HARV\ 10-K\ TONE$ are negative and significant. In the last three columns, we present results using the conference call sample. We find that the coefficient on the $RF\ FUT\ EARN\ SURP$ measure is positive and significant, but the coefficients on the $LM\ CC\ TONE$ and $HARV\ CC\ TONE$ measures are insignificant. The adjusted- R^2 of the model including $RF\ FUT\ EARN\ SURP$ is equal to 5.791%. Overall, these results suggest that only the machine-learning sentiment measures are positively correlated with information that is unexpected at the subsequent earnings announcement date. We therefore further explore whether this predictive ability results in significant hedge portfolio returns.

We first create equal-weighted $FUT\ EA\ CAR[0,1]_{i,t}$ portfolio returns for quintiles based on each 10-K disclosure measure ($LM\ 10-K\ TONE_{i,t}$, $HARV\ 10-K\ TONE_{i,t}$, and $RF\ FUT\ EA\ CAR[0,1]_{i,t}$). To avoid look-ahead bias, we assign observations to quintiles based on rankings relative to all other observations in the sample over the prior 365 days. We expect the average return in each quintile to increase monotonically from the first to fifth quintile for each disclosure sentiment

Table 5. Predicting Future Earnings Surprises

	10-K sample			Conference call sample		
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Intercept</i>	0.001* (1.715)	0.001** (2.422)	0.001** (2.110)	0.000 (0.083)	−0.000 (−0.202)	0.000 (0.275)
<i>LM (10-K OR CC) TONE_{i,t}</i>	−0.001*** (−3.509)			0.000 (1.256)		
<i>HARV (10-K OR CC) TONE_{i,t}</i>		−0.001** (−2.046)			0.001 (1.043)	
<i>RF FUT EARN SURP_{i,t}</i>			0.324*** (7.637)			0.494*** (10.067)
<i>EARN SURP_{i,t}</i>				0.182*** (14.130)	0.182*** (14.140)	0.175*** (13.819)
<i>ln(MVE)_{i,t}</i>	−0.000 (−0.402)	−0.000 (−0.443)	−0.000 (−0.322)	0.000 (0.912)	0.000 (0.981)	0.000 (0.880)
<i>BTM_{i,t}</i>	0.000 (0.504)	0.000 (0.721)	0.000 (0.848)	−0.000 (−0.616)	−0.000 (−0.603)	−0.000 (−0.500)
<i>TURNOVER_{i,t}</i>	0.000* (1.876)	0.000* (1.944)	0.000* (1.913)	0.000*** (3.266)	0.000*** (3.215)	0.000*** (3.473)
<i>PRE_FFALPHA_{i,t}</i>	0.114** (2.309)	0.116** (2.346)	0.109** (2.214)	0.088* (1.914)	0.089* (1.944)	0.090** (1.963)
<i>INSTOWN_{i,t}</i>	0.000 (0.674)	0.000 (0.520)	0.000 (0.181)	0.000 (1.003)	0.000 (1.122)	0.000 (0.315)
<i>NASDAQ_i</i>	−0.000** (−2.135)	−0.000** (−2.001)	−0.000** (−1.985)	0.000 (0.380)	0.000 (0.311)	−0.000 (−0.448)
<i>DISPERSION_{i,t}</i>	−0.403*** (−10.985)	−0.400*** (−10.942)	−0.370*** (−10.238)	−0.249*** (−7.527)	−0.250*** (−7.556)	−0.234*** (−7.187)
<i>REVISIONS_{i,t}</i>	0.058*** (3.588)	0.059*** (3.633)	0.058*** (3.634)	0.039*** (2.965)	0.039*** (2.976)	0.043*** (3.255)
#OBS	53,749	53,749	53,749	90,518	90,518	90,518
Adjusted R ²	5.191%	5.181%	5.775%	5.416%	5.416%	5.791%

Notes. This table reports results of regressions where the dependent variable is equal to the earnings surprise for the quarter immediately following the 10-K filing or conference call date (*FUT EARN SURP_{i,t}*). Standard errors are clustered by firm. All variables are defined in the appendix. All continuous variables are winsorized at the 1% and 99% levels.

*, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

measure. We report the results in columns (1)–(3) of Table 6. We graph the average return for each quintile in Figure 3 to more easily detect monotonic increases in average returns across quintiles. We do not find evidence of a monotonic increase in average returns from the first to fifth quintiles of the *LM 10-K TONE_{i,t}* and *HARV 10-K TONE_{i,t}* measures. When we subtract the average return in quintile one from the average return

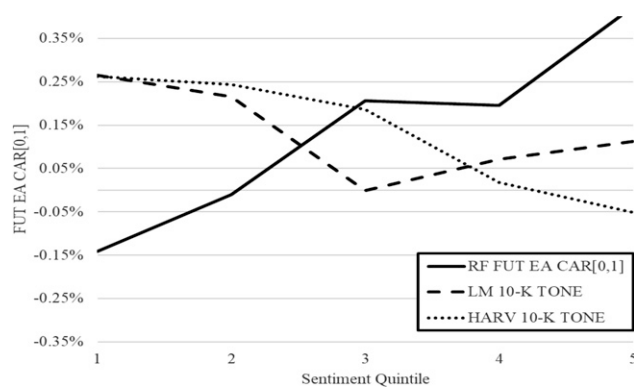
in quintile five, we find that the change in portfolio returns for the *LM 10-K TONE_{i,t}* and *HARV 10-K TONE_{i,t}* measures are negative, which is opposite expectations. These results suggest that neither the LM nor Harvard Dictionaries yield positive hedge portfolio returns. In contrast, consistent with expectations, we find that *RF FUT EA CAR[0,1]_{i,t}* is monotonically increasing from the first to fifth quintile in Figure 3. The change in the

Table 6. Predicting Future Earnings Announcement Returns

Quintile	10-K sample			Conference call sample		
	<i>LM 10-K TONE</i>	<i>HARV 10-K TONE</i>	<i>RF FUT EA CAR[0,1]</i>	<i>LM CC TONE</i>	<i>HARV CC TONE</i>	<i>RF FUT EA CAR[0,1]</i>
1	0.266%	0.262%	−0.142%	−0.083%	0.073%	−0.634%
2	0.216%	0.243%	−0.010%	−0.160%	0.022%	−0.009%
3	−0.001%	0.186%	0.207%	−0.056%	0.009%	0.043%
4	0.072%	0.017%	0.196%	0.015%	−0.052%	0.138%
5	0.112%	−0.052%	0.425%	0.178%	−0.149%	0.396%
Quintile 5–quintile 1	−0.153%	−0.314%***	0.566%***	0.261%***	−0.222%**	1.030%***
(<i>t</i> -statistic)	(−1.59)	(−3.42)	(5.31)	(2.98)	(−2.52)	(10.94)

Notes. This table reports mean values of the cumulative abnormal return during the [0, +1] trading window surrounding the next earnings announcement date immediately following the 10-K filing or conference call date (*FUT EA CAR[0,1]_{i,t}*) based on quintiles of each sentiment measure. All variables are defined in the appendix.

*, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

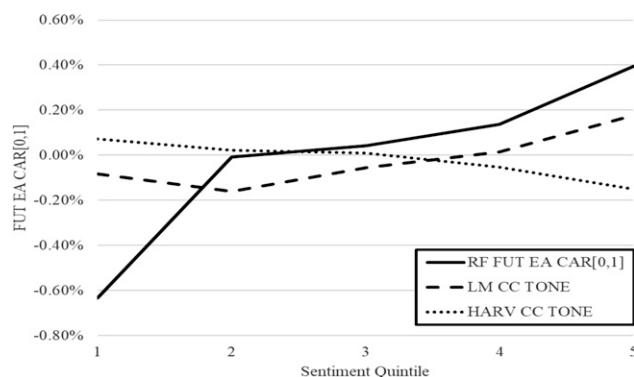
Figure 3. Future Earnings Announcement Returns and 10-K Sentiment

Note. Figure 3 plots the average value of $FUT EA CAR[0,1]$ for quintiles based on $RF 10-K CAR[0,1]$, $LM 10-K TONE$, and $HARV 10-K TONE$.

average return from the first to fifth quintile is equal to 0.566% and is significant at the 1% level, which suggests that taking a long position in the fifth quintile and a short position in the first quintile of $RF FUT EA CAR[0,1]_{i,t}$ yields a positive hedge portfolio return.

We perform a similar analysis with the conference call sample in columns (4)–(6) of Table 6 and Figure 4. We report and plot the average return for each quintile of $LM CC TONE_{i,q}$, $HARV CC TONE_{i,q}$, and $RF FUT EA CAR[0,1]_{i,q}$. Figure 4 provides graphical evidence that $RF FUT EA CAR[0,1]_{i,q}$ yields a monotonic increase in average returns across quintiles and also represents the steepest line. The line for $LM CC TONE_{i,q}$ is increasing nearly monotonically but is less steep, and the line for $HARV CC TONE_{i,q}$ is nearly flat or somewhat decreasing.

The change in average returns from the first to fifth quintile for $RF FUT EA CAR[0,1]_{i,q}$ is the largest among the three measures in columns (4)–(6) of Table

Figure 4. Future Earnings Announcement Returns and Conference Call Sentiment

Note. Figure 4 plots the average value of $FUT EA CAR[0,1]$ for quintiles based on $RF CC CAR[0,1]$, $LM CC TONE$, and $HARV CC TONE$.

6. Taking a long position in the fifth quintile and a short position in the first quintile of the $RF FUT EA CAR[0,1]_{i,q}$ quintile yields a positive and significant (1% level) hedge portfolio return of 1.030% per quarter, which would be an annual abnormal return of 4.121%. The change in the average return from the first to fifth quintile for $LM CC TONE_{i,q}$ is also significantly positive (1% level) and equal to 0.261%, which is 3.94 times smaller than that for $RF FUT EA CAR[0,1]_{i,q}$. Inconsistent with expectations, the change in average returns from the first to fifth quintile for $HARV 10-K TONE_{i,t}$ is significantly negative at the 5% level. Overall, our results suggest that the RF method better predicts future earnings surprises and returns than the dictionary-based measures used in prior research.

5.2. Alternative Disclosures—Earnings Announcement Press Releases

As an additional test, we extend our analysis to earnings announcement press releases to examine whether our inferences extend to other disclosures. We apply the same procedure we use in estimating the dictionary-based sentiment measures and the machine-learning models in the conference call setting to the earnings announcement press release setting. We identify 177,900 firm-quarter earnings announcement 8-K press releases between 2004 and 2019 with sufficient data to calculate the dependent and independent variables. The coefficients on the sentiment measures based on the LM dictionary, Harvard Dictionary, and the RF model are positive and significant, suggesting that the dictionary-based and machine-learning measures capture earnings announcement press release sentiment. The adjusted- R^2 , when the RF measure is the independent variable, yields the highest adjusted- R^2 . We more thoroughly describe and report these results in the online appendix.

6. Conclusion

We compare the ability of dictionary-based and machine-learning measures to capture the sentiment of 10-Ks and conference calls. Our study is spurred by Loughran and McDonald (2011), who find that their dictionary better explains 10-K filing returns than the Harvard Psychosociological Dictionary. We add to prior research in four key ways. First, we provide evidence that machine-learning methods produce more powerful and robust disclosure sentiment measures in conference calls, 10-K filings, and earnings-announcement press releases than dictionary-based measures. Our results guide future research that seeks to capture disclosure sentiment. Given the abundance

of research examining disclosure sentiment, reducing measurement error is likely to improve inferences.

Second, the sentiment measures based on the Loughran and McDonald (2011) and Harvard Dictionaries do not consistently correlate with market reactions to 10-K filings. Specifically, the dictionary-based sentiment measures are unassociated with 10-K filing returns in recent years. In contrast, machine-learning measures consistently capture the sentiment of 10-K filings across time. Third, our results suggest that the random-forest-regression-tree method produces measures that capture sentiment with the least measurement error relative to measures based on alternative machine-learning methods (i.e., support-vector regression and supervised-latent-Dirichlet allocation). Because of the costs to researchers of estimating multiple machine-learning models (e.g., learning costs, computing time, etc.), our results suggest that future researchers can focus on the random-forest method to capture sentiment with the least amount of error. Fourth, we provide evidence that machine-learning methods capture information consistent with investor reactions to disclosures and capture information initially overlooked by investors. Measures based on the random-forest-regression-tree method predict future earnings surprises and yield positive future hedge-portfolio returns.

Our results highlight several guiding principles for future researchers. First, because dictionary construction requires significant researcher judgment, dictionaries often miss important disclosure features. Machine-learning methods, in contrast, use words and phrases that may not be apparent to the researcher. Thus, machine-learning methods may capture disclosure signals with significantly less measurement error, as we have found to be the case with disclosure sentiment. However, machine-learning methods are subject to the criticism that they extract signals without reliance on intuition. In contrast, using intuition in dictionary construction is thought to limit overfitting or spurious relations. However, this spurious correlation problem in machine learning can be reduced by use of out-of-sample tests. Using any research method, researchers should avoid iterative out-of-sample tests based on results produced. Iterative testing, not machine-learning methods per se, produces overfitting. Dictionaries can also be improperly adjusted by iterative testing. Therefore, researchers constructing dictionaries are subject to the same pitfalls as users of machine-learning methods. Based on these conclusions, we recommend researchers use machine-learning methods to reduce measurement error, thus limiting the likelihood of both type I and type II errors (Jennings et al. 2021).

Second, dictionary-based methods assume that word lists created for financial contexts can be used in any disclosure setting (e.g., the word “question” from

the LM dictionary is considered a negative word in the 10-K setting but is unlikely to convey a similar meaning in the conference call setting) or time period. In contrast, machine-learning methods use rolling training windows within a specific setting thus adapting to differences in disclosures across time and setting. We therefore recommend that future researchers either use machine-learning methods or update their dictionaries over time and for each disclosure setting.

Third, we recommend that implementation difficulty should not influence a researcher’s decision regarding the choice between machine-learning and dictionary-based methods. Although machine-learning methods initially require a greater upfront investment, once the researcher obtains the required skill set, the difference in execution difficulty between machine-learning and dictionary-based methods is negligible. Fourth, in some instances, machine-learning models can be difficult to replicate because of the many parameter choices that go into these models (Loughran and McDonald 2016). For this reason, we provide a detailed online appendix discussing the decisions that we made when setting the parameters used in each machine-learning model. We recommend future research use complete transparency with detailed descriptions of each parameter used in model estimation to enhance replicability.

Despite the advantages of machine-learning methods, dictionary methods remain useful when the researcher seeks to examine a specific narrative attribute during the time period in which the dictionary is built and when the attribute has no observable measure beyond the words themselves to use in training a machine-learning model. For example, Bochkay et al. (2020) use a dictionary to test whether *extreme language* in earnings conference calls is associated with trading volume and stock returns. Another example is Suslava (2021) who examines whether the use of *euphemisms* causes investors to underreact to negative information in the earnings announcement.

Overall, we conclude that machine-learning methods produce a more powerful measure of disclosure sentiment than measures based on dictionaries across multiple disclosure settings and across time. We, therefore, encourage future research to consider machine learning as a way to reduce measurement error in capturing disclosure attributes.

Acknowledgments

This paper has benefited significantly from comments by seminar participants at the Brigham Young University Accounting Research Symposium, Brigham Young University, Dopuch Accounting Conference (Washington University), University of Houston, University of Minnesota, University of Michigan, Northwestern University, University of Southern California, and Washington University in St. Louis. The authors thank Julie Steiff for editing.

Appendix. Variable Definitions

Variable	Definition
<i>10-K CAR</i> [0,1]	The cumulative market-adjusted return for the firm in the [0,1] trading window surrounding the current-year 10-K filing date
<i>BTM</i>	Book to market ratio calculated as the firm's book value of common equity at the end of quarter or year divided by <i>MVE</i>
<i>CC CAR</i> [0,1]	The cumulative market-adjusted return for the firm in the [0,1] trading window surrounding the current-quarter conference call date
<i>DICT DISC NEG TONE</i>	The number of negative words divided by the total number of words in the <i>DICT</i> disclosure, where the negative words are determined based on the <i>DICT</i> dictionary. <i>DICT</i> is either <i>HARV</i> (Harvard Psychosociological Dictionary) or <i>LM</i> (Loughran and McDonald (2011) dictionary) and <i>DISC</i> is either <i>10-K</i> (10-K filing) or <i>CC</i> (conference call)
<i>DICT DISC POS TONE</i>	It is the number of positive words divided by the total number of words in the <i>DICT</i> disclosure, where the positive words are determined based on the <i>DICT</i> dictionary. <i>DICT</i> is either <i>HARV</i> (Harvard Psychosociological Dictionary) or <i>LM</i> (Loughran and McDonald (2011) dictionary), and <i>DISC</i> is either <i>10-K</i> (10-K filing) or <i>CC</i> (conference call).
<i>DICT DISC TONE</i>	It is the number of positive words minus the number of negative words divided by the total number of positive and negative words in the <i>DISC</i> disclosure, where the positive and negative words are determined based on the <i>DICT</i> dictionary. <i>DICT</i> is either <i>HARV</i> (Harvard Psychosociological Dictionary) or <i>LM</i> (Loughran and McDonald (2011) dictionary), and <i>DISC</i> is <i>10-K</i> (10-K filing), <i>CC</i> (conference call), or <i>EA</i> (earnings announcement press release).
<i>DISPERSION</i>	The standard deviation of analyst earnings per share forecasts used in the computation of <i>FUT EARN SURP</i> scaled by the firm's stock price at the end of the period
<i>EA CAR</i> [0,1]	The cumulative market-adjusted return for the firm in the [0,1] trading window surrounding the current-quarter earnings announcement date
<i>EARN SURP</i>	It is the earnings per share for the firm in the current quarter less the mean earnings per share forecast for the firm made prior to the current-quarter earnings announcement date scaled by the firm's stock price at the end of the quarter. For each analyst, we use the latest forecast prior to the current-quarter earnings announcement date, removing any forecasts made more than 90 days prior to the earnings announcement date.
<i>FUT EA CAR</i> [0,1]	The cumulative market-adjusted return for the firm in the [0,1] trading window surrounding the earnings announcement date in the next quarter following the current-quarter conference call date or current-year 10-K filing date
<i>FUT EARN SURP</i>	It is the earnings per share for the firm in the next quarter following the current-quarter conference call date or current-year 10-K filing date less the mean earnings per share forecast for the firm made prior to the next earnings announcement date scaled by the firm's stock price at the end of the next quarter. For each analyst, we use the latest forecast prior to the next-quarter earnings announcement date, removing any forecasts made more than 90 days prior to the next earnings announcement date.
<i>INSTOWN</i>	The percentage of shares of the firm held by institutional investors
<i>ML 10-K CAR</i> [0,1]	The factor obtained when estimating a factor analysis on <i>SVR 10-K CAR</i> [0,1], <i>sLDA 10-K CAR</i> [0,1], and <i>RF 10-K CAR</i> [0,1]
<i>ML CC CAR</i> [0,1]	The factor obtained when estimating a factor analysis on <i>SVR CC CAR</i> [0,1], <i>sLDA CC CAR</i> [0,1], and <i>RF CC CAR</i> [0,1]
<i>MLMETHOD DISC CAR</i> [0,1]	The out-of-sample <i>MLMETHOD</i> prediction of <i>CAR</i> [0,1] for the firm in the current quarter or year using the counts of one- and two-word phrases of the <i>DISC</i> disclosure, where <i>MLMETHOD</i> is <i>SVR</i> (support-vector regression), <i>sLDA</i> (supervised latent-Dirichlet-allocation), or <i>RF</i> (random forest regression trees) and <i>DISC</i> is either <i>10-K</i> (10-K), <i>CC</i> (conference call), or <i>EA</i> (earnings announcement press release)
<i>MVE</i>	Market value of equity for the firm in the current quarter or year calculated as the firm's stock price multiplied by the number of shares outstanding at the end of the quarter or year
<i>NASDAQ</i>	An indicator variable equal to the one if the firm is traded on the NASDAQ and equal to zero otherwise
<i>PRE_FFALPHA</i>	It is the Fama–French alpha based on a regression of their three-factor model using trading days [−252, −6] relative to the conference call or 10-K filing date. At least 60 observations of daily returns must be available to be included in the sample.
<i>REVISIONS</i>	It is the mean revision in analyst earnings per share forecasts for the next quarter following the current-quarter conference call date or current-year 10-K filing date scaled by the firm's stock price. The revision is based on forecasts outstanding four months prior to the next quarter earnings announcement date relative to forecasts outstanding just prior to the next quarter earnings announcement date.
<i>RF FUT EA CAR</i> [0,1]	The out-of-sample random-forest-regression-tree prediction of <i>FUT EA CAR</i> [0,1] using the counts of one- and two-word phrases of the 10-K or conference call
<i>RF FUT EARN SURP</i>	The out-of-sample random-forest-regression-tree prediction of <i>FUT EARN SURP</i> using the counts of one- and two-word phrases of the 10-K or conference call
<i>TURNOVER</i>	The number of shares traded for the firm in the trading days [−252, −6] relative to the conference call or 10-K filing date divided by the firm's shares outstanding at the conference call or 10-K filing date

Endnotes

¹ The meaning of “sentiment” is unclear. Loughran and McDonald (2011, p. 36) do not define sentiment but link it to “negative implications in a financial sense.” Their use of the term stems from its use by Tetlock (2007) and DeLong et al. (1990) who associate sentiment with “noninformation” or noise trading. We follow Loughran and McDonald’s (2011) sense that sentiment reflects positive or negative attitudes conveyed by the text regarding a company’s prospects. Whether these expectations are grounded in economic substance or mood, Loughran and McDonald (2011), Tetlock (2007), and DeLong et al. (1990) agree that sentiment can influence price.

² See, for example, Chen et al. (2018a), Chen et al. (2018b), Brockman et al. (2015), Davis et al. (2015), Price et al. (2012), Huang et al. (2018), Allee and DeAngelis (2015), Blau et al. (2015), Law and Mills (2015), Rogers et al. (2011), D’Augusta and DeAngelis (2020), Feldman et al. (2010), and Davis and Tama-Sweet (2012).

³ Other papers manually classify documents to measure disclosure trends (Hooks and Moon 1993) and the favorability of the MD&A (Callahan and Smith 2004, Sun 2010).

⁴ For example, Feldman et al. (2010) associate the tone of the MD&A with the market reaction to the 10-K and portfolio drift returns measured after the 10-K’s filing. In addition, Loughran and McDonald (2011) associate the tone of the MD&A with the market reaction to the 10-K, trading volume, and return volatility. There are many other papers that use a dictionary method to estimate the tone of a financial disclosure (Tetlock 2008, Kothari et al. 2009, Rogers et al. 2011, Davis et al. 2012, Price et al. 2012, Allee and DeAngelis 2015, Huang et al. 2018). Prior research has also used the dictionary method to capture cognitive dissonance (Hobson et al. 2012), deception (Larcker and Zakolyukina 2012), changes in risk disclosures (Kravet and Muslu 2013), and competition (Li et al. 2013).

⁵ Li (2008) was among the first in the accounting and finance literature to use the Gunning FOG index to assess the readability of the 10-K and MD&A. Other research has used the Gunning FOG index to measure financial disclosure readability (e.g., Leavy et al. 2011) and attempted to improve readability proxies (e.g., Loughran and McDonald 2014, Bonsall et al. 2017). Brown and Tucker (2011) used a cosine similarity score to assess the similarity of documents over time or across firms. Other papers have also used the cosine similarity score to assess the similarity of disclosures. Lang and Stice-Lawrence (2015) use the cosine-similarity score to measure how similar annual disclosures are for non-U.S. companies. Peterson et al. (2015) also use the cosine-similarity score to measure the similarity of accounting footnotes over time and across firms.

⁶ For example, see Bao and Datta (2014), Dyer et al. (2017), Hoberg and Lewis (2017), Huang et al. (2018), and Brown et al. (2020).

⁷ For all conference call transcripts in calendar year t , the training sample consists of all conference call transcripts from the fourth quarter of year $t-5$ to the third quarter of year $t-1$. For all 10-Ks in calendar year t , the training sample consists of all 10-K transcripts from year $t-2$ to year $t-1$. Using these different training windows provides a similar number of one- and two-word phrases in each disclosure setting (i.e., 365,643 words on average across all training samples in the 10-K setting and 383,674 words across all training samples in the conference call setting). Using larger training samples increases the number of phrases included in the models and increases computing time.

⁸ Dyer et al. (2017) use 150 topics when estimating LDA, and Bao and Datta (2014) use 30 topics. We choose 200 topics to give sLDA even greater flexibility. We estimate robustness tests using 50 topics. The explanatory power of the sLDA sentiment using 50 topics is similar to the explanatory power using 200 topics in the 10-K setting. Explanatory power is slightly diminished using 50 topics in the conference call setting.

⁹ Random forests have been used in the economics literature for demand estimation (Bajari et al. 2015); however, they have not previously been applied to assessing the narrative content of financial disclosures.

¹⁰ We estimate robustness tests using 1,000, 4,000, and 6,000 regression trees. The explanatory power of the random forest measure is diminished only slightly using 1,000 trees and is similar when using 4,000 or 6,000 trees in both the conference call and 10-K settings.

¹¹ Another way to evaluate economic significance is by examining standardized coefficients. Because of the differences in the standard deviation of the independent variables, it is difficult to directly compare one coefficient to another. We calculate standardized coefficients by multiplying the coefficient by the standard deviation of the independent variable, then dividing the product by the standard deviation by the dependent variable. In Panel B of Table 3, we find that the standardized coefficient on $LM\ 10-K\ NEG\ TONE_{i,q}$ is equal to -0.0195 and $RF\ 10-K\ CAR[0,1]_{i,q}$ is equal to 0.0303 . None of the Harvard Dictionary measures are significant in the predicted direction.

¹² As an additional untabulated robustness test, we follow Loughran and McDonald (2011) and re-estimate Equation (1) with industry fixed effects and using Fama-MacBeth quarterly regressions with Newey-West standard errors with one lag. We find that all dictionary-based sentiment measures are insignificantly related to the cumulative abnormal 10-K filing return in both the early and full samples. In contrast, although the machine-learning measures are insignificant in the early 10-K sample, the SVR, sLDA, and machine learning factor are significant and positively related to the return using the full 10-K sample. We conclude that the machine-learning methods more consistently capture 10-K sentiment than dictionary-based methods using different regression specifications.

¹³ In Panel A of Table 4, we find that the standardized coefficient on $LM\ CC\ TONE_{i,q}$ is equal to 0.153 , $HARV\ CC\ TONE_{i,q}$ is equal to 0.099 , and $RF\ CC\ CAR[0,1]_{i,q}$ is equal to 0.293 .

¹⁴ In additional untabulated robustness tests, we use a two-year window to train $RF\ CC\ CAR[0,1]_{i,t}$ and a four-year window to train $RF\ 10-K\ CAR[0,1]_{i,t}$. We find qualitatively similar results using these different training windows.

References

- Allee K, DeAngelis M (2015) The structure of voluntary disclosure narratives: Evidence from tone dispersion. *J. Accounting Res.* 53 (2):241–274.
- Bajari P, Nekipelov D, Ryan S, Yang M (2015) Machine learning methods for demand estimation. *Amer. Econom. Rev.* 105(5):481–485.
- Ball R, Brown P (1968) An empirical evaluation of accounting income numbers. *J. Accounting Res.* 6(2):159–178.
- Bao Y, Datta A (2014) Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Sci.* 60(6):1371–1391.
- Beyer A, Cohen DA, Lys TZ, Walther BR (2010) The financial reporting environment: Review of the recent literature. *J. Accounting Econom.* 50(2–3):296–343.
- Blau BM, DeLisle JR, Price SM (2015) Do sophisticated investors interpret earnings conference call tone differently than investors at large? Evidence from short sales. *J. Corporate Finance* 31:203–219.
- Blei D, McAuliffe J (2007) Supervised topic models. Platt JC, Koller D, Singer Y, Roweis ST, eds. *Proc. 20th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 121–128.
- Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3:993–1022.
- Bochkay K, Hales J, Chava S (2020) Hyperbole or reality? Investor response to extreme language in earnings conference calls. *Accounting Rev.* 95(2):31–60.
- Bonsall S, Leone A, Miller B, Rennekamp K (2017) A plain English measure of financial reporting readability. *J. Accounting Econom.* 63(2):329–357.

- Botosan C (1997) Disclosure level and the cost of equity capital. *Accounting Rev.* 72(3):323–349.
- Breiman L (2001) Random forests. *Machine Learn.* 45:5–32.
- Brockman P, Li X, Price SM (2015) Differences in conference call tones: Managers vs. analysts. *Financial Anal. J.* 71(4):24–42.
- Brown S, Tucker J (2011) Large-sample evidence on firms' year-over-year MD&A modifications. *J. Accounting Res.* 49(2):309–346.
- Brown N, Crowley R, Elliott B (2020) What are you saying? Using topic to detect financial misreporting. *J. Accounting Res.* 58(1):237–291.
- Bryan S (1997) Incremental information content of required disclosures contained in management discussion and analysis. *Accounting Rev.* 72(2):285–301.
- Callahan C, Smith R (2004) Firm performance and management's discussion and analysis disclosures: An industry approach. Working paper, University of Louisville, Louisville, KY.
- Campbell J, Chen H, Dhaliwal D, Lu H, Steele L (2014) The information content of mandatory risk factor disclosures in corporate filings. *Rev. Accounting Stud.* 19:396–455.
- Cao S, Jiang W, Yang B, Zhang A (2021) How to talk when a machine is listening: Corporate disclosure in the age of AI. Working paper, Georgia State University, Atlanta.
- Chen J, Demers E, Lev B (2018a) Oh what a beautiful morning! Diurnal influences on executives and analysts: Evidence from conference calls. *Management Sci.* 64(12):5899–5924.
- Chen JV, Nagar V, Schoenfeld J (2018b) Manager-analyst conversations in earnings conference calls. *Rev. Accounting Stud.* 23(4):1315–1354.
- Core J (2001) A review of the empirical literature. [Discussion] *J. Accounting Econom.* 31:441–456.
- D'Augusta C, DeAngelis MD (2020) Does accounting conservatism discipline qualitative disclosure? Evidence from tone management in the MD&A. *Contemporary Accounting Res.* 37(4):2287–2318.
- Davis AK, Tama-Sweet I (2012) Managers' use of language across alternative disclosure outlets: Earnings press releases vs. MD&A. *Contemporary Accounting Res.* 29(3):804–837.
- Davis A, Piger J, Sedor L (2012) Beyond the numbers: Measuring the information content of earnings press release language. *Contemporary Accounting Res.* 29(3):845–868.
- Davis AK, Ge W, Matsumoto D, Zhang JL (2015) The effect of manager-specific optimism on the tone of earnings conference calls. *Rev. Accounting Stud.* 20(2):639–673.
- DeLong JB, Shleifer A, Summers LH, Waldmann RJ (1990) Noise trader risk in financial markets. *J. Political Econom.* 98(4):703–738.
- Donovan J, Koharki K, Jennings J, Lee J (2021) Measuring credit risk using qualitative disclosure. *Rev. Accounting Stud.* 26:815–863.
- Dyer T, Lang M, Stice-Lawrence L (2017) The evolution of 10-K textual disclosure: Evidence from latent Dirichlet allocation. *J. Accounting Econom.* 64(2–3):221–245.
- Feldman R, Govindaraj S, Livnat J, Segal B (2010) Management's tone change, post earnings announcement drift and accruals. *Rev. Accounting Stud.* 15(4):915–953.
- Frankel R, Jennings J, Lee J (2016) Using unstructured and qualitative disclosures to explain accruals. *J. Accounting Econom.* 45(2):209–227.
- Gentzkow M, Kelly B, Taddy M (2019) Text as data. *J. Econom. Literature* 57(3):535–574.
- Hoberg G, Lewis C (2017) Do fraudulent firms produce abnormal disclosure? *J. Corporate Finance* 43:58–85.
- Hobson J, Mayew B, Venkatachalam M (2012) Analyzing speech to detect financial misreporting. *J. Accounting Res.* 50(2):349–392.
- Hooks K, Moon J (1993) A classification scheme to examine management discussion and analysis compliance. *Accounting Horizons* 7(2):41–59.
- Huang A, Zang A, Zheng R (2014) Evidence on the information content of text in analyst reports. *Accounting Rev.* 89(6):2151–2180.
- Huang A, Leavy R, Zang A, Zheng R (2018) Analyst information discovery and interpretation roles: A topic modeling approach. *Management Sci.* 64(6):2833–2855.
- Jennings JN, Kim JM, Lee JA, Taylor DJ (2021) Measurement error and bias in causal models in accounting research. Preprint, submitted January 8, <https://dx.doi.org/10.2139/ssrn.3731197>.
- Kothari SP, Li X, Short J (2009) The effect of disclosures by management, analysts, and financial press on the equity cost of capital: A study using content analysis. *Accounting Rev.* 84(5):1639–1670.
- Kravit T, Muslu V (2013) Textual risk disclosures and investors' risk perceptions. *Rev. Accounting Stud.* 18:1088–1122.
- Law KK, Mills LF (2015) Taxes and financial constraints: Evidence from linguistic cues. *J. Accounting Res.* 53(4):777–819.
- Lang M, Stice-Lawrence L (2015) Textual analysis and international financial reporting: Large sample evidence. *J. Accounting Econom.* 60:110–135.
- Larcker D, Zakolyukina A (2012) Detecting deceptive discussions in conference calls. *J. Accounting Res.* 50:495–540.
- Leavy R, Li F, Merkley K (2011) The effect of annual report readability on analyst following and the properties of their earnings forecasts. *Accounting Rev.* 86(3):1087–1115.
- Li F (2008) Annual report readability, current earnings, and earnings persistence. *J. Accounting Econom.* 45:221–247.
- Li F (2010) The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *J. Accounting Res.* 48(5):1049–1102.
- Li F, Lundholm R, Minnis M (2013) A measure of competition based on 10-K filings. *J. Accounting Res.* 51(2):399–436.
- Li E, Ramesh K (2009) Market reaction surrounding the filing of periodic SEC reports. *Accounting Rev.* 84(4):1171–1208.
- Loughran T, McDonald B (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Finance* 66(1):35–65.
- Loughran T, McDonald B (2014) Measuring readability in financial disclosures. *J. Finance* 69(4):1643–1671.
- Loughran T, McDonald B (2016) Textual analysis in accounting and finance: A survey. *J. Accounting Res.* 54(4):1187–1230.
- Manela A, Moreira A (2017) News implied volatility and disaster concerns. *J. Financial Econom.* 123(1):137–162.
- Matsumoto D, Pronk M, Roelofsens E (2011) What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions. *Accounting Rev.* 86(4):1384–1414.
- Peterson R, Schmardebeck R, Wilks J (2015) The earnings quality and information processing effects of accounting consistency. *Accounting Rev.* 90(6):2483–2514.
- Price SM, Doran JS, Peterson DR, Bliss BA (2012) Earnings conference calls and stock returns: The incremental informativeness of textual tone. *J. Banking Finance* 36(4):992–1011.
- Rogers J, Van Buskirk A, Zechman S (2011) Disclosure tone and shareholder litigation. *Accounting Rev.* 86:2155–2183.
- Sun Y (2010) Do MD&A disclosures help users interpret disproportionate inventory increases? *Accounting Rev.* 85(4):1411–1440.
- Suslava K (2021) “Stiff business headwinds and uncharted economic waters”: The use of euphemisms in earnings conference calls. *Management Sci.*, ePub ahead of print, <https://doi.org/10.1287/mnsc.2020.3826>.
- Tetlock PC (2007) Giving content to investor sentiment: The role of media in the stock market. *J. Finance* 62(3):1139–1168.
- Tetlock PC, Saar-Tsechansky M, Macskassy S (2008) More than words: Quantifying language to measure firms' fundamentals. *J. Finance* 63(3):1437–1467.