Using Natural Language Processing to Assess Text Usefulness to Readers: The Case of Conference Calls and Earnings Prediction

Richard Frankel
Olin Business School
Washington University in St. Louis
St. Louis, MO 63130-6431
frankel@wustl.edu

Jared Jennings
Olin Business School
Washington University in St. Louis
St. Louis, MO 63130-6431
jaredjennings@wustl.edu

Joshua Lee
Terry College of Business
University of Georgia
Athens, GA 30602-6269
joshlee@uga.edu

January 2017

We examine whether support vector regressions (SVR), supervised LDA (sLDA), random forest regression trees (RF), and 'tone' extract narrative content from conference calls that correlates with useful information that a human reader would identify. We find that each narrative-content measure (along with a composite measure) explains a portion of analyst-forecast revisions for quarter $q+1$ issued after the conference call in quarter q. Correlation with analyst-forecast revisions improves when the composite measure adapts to context (positive/negative returns; high variance/low variance) and ignores sparse words. The correlation is comparable and incremental to that of financial signals (cash-flow changes, earnings surprises, and management forecasts), which suggests that the narrative content of conference calls as extracted by readers is economically significant. Our results suggest that models of narrative content have reasonable construct validity and that this validity is likely to be improved by further thought on the unique characteristics of text.

JEL codes: M40, M41
Keywords: Textual Analysis, Machine Learning, Disclosure, Conference Calls

We are thankful to Olin Business School and the Terry College of Business for financial support. This paper has benefited significantly from comments by seminar participants at the BYU Accounting Research Symposium, 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. We thank Julie Steiff for editing.

1. Introduction

We correlate conference call narrative-content measures generated by three machine-learning methods and one dictionary method with analyst-forecast revisions. Our motive is to understand whether statistical estimates of narrative content correspond to insight gained by readers and get a sense of its relative importance. We compare the information content of the narrative content measures to that of accounting measures (e.g., cash flow changes, earnings surprises, and management guidance). Linking statistical methods to analyst revisions allows us to draw conclusions regarding disclosure *usefulness* from statistical measures.[1] Our assumption is that the effect of text-disclosure on readers' actions proxies for the usefulness of the disclosure. Our findings are as follows:

1) The four measures explain analyst-forecast revisions surrounding the conference call, and a combined measure produces higher explanatory power than any single measure.
2) The explanatory power of the combined measure is higher when unusual (i.e., sparse) words are excluded from the estimation process.
3) Adapting model estimation to context (i.e., positive/negative and more-/less-volatile quarter returns) significantly improves the explanatory power of the combined measure.
4) The combined measure explains analyst revisions far better than cash flows changes, but only about 81% as well as the current earnings surprise.
5) Approximately two-thirds of the combined measure's explanatory power is incremental to the current earnings surprise.

These results suggest that statistical methods capture an economically meaningful proportion of the content of the conference call gained by readers. They also suggest future research can improve inferences by combining statistical measures, adapting models to contexts, and removing unusual words. Future research can use these methods to convert qualitative narrative into quantitative

[1] We define "useful" disclosures as those that contain information that 1) is used by analysts to update expectations and 2) accurately predicts future earnings. We do not consider information in a disclosure that biases analyst forecasts to be useful as it is unlikely to accurately predict future earnings.

forecasts.

Given the overwhelming volume of narrative disclosure, our objective is to provide guidance for research using machine-based estimates to measure the quality or usefulness of disclosures to readers (e.g., Li, 2008; Dyer et al., 2017). This objective corresponds to regulators' stated interest in assessing or improving the effectiveness of textual disclosure (FASB, 2012; SEC, 1998, 2013). For example, the FASB's invitation to comment on the Disclosure Framework begins, "The objective and primary focus of this project is to improve the effectiveness of disclosures in notes to financial statements by clearly communicating the information that is most important to *users* of each entity's financial statements" (italics added). With this objective in mind, we identify methods that future research can use to assess the usefulness of disclosures to readers by directly predicting earnings using the text of disclosures.

We adapt three machine-learning methods (i.e., support vector regression, supervised latent Dirichlet allocation, and random forest regression trees) and one dictionary method (i.e., tone) to earnings prediction using conference call transcripts.[2] Each method transforms the words in a given disclosure into a scalar designed to capture the narrative content of the disclosure with respect to a variable the firm might wish to convey (i.e., *q+1* earnings expectations). To construct our measures, we use a sample of conference call transcripts between 2004 and 2015 to predict the earnings surprise for quarter *q+1* (earnings for quarter *q+1* less the consensus analyst expectation available *before* the quarter *q* conference call) using one- and two-word phrases from the quarter *q* conference call. Note, we fit these models to the *earnings surprise* rather than the *analyst revision* so that our text content measure captures language that is useful in predicting the future earnings surprise. We then take these fitted values and predict analyst revisions to measure the usefulness

[2] We use supervised machine-learning methods to capture narrative content. Unlike unsupervised methods (e.g., LDA), supervised methods use a dependent variable to identify predictive words and phrases in a disclosure.

of text to humans.

The three machine-learning methods use different processes to extract meaning from text. Support vector regression places weights on individual words and phrases to explain a dependent variable. Supervised Latent Dirichlet Allocation groups words and phrases into topics that are predictive of a dependent variable. Random forest creates decision trees based on the words and phrases in text that best explain a dependent variable. As such, random forest identifies interactions among words and phrases to assist in explaining a dependent variable. Because humans likely extract meaning from text by placing weight on specific words/phrases, identifying topics, and identifying interactions among words/phrases, each method likely captures a different aspect of the human interpretation process.

We use rolling training samples to estimate each model using conference calls between the fourth quarter of year *t-5* and the third quarter of year *t-1*. We then apply these models to conference-call transcripts in year *t* to obtain out-of-sample predictions for the earnings surprise in quarter *q+1*.[3] We label these out-of-sample predictions *SVR* $Pred_{i,q}$, *sLDA* $Pred_{i,q}$, *RF* $Pred_{i,q}$, and *Tone* $Pred_{i,q}$. We also use factor analysis to estimate a composite measure (*FACTOR* $Pred_{i,q}$) from the four narrative content variables. We find that each narrative-content measure positively correlates with consensus analyst-forecast revisions, which suggests that these measures identify narrative content that a human reader would identify. Moreover, a composite measure (*FACTOR* $Pred_{i,q}$) explains the highest percentage of the variation in the average analyst-forecast revision at 7.9%.

Drawing on linguistics research (e.g., Manning and Schütze, 1999), we consider two salient

[3] We follow Loughran and McDonald (2011) to calculate a summary measure for conference call tone, and we use OLS to predict the future earnings surprise. SVR, sLDA, and RF use all one- and two-word phrases to estimate future earnings surprises.

linguistic features and adapt our implementation of machine-learning methods to incorporate them. First, text often contains a preponderance of sparse phrases. For instance, 92% of the 1.667 million unique phrases included in conference calls in 2015 are mentioned in fewer than 10 conference calls, yet on average, these unique phrases constitute only 8.7% of the total phrases included in the conference call. When used in model training, these infrequent phrases can induce unreliable predictions. We test this conjecture by estimating the combined narrative-content measure with and without sparse words (i.e., words that are used in fewer than 10 conference calls). We find that out-of-sample explanatory power decreases from 7.9% to 7.0% when including sparse words in the estimation model, and the difference in the R^2 is statistically significant at the 1% level based on the Vuong (1989) likelihood ratio test. These results suggest that future research can improve statistical estimates by removing sparse words from the estimation process or finding ways to reduce sparsity (e.g., by combining synonyms).

Second, linguistics theory suggests that context often determines meaning (Portner, 2005). Words can have multiple dictionary definitions (lexical semantics), and their meaning can vary with sentence structure (compositional semantics). By using dictionaries and statistical associations between word counts and outcome variables, researchers attempt a rough depiction of the meaning conveyed by these semantics. *Literal meaning* is invariant across uses, users, and situations, but *meaning* is a combination of semantics and pragmatics (context) (Stalnaker, 1970). For example, the phrase "increased demand" can have differing implications for future earnings if a manager makes this statement with regard to raw materials or the products that a firm sells. The effect can be interactive, depending on whether the firm is in the midst of positive outcomes or negative outcomes. The statement could be vacuous or evasive, if the answer is a response to an analyst asking why sales have increased. A moment's reflection produces many possible contexts.

We allow for context by estimating our models within subsamples of the data. The benefits of finer, context-based data partitions are offset by the cost of smaller samples. Moreover, attempting several partitions and choosing those that maximize "out-of-sample" explanatory power results in overfitting and calls into question the validity of the out-of-sample predictions. We allow the narrative content measures to adapt to context by estimating the models within two subsamples of the data – high/low return volatility and positive/negative news. We assume that word meanings differ when the firm experiences high or low volatility or positive or negative news in the previous quarter. When estimating the models within the relevant subsamples, we find that the explanatory power of our narrative-content measure increases to 9.5%, which represents a 20% improvement.

To assess the economic magnitude of the measure's explanatory power (9.5%), we compare it to the explanatory power of three other accounting signals that analysts likely incorporate into their forecasts. We find that the explanatory power of the narrative-content measure is 81% of the explanatory power of the concurrent earnings surprise, 28% of the explanatory power of the managerial guidance surprise, and 19.0 times larger than the explanatory power of the change in cash flows. The machine-based measure identifies economically significant earnings-relevant information. We also assess whether the narrative-content measure captures information that overlaps with these alternative signals. We find that approximately 61% of the explanatory power of the narrative-content measure is incremental to the earnings surprise and 27% is incremental to the combined explanatory power of the earnings surprise, change in cash flows, and managerial guidance surprise, reinforcing the inference that the narrative contains significant information.

Finally, fitting our models to earnings surprises rather than analyst revisions creates the

possibility that the fitted values contain information missed by analysts. Therefore, we check the extent to which these fitted values predict future returns. Our results fail to provide strong evidence that the combined narrative content measure consistently captures information that humans overlook at the time of the conference call. Decile hedge strategies yield positive returns over the next quarter and during the subsequent earnings announcement in most years. However, these returns are dramatically negative during the financial crisis in 2009. This negative return increases the standard error and reduces the mean sufficiently to make the average hedge portfolio returns during the next quarter and earnings announcement insignificant at conventional levels.

These methods can advance research in two ways. First, we introduce several methods, two of which are new (random forest regression trees and supervised LDA), that allow investors and researchers to convert qualitative narrative disclosures into quantitative forecasts.[4] Managers can provide information about future earnings through narrative disclosures that is separate from formal quantitative forecasts (e.g., managerial guidance). Thus, only focusing on the information that managers convey in guidance likely provides an incomplete picture of managerial disclosure. Future research can use the methods in this paper to 1) estimate the information content of narrative disclosure (especially when managers provide no formal guidance), 2) measure the incremental informativeness of qualitative disclosures to readers, and 3) to assess the different roles (e.g., information/verification) of qualitative disclosure. Prior research measures text characteristics

[4] Supervised latent Dirichlet allocation (sLDA) categorizes the words and phrases of a corpus into a predetermined number of categories with respect to a dependent variable (Blei and McAuliffe, 2007). Supervised LDA differs from unsupervised LDA, which has been previously used in the accounting literature (e.g., Bao and Datta, 2014; Campbell et al., 2014; Dyer et al., 2016), in that unsupervised LDA identifies topics without respect to a dependent variable. Blei and McAuliffe (2007) suggest that sLDA is a better choice than unsupervised LDA (introduced by Blei et al., 2003) when the goal is predicting a response variable, which is our goal in this paper. Random forest is a regression tree method that creates decision "trees" based on words and phrases in a corpus to predict a specific dependent variable (e.g., future earnings surprises). 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.

(e.g., readability, similarity, deception, length, and tone) and associates those characteristics with fundamentals.[5] We move the literature forward by identifying methods that can be used to covert narrative disclosures into quantitative forecasts of earnings, which can be used in future empirical analyses.

Second, we identify two ways to improve estimates of narrative content: eliminate sparse words and adapt the methods to specific contexts. We expect future research to continue to improve statistical measures to better handle issues of sparsity and context. For example, rather than deleting sparse words from the estimation of the models, perhaps researchers could develop a method of grouping sparse words to improve statistical power. In addition, rather than estimating the statistical models within broad subsamples to account for context, researchers could use greater creativity to account for context and improve the construct validity of the measures. We encourage future research to continue to identify and improve upon empirical methods that can be used to measure narrative content.

2. Prior Literature

Managers use quantitative and qualitative disclosures to provide information to investors about the past, current, or future performance of the firm (Core, 2001). Unlike quantitative disclosures, qualitative disclosures are not easily transformed into quantitative measures for use by researchers or market participants (Loughran and McDonald, 2011). Prior literature uses textual analysis techniques to develop quantitative measures from qualitative disclosures.

[5] For example, Bonsall, Leone, and Miller (2015); Brown and Tucker (2011); Lang and Stice-Lawrence (2014); Larker and Zakolyukina (2012); Li (2008); and Li, Minnis, Nagar, and Rajan (2014); Petersen, Schmardebeck, and Wilks (2015). Research proposes various disclosure-characteristic measures, such as tone, forward-looking statements, FOG, length, etc., and offers methodological refinements. These refinements include adjustments to tone dictionaries (Loughran and McDonald, 2011; Henry and Leone, 2016) and readability measures (Loughran and McDonald, 2014; Bonsall, Leone, Miller, and Rennekamp, 2017).

Li (2008) is among the first to employ textual analysis to measure disclosure readability. Specifically, Li (2008) uses the Gunning FOG index, which is a function of the number of words in a sentence, the number of syllables per word, and the length of the annual report, to measure its readability. Several papers use similar measures or develop other measures of disclosure readability.[6] Other papers use or develop dictionaries to count words to measure firm characteristics, such as tone, competition, and risk.[7]

Research also uses unsupervised machine-learning techniques to parse the language in a corpus. Unsupervised machine-learning techniques model the words in the document without fitting to a pre-determined dependent variable. Unsupervised Latent Dirichlet Allocation (LDA) is a common unsupervised method used in the prior literature to "construct features for classification" with the goal of reducing data dimensionality (Blei and McAuliffe, 2007). Unsupervised LDA categorizes the language in a corpus into a pre-determined number of topics. Researchers often examine the words in each topic and define an intuitive construct that they believe most appropriately matches the words in each topic (e.g., Bao and Datta, 2014). Several studies use unsupervised LDA to assess disclosure similarities (Dyer et al., 2017) and to classify disclosures (Bao and Datta, 2014; Campbell et al., 2014). Unsupervised methods are used primarily for identifying topics rather than prediction (Blei and McAulliffe, 2007).

Unlike unsupervised techniques, supervised machine-learning techniques parse text with respect to a response (i.e., dependent) variable, making these methods particularly useful for prediction (Blei and McAuliffe, 2007). For example, Frankel, Jennings, and Lee (2016) use

[6] Papers that use FOG and disclosure length to measure readability include Dyer et al. (2017) and Loughran and McDonald (2014).

[7] Papers that examine tone include Tetlock et al. (2008), Kothari et al. (2009), Feldman et al. (2010), Loughran and McDonald (2011), Rogers et al. (2011), Davis et al. (2012), Price et al. (2012), and Allee and DeAngelis (2015). Kravet and Muslu (2013) use a dictionary to measure risk. Larcker and Zakolyukina (2012) use a dictionary to measure deception. Li et al. (2013) count the uses of the word "competition" in the MD&A.

support vector regressions (SVR) to identify words in the management's discussion and analysis (MD&A) section of the 10-K to explain accruals and predict future cash flows. They also use SVR to identify words in the conference call that is to explain accruals. The results in Frankel et al. (2016) suggest that machine-learning methods identify narrative content in firm disclosures; however, they do not provide evidence that narrative-content measures capture language that is associated with what human readers identify at the time of the disclosure. SVR has also been used in finance to estimate macro-economic uncertainty (Manela and Moreira (2016) and stock return volatility (Kogan et al., 2009). A criticism of the SVR method is its lack of interpretability—it uses optimization rather than intuition to identify relevant words and phrases in a document (see Frankel et al., 2016; Manela and Moreira, 2016).

Although the majority of the prior research in this area focuses on counting relevant words to extract information from firm disclosures, humans likely use a more sophisticated process to extract meaning. Linguistics research suggests that meaning emerges from a combination of three factors. First, readers use lexical meanings (i.e., word definitions) to understand the actions, concepts, and words in the sentence (Johnson, 2008). Second, readers gather meaning through the syntactic structure (i.e., grammar) of a sentence (Dowty, 2007). If the structure of a sentence changes, its meaning can either change or become unintelligible.[8] These first two elements of linguistic meaning have their roots in linguistic theory dating back to the late 19th century. The Principle of Compositionality, which is commonly attributed to the late 19th century German mathematician Gottlob Frege, suggests that "the meaning of a sentence is a function of the meanings of the words in it and the way that they are combined syntactically" (Dowty, 2007). The

[8] For example, take the following sentence: *The police officer gave a citation to the motorcyclist*. If the sentence were written as *the motorcyclist gave a citation to the police officer* or *the citation a given to the was motorcyclist police officer gave*, then meaning of the sentence either changes or becomes unintelligible.

final factor that determines meaning is the context of the sentence. The context of a sentence can be derived from references, implications, presuppositions, deictic words, and deictic expressions (Huang, 2014). For example, the meaning of the phrase "we expect a large increase in these items" changes if the preceding sentences refer to income-increasing items rather than income-decreasing items.

The statistical methods in this paper make some headway into better capturing the meaning of a disclosure along two dimensions. First, we train statistical models to identify common language used in disclosures to discuss future earnings. We use a number of statistical techniques to measure lexical meanings and syntactic structure. These statistical models estimate the importance of individual words and phrases using word counts. Second, we allow the models to adapt to context by taking into consideration the broad circumstances of the firm at the time of the disclosure (e.g., high vs. low returns during the quarter). The importance of unique words can vary based on the circumstances or situation of the firm. Research can continue to improve upon these methods. For example, research can consider parts of speech (e.g., nouns, verbs, etc.) or sentence grammar (e.g., subjects, direct objects, etc.). Researchers can also examine the context of sentences within a disclosure relative to preceding sentences or identify other relevant broad contexts such as those involving the incentives of the managers providing the disclosures.

Statistical methods should also account for the salient features of text data. One feature that is so outstanding that it was noted by early research is 'sparsity' (e.g., Manning and Schütze, 1999). Many words in a corpus (a body of text under analysis) occur at very low rates, and a few words with seemingly little meaning (e.g., *the, and, a, to, of*) occur at very high rates. The novel *Tom Sawyer* has 71,370 words and 8,018 different words. While these numbers suggest that each word is used about 8.9 times, this average frequency is skewed. Almost half of the 8,018 words in *Tom*

Sawyer are only used once. In addition to removing words like *the*, *to*, and *a* (i.e., stop words) from our estimations, we consider the effects of the inclusion of sparse words.

3. Measures of Narrative Content

To determine whether text-based estimates of disclosure content correlate with insights gained by a reader, we first use machine-learning techniques to predict earnings surprises in quarter $q+1$ using the transcripts of conference calls following the quarter q earnings announcement. Earnings surprises ($EARN\ SURP_{i,\ q+1}$) are defined as the *IBES* actual earnings per share for firm i in quarter $q+1$ less the mean earnings per share forecast for quarter $q+1$ issued prior to the conference call date in quarter q scaled by the firm's stock price at the end of quarter q.[9] Our goal is to produce a text-based narrative-content measure that captures news about future earnings discussed in the conference call. We employ four text-based methods to predict earnings surprises: support vector regression (SVR), supervised latent Dirichlet allocation (sLDA), random forest regression trees (RF), and ordinary least squares using tone (OLS). The machine-learning methods (i.e., all but OLS) take as inputs the counts of all one- and two-word phrases from conference calls in quarter q to predict $EARN\ SURP_{i,q+1}$.

To reduce the influence of sparsity on the machine-learning techniques, we stem all words using the Porter Stemmer algorithm and remove any one- or two-word phrase that is used in fewer than 10 conference calls 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 $EARN\ SURP_{i,q+1}$ using rolling training samples for each calendar year in our sample. We use rolling training samples to estimate the parameters for each of our

[9] We use the last forecast issued by each analyst prior to the conference call date and remove stale forecasts made more than 90 days prior to the conference call.

statistical methods, because language that is useful in determining the narrative content of the conference call could change over time (Brown, Crowley, and Elliott, 2016; Frankel et al., 2016). 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*. Figure 1 shows a timeline. Since the out-of-sample estimation method is slightly different for SVR, sLDA, RF, and OLS using tone, we describe each method in more detail below. We also discuss the relative strengths of each method and provide information on how to implement each method using software that can be imported into well-known text analysis programs (e.g., Python).

3.1. Support Vector Regressions

We first use SVR to predict $EARN\,SURP_{i,q+1}$. SVR allows the estimation of a unique weight for each one- and two-word phrase count that is included in the conference call. Ordinary least squares cannot generate weights for each unique one- and two-word phrase because the number of phrases (316,399 phrases on average) is greater than the number of observations in the training samples (15,858 observations on average). 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, Jennings, and Lee (2016) and Manela and Moreira (2016) provide a more in-depth discussion of the SVR estimation procedure. Using our rolling training samples, we first 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 conference calls in year *t* to estimate an out-of-sample prediction of the future earnings surprise, which we label $SVR\ Pred_{i,q}$.

SVR is effective if weights on words and phrases can predictive of a dependent variable. SVR can be implemented in Python using the SVM-light implementation of Support Vector Machines in C created by Thorsten Joachims (accessible at http://svmlight.joachims.org/).

3.2. Supervised Latent Dirichlet Allocation

We also use sLDA to predict *EARN SURP*$_{i,q+1}$. sLDA jointly models the language in the disclosure and a response variable to find latent topics that best predict the responses for out-of-sample documents (Blei and McAuliffe, 2007). sLDA is similar to unsupervised LDA in that it categorizes the words and phrases in the disclosure into a set of latent (i.e., unknown) topics. The algorithm assumes that all disclosures share the same set of topics; however, the mix of each topic varies by disclosure. The words from the documents are then sorted into topics based on the probability of words co-occurring within documents (Dyer et al., 2017). sLDA adds another layer to the unsupervised LDA model by choosing topics with respect to a response variable.

Similar to our SVR estimation and using the earnings surprise in quarter *q+1* (*EARN SURP*$_{i,q+1}$) as the response variable, we use the counts of all one- and two-word phrases found in the conference calls in each training sample to identify 200 topics and the associated weights on the empirical topic frequencies.[10] We then apply these estimated weights to the topic frequencies for all conference calls in year *t* to estimate an out-of-sample prediction for *EARN SURP*$_{i,q+1}$, which we label *sLDA Pred*$_{i,q}$. Unsupervised LDA, which has been used in the prior accounting literature (e.g., Bao and Datta, 2014; Campbell et al., 2014; Dyer et al., 2017), differs from sLDA in that it

[10] 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 more flexibility. In untabulated results, we find that the explanatory power of the sLDA prediction is only slightly diminished if we use 100 topics, which suggests that our choice of 200 topics is reasonable.

does not fit topics to a dependent variable. The goal of unsupervised LDA is text categorization, while the goal of sLDA is prediction.[11]

Due to 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 to explain a dependent variable. Topics likely include words with similar meanings (e.g., income vs. earnings vs. EPS) as well as words that are used to discuss a particular topic (e.g., covenant, debt, threshold). sLDA can be implemented using the sLDA package in Python. More information can be found at https://pypi.python.org/pypi/slda/0.1.4.

3.3. Random Forecast Regression Trees

In addition, we use random forests (RF), a regression tree method, to predict *EARN SURP*$_{i,q+1}$. 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 dataset'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 pre-specified 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

[11] Blei and McAuliffe (2007) provide an example. "When the goal is prediction, fitting unsupervised topics may not be a good choice. Consider predicting a movie rating from the words in its review. Intuitively, good predictive topics will differentiate words like "excellent," "terrible," and "average," without regard to genre. But topics estimated from an unsupervised model may correspond to genres, if that is the dominant structure in the corpus" (page 1). In our setting, an unsupervised approach would likely group words into topics relating to industry, if industry is the dominant structure in the conference call corpus.

terminal (i.e., final) node represents the predicted value of the response given the preceding binary partitions of the data.

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.

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 earnings surprise in quarter $q+1$, which we label $RF\ Pred_{i,q}$.

The strength of the RF method is that it allows for interactions between words in its algorithm. For example, consider the simple regression tree labeled Figure 2. The first split of the data is based on the word count of word 1 (WC1). No further splits of the data minimize the SSE when WC1 is equal to 0, but when WC1 > 0, an additional split of the data based on the word count of word 2 (WC2) further minimizes the SSE within the resulting observations. The RF method is interactive in that it provides a different prediction for disclosures that have or do not

have word 2 **conditional upon having** word 1. RF can be implemented in Python using the scikit-learn package. More information can be found at http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

4.4. Tone

Lastly, we use ordinary least squares to predict $EARN\ SURP_{i,q+1}$ based on the tone ($TONE_{i,q}$) of the conference call in quarter q. We calculate $TONE_{i,q}$ as the number of positive words less the number of negative words (Loughran and McDonald, 2011) scaled by the total number of positive and negative words in the conference call of firm i in quarter q. For each rolling training sample, we then use OLS to regress $EARN\ SURP_{i,q+1}$ on $TONE_{i,q}$ to identify the portion of tone that is related to the earnings surprise in quarter $q+1$. We then multiply the coefficient from the training sample estimation with the $TONE_{i,q}$ of the conference calls in year t to predict the earnings surprise in quarter $q+1$, which we label $TONE\ Pred_{i,q}$.

The primary strength of the tone measure is that it is the most intuitive. However, its ability to capture narrative content is limited to the quality of the dictionary, which is researcher generated. In addition, the words that are used to construct the tone measure are weighted equally, which may or may not provide a realistic view of how humans weight the words when assessing the narrative content of the disclosure.

4.5. Word Lists

We list important one- and two-word phrases identified by the random forest and support vector regression methods and topics identified by supervised LDA in Appendix A. For random forest and support vector regression, we compute the average importance of each phrase across all years and report the top 100 phrases. For supervised LDA, we report the 5 most negatively predictive topics and the 5 most positively predictive topics in 2015 since topics cannot be easily

aggregated across years.

Many of the phrases included in the lists intuitively relate to firm performance (e.g., *congratul*, *great quarter*, *disappoint*, *deterior*, *improv*, etc.), which suggests that the machine-learning methods identify predictive language consistent with that identified by humans. The most positively and negatively predictive topics identified by sLDA also appear to capture human interpretation. Negatively predictive topics discuss items such as credit facilities, inventory reductions, restructuring charges, and economic recovery. Positively predictive topics discuss strong results, dividends, and positive guidance. Some phrases, admittedly, lack intuitive appeal (e.g., *fourth quarter*, *talk about*, *good morn*, etc.). The in-sample nature of the machine-learning methods allows for less intuitive phrases to influence the estimation process. However, because the statistical methods are then applied out-of-sample, less useful phrases should simply add noise to the out-of-sample predictions. Readers should not to place too much weight on any one particular phrase because each phrase represents only a small proportion of the total importance across all phrases. For example, the top 100 random forest (support vector regression) phrases comprise only 3.7% (8.1%) of the total importance across all phrases. The large dispersion of phase importance is a benefit of the machine learning methods. The algorithms identify many aspects of the disclosure that explain future earnings surprises. We view these phrase lists as providing some, albeit noisy, qualitative evidence that the machine learning methods identify information useful to humans.

4.5. Composite Measure

Humans likely extract meaning from a disclosure by placing weight on specific words/phrases, identifying topics, and identifying interactions among words/phrases. Therefore, each method likely captures a different aspect of the human interpretation process. In addition, the

primary cost to the researcher in estimating these models is setting up the data in the format required by the pre-packaged software. Fortunately, the data inputs into each method are similar, and the incremental cost to estimate an alternative model is low.

Given the above, we calculate a composite variable of all the text-based predictions of the earnings surprise in quarter *q+1* using factor analysis. To avoid look-ahead bias (i.e., using conference calls in the factor analysis that are unavailable at the time of each individual conference call), we estimate the factor analysis separately for each conference call including all other conference calls in the sample over the prior 365 days. We label the composite variable as $FACTOR\ Pred_{i,q}$.

4. Sample and Main Results

4.1. Sample Selection and Descriptive Statistics

We obtain a sample of 110,081 conference call transcripts from Factiva's FD Wire between 2004 and 2015. We drop 36,559 observations with insufficient data to calculate the analyst-forecast-revision and earnings-surprise variables. We also drop 1,629 observations with insufficient data to calculate future stock-market returns. We also require prediction data in the previous calendar year to construct the factor variables. Thus, our final sample consists of 67,370 conference call transcripts for 23,084 unique firm-years from 2005 to 2015.

Table 1 presents the descriptive statistics for the variables used in our primary analyses. The analyst-forecast revision for quarter *q+1* measured around the conference call in quarter *q* ($CH\ FOR_{i,q}$) is equal to the mean EPS forecast for quarter *q+1* issued within the 10 trading days after the conference call in quarter *q* less the mean EPS forecast for quarter *q+1* issued prior to the

conference call in quarter q, scaled by the stock price at the end of quarter q.[12] The mean value of $CH\ FOR_{i,q}$ is equal to -0.002. The mean values of $SVR\ Pred_{i,q}$, $sLDA\ Pred_{i,q}$, $RF\ Pred_{i,q}$, and $TONE\ Pred_{i,q}$ (previously defined) are equal to -0.002, -0.002, -0.002, and -0.005, respectively. These mean values are comparable in sign and magnitude to the analyst-forecast revision ($CH\ FOR_{i,q}$). These results are consistent with optimism in initial analyst forecasts followed by downward revisions (Richardson, Teoh, and Wysocki, 2004). The $EARN\ SURP_{i,q}$ variable is equal to the actual *IBES* earnings per share for firm i in quarter q less the mean analyst consensus forecast prior to the earnings announcement date in quarter q and scaled by firm i's stock price at the end of quarter q. The mean value of $EARN\ SURP_{i,q}$ is equal to 0.000.

Table 2 presents the Pearson and Spearman correlations among these variables. All correlations are significant at the 1% level. The Spearman (Pearson) correlations between $FACTOR\ Pred_{i,q}$ and each of the other prediction variables range from 0.48 (0.44) to 0.76 (0.80) and suggest that the factor analysis identifies common variation among the individual narrative content variables, but that the variation in $FACTOR\ Pred_{i,q}$ is not dominated by any one measure.

4.2. Analyst-Forecast Revisions

Our objective is to test whether narrative content measures (i.e., $SVR\ Pred_{i,q}$, $sLDA\ Pred_{i,q}$, $RF\ Pred_{i,q}$, $TONE\ Pred_{i,q}$, and $FACTOR\ Pred_{i,q}$) correlate with information gained by sophisticated human readers from the conference call. We use analyst-forecast revisions for quarter $q+1$ after the conference call in quarter q ($CH\ FOR_{i,q}$) to proxy for what a reader would learn about future earnings after reading or hearing the conference call. In Table 2, the Spearman and Pearson correlations between $CH\ FOR_{i,q}$ and the narrative content variables are positive and significant at

[12] We use the last forecast issued by each analyst to calculate the mean forecast prior to the conference call and the first forecast issued by each analyst to calculate the mean forecast following the conference call. We delete all forecasts issued more than 90 days prior to or 10 days following the conference call. We include analyst forecasts only if the analyst issued a forecast both before and after the conference call.

the 1% level, providing preliminary evidence that the machine-learning techniques that we employ identify changes to the firm's fundamentals that are consistent with human readers' updated expectations about firm fundamentals following the conference call. We further examine the explanatory power of the narrative content variables when explaining analyst forecast revisions for quarter $q+1$ using the following equation.[13]

$$CH\ FOR_{i,q} = \alpha + \alpha_1\ Pred_{i,q} + \varepsilon \quad (1)$$

We present the results from estimating Equation 1 in Table 3. The coefficients on *SVR* $Pred_{i,q}$, *sLDA* $Pred_{i,q}$, *RF* $Pred_{i,q}$, *TONE* $Pred_{i,q}$, and *FACTOR* $Pred_{i,q}$ are equal to 0.063, 0.512, 0.872, 0.136, and 0.002, respectively, and are significant at the 1% level. This evidence suggests that the narrative content measures are associated with information that analysts use to update their earnings per share forecasts after the conference call. The relative explanatory power of the narrative content measures varies. *TONE* $Pred_{i,q}$ and *SVR* $Pred_{i,q}$ have the lowest explanatory power with adjusted R^2s equal to 1.9% and 2.1%, respectively. *sLDA* $Pred_{i,q}$ and *RF* $Pred_{i,q}$ have the highest explanatory power with adjusted R^2s equal to 5.2% and 5.8%, respectively.[14] Not surprisingly, *FACTOR* $Pred_{i,q}$ explains the highest percentage of the variation in analyst revisions at 7.9%. This evidence suggests that each of the four methods identifies distinct narrative content. Thus, the remainder of our tests exclusively focus on *FACTOR* $Pred_{i,q}$ since it best captures the narrative content of the conference call. Our next set of tests provides evidence on how to improve the estimation of the narrative content measures generated by machine-learning methods.

[13] We do not include control variables in our primary analysis because our primary purpose is not to determine whether the narrative content measures are incremental to other quantitative forecasts or firm characteristics that are known at the time of the conference call. Rather, our purpose is to examine whether the statistical methods identify narrative content that a human would also identify.

[14] Although our results suggest that the Random Forest method produces the highest explanatory power in Table 3, this method may not dominate in other disclosure settings (e.g., MD&A) or when using alternative dependent variables (e.g., accruals). Therefore, we recommend future research to use all methods to estimate the narrative content of a disclosure.

4.3. Sparsity

Our first test examines how sparse (i.e., unusual) phrases affect the estimation of the narrative-content measures. The English language allows for significant diversity in the words and phrases that speakers use to communicate concepts and ideas. For example, one speaker might say "sales from division X have increased this year," while another might say "division X has ratcheted up revenues relative to last year." A human would likely infer that the two sentences have approximately the same meaning; however, a statistical model may be unable to infer the effect of a "sales increase" on future earnings if firms describe the increase in many different ways. Sparsity can be particularly acute in textual data relative to other types of data such as financial accounting data. The accounting system is designed to summarize diverse data points into multiple pre-defined constructs (e.g., sales, assets, etc.), while language has seemingly endless ways to describe similar constructs.

To illustrate (untabulated), the 6,779 conference calls in our sample in 2015 contain 1.667 million unique one- and two-word stemmed phrases after we remove 'stop' words. Of these phrases, 990,617 (59.4%) are included in only one conference call, and an additional 546,387 (32.8%) are included in fewer than 10 conference calls. We label these 1.537 million phrases as "sparse." Interestingly, the remaining 130,118 "non-sparse" phrases (7.8%) constitute the vast majority of the total language used in the conference call (i.e., approximately 91.3% of the total phrases in the conference call on average across all conference calls in 2015).

We exclude sparse phrases in the primary estimation of our models because their inclusion could lead to high fit within the training sample but produce a model that is less relevant for out-of-sample prediction. To provide direct evidence on the effect of sparse phrases on the machine-learning models and to guide future researchers who use these methods, we re-estimate the models

including both sparse and non-sparse phrases. We then combine the resulting SVR, sLDA, RF, and tone predictions using factor analysis (as described previously) and label the prediction *FACTOR Pred (KEEP SPARSE)*$_{i,q}$. We then re-estimate Equation 1 with *FACTOR Pred (KEEP SPARSE)*$_{i,q}$ as the independent variable and present the results in Table 4. For ease of comparison, in Column 1 we repeat the result of Equation 1 with *FACTOR Pred*$_{i,q}$ as the independent variable (also reported in Table 3, Column 5), which yields an adjusted R^2 equal to 7.9%. In Column 2, we report the result with *FACTOR Pred (KEEP SPARSE)*$_{i,q}$ as the independent variable and find that the adjusted R^2 is equal to 7.0%, which represents an 11.4% reduction relative to the model in Column 1. Using the Vuong (1989) likelihood ratio test, we find that the adjusted R^2s are significantly different at the 1% level. These results suggest that sparse words slightly weaken the ability of the machine-learning models to identify the narrative content of the conference call.[15]

4.4. Context

Our second test examines whether the explanatory power of the narrative-content measures improves when we estimate the models within different contexts. We estimate each model (i.e., SVR, sLDA, RF, and tone) within subsamples of positive/negative news, which we define as the cumulative size-adjusted returns during quarter *q* greater or less than zero, and high/low return volatility, which we define as daily return volatility during quarter *q* greater or less than the sample median. We apply the positive or negative news model to the current quarter conference call based on whether the current quarter size-adjusted return is positive or negative. We apply the high or low return volatility model to the current quarter conference call based on whether the current

[15] We note that the weaker correlation between analyst-forecast revisions and the combined narrative-content measure when including sparse phrases in the models could be attributed either to 1) measurement error in the models, or 2) analysts using the language of calls with higher sparsity less to determine their forecast revisions. To address this issue, we re-estimate Equation 1 with the actual realized future earnings surprise as the dependent variable and find that the adjusted R^2 is lower when the models use sparse language. Thus, the weaker explanatory power of the combined narrative-content measure is likely due to measurement error induced by sparse language.

quarter return volatility is in the top or bottom half of the distribution of quarterly return volatility measured over the previous calendar year. Similar to the procedure we use to calculate *FACTOR Pred*$_{i,q}$, we use factor analysis to compute combined variables, which we label *FACTOR Pred (QTR CAR)*$_{i,q}$ and *FACTOR Pred (RET VOL)*$_{i,q}$, respectively. We also use factor analysis to combine all of the context-specific predictions and the full-sample predictions into one measure that we label *FACTOR Pred (CONTEXT)*$_{i,q}$.

We then re-estimate Equation 1 with these context-specific measures as the independent variables of interest and present the results in Table 5. In Columns 1 and 2, the coefficients on *FACTOR Pred (RET VOL)*$_{i,q}$ and *FACTOR Pred (QTR CAR)*$_{i,q}$ are positive and significant, and the adjusted R^2s are equal to 7.5% and 8.0%, respectively. In Column 3, the coefficient on *FACTOR Pred (CONTEXT)*$_{i,q}$ is positive and significant, and the adjusted R^2 is equal to 9.5%, which is 20% higher than the adjusted R^2 when *FACTOR Pred*$_{i,q}$ is the independent variable. Using the Vuong (1989) likelihood ratio test, we find that the R^2 from the *FACTOR Pred (CONTEXT)*$_{i,q}$ regression is significantly larger at the 1% level than the R^2 from the *FACTOR Pred*$_{i,q}$ regression. This evidence suggests that when we allow the narrative-content models to adapt to relevant contexts, the output better explains insights gained by human readers.[16] In the remaining tests, we focus on the *FACTOR Pred (CONTEXT)*$_{i,q}$ variable since it best captures the narrative content of the conference call.

We note that attempting several partitions and choosing those that maximize "out-of-sample" explanatory power results in overfitting and calls into question the validity of the out-of-

[16] We estimate the explanatory power of *FACTOR Pred (CONTEXT)*$_{i,q}$ by year and by firm to examine whether the predictive power is driven by specific firms or years. In untabulated results, we find that the yearly adjusted R^2s are stable with a slight increase during the financial crisis of 2007, 2008, and 2009. The average adjusted R^2 for the yearly regressions is equal to 9.55%. We require at least 12 observations for each firm in the firm-specific regressions. The average adjusted R^2 of the firm-specific regressions is equal to 9.54%. These results are consistent with the overall explanatory power of the pooled regression model and suggest that time and firm effects do not significantly affect the explanatory power of our models.

sample predictions. As a result, we recommend that future research identify relevant subsamples prior to estimating the models.

4.5. Narrative-Content Measures Relative to Other Earnings Signals

One outstanding question is how to interpret the economic significance of the narrative-content measures. Is the 9.5% explanatory power of *FACTOR Pred (CONTEXT)*$_{i,q}$ high or low relative to other quantitative earnings signals provided at the time of the earnings announcement and conference call? To provide evidence related to this question, we re-estimate Equation 1 with other quantitative signals as the independent variables of interest. We consider three signals: the concurrent earnings surprise (*EARN SURP*$_{i,q}$), the change in operating cash flows ($\Delta CF_{i,q}$), and the management earnings guidance surprise (*GUID SURP*$_{i,q}$). We expect that analysts revise their forecasts of future earnings after the conference call at least partially based on information they obtain from these signals.[17]

We define *EARN SURP*$_{i,q}$ as earnings per share for firm *i* in quarter *q* less the mean quarter *q* earnings per share forecast for firm *i* made prior to the earnings announcement date in quarter *q* scaled by firm *i*'s stock price at the end of quarter *q*. For each analyst, we use the latest forecast prior to the earnings announcement date, removing any forecasts made more than 90 days prior to the earnings announcement date. We define $\Delta CF_{i,q}$ as operating cash flows for firm *i* in quarter *q* less operating cash flows for firm *i* in the same quarter of the prior year divided by total assets for firm *i* at the end the same quarter of the prior year. We define *GUID SURP*$_{i,q}$ as the manager's earnings per share guidance for quarter *q+1* given in the [-1, +1] window surrounding the conference call date less the analyst consensus earnings per share forecast for quarter *q+1* made prior to the earnings announcement date in quarter *q* and scaled by firm *i*'s stock price at the end

[17] Information used by analysts outside of the call also affects revisions. We use these alternative earnings measures for comparison, because they provide a basis for comparison that is similarly affected by this other information.

of quarter q.

We present the results of these other quantitative signals in Table 6. For ease of comparison, we report the model results with *FACTOR Pred (CONTEXT)*$_{i,q}$ as the independent variable in Column 1 of Panel A (also reported in Table 5, Column 3). In Column 2, we present the results of the models with *EARN SURP*$_{i,q}$ as the independent variable of interest and find that the adjusted R^2 is equal to 11.8%, which suggests that the combined narrative-content measure captures 81% (9.5% / 11.8%) of the variation in analyst-forecast revisions relative to the earnings surprise. This result suggests that the combined narrative-content measure captures economically significant information from the conference call that approaches the explanatory power of the concurrent earnings surprise. In Column 3, we find that $\Delta CF_{i,q}$ explains approximately 0.5% of the variation in analyst-forecast revisions, and the explanatory power of *FACTOR Pred (CONTEXT)*$_{i,q}$ is 19.0 times larger than the variation captured by the change in cash flows.

In Columns 4 and 5, we examine whether the information provided by the narrative-content measure overlaps with or is incremental to these other signals. In Column 4, we include *EARN SURP*$_{i,q}$ and $\Delta CF_{i,q}$ in the model as the independent variables and find an adjusted R^2 equal to 12.0%. In Column 5, we add *FACTOR Pred (CONTEXT)*$_{i,q}$ to the model from Column 4 and find that the adjusted R^2 is equal to 17.8%. Thus, the incremental R^2 from Columns 4 to 5 is approximately 61% ([17.8% - 12.0%] / 9.5%) of the explanatory power of *FACTOR Pred (CONTEXT)*$_{i,q}$ in Column 1. These results suggest that the combined narrative-content measure captures some overlapping earnings- and cash-flow-related information from the conference call, but also captures a significant amount of information beyond the concurrent earnings surprise and change in cash flows.

In Panel B, we compare the explanatory power of the combined narrative-content measure

to the explanatory power of the concurrent earnings surprise, the managerial guidance surprise, and the change in cash flows within the subsample of firms that provide earnings guidance at the time of the conference call. We note that the sample size decreases to 14,044 observations when we require quarterly managerial earnings forecasts. In Column 1, the adjusted R^2 is equal to 14.1% when *FACTOR Pred (CONTEXT)*$_{i,q}$ is the independent variable. Thus, within the subsample of guidance firms, the combined narrative-content measure better captures information that analysts use to update their earnings expectations. This result suggests that managers are more likely to discuss forward-looking information that is related to the future earnings surprise if they are willing and able to provide earnings guidance on the conference call date. We then include *EARN SURP*$_{i,q}$ as the sole independent variable in Column 2 and find the R^2 to be equal to 2.1%. The explanatory power of *EARN SURP*$_{i,q}$ is significantly lower in the guidance sample (Panel B) than in the full sample (Panel A). Nevertheless, these results continue to suggest that the combined narrative-content measure captures an economically significant amount of information in the conference call relative to the concurrent earnings surprise. In Column 3, the adjusted R^2 is equal to 0.9% when $\Delta CF_{i,q}$ is the independent variable. In Column 4, the adjusted R^2 is equal to 49.9% when *GUID SURP*$_{i,q}$ is the independent variable of interest. The high adjusted R^2 in Column 4 is not surprising given that analysts likely focus on and use managerial forecasts to revise their own forecasts. Interestingly, the explanatory power of *FACTOR Pred (CONTEXT)*$_{i,q}$ is still 28% of that of *GUID SURP*$_{i,q}$.

In Column 5, we include *EARN SURP*$_{i,q}$, $\Delta CF_{i,q}$, and *GUID SURP*$_{i,q}$ as independent variables and find an adjusted R^2 equal to 50.8%. In Column 6, we include the variables from Column 5 and *FACTOR Pred (CONTEXT)*$_{i,q}$ to find an adjusted R^2 of 54.6%, which suggests an incremental R^2 of 3.8%. Approximately 27% of the explanatory power of *FACTOR Pred*

(CONTEXT)$_{i,q}$ does not overlap with managerial earnings forecasts, operating cash flows, or the earnings surprise. While the conference call appears to reiterate some of the information that can be found in earnings, these results also suggest that the conference call provides an economically significant amount of information beyond that provided in the earnings numbers.

4.6. Future Abnormal Returns

Our methods fit conference calls to the *earnings surprise* rather than the *analyst revision* so that the text content measure captures information that is relevant to future earnings and is unaffected by conference call statements that mislead analysts. However, fitting our models to earnings surprises rather than analyst revisions creates the possibility that the fitted values contain information missed by analysts and that this missed information could predict returns. Moreover, prior research suggests that humans identify some but not all of the information contained in the conference call (Mayew and Venkatachalam, 2012; Lee, 2016; Davis, Piger, and Sedor, 2012; Demers and Vega, 2010). Thus, it is possible that the combined narrative-content measure identifies information that investors and analysts both capture and miss at the time of the conference call.

To test this conjecture, we measure returns over two periods: 1) the size-adjusted abnormal return from two days following the conference call to two days before the quarter *q+1* earnings announcement (*QTR CAR*$_{i,q+1}$), and 2) the size-adjusted abnormal return from the day prior to the day after the earnings announcement in quarter *q+1* (*EA CAR*$_{i,q+1}$). For each conference call we decile rank the *FACTOR Pred (CONTEXT)*$_{i,q}$ relative to all other firms included in the sample over the previous 365 days. We then form portfolios by year and calculate the hedge portfolio return for each year. In Table 7, we find that the average abnormal return from shorting firms in the 1st and holding firms in the 10th decile is equal to 2.36% (Panel A) for *QTR CAR*$_{i,q+1}$ and 0.74% (Panel

B) for $EA\ CAR_{i,q+1}$. However, in 2009, the return is highly negative and the time-series test statistic is insignificant for both returns, which suggests that the trading strategy is not riskless and is not implementable when tracking error (e.g., positive hedge portfolio returns are not consistent from year to year) is a concern. Overall, these results suggest that the combined narrative-content measure does not consistently identify information missed by humans.

5. Additional Analyses – Cross-sectional Tests of Analyst-Forecast Revisions

We estimate several untabulated cross-sectional tests to provide additional evidence that the narrative-content measure captures information that a human reader would identify from the conference call. The three measures we consider are 1) the absolute intra-day conference call return, 2) cosine similarity, and 3) the percentage of forward-looking statements on the conference call. Our assumption is that these measures proxy for information that readers obtain from the call. We expect the strength of the association between the combined narrative-content measure and analyst-forecast revisions to increase when these alternative measures suggest greater narrative content. For each cross-sectional test, we divide the sample into terciles based on the cross-sectional variable and re-estimate Equation 1 for each tercile subsample.

The intra-day investor reaction is perhaps the strongest alternative proxy for the narrative content of the conference call. However, since conference calls often occur outside of normal trading hours, its general application is limited. For instance, in our sample, we identify a subset of 17,769 conference calls (approximately 26%) with start times that occur during trading hours and are at least 30 minutes following the earnings announcement time.[18] We calculate the investor

[18] We obtain start times from the ThomsonOne database. We follow Matsumoto et al. (2011) to approximate call end times by dividing the conference call word count by 160, an estimate of words spoken per minute

reaction to the conference call as the absolute value of the return between its start and end times.[19] We find that the coefficient on *FACTOR Pred (CONTEXT)*$_{i,q}$ in the lowest tercile of the absolute conference call return is statistically (1% level) less than the coefficient on *FACTOR Pred (CONTEXT)*$_{i,q}$ in the highest tercile. In addition, the adjusted R^2 is equal to 6.89% in the lowest tercile and 9.06% in the highest tercile. These results suggest that *FACTOR Pred(CONTEXT)*$_{i,q}$ better identifies information in the call when investors also respond to information within the call.

The second measure we consider is cosine similarity. Cosine similarity inversely proxies for 'new' information available in the conference call that was not available in prior calls (Brown and Tucker 2011). We calculate cosine similarity using the word count vector of the conference call for firm *i* in quarter *q* and the word count vector of the conference call for firm *i* in the same quarter of the previous calendar year. Higher cosine similarity suggests that the disclosures are similar and indicates less narrative content. We find that the coefficient on *FACTOR Pred (CONTEXT)*$_{i,q}$ in the lowest tercile of cosine similarity is significantly (1% level) higher than that in the highest tercile. The adjusted R^2 also drops from 10.81% in the lowest tercile to 7.48% in the highest tercile.

The final measure we consider is the percentage of forward-looking statements in the conference call, where a higher percentage indicates greater narrative content. We calculate the percentage of forward-looking sentences in the conference call as the percentage of sentences containing words such as "anticipate," "expect," "will," etc. (Li, 2010). We find that the coefficient on *FACTOR Pred (CONTEXT)*$_{i,q}$ in the lowest tercile is significantly (1% level) lower than that in the highest tercile, but that the adjusted R^2 is equal to 9.30% in the bottom tercile and equal to

[19] More specifically, the absolute return during the conference call is equal to the absolute value of the final trade price at the end of the conference call less the final trade price prior to the conference call scaled by the final trade price prior to the conference call.

8.25% in the top tercile. However, we find stronger results when we split the sample based on the percentage of forward-looking statements with accounting-related terms, likely because forward-looking accounting statements are more useful to analysts in communicating information about the earnings surprise in quarter $q+1$.[20] The adjusted R^2 increases from 7.50% in the lowest tercile to 10.08% in the highest tercile. Taken together, these results provide additional evidence that $FACTOR\ Pred\ (CONTEXT)_{i,q}$ captures information that a human reader would identify from the conference call.

6. Conclusion

We examine whether support vector regressions, supervised LDA, random forest regression trees, and tone identify narrative content in conference calls that is correlated with what a human reader would glean from the conference call. We also create a combined measure of the four narrative content measures using factor analysis. While the individual narrative content measures are associated with analyst-forecast revisions for quarter $q+1$ made after the conference call for quarter q, we find that the combined measure explains more of the variation in analyst-forecast revisions than any single measure. This finding supports the notion that the measure identifies narrative content that a human reader would identify. This evidence also suggests that each narrative-content measure identifies narrative content that other measures likely do not. Future research should consider combining several individual measures to develop a measure for narrative content.

We provide evidence for future research on how to improve the estimation of the narrative content measures generated by machine-learning methods. The explanatory power of the

[20] We define accounting-related terms using the Oxford Reference Dictionary of Accounting.

combined measure is higher when unusual (i.e., sparse) words are excluded from the estimation process and when the model estimation is adapted to specific disclosure contexts (i.e., positive/negative and more-/less-volatile quarterly returns). Therefore, future research should consider removing sparse language and estimating the narrative-content measures within context when developing narrative-content measures.

We also provide evidence that the combined narrative-content measure identifies a significant amount of the variation in analyst-forecast revisions relative to the current quarter's earnings announcement, cash flow changes, and managerial earnings surprises. Approximately 28% of the explanatory power of the combined narrative-content measure is incremental to these three figures, which suggests that the conference call provides information beyond that provided by earnings or future managerial forecasts. Lastly, we find that the combined narrative-content measure does not consistently predict abnormal returns during the subsequent quarter and during the subsequent earnings announcement.

One previous criticism of machine-learning techniques in estimating the narrative content of disclosures is that we did not know whether these techniques approximate what a reader gathers from the disclosure. The results in this paper provide evidence that machine-learning techniques identify a portion of the narrative content that a sophisticated reader would gather from a disclosure. These results aid future research by 1) providing a method to convert qualitative narrative disclosures into quantitative forecasts and 2) providing a foundation that future research can build on to identify better measures for narrative content.

References

Allee, K., DeAngelis, M., 2015. The structure of voluntary disclosure narratives: Evidence from tone dispersion. Journal of Accounting Research 53(2), 241-274.

Bajari, P., Nekipelov, D., Ryan, S., Yang, M., 2015. Machine learning methods for demand estimation. NBER working paper.

Bao, Y., Datta, A., 2014. Simultaneously discovering and quantifying risk types from textual risk disclosures. Management Science 60(6), 1371-1391.

Blei, D., McAuliffe, J., 2007. Supervised topic models. Neural Information Processing Systems.

Blei, D., Ng, A., Jordan, M., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993-1022.

Bonsall, S., A. Leone, B. Miller, and K. Rennekamp. 2017. A plain English measure of financial reporting readability. Journal of Accounting and Economics 63, 329-357.

Breiman, L., 2001. Random forests. Machine Learning 45, 5-32.

Brown, S., Tucker, J., 2011. Large-sample evidence on firms' year-over-year MD&A modifications. Journal of Accounting Research 49(2), 309-346.

Brown, N., Crowley, R., Elliott, B. 2016. What are you saying? Using topic to detect financial misreporting, Working paper

Campbell, J., Chen, H., Dhaliwal, D., Lu, H., Steele, L., 2014. The information content of mandatory risk factor disclosures in corporate filings. Review of Accounting Studies 19, 396-455.

Core, J., 2001. A review of the empirical literature: Discussion. Journal of Accounting and Economics 31, 441-456.

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

Demers, E., Vega, C. 2010. Soft information in earnings announcements: News or noise? Working Paper.

Dowty. D., 2007. Compositionality as an Empirical Problem. *Direct Compositionality*. Oxford University Press, 23-101.

Dyer, T., Lang, M., Stice-Lawrence, L., 2017. The evolution of 10-K textual disclosure: Evidence from latent Dirichlet allocation. Journal of Accounting and Economics, Forthcoming.

Feldman, R., Govindaraj, S., Livnat, J., Segal, B., 2010. Management's tone change, post earnings announcement drift and accruals. Review Accounting Studies 15, 915–953.

Financial Accounting Standards Board (FASB), 2012. Disclosure Framework: Invitation to Comment. Norwalk, CT.

Frankel, R., Jennings, J., Lee, J., 2016. Using unstructured and qualitative disclosures to explain accruals. Journal of Accounting and Economics 45(2), 209-227.

Frankel, R., Johnson, M., Skinner, D., 1999. An empirical examination of conference calls as a voluntary disclosure medium. Journal of Accounting Research 37(1), 133-150.

Henry, E., Leone, A., 2016. Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. The Accounting Review 91(1), 153-178.

Huang, Y. 2014. Pragmatics. Oxford University Press.

Johnson, K. 2008. The overview of lexical semantics. Philosophy Compass 3, 119-134.

Kogan, S., Levin, D., Routledge. B., Sagi, J., Smith, N., 2009. Predicting risk from financial reports with regression. NAACL-HLT 2009, Boulder, Colo., May–June 2009.

Kothari, S. P., 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 Review 84, 1639-70.

Kravet, T., Muslu, V., 2013. Textual risk disclosures and investors' risk perceptions. Review of Accounting Studies 18, 1088–1122.

Larker, D., Zakolyukina, A., 2012. Detecting deceptive discussions in conference calls. Journal of Accounting Research 50, 495-540.

Lee, J. 2016. Can Investors Detect Managers' Lack of Spontaneity? Adherence to Predetermined Scripts During Conference Calls. The Accounting Review, 91(1), 229-250.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics 45, 221-247.

Li, F., 2010. The information content of forward-looking statements in corporate filings – A naïve Bayesian machine learning approach. Journal of Accounting Research 48(5), 1049-1102.

Li, F., Lundholm, R., Minnis, M., 2013. A measure of competition based on 10-K filings. Journal of Accounting Research 51(2), 399-436.

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

Loughran, T., McDonald, B. 2014. Measuring readability in financial disclosures. Journal of Finance 69, 1643-1671.

Manela, A., Moreira, A., 2016. News implied volatility and disaster concerns. Journal of Financial Economics 123: 137-162.

Manning, C., and Schutze, H., 1999, *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachuesetts.

Matsumoto, D., Pronk, M., Roelofsen, E., 2011. What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions. The Accounting Review 86(4), 1384-1414.

Mayew, B., and Venkatachalam, M, 2012, The power of voice: Managerial affective states and future firm performance, Journal of Finance 67: 1-43.

Portner, P. 2005. What is Meaning? Blackwell Publishing. Malden, MA.

Price, S., Doran, J., Peterson, D., Bliss, B., 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. Journal of Banking & Finance 36, 992-1011.

Richardson, S., Teoh, S., Wysocki, P. 2004. The walk-down to beatable analyst forecasts: The role of equity issuance and insider trading incentives. Contemporary Accounting Research 21: 885-924.

Rogers, J., Van Buskirk, A., Zechman, S., 2011. Disclosure tone and shareholder litigation. The Accounting Review 86, 2155-2183.

Securities and Exchange Commission (SEC), 1998. A Plain English Handbook: How to Create Clear SEC Disclosure Documents. https://www.sec.gov/pdf/handbook.pdf. SEC Offices, Washington, D.C.

Securities and Exchange Commission (SEC), 2013. Report on Review of Disclosure Requirements in Regulation S-K. http://www.sec.gov/news/studies/2013/reg-skdisclosure-requirements-review.pdf. SEC Offices, Washington D.C.

Stalnaker, R., 1970. Pragmatics. Synthese 22, 272-289.

Tetlock, P., Saar-Tsechansky, M., Sofus, M., 2008. More than words: Quantifying language to measure firms' fundamentals. The Journal of Finance 63(3), 1437-1467.

Vuong, Q., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica 57(2), 307-333.

Appendix A – Important Phrases and Topics

Random Forest (Top 100 Phrases Sorted by Average Importance Across All Years)

coven	receiv	credit facil	last question
congratul	increas	strateg	equival
net loss	came	third	up
great quarter	earn	consist	bit
loss	compon	loss per	ye
strong	reduc	place	market
liquid	new	ebitda coven	really
disappoint	percent	those two	over year
amend	difficult	breakeven	crude
line avail	not go	conclud today	first quarter
unrestrict cash	expect	boulder	preclin
deterior	under	anyth	due
phase	mine	need	restrict
nice quarter	about	rate	ep
stock split	posit ebitda	continu	complet
growth	not	call	come
good quarter	now disconnect	previous	oper loss
chang	exclud	guidanc	financi coven
ford	restrict cash	over	cash balanc
realiz	impact	delay	probabl
year	exchang	away	cash burn
detail	very strong	inventory	presid
cash equival	more	first	cash posit
peopl	way	back over	quick
apart	significantly	late	close remark

Support Vector Regression (Top 100 Phrases Sorted by Average Importance Across All Years)

fourth quarter	second half	game	station
fourth	bank	technic	okay
last year	first quarter	turn	power
coven	retail	commerc	complet
last	per	phase	mine
improv	difficulty	digit	full year
second	declin	aggress	fleet
revenu	next question	wafer	period
non gaap	foreign	store	brad
liquid	sale	barrel	growth
pleas	tanker	capac	littl
talk about	morn	proceed	televis
cash	gross margin	constant	gross
partner	impair	magazin	phase iii
third quarter	richard	net revenu	ladi
custom	quarter	pulp	prior year
good morn	raw	year over	real
million	over year	cancer	drill
crude	next	perform	exchang rate
second quarter	ebitda	compar	depreci
earn	raw materi	array	incom
ahead	congratul	third	go ahead
margin	aircraft	increas	profit
reduct	strong	very	month
talk	year	led	per day

sLDA Topics (Top 5 Most Negative Topics and Top 5 Most Positive Topics in 2015):

Negative Topics	
Topic 1	million, facil, credit, revolv, credit facil, under, capit, liquid, current, borrow, agreement, end, avail, complet, note, no, dure, senior, outstand, close
Topic 2	inventory, level, product, down, quarter, sale, end, sell, through, inventory level, balanc, work, lower, ship, shipment, reduc, time, not, day, decreas
Topic 3	solar, project, modul, cost, market, percent, megawatt, china, expect, thank, quarter, cell, think, shipment, now, wafer, question, see, effici, total
Topic 4	cost, million, restructur, charg, save, reduct, reduc, action, oper, loss, down, impair, cost reduct, cost save, benefit, structur, plan, exclud, profit, includ
Topic 5	see, down, not, back, still, slow, seen, economy, recovery, come, saw, environ, happen, up, issu, pick, weak, pick up, econom, month

Positive Topics	
Topic 1	very, strong, see, well, great, continu, very strong, really, again, good, up, very very, pleas, thank, obvious, very well, nice, very good, all, excit
Topic 2	tax, rate, tax rate, earn, adjust, share, per share, per, impact, benefit, item, effect, incom, earn per, effect tax, year, expect, percent, exclud, relat
Topic 3	billion, dividend, return, capit, sharehold, invest, percent, valu, cash, earn, year, company, balanc, alloc, look, share, investor, record, thank, over
Topic 4	million, net, incom, dilut, share, per, net incom, compar, dilut share, per dilut, compani, tax, oper, increas, result, approxim, base, end, cash, statement
Topic 5	year, guidanc, expect, full, percent, full year, rang, approxim, increas, provid, end, outlook, addit, dure, continu, base, total, reflect, includ, perform

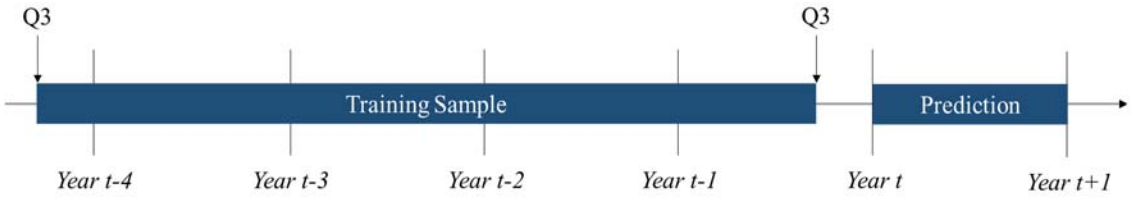
Appendix B
Variable Definitions

Variable	Definition
$BTM_{i,q}$	Book-to-market ratio of firm *i* in quarter *q* calculated as firm *i*'s book value of common equity at the end of quarter *q* divided by $MVE_{i,q}$.
$\Delta CF_{i,q}$	Operating cash flows for firm *i* in quarter *q* less operating cash flows for firm *i* in the same quarter of the prior year divided by total assets for firm *i* at the end the same quarter of the prior year.
$CH\ FOR_{i,q}$	The mean revision in analysts' quarter *q+1* earnings per share forecasts for firm *i* for all analysts who update their forecasts within 10 trading days on or following firm *i*'s conference call date in quarter *q* scaled by firm *i*'s stock price at the end of quarter *q*. For each analyst, we use the earliest forecast following the conference call date and the latest forecast prior to the conference call date, removing any forecasts made more than 90 days prior to the conference call date of firm *i* in quarter *q*.
$EA\ CAR_{i,q+1}$	The cumulative size-adjusted return for firm *i* for the trading-day window from one trading day before to one trading day after the earnings announcement in quarter *q+1*.
$EARN\ SURP_{i,q}$	Earnings per share for firm *i* in quarter *q* less the mean quarter *q* earnings per share forecast for firm *i* made prior to the earnings announcement date for firm *i* in quarter *q* scaled by firm *i*'s stock price at the end of quarter *q*. For each analyst, we use the latest forecast prior to the earnings announcement date, removing any forecasts made more than 90 days prior to the earnings announcement date of firm *i* in quarter *q*.
$EARN\ SURP_{i,q+1}$	Earnings per share for firm *i* in quarter *q+1* less the median quarter *q+1* earnings per share forecast for firm *i* made prior to the conference call date for firm *i* in quarter *q* scaled by firm *i*'s stock price at the end of quarter *q*. For each analyst, we use the latest forecast prior to the conference call date, removing any forecasts made more than 90 days prior to the conference call date of firm *i* in quarter *q*.
$FACTOR\ Pred_{i,q}$	The factor obtained when estimating a factor analysis on $SVR\ PRED_{i,q}$, $SLDA\ PRED_{i,q}$, $RF\ Pred_{i,q}$, and $TONE\ Pred_{i,q}$. The factor analysis is run separately for each conference call observation using all conference calls in the sample over the previous 12 months.
$FACTOR\ Pred\ (KEEP\ SPARSE)_{i,q}$	The factor obtained when estimating a factor analysis on the SVR, sLDA, RF, and tone predictions of $EARN\ SURP_{i,q+1}$, where the models include both sparse and non-sparse one- and two-word phrases in the estimation processes. The factor analysis is run separately for each conference call observation using all conference calls in the sample over the previous 12 months.
$FACTOR\ Pred\ (QTR\ CAR)_{i,q}$	The factor obtained when estimating a factor analysis on the SVR, sLDA, RF, and tone predictions of $EARN\ SURP_{i,q+1}$, where the models are estimated within subsamples of positive and negative cumulative size-adjusted returns for firm *i* over the previous quarter. The positive or negative return models are applied to the current

	quarter conference call based on whether the current quarter cumulative size-adjusted return is positive or negative. The factor analysis is run separately for each conference call observation using all conference calls in the sample over the previous 12 months.
FACTOR Pred (RET VOL)$_{i,q}$	The factor obtained when estimating a factor analysis on the SVR, sLDA, RF, and tone predictions of *EARN SURP*$_{i,q+1}$, where the models are estimated within subsamples of above and below median daily return volatility for firm *i* over the previous quarter. The high or low return volatility model is applied to the current quarter conference call based on whether the current quarter return volatility is in the top or bottom half of the distribution of return volatility measured over the previous calendar year. The factor analysis is run separately for each conference call observation using all conference calls in the sample over the previous 12 months.
FACTOR Pred (CONTEXT)$_{i,q}$	The factor obtained when estimating a factor analysis on the set of SVR, sLDA, RF, and tone predictions of *EARN SURP*$_{i,q+1}$ when estimated within 1) the full sample, 2) subsamples based on high and low daily return volatility during the quarter, and 3) subsamples based on positive and negative cumulative size-adjusted abnormal returns over the quarter. The factor analysis is run separately for each conference call observation using all conference calls in the sample over the previous 12 months.
GUID SURP$_{i,q}$	The manager's earnings per share guidance for quarter *q+1* given in the [-1, +1] window surrounding the conference call date less the analyst consensus earnings per share forecast for quarter *q+1* made prior to the earnings announcement date in quarter *q* and scaled by firm *i*'s stock price at the end of quarter *q*. If the manager does not provide quarterly guidance, but provides annual guidance, the variable is equal to the earnings per share guidance for the next annual period less the annual analyst consensus earnings per share forecast made prior to the earnings announcement date and scaled by firm *i*'s stock price at the end of quarter *q*.
MOM$_{i,q}$	The cumulative size-adjusted return for firm *i* for the [-131, -5] trading-day window prior to the conference call date in quarter *q*.
MVE$_{i,q}$	Market value of equity for firm *i* in quarter *q* calculated as firm *i*'s stock price multiplied by the number of shares outstanding at the end of quarter *q*.
QTR CAR$_{i,q+1}$	The cumulative size-adjusted return for firm *i* for the trading-day window from two trading days after the conference call in quarter *q* to two days before the earnings announcement in quarter *q+1*.
RF Pred$_{i,q}$	The out-of-sample random forest prediction of *EARN SURP*$_{i,q+1}$ using the counts of one- and two-word phrases of the conference call for firm *i* in quarter *q*. The training data include all conference calls made from Q4 of year *t-5* to Q3 of year *t-1*.
sLDA Pred$_{i,q}$	The out-of-sample supervised-latent-Dirichlet-allocation prediction of *EARN SURP*$_{i,q+1}$ using the counts of one- and two-word phrases of the conference call for firm *i* in quarter *q*. The training data include all conference calls made from Q4 of year *t-5* to Q3 of year *t-1*.

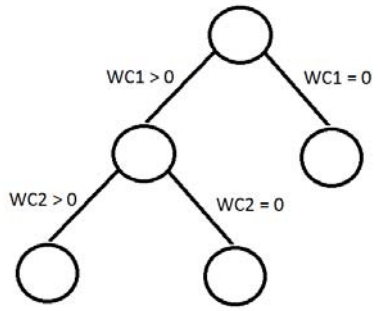
SVR Pred$_{i,q}$	The out-of-sample support-vector-regression prediction of *EARN SURP*$_{i,q+1}$ using the counts of one- and two-word phrases of the conference call for firm *i* in quarter *q*. The training data include all conference calls made from Q4 of year *t-5* to Q3 of year *t-1*.
TONE Pred$_{i,q}$	The out-of-sample tone prediction of *EARN SURP*$_{i,q+1}$ using the net tone of the conference call for firm *i* in quarter *q*, where net tone is defined as the number of positive words less the number of negative words divided by the sum of the number of positive and negative words. The training data include all conference calls made from Q4 of year *t-5* to Q3 of year *t-1*.

Figure 1
Timeline of Training and Test Samples



The above figure presents the timeline for how we estimate each of the statistical methods. The "Training Sample" is the period in which we estimate the relevant parameters. The "Prediction" period is when we apply the estimated parameters from the "Training Sample" to the out-of-sample conference calls.

Figure 2
Random Forest Regression Tree Example



The above figure represents an example of a simple regression tree, which splits data based on word counts to minimize the sum of squared errors (SSE) in each split of the data. WC1 (WC2) represents the word count of word 1 (2).

TABLE 1
Descriptive Statistics

This table presents the descriptive statistics for the variables used in the main empirical analyses. All variables are defined in Appendix B. All continuous variables are winsorized at the 1st and 99th percentiles. The sample spans 2005 to 2015 and includes 67,370 observations.

Variable	Mean	Std. Dev.	Min	Q1	Median	Q3	Max
$CH\ FOR_{i,q}$	-0.002	0.007	-0.038	-0.002	-0.001	0.001	0.020
$SVR\ Pred_{i,q}$	-0.002	0.015	-0.045	-0.011	-0.002	0.008	0.036
$sLDA\ Pred_{i,q}$	-0.002	0.003	-0.012	-0.003	-0.001	0.000	0.005
$RF\ Pred_{i,q}$	-0.002	0.002	-0.010	-0.003	-0.001	-0.001	0.002
$TONE\ Pred_{i,q}$	-0.005	0.007	-0.035	-0.008	-0.004	0.000	0.022
$FACTOR\ Pred_{i,q}$	0.000	1.000	-5.550	-0.534	0.130	0.684	3.024
$EARN\ SURP_{i,q}$	0.000	0.011	-0.061	-0.001	0.001	0.002	0.041
$\Delta CF_{i,q}$	0.003	0.037	-0.128	-0.012	0.002	0.017	0.141
$GUID\ SURP_{i,q}$	-0.003	0.010	-0.053	-0.003	-0.001	0.001	0.037

TABLE 2
Correlations

This table presents the Pearson and Spearman correlations for the variables used in the main empirical analyses. Pearson (Spearman) correlations are above (below) the diagonal. All correlations are significant at the 1% level. All variables are defined in Appendix B. All continuous variables are winsorized at the 1st and 99th percentiles.

		I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.
I.	$CH\ FOR_{i,q}$	1.00	0.14	0.23	0.24	0.14	0.27	0.34	0.07	0.48
II.	$SVR\ Pred_{i,q}$	0.13	1.00	0.29	0.34	0.14	0.57	0.10	0.03	0.05
III.	$sLDA\ Pred_{i,q}$	0.21	0.26	1.00	0.60	0.29	0.77	0.14	0.05	0.10
IV.	$RF\ Pred_{i,q}$	0.24	0.28	0.56	1.00	0.26	0.80	0.18	0.07	0.11
V.	$TONE\ Pred_{i,q}$	0.17	0.14	0.29	0.29	1.00	0.44	0.10	0.06	0.05
VI.	$FACTOR\ Pred_{i,q}$	0.28	0.54	0.76	0.75	0.48	1.00	0.20	0.08	0.12
VII.	$EARN\ SURP_{i,q}$	0.32	0.09	0.09	0.17	0.13	0.19	1.00	0.07	0.13
VIII.	$\Delta CF_{i,q}$	0.09	0.03	0.07	0.09	0.07	0.09	0.09	1.00	0.03
IX.	$GUID\ SURP_{i,q}$	0.57	0.08	0.12	0.17	0.09	0.18	0.23	0.05	1.00

TABLE 3
Forecast Revisions and Narrative Content Measures

This table includes all firm-quarter observations from 2005 to 2015 with sufficient data to calculate the dependent and independent variables. The dependent variable is the change in the analyst-forecast revision of $EPS_{i,q+1}$ following firm *i*'s conference call in quarter *q* ($CH\ FOR_{i,q}$). Standard errors are clustered by firm and year-quarter. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]	[2]	[3]	[4]	[5]
Intercept	-0.002***	-0.001***	-0.000**	-0.001***	-0.002***
	(-39.380)	(-26.857)	(-2.262)	(-26.988)	(-44.205)
SVR Pred$_{i,q}$	0.063***				
	(22.106)				
sLDA Pred$_{i,q}$		0.512***			
		(28.162)			
RF Pred$_{i,q}$			0.872***		
			(26.002)		
TONE Pred$_{i,q}$				0.136***	
				(23.823)	
FACTOR Pred$_{i,q}$					0.002***
					(33.767)
#OBS	67,370	67,370	67,370	67,370	67,370
Adjusted R^2	0.019	0.052	0.058	0.021	0.079

TABLE 4
Sparsity and Narrative Content

This table includes all firm-quarter observations from 2005 to 2015 with sufficient data to calculate the dependent and independent variables. The dependent variable is the change in the analyst-forecast revision of $EPS_{i,q+1}$ following firm i's conference call in quarter q ($CH\ FOR_{i,q}$). Standard errors are clustered by firm and year-quarter. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]	[2]
Intercept	-0.002***	-0.002***
	(-44.205)	(-44.518)
FACTOR Pred$_{i,q}$	0.002***	
	(33.767)	
FACTOR Pred (KEEP SPARSE)$_{i,q}$		0.002***
		(31.781)
#OBS	67,370	67,370
Adjusted R^2	0.079	0.070

TABLE 5
Context and Narrative Content

This table includes all firm-quarter observations from 2005 to 2015 with sufficient data to calculate the dependent and independent variables. The dependent variable is the change in the analyst-forecast revision of $EPS_{i,q+1}$ following firm *i*'s conference call in quarter *q* ($CH\ FOR_{i,q}$). Standard errors are clustered by firm and year-quarter. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

	[1]	[2]	[3]
Intercept	-0.002***	-0.002***	-0.002***
	(-45.829)	(-43.729)	(-45.358)
FACTOR Pred (RET VOL)$_{i,q}$	0.002***		
	(34.893)		
FACTOR Pred (QTR CAR)$_{i,q}$		0.002***	
		(36.661)	
FACTOR Pred (CONTEXT)$_{i,q}$			0.002***
			(37.011)
#OBS	67,370	67,370	67,370
Adjusted R^2	0.075	0.080	0.095

TABLE 6
Narrative Content Relative to Other Earnings Signals

This table includes all firm-quarter observations from 2005 to 2015 with sufficient data to calculate the dependent and independent variables. The dependent variable is the change in the analyst-forecast revision of $EPS_{i,q+1}$ following firm *i*'s conference call in quarter *q* ($CH\ FOR_{i,q}$). Standard errors are clustered by firm and year-quarter. All variables are defined in Appendix B. All continuous variables are winsorized at the 1% and 99% levels. *, **, and *** represent significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Narrative Content, Earnings Surprise, and Change in Cash Flows

	[1]	[2]	[3]	[4]	[5]
Intercept	-0.002***	-0.002***	-0.002***	-0.002***	-0.002***
	(-45.358)	(-39.735)	(-37.850)	(-40.083)	(-45.967)
$FACTOR\ Pred\ (CONTEXT)_{i,q}$	0.002***				0.002***
	(37.011)				(29.095)
$EARN\ SURP_{i,q}$		0.219***		0.217***	0.185***
		(30.679)		(30.422)	(26.664)
$\Delta CF_{i,q}$			0.013***	0.009***	0.006***
			(13.207)	(9.456)	(6.256)
#OBS	67,370	67,370	67,370	67,370	67,370
Adjusted R^2	0.095	0.118	0.005	0.120	0.178

Panel B: Narrative Content and Earnings Guidance

	[1]	[2]	[3]	[4]	[5]	[6]
Intercept	-0.002***	-0.002***	-0.002***	-0.000***	-0.001***	-0.001***
	(-22.933)	(-19.954)	(-19.071)	(-4.725)	(-5.848)	(-8.439)
FACTOR Pred (CONTEXT)$_{i,q}$	0.002***					0.001***
	(19.201)					(11.415)
EARN SURP$_{i,q}$		0.142***			0.074***	0.062***
		(5.210)			(3.992)	(3.630)
ΔCF$_{i,q}$			0.017***		0.009***	0.006***
			(7.601)		(5.905)	(4.265)
GUID SURP$_{i,q}$				0.480***	0.474***	0.440***
				(15.954)	(15.590)	(14.436)
#OBS	14,044	14,044	14,044	14,044	14,044	14,044
Adjusted R^2	0.141	0.021	0.009	0.499	0.508	0.546

TABLE 7
Future Abnormal Returns

This table reports average size-adjusted cumulative abnormal returns by decile of *FACTOR Pred (CONTEXT)*$_{i,q}$ and by year for all firm-quarter observations from 2005 to 2015. Panel A reports the average size-adjusted cumulative abnormal return from two days after earnings conference call for firm *i* in quarter *q* to two days before the earnings announcement for firm *i* in quarter *q+1* (*QTR CAR*$_{i,q+1}$). Panel B reports the size-adjusted cumulative abnormal return from one day before to one day after the earnings announcement of firm *i* in quarter *q+1* (*EA CAR*$_{i,q+1}$). All variables are defined in Appendix B. The *CAR* variables are unwinsorized.

Panel A: *QTR CAR*$_{i,q+1}$

FACTOR Pred (CONTEXT)$_{i,q}$ *Decile*	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
1	0.9%	0.9%	-6.9%	-5.8%	20.7%	0.1%	-2.5%	-2.9%	-0.5%	-5.6%	-8.7%	-0.93%
2	-2.1%	0.2%	-4.1%	-2.3%	7.2%	3.6%	-1.9%	-0.1%	1.4%	-2.4%	-5.2%	-0.51%
3	-0.1%	2.0%	-1.5%	-1.6%	3.1%	2.3%	-1.1%	-0.8%	-0.9%	-2.9%	-3.6%	-0.47%
4	0.8%	0.9%	-1.3%	0.8%	1.6%	1.4%	0.0%	-0.5%	0.8%	0.1%	-0.5%	0.37%
5	0.3%	-0.4%	-1.8%	-1.3%	0.4%	1.8%	-0.7%	-0.6%	1.1%	-0.8%	-0.9%	-0.25%
6	0.7%	-1.5%	-0.8%	0.7%	0.5%	0.2%	0.9%	0.2%	1.2%	0.6%	0.6%	0.28%
7	-0.3%	-0.8%	0.4%	2.4%	-0.3%	1.0%	0.9%	-0.7%	1.4%	1.3%	0.6%	0.54%
8	2.0%	-0.8%	0.5%	2.6%	0.8%	0.8%	0.7%	0.3%	2.1%	1.4%	1.7%	1.11%
9	1.1%	0.0%	0.4%	2.0%	-1.6%	0.7%	0.8%	0.4%	1.6%	1.7%	1.5%	0.78%
10	2.6%	1.2%	2.0%	1.9%	-1.6%	1.5%	1.1%	1.0%	2.8%	1.8%	1.6%	1.43%
High - Low	1.7%	0.3%	8.9%	7.7%	-22.3%	1.5%	3.5%	3.9%	3.2%	7.4%	10.3%	2.36%
												t-stat = 0.89

Panel B: *EA CAR$_{i,q+1}$*

FACTOR Pred (CONTEXT)$_{i,q}$ *Decile*	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Average
1	-0.6%	-0.8%	-0.9%	0.4%	4.4%	-0.7%	-0.9%	-1.0%	-0.9%	-1.4%	0.0%	-0.21%
2	-0.1%	-0.6%	-0.7%	-0.1%	1.5%	-0.7%	-0.8%	0.0%	-0.1%	-0.2%	0.6%	-0.10%
3	0.2%	0.8%	0.1%	-0.2%	1.2%	-0.1%	-0.3%	0.2%	-0.2%	-0.9%	-0.7%	0.01%
4	-0.5%	-0.5%	-0.5%	-0.2%	1.5%	0.2%	-0.2%	0.1%	0.2%	0.0%	-0.3%	-0.03%
5	-0.1%	-0.4%	0.4%	-0.6%	0.4%	0.0%	0.1%	-0.1%	0.5%	0.0%	-0.3%	-0.01%
6	0.6%	-0.1%	0.5%	0.9%	-0.2%	-0.1%	-0.3%	-0.5%	0.1%	-0.3%	-0.2%	0.04%
7	0.1%	0.2%	-0.1%	-1.0%	0.9%	0.0%	0.1%	-0.1%	0.2%	0.2%	-0.2%	0.04%
8	-0.1%	-0.1%	0.3%	-0.1%	0.5%	0.7%	-0.2%	-0.1%	0.3%	0.2%	-0.4%	0.08%
9	0.6%	0.4%	0.6%	1.0%	0.2%	0.5%	-0.2%	0.5%	0.4%	0.0%	0.3%	0.40%
10	0.2%	0.5%	0.6%	0.3%	1.1%	0.7%	0.5%	0.4%	0.6%	1.1%	0.0%	0.54%
High - Low	0.8%	1.3%	1.5%	-0.1%	-3.4%	1.3%	1.4%	1.5%	1.5%	2.5%	0.0%	0.74%
												t-stat = 1.60