The Informativeness of Text, the Deep Learning Approach*

Allen H. Huang [a]
Hong Kong University of Science and Technology

Hui Wang [b]
Hong Kong University of Science and Technology

Yi Yang [c]
Hong Kong University of Science and Technology

Preliminary, please do not circulate or quote without permission.

Abstract

This paper uses a deep learning natural language processing approach (Google's Bidirectional Encoder Representations from Transformers, hereafter BERT) to comprehensively summarize financial texts and examine their informativeness. First, we compare BERT's effectiveness in sentiment classification in financial texts with that of a finance specific dictionary, the naïve Bayes, and Word2Vec, a shallow machine learning approach. We find that first, BERT outperforms all other approaches, and second, pre-training BERT with financial texts further improves its performance. Using BERT, we show that conference call texts provide information to investors and that other less accurate approaches underestimate the economic significance of textual informativeness by at least 25%. Last, textual sentiments summarized by BERT can predict future earnings and capital expenditure, after controlling for financial statement based determinants commonly used in finance and accounting research.

JEL classifications: G31, G35, M41, M52

Keywords: Natural Language Processing, Machine Learning, Deep Learning, Textual Analysis, Informativeness, Earnings, Capital Investment

* We thank Tim Loughran, Diego Garcia (discussant) and workshop participants in HKUST, Ernst & Young and Bergen Fintech conference for comments. We gratefully acknowledge the financial support from the Hong Kong Research Grant Council (T31-604/18-N).
[a] *Email:* allen.huang@ust.hk, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.
[b] *Email:* hwangcr@connect.ust.hk, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.
[c] *Email:* imyiyang@ust.hk, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.

1. Introduction

There is a burgeoning literature in finance and accounting that use natural language processing (hereafter NLP) algorithms to conduct textual analysis (see reviews by Li [2010a], Das [2014], Kearney and Liu [2014], Loughran and McDonald [2016] and El-Haj, Rayson, Walker, Young and Simaki [2019]). Most, if not all, of these studies rely on NLP that assume a bag-of-words structure. That is, these models treat words as independent (Wallach [2006]) and disregard grammar and word order, i.e., the context of words (Loughran and McDonald [2016]), and represent a text as a bag of its words.[1] Research in finance and accounting usually cite early studies in NLP (e.g., Lewis [1998], Manning and Schutze [1999], Hastie, Tibshirani and Friedman [2001]), which find that despite the bag-of-words assumption, these algorithms yield results that are as effective as other contemporaneous algorithms that incorporate the internal structures of documents (Li [2010b], Huang, Zang and Zheng [2014], Buehlmaier and White [2018]).

In recent years, NLP researchers have introduced deep-learning based algorithms, such as Word2Vec (Mikolov et al. [2013]), Embedding from Language Model (ELMo, Peters et al. [2018]), Open AI Generative Pre-Training (OpenAI GPT, Radford et al. [2018]) and Bidirectional Encoder Representations from Transformers (hereafter BERT, Devlin et al. [2018]), that take into account word contexts, e.g., other words in the same text, and word sequences in summarizing texts. They show that these new algorithms can significantly

[1] These algorithms include dictionary approaches (Li, Lundholm and Minnis [2013], Loughran and McDonald [2011], Bodnaruk, Loughran and McDonald [2015]), naïve Bayes classifications (Li [2010b], Huang, Zang and Zheng [2014], Buehlmaier and White [2018]), topic modelling techniques, e.g., Latent Dirichlet Allocation algorithms (Blei, Ng and Jordan [2003], Lang and Stice-Lawrence [2015], Huang, Lehavy, Zang and Zheng [2018]), and others that measure textual features such as readability (Li [2008], Loughran and McDonald [2014], Bonsall, Leone and Miller [2018]), salience or concreteness (Lundholm, Rogo and Zhang [2014], Elliott Rennekamp and White [2015], Huang, Nekrasov and Toeh [2018]), and similarity (Brown and Tucker [2011], Hoberg and Phillips [2015], and Lang and Stice-Lawrence [2015])

outperform NLP assuming a bag-of-words structure in tasks such as sentiment classification in general texts, language translation, and answering a question (Devlin et al. [2018]).

However, compared to earlier and simpler approaches, these new NLP algorithms have several disadvantages. First, because these algorithms are based on deep neural network models with millions or even billions of parameters, they require a huge amount of textual data, substantial computing resources (e.g., large storage and graphic processing unit (GPU) or cloud servers), and more time to train and apply in research.[2][3] These constraints can impose considerable costs on researchers, especially those in finance and accounting, who may need to master a different programming language such as Python. Even computer science researchers question whether some recently developed deep learning algorithms are too complex and achieve too little improvements to justify their increase in implementation costs (Strubell, Ganesh and McCallum [2019]). Second, deep learning algorithms are notoriously opaque. Even though the algorithms' general intuitions are straightforward, it is often difficult, if not impossible, to know precisely how their trained model turn inputs into outputs or which inputs carry the most significant weights in predicting the outputs, leading many to refer to deep learning as black boxes (Castelvecchi [2016]). The lack of interpretability poses an obstacle to researchers who want to use these models to test economic theories. In this paper, we seek to quantify these algorithms' advantages over simpler, less costly, and more interpretable approaches in NLP tasks in financial texts to help finance and accounting researchers to decide whether to adopt these models.

[2] Neural network is a class of machine learning algorithms that includes an input layer, an output layer, and layers between them which are referred to as hidden layers. Deep learning, or deep neural network, refers to a subset of neural network with more than one hidden layers. See Section 2 for more details on their intuition and differences.
[3] For example, BERT (first released in 2018) has 345 million parameters and OpenAI's GPT-3 (released in 2020) has 175 billion parameters.

In ex-ante, it is not clear whether deep learning algorithms can substantially outperform simpler NLP algorithms in financial texts for several reasons. First, although computer scientists document the outperformance of deep learning algorithms, they typically compare algorithms' performance in general texts. Economics and finance researchers have introduced domain-specific word lists (i.e., finance and accounting) and demonstrated that they outperform general bag-of-words approaches (see review in Loughran and McDonald [2016]).[4] Thus, even if deep learning NLP algorithms outperform generic bag-of-words approaches in general texts, it remains to be seen how they fare against those that incorporate financial knowledge in summarizing financial texts such as annual reports or conference calls.

Second, studies in computer science usually do not examine NLP algorithms' performance in outcomes of interests to finance and accounting researchers, such as informativeness to investors or corporate decisions. Instead, they focus on tasks that have immediate real-world applications and thus more commercial appeals such as formulating an answer to a question, identifying entities from texts, and translating from one language to another. Whether newer and more complex NLP algorithms have an advantage over the simpler ones in tasks most relevant to finance and accounting researchers is an empirical question.

Third, comparisons of NLP algorithms in computer science studies are usually based on how each algorithm's output performs in isolation in the prediction tasks and do not control for other useful information. For example, they do not examine how an algorithm performs after controlling for determinants that finance or accounting research frequently use in the same prediction task, e.g., those from financial statements or market trading data. Thus, it is possible

[4] Accounting and finance researchers have constructed word- and phrase-lists that capture concepts such as forward-looking statements (Li [2010a], Muslu et al. [2015]), ethics (Loughran, McDonald and Yun [2009]), M&A and restructuring (Matsumoto, Pronk and Roelofsen [2011]), competition, and sentiments (Henry [2006; 2008], Loughran and McDonald [2016]).

that a new algorithm outperforms older ones in a prediction task simply because its summary of textual information correlates more with a financial statement ratio well documented in finance or accounting research to be related to the predicted variable. While this result may be interesting of its own right, it has limited implications for accounting and finance researchers whose goal is to measure incremental textual informativeness or improve their prediction model.

In this paper, we examine the above empirical question using the state-of-the-art NLP model: Google BERT. In particular, we use two BERT models. The first one is the original version that Google has pre-trained based on general texts and released in 2018, which has 345 million parameters. The second one is a financial-domain BERT that we further pre-train with financial texts (hereafter, Finance-BERT). During our pre-train process, we use the original BERT parameters as the initial value and allow the algorithm to update the parameters to better represent financial texts (for a detailed discussion, see Section 2.4). The additional pre-train process with BERT parameters as initial value allows the model to retain most of what it has learned about general texts and learn characteristics of texts in financial texts.[5]

We compare BERT and Finance-BERT with other approaches popular in finance and accounting including the Loughran and McDonald financial dictionary, the naïve Bayes, and Word2Vec (a shallow machine learning algorithm) in summarizing the sentiments of financial texts. First, we find that BERT significantly outperforms Loughran and McDonald's finance domain specific word list and simpler machine learning algorithms (the naïve Bayes and Word2Vec) in sentiment classification. Specifically, using a sample of 10,000 pre-labeled

[5] Recent studies have further pre-trained BERT with domain specific texts and find that the additional process improves BERT's performance in NLP task in these domains (Alsentzer et al. [2019], Lee et al. [2019], Beltagy, Lo and Cohan [2019]). See Section 2 for more details.

sentence from financial text as the training dataset, BERT achieves a classification accuracy rate of 85.5%, whereas Loughran and McDonald's finance domain specific word list, the naïve Bayes, and Word2Vec achieve accuracies of 61.7%, 82.7%, and 50.9%, respectively. Further analyses show that BERT has significantly less errors in sentences mislabeled by other algorithms as neutral, suggesting that incorporating contextual information helps uncover sentiments that bag-of-words approaches do not identify. Second, we find that incorporating finance domain knowledge further improves the performance of BERT. That is, Finance-BERT classifies sentiments with even higher accuracy (88.4% of the pre-labeled sentences).

Next, we show that these improvements carry out of sample and result in a more accurate measure of how investors interpret financial text sentiments. We first estimate the sentiments of the presentation part of 18,607 earnings conference call transcripts from 2003 to 2012 using BERT, Finance-BERT and other NLP approaches, and then use these sentiments to conduct a short-window event study to test their market reactions, measured with three-day abnormal returns. We find that while market reaction is positively associated with textual sentiments measured with all NLP algorithms, the economic significance of this association is the highest when sentiments are measured with BERT and Finance-BERT. Specifically, using Loughran and McDonald's financial dictionary results in an underestimation of the economic magnitude by 24.5% and 27% compared to BERT and Finance-BERT respectively.

We conduct two additional analyses to ascertain the outperformance of Finance-BERT. First, we compare the algorithms' performance in each industry and find that Finance-BERT outperforms Loughran and McDonald's financial dictionary in all industries except a few industries with relatively stable operational environment, namely Energy, Telecommunication

and Utilities. Next, we compare the algorithms' performance in a small sample when, due to low power, accurately measuring sentiment is especially important. Specifically, we randomly select 1,000 observations from our sample and repeat the test. We iterate this process 400 times and find that sentiments captured by BERT or Finance BERT is significant in 395 or 397 times while those captured by Loughran and McDonald's financial dictionary, the naïve Bayes, and Word2Vec are only significant in 319, 267 and 112 times respectively. This highlights that an advantage of using BERT and Finance-BERT is that their higher accuracy increases the power of empirical tests relying on them and allows researchers to test hypothesis with smaller samples.

Last, we show that conference call sentiments captured by BERT and Finance-BERT can better forecast firms future earnings and capital expenditures than those by other algorithms, after controlling for determinants from financial statements that are commonly used in finance and accounting research. Specifically, we find that Finance-BERT predicts future earnings in the subsequent two years and capital expenditures in subsequent four years with larger economic magnitudes than Loughran and McDonald's finance domain specific word list.

Our paper primarily contributes to the growing literature that examines informativeness of financial texts (e.g., Li [2010b], Larcker and Zakolyukina [2012], Huang, Zang and Zheng [2014]). We are the first to introduce BERT and Finance-BERT, two state-of-the-art unsupervised deep learning NLP algorithms to the finance and accounting studies. By documenting BERT and Finance-BERT's superior performance over other algorithms, especially Loughran and McDonald word list which is arguably the gold standard in finance and accounting, we showcase the strength and potential of deep learning NLP algorithms in financial economics research. We quantify the benefit of using BERT and Finance-BERT with economic

magnitudes. We also demonstrate that when the empirical tests lack power, e.g., when the sample size is small, using BERT and Finance-BERT reduces the chance of Type II errors. To facilitate researchers who want to use Finance-BERT, we make the pre-trained Finance-BERT available to public.

Our study also adds to NLP studies by documenting the performance of BERT in financial texts and compare them with algorithms commonly used in the literature. Prior state-of-the-art deep learning NLP algorithms such as Word2Vec tend to underperform domain specific word lists or simpler machine learning approaches such as the naïve Bayes in financial texts. By documenting BERT's superior performance over other algorithms, we show that deep learning NLP algorithms can perform well in domain-specific tasks that typical computer scientists do not focus on.

Our results have implications for the choice of NLP algorithms in financial economic research. Specifically, we show that simple bag-of-words approaches that do not require substantial computing resources can produce satisfactory results when the empirical tests have sufficient power. However, when researchers are concerned with Type II errors, they should use newer and more accurate deep learning NLP algorithms.

2. Neural Network and Deep Learning Natural Language Processing Algorithms

2.1 Neural Network

Deep learning algorithms belong to neural network, which is a class of machine learning algorithms that includes an input layer, an output layer, and hidden layer(s) between them which connects the input and output layers (see Figure 1 for a graphic demonstration of neural network algorithm). The input layer takes in the initial raw data, i.e., the texts, and outputs to the hidden

layers(s). The hidden layer(s), which are connected to the input layer, process data from the input layer.[6] The output layer is connected to the last hidden layer and presents the prediction outcome, e.g., positive, negative or neutral for sentiment classification. Specifically, each hidden layer uses the previous layer's output as inputs, multiply them by a weight matrix, add intercepts, and apply a non-linear function often referred to as an activation function.[7] That is,

$$Output^{[n]} = Activation\ function\left(W^{[n]T} Output^{[n-1]} + Intercept^{[n]}\right),$$

where $W^{[n]}$ is the weight matrix, $Intercept^{[n]}$ is the intercepts and $Output^{[n]}$ is the output, all of layer *n*. A neural network algorithm usually specifies an activation function, and an initial weight matrix and intercept for each layer, and uses the training data to find the weight matrixes and the intercepts that minimize prediction errors, i.e., the difference between the prediction from the output layer and the actual value. A logistic regression can be considered as the simplest neural network with only input and output layers, i.e., no hidden layer, and a sigmoid activation function. Neural networks with only three layers, i.e., only one hidden layer, are called shallow neural networks.

2.2 Deep Learning Algorithms

Deep learning, or deep neural network, refers to a subset of neural network with more than one hidden layers (LeCun et al. [2015]). In fact, many recent deep learning algorithms have tens or hundreds of hidden layers, which can capture more complex relations, but require larger amount of training data to estimate the coefficients.

[6] When there are multiple hidden layers, only the first hidden layer is connected to the input layer while others are connected with the previous hidden layer, i.e., the n^{th} layer is connected to the $n\text{-}1^{th}$ hidden layer, etc. The output layer is connected to the last hidden layer.

[7] The most popular activation functions are Sigmoid, Hyperbolic tangent (Tanh), and Rectified linear units (Relu). A neural network without activation function is a linear regression model.

Researchers show that deep learning has superior performance in supervised learning tasks compared with traditional shallow learning approaches including logistic regression and naïve Bayes, in many domains such as computer vision (Krizhevsky et al. [2012]), speech recognition (Hinton et al. [2012]), bioinformatics (Chen et al. [2016]), high-energy physics (Baldi et al. [2014]). In recent years, deep learning based methods have been shown to perform very well on NLP tasks such as document classification (Socher et al. [2012]) and machine translation. The advantages of deep learning in NLP tasks origin from its ability to handle a large amount of data and its deep structure that can capture textual documents' semantic meanings.

2.3 Using Neural Word Embedding to Represent Texts

Most deep learning NLP algorithms use neural word embedding (also known as word representations), which represents words with vectors and use the vectors as inputs to the deep learning NLP algorithms. Conceptually, neural word embedding is a type of language model, which predicts the probability of word occurrence given its context (surrounding words and sentences). Neural word embedding differs from previous NLP approaches such as word lists and one-hot encoding in several ways.[8] First and most importantly, neural word embedding learns the relation among words from existing corpus and thus can represents words using vectors that capture both semantic and syntactic information of words. That is, unlike traditional one-hot encoding where each word is independent of each other (i.e., each pair of words have the same distance in the vector space), words with similar meanings have vectors that are closer in

[8] One-hot encoding records whether a word (token) exists in a document. In one-hot encoding, each document is represented by a V x 1-dimension vector where V is the number of distinct words in the corpus vocabulary and each element in the vector corresponds to a unique word (token). If a word exists in a document, the vector that represents the document will take a value of one in the element that correspond to the word, and zero otherwise.

neural word embedding. For example, word pair "Japan" and "sushi" will be closer in the vector space than word pair "Japan" and "pizza" (Mikolov et al. [2013]).[9] Second, compared to the one-hot encoding, vectors in neural word embedding are usually of a much smaller dimension such as 100 or 300.[10] A smaller vector dimension reduces the number of parameters that the machine learning algorithms estimate and improves their performance. Last, unlike dictionary approaches where researchers manually group similar words, e.g., produce a list of positive and negative words, neural word embedding algorithms learn the relation among words by training on a large amount of unlabeled texts, i.e., texts that have not been pre-labeled by researchers. Thus, neural word embedding reduces the time requirement from researchers and is more objective.

Neural word embedding determines each word's vector representation by training on unlabeled texts and learning similarities in words, i.e., word relations, from their co-occurrences in the texts.[11] For example, Word2Vec, one of the earliest and most popular neural word embedding, is trained on Google News text that includes 6.6 billion words (tokens) (Mikolov et al. [2013]).[12] Since the inputs to these algorithms do not need to be labeled, they are also referred to as unsupervised machine learning algorithms.

Neural word embedding models differ in their training's objective function. For example, in Word2Vec, the algorithm predicts words immediately adjacent to a target word in a sentence based on the target word (Mikolov et al. [2013]). That is, given a sentence "Apple sues Huawei,"

[9] As another example, using Word2Vec's vector, "King – Man + Women" results in a vector very close to "Queen" (Mikolov, Yih and Zweig [2013]).
[10] As a reference, Loughran and McDonald's 2018 master dictionary (available at http://sraf.nd.edu), which includes all words in all 10-K/Qs and earnings calls from 1994-2018, has 86,486 unique words. Other more general word lists can include hundreds of thousands of word tokens.
[11] The vector size is pre-set by researchers prior to training.
[12] Word2Vec can be trained using Skip-gram model or Continuous Bag of Words (CBOW) model (Mikolov et al. [2013]). Other neural word embedding methods include Stanford's GloVe (Pennington, Socher and Manning [2014]) and Facebook's fastText model (Joulin et al. [2017]).

Word2Vec aims to predict the context word “Apple” and “Huawei” based on the target word “sues.”

Due to the complexity of human language, e.g., the large number of words, and the difficulty in the prediction task, these algorithms require a massive amount of training data and can take a long time to train. However, as most texts contain some general semantic and synthetic information such as grammar and common word meanings, many word embeddings can be first *pre-trained* on a large amount of text. Researchers who want to use word embeddings for NLP tasks such as sentiment analysis can then fine-tune the pre-trained algorithms using smaller and labeled texts, substantially reducing the implementation costs.

Earlier word embedding models, such as Word2Vec, Stanford’s GloVe (Pennington, Socher and Manning [2014]) and Facebook’s fastText (Joulin et al. [2017]), are context independent embedding, i.e., each word is represented by a static vector, regardless of its surrounding words. Recent word embedding models such as ELMo (Peters et al. [2018]), OpenAI GPT (Radford et al. [2018]) and BERT, are contextualized embeddings, i.e., words have different vectors depending on their contexts. For example, in the BERT model, the word “bank” has different embeddings in sentences such as “I went to the bank to deposit a check” and “We walk along the river bank.” Incorporating the context in word embedding has yielded significant improvement in NLP tasks (Peters et al. [2018], Devlin et al. [2018]). In this paper, we use a contextualized word embedding, Google’s BERT.

2.4 Google’s BERT

In training the original Google’s BERT, the training function is to predict masked words (15% of words in a sentence) based on the remaining words in the sentence, and to predict the next sentence (masked sentence) based on the previous sentence, and to minimize the combined

loss function of the two prediction tasks. Google's BERT is trained on general text corpus including Wikipedia and BooksCoprus with a total of 3.3 billion word tokens. Training BERT on such large scale text data requires extensive computing resources. According to Google, the training is performed on 16 Google Cloud TPUs and takes four days to complete. The trained BERT has 340 million parameters. BERT has achieved state-of-the-art results in a wide variety of NLP tasks, and Google uses this technology to improve its search engine results.

Since BERT is pre-trained by Google using general texts including Wikipedia and BooksCorpus, BERT's word embedding may not capture domain specific knowledge. For example, the pre-trained word embedding for the word "allowance" may include multiple meanings such as the money given by parents to a child and an amount of something that someone is allowed to have, while "allowance" may predominantly refer to an amount that is planned for future cost in financial texts. One way to incorporate more domain specific knowledge into the pre-trained word embedding is to pre-train it with texts from this domain text. That is, adjust the pre-train BERT word embedding using domain specific texts such that words with similar meanings in this domain will be closer in the embedding vector space, e.g., "bank" will be closer with words such as "lending" and further with words such as "fish" or "river."[13] Recent studies have shown that fine-tuning improves the performance of BERT in NLP tasks in biomedical text and computer science text (Alsentzer et al. [2019], Lee et al. [2019], Beltagy, Lo and Cohan [2019]). Therefore, we further pre-train BERT with financial texts including corporate filings, analyst reports and conference calls and refer to this version as the Finance-BERT.

[13] During our pre-train process, we use the original BERT's parameters as the initial value and allow all parameters to change based on the information it learns from the financial texts.

2.5 Finance BERT

We obtain three types of financial textual data to pre-train a finance-specific BERT model: 10-K and 10-Q, Earnings Conference Call transcripts and Analyst Reports. First, we obtain 60,490 Form 10-Ks and 142, 622 Form 10-Qs of Russell 3000 firms during 1994 and 2019 from SEC's EDGAR website. We further parse 10-Ks and 10-Qs into different items to only include items that are most relevant for investors and contain the least amounts of tables. For Form 10-Ks, we include three sections: Item 1 (Business), Item 1A (Risk Factors) and Item 7 (Management's Discussion and Analysis). For Form 10-Qs, we include two sections: Item 2 (Management's Discussion and Analysis) and Item 1A (Risk Factors). Second, we obtain 136,578 earnings conference call transcripts of 7,740 public firms between 2004 and 2019 from SeekingAlpha website. During an earnings conference call, company executives discuss the latest firm developments and financial results with investors and analysts. We include both the presentation section and the question and answer section of the conference calls. The last financial text we use is analyst reports. Financial analysts are the most important information intermediary in the financial market and prior research consistently find that their written reports provide information to investors (Huang, Zang and Zheng [2014], Huang et al. [2018]). We obtain 476,633 analyst reports issued for S&P 500 firms between 2003 and 2012 from Thomson Investext database.

We follow the suggestion in Devlin et al. [2018], the original BERT paper, to pre-train the Finance-BERT. Specifically, we use the parameters in the BERT model released by Google as the initial value, and pre-train the Finance-BERT model on all three sources of financial texts discussed previously. Our financial texts include a total of 4.9 billion tokens, comparable in size to the 3.3 billion tokens from English Wikipedia and BookCorpus that Google used to pre-train

the original BERT. We use the original BERT code to train our finance-specific BERT with the same configuration as BERT model.

We pre-train Finance-BERT using an Nvidia DGX-1 server with four Tesla P100 GPUs and 128 GB of GPU memory. To utilize all four GPUs at the same time, we use Horovord, Uber's distributed training framework, which allows for multi-GPU training. Overall, the total time taken to perform the additional pre-training is approximately two days.

3. Variable Measurement and Research Design

3.1 Tone Measurement

We use five approaches to measure of tone of conference calls, including two neural word embedding (BERT and Finance-BERT) and three other NLP approaches (Loughran and McDonald's financial dictionary, the naïve Bayes and Word2Vec). When applying the Loughran and McDonald's financial dictionary, we define the sentence to be negative when it contains at least one negative word, positive when it contains only positive words, and neutral otherwise.[14] Each approach yields the number of positive (N_{pos}), negative (N_{neg}), and neutral (N_{neu}) sentences. To measure the overall sentiment in the presentation part of an earnings conference call transcript, we use the metric: $Tone_j = Pos_j - Neg_j$, where Pos_j (Neg_j) is the percentage of positive (negative) sentences and *j* represents the approach we use.

3.2 Research Design

First, we examine whether BERT and Finance-BERT can better capture conference call tone compared to other NLP approaches. Specifically, we use investors' reaction to the

[14] In a sensitivity test, we define a sentence as negative when it has more negative words than positive words, positive when it has more positive words than negative words and neutral otherwise. We find similar results (untabulated).

conference call to represent the "true" underlying sentiment and compare its correlation with tones measured by different approaches. A higher correlation indicates that an approach offers a closer approximation of how investors interpret the conference calls, and thus better performance. Compared to solely using NLP algorithms' performance in the human labeled sample, this method is more reliable and objective (Jegadeesh and Wu 2013).

We use two measures of market reaction, the cumulative abnormal returns and the abnormal trading volume, both in the three-day window surrounding the conference call date. We estimate following regression for abnormal returns:

$$CAR = \alpha + \beta_1 Tone + \beta_2 Earn + \beta_3 UE + \beta_4 Size + \beta_5 Loss + \varepsilon \qquad (1)$$

CAR is calculated as cumulative abnormal returns in three-day window around the conference call hosting date. Following prior literature (e.g., Henry and Leone, 2016), we control for the current quarter earnings (*Earn*), unexpected earnings (*UE*, the difference between the actual earning and the consensus analyst earnings forecast immediately prior to the conference call), firm size (*Size*), and an indicator of loss (*Loss*). We also include year fixed effect to control for macroeconomic trends. In Equation (1), each tone measure is normalized to have a mean of zero and a standard deviation of one. All standard errors are clustered at the firm level.

Following Bamber et al. (1997), we estimate the following regression for the abnormal trading volume:

$$AbVol = \alpha + \beta_1 Pos_j + \beta_2 Neg_j + \beta_3 |UE| + \beta_4 Size + \beta_5 MTB + \varepsilon \qquad (2)$$

where $AbVOL$ is the cumulative abnormal trading volumes in the three-day window surrounding the conference call date and abnormal daily trading volume is the difference between a firm's trading volume and the mean volume during the 60 days before the conference call date, scaled by the standard deviation of volume during the same window. We separately include positive

and negative sentiments because volume is non-directional. We include firm size, absolute values of unexpected earnings (|*UE*|), and market to book ratio (*MTB*). As in Eq. (1), both tone measures are normalized (rescaled to have a mean of zero and a standard deviation of one) to facilitate interpreting coefficient magnitudes. All standard errors are clustered at the firm level.

Last, we examine whether the tone measures generated by BERT or Finance-BERT can predict future earnings and capital expenditures. We use a similar regression model from Li (2010) and test whether the model using sentiment measures generated by Finance-BERT has higher explanatory power than that incorporating tone measures from Loughran and McDonald's financial dictionary.

$$Earn(Capex) = \alpha + \beta_1 Tone_j + \gamma Controls + \varepsilon \qquad (3)$$

Earnings (*Earn*) are annual earnings in the subsequent four years scaled by total assets. Capital expenditures (*CAPEX*) are the annual capital expenditures in the subsequent four years scaled by total assets. Following Li (2010), we control for current performance (*Earn, Return*), firm size (*Size*), growth opportunities (*MTB*), firm age (*Age*), accruals (*Accruals*), the volatility of operations (*EarnVol, ReturnVol*), complexity of operations (*#BusSeg*, *#GeoSeg*), firm events (*SEO*, *M&A*) and special items (*SI*).

4. Empirical Results

4.1 Performance of BERT in Sentiment Classification

We use the 10,000 sentences randomly selected from analyst reports, used in Huang, Zang and Zheng (2014), as the training dataset BERT as well as naïve Bayes and Word2Vec. The authors in that study read the 10,000 sentences and label them as positive, negative and neutral based on the sentiments. The training dataset has a total of 3,577 positive, 4,586 neutral,

and 1,837 negative sentences. In Table 1, we compare the sentiment classification of NLP algorithms and human. Specifically, in Panel A, we compare the classification of the Finance-BERT with that of human and in Panels B to E, we separately compare the remaining approaches, including BERT, LM word list, the naïve Bayes and Word2Vec with human and Finance-BERT respectively.

In sum, Table 1 shows that when using human-labeled sentiment as the benchmark, BERT and Finance-BERT outperform other NLP models sentiment classification. First, Finance-BERT achieves the highest overall accuracy (88.4%) among all NLP algorithms in our study while the original BERT from Google achieves a classification accuracy rate of 85.5%. The remaining approaches result in accurate rate ranges from 61.7% for LM word list, 82.7% for the naïve Bayes, and 50.9% for Word2Vec. The results suggest that first, machine learning algorithms that consider word contexts (i.e., BERT and Finance-BERT) do out-perform bag-of-words approaches. Second, incorporating domain specific knowledge improves NLP algorithms, e.g., comparing Finance-BERT with BERT and Word2Vec.

Next, we explore the accuracy of NLP models in different categories of sentiments. We observe that Finance-BERT and BERT obtain consistent results across sentiment types. In sentences labeled by researchers as positive, neutral and negative, the two algorithms correctly identify them in 89.6% and 85.2%, 88.8% and 88.4%, and 84.9% and 78.7% of the cases respectively. On the other hand, other algorithms are far less consistent. For example, although the LM word list's classification is consistent with researchers in 82.1% of neutral sentences, it classifies only 39.3% (54.5%) of researcher-labeled positive (negative) sentences as such, with the majority classified as neutral. That is, LM significantly understates the proportion of sentences with sentiments in financial text. We observe different patterns for the remaining two

algorithms. The naïve Bayes' performance is comparable to BERT and Finance-BERT in positive and neutral sentences (83.2% and 91.3% respectively), but it only correctly identifies negative sentences 60.2% of the time; while Word2Vec performs reasonably well in neutral statements (83.4%), but it only correctly identify a minority of sentences with positive and negative tones. In fact, it almost never correctly identifies negative sentences (3.1% accuracy).

Overall, the results show that BERT can mimic humans in classifying sentiments in financial text, and incorporating finance domain knowledge, through pre-training in financial texts, further improves its performance. The outperformance of BERT and Finance-BERT compared to the Loughran and McDonald's financial dictionary, the naïve Bayes and Word2Vec, is primarily in the former's ability in identifying sentences with positive and negative tones, likely due to their incorporation of contextual information.

To provide more specific details about the outperformance of Finance-BERT, Table 2 shows the capability of Finance-BERT to correct sentences mislabeled by the Loughran and McDonald's financial dictionary. In Panel A, for all positive sentences labeled by human, only 39.3% of them are identified as positive by Loughran and McDonald's financial dictionary, but Finance-BERT improves the accuracy by identifying positive sentences that are mislabeled by Loughran and McDonald's financial dictionary as neutral (85.6% of them) and negative (82.6% of them). Panel B shows that for all neutral sentences mislabeled as positive (negative) by Loughran and McDonald's financial dictionary, 70.1% (82%) of them are amended by Finance-BERT. Panel C reports that only 54.5% negative sentences are identified by Loughran and McDonald's financial dictionary and Finance-BERT correctly define sentences mislabeled as positive (74.3%) and neutral (78.5%). In sum, Finance-BERT has higher accuracy than Loughran

and McDonald's financial dictionary because more sentiments are identified by BERT when considering the context.

4.2 Sample Description and Descriptive Statistics

To examine whether the superior performance of BERT and Finance-BERT in sentiment classification in the training dataset results in a more accurate measure of how investors interpret financial text sentiments, we implement the NLP algorithms on the presentation sections of earnings conference call transcripts. We first download all transcripts from the Thomson Reuters StreetEvents database from 2003 to 2012 and then analyze the sentiments of presentation sections using five NLP approaches (BERT, Finance-BERT, Loughran and McDonald's financial dictionary, the naïve Bayes, and Word2Vec). As shown in Table 3, our sample consists of 18,607 transcripts from 690 companies. We require each company to have a matching GVKEY as well as PERMNO and obtain necessary variables from Compustat and CRSP. Finally, we construct a sample of 16,840 transcripts from 632 companies.

Table 4, Panel A shows the descriptive statistics for tone scores of presentation parts from earnings conference call transcripts. The overall sentiments of presentation sections tend to be positive because both the average and median values of all tone measures are larger than zero. Although tone measures from different approached show consistency in net positive tone of presentation sections, they have different levels of specific scores. The mean and median values of $Tone_{BERT}$ (mean=0.38 median=0.38), $Tone_{FinBERT}$ (mean=0.39 median=0.4), $Tone_{NB}$ (mean=0.43 median=0.43) and $Tone_{W2V}$ (mean=0.26 median=0.25) are significantly larger than the mean and median values of $Tone_{LM}$ (mean=0.09 median=0.1). The mean and median values of $Tone_{BERT}$ and $Tone_{FinBERT}$ are significantly larger than those of $Tone_{W2V}$ but significantly smaller than those of $Tone_{NB}$. Table 4, Panel B shows the correlation among these tone scores.

The BERT and Finance-BERT have very similar sentiment classifications because the correlation is nearly 97%, while the correlation between BERT and other approaches ranges from 47% to 77%. Compared with other tone measures, $Tone_{BERT}$ and $Tone_{FinBERT}$ have the largest magnitude of coefficients with *CAR (-1, +1)*.

4.3 Tone and Market Reaction

First, we employ Equation (1) and (2) to examine whether the improvements of BERT and Finance-BERT result in a more accurate measure of how investor interpret financial text sentiments. Table 5, Panel A provides the results of Equation (1). Column (1) to (5) use tone measures from Finance-BERT, BERT, Loughran and McDonald's financial dictionary, the naïve Bayes, and Word2Vec, respectively. All tone measures are significantly and positively related to three-day cumulative abnormal returns, which means that the net positive tone is correlated with a higher abnormal return around the conference call hosting date. Further, $Tone_{FinBERT}$ has the largest economic significance (e.g., one standard increase in tone increase the three-day CAR by 73.4%). For example, using Loughran and McDonald's financial dictionary results in an underestimation of the economic magnitude by 24.5% and 27% compared to BERT and Finance-BERT respectively. Then, we employ Vuong (1989) tests to compare the explanatory power of models incorporating different tone measures. In Panel B of Table 5, the results suggest that the model incorporating the tone measure from Deep NLP approaches (e.g., BERT or Finance-BERT) has greater explanatory power than other models.

Table 5, Panel C represents the results of Equation (2) and shows that all positive and negative tone measures are positively associated with abnormal trading volumes. Meanwhile, the positive tone has a larger coefficient in all models, which means that investors react more actively to positive sentiment signals. In Column (1), both $Neg_{FinBERT}$ and $Pos_{FinBERT}$ have the

largest economic significance than other NLP approaches. For example, in Column (3), positive (negative) tone measured with Loughran and McDonald's financial dictionary is significant with an underestimation of the economic magnitude by 18.7% (28%) and 16.8%(22.7%) compared with Finance-BERT and BERT. Further, Panel D of Table 5 shows that the model using tone measures from Finance-BERT outperforms models using tone measures from other approaches in explaining trading volumes except the naïve Bayes model. In sum, while the market reaction (e.g., three-day car, abnormal trading volume) is positively related to textual sentiments of conference calls measured with all NLP algorithms, the economic significance is the highest when textual sentiments are measured with Finance-BERT. Finance-BERT considers the context and is fine-tuned with financial texts when determining the textual sentiments, thus it gives us better measurements for investors' perceived sentiments.

Second, we conduct another two analyses to ascertain the outperformance of Finance-BERT. To begin with, we compare the algorithms' performance in small samples when accurately measuring sentiment is important. We randomly select 1,000 observations from our sample and estimate Equation (1) and (2) respectively. We iterate this process for 400 times and the summary statistics of the coefficients of main tone measures are reported in Table 6, including the mean of coefficients (*Mean*), the frequency of sentiment effects that are in the positive direction (*#Positive sign*), and the frequency of sentiment effects that are in the positive direction and significant at either 10%, 5%, 1% level (*#Positively significant*). Table 6, Panel A replicates regressions of abnormal return on textual sentiments and shows that 100% (400 times) of $Tone_{FinBERT}$ coefficients are positive, with 99.25% (397 times), 98% (392 times), and 92% (369 times) of them significantly positive at the 10%, 5%, and 1% level. Sentiments captured by BERT have similar performance with Finance-BERT, but those captured by Loughran and

McDonald's financial dictionary, the naïve Bayes, and Word2Vec are only significant at a 10% level in 319, 267 and 112 times respectively. Table 6, Panel B replicates the regressions of abnormal trading volume on textual sentiments using small samples. Consistent with results shown in Panel C of Table 5, coefficients of positive tone measures have a larger magnitude than those of negative tone measures, which means that investors respond more actively to positive sentiments. Panel B of Table 6 shows that nearly 100% of $Pos_{FinBERT}$ coefficients are positive with 81.25% (325 times), 71% (284 times), and 47.25% (189 times) of them significantly positive at the 10%, 5%, and 1% level. Although positive tone labeled by the naïve Bayes has the highest frequency to be positively significant, negative tone labeled by the naïve Bayes has a much weaker performance than BERT or Finance-BERT. In short, both Finance-BERT and BERT outperform other algorithms in small samples when summarizing investors' perception of financial texts. This highlights that an advantage of using BERT and Finance-BERT is that their higher accuracy increases the power of empirical tests and allows researchers to test hypotheses with smaller samples.

In the second additional analysis, we compare the algorithms' performance in each industry. To simplify the analysis, we compare Finance-BERT to Loughran and McDonald's financial dictionary based on all sentences from presentation parts of conference call transcripts. We do the comparison across two-digit GIC industries and the results are reported in Table 7. Panel A of Table 7 tabulates the statistics of discrepancy in the Loughran and McDonald's financial dictionary labeled tone and the Finance-BERT labeled tone across industries. The statistics are the number of sentences labeled by the two algorithms as a percentage of the total number of sentences in presentation sections of all firms within a given industry. For each industry, more than 13% (4%) of sentences labeled as positive (negative) by Finance-BERT are

defined as neutral by Loughran and McDonald's financial dictionary and more than 4% (2.3%) of sentences labeled as neutral by Finance-BERT are classified as negative (positive) by Loughran and McDonald's financial dictionary. Because Finance-BERT considers the context (e.g., the sequence of words) of each sentence rather than searching pre-defined negative or positive words, the discrepancy in the Finance-BERT labeled tone and the Loughran and McDonald's financial dictionary labeled tone is more prominent when the context of a sentence is complex. For example, after summing the statistics of each row in Panel A of Table 7, we find that the difference between Finance-BERT and Loughran and McDonald's financial dictionary is lowest in the industry with stable operational environment, such as Energy (GIC 10) and Utilities (GIC 55), while the difference is largest in industries with developing technology and professional knowledge, such as Financials (GIC 40), Industrials (GIC 20) and Health Care (GIC 35)[15]. We also compare the performance of Finance-BERT with Loughran and McDonald's financial dictionary in Equation (1) and (2) across all two-digit GIC industries and report the results in Panel B and C of Table 7, respectively. For the relation between textual sentiment and abnormal return in a short window around conference call hosting date, Panel B of Table 7 indicates that $\text{Tone}_{\text{FinBERT}}$ has higher economic significance than Tone_{LM} in all industries except those industries (including Energy, Communication Services and Utilities) with the lowest discrepancy in Finance-BERT labeled tone and Loughran and McDonald's financial dictionary labeled tone. Similarly, for the relation between textual sentiment and abnormal trading volume, Panel C of Table 7 shows that tone measures from Finance-BERT have more significant coefficients and outperform those from Loughran and McDonald's financial dictionary in Industrials (GIC 20) which has a second largest discrepancy in sentiment

[15] Details of GIC industry classification can be find in Wiki.
https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard

classifications. These results are consistent with that Finance-BERT achieves dramatic improvements over Loughran and McDonald's financial dictionary in summarizing sentiments especially when the context is complex.

To detect in which situation Finance-BERT outperforms Loughran and McDonald's financial dictionary in summarizing sentiment, we re-estimate Equation (1) and (2) using the subsamples that below or over the median of given firm characteristics. We use three measures to partition the full sample, respectively, including firm performance (*Earn*), firm size (*Size*), and tone difference between Finance-BERT and Loughran and McDonald's financial dictionary (*ToneDiff*). Table 8, Panel A provides the results of Equation (1), which shows that Finance-BERT outperforms Loughran and McDonald's financial dictionary in all situations. This conclusion also holds for results of Equation (2) reported in Panel B of Table 8.

4.4 Tone and Firm's Future Earnings and Investments

To further examine the information content of tone measures, we employ Equation (3) to detect whether tone measures generated by Finance-BERT can forecast the firm's future earnings and investments and use Vuong tests to compare the explanation power of models. In the beginning, we examine the capability of the Finance-BERT's tone measure to predicting future earnings. Table 9 tabulates the regression results of earnings in the subsequent four years and reports the results of Vuong tests in the last two rows. From Column (1) and (3), we find that $\text{Tone}_{\text{FinBERT}}$ is positively associated with future earnings and the relation lasts for two years. These results indicate that the sentiment of management presentation predicts performance in the future. Moreover, Vuong tests support that predicting ability of tone measure generated by Finance-BERT outperforms that by Loughran and McDonald's financial dictionary in the subsequent two years.

Next, we assess whether the sentiment in presentation sections can predict future capital expenditures. In Table 10, column (1) to (8) represent the results of the future four years' capital expenditures and the last two rows provide the results of Vuong tests. From Column (3), (5), and (7), we observe that $\text{Tone}_{\text{FinBERT}}$ is positively and significantly related to future capital expenditures. The results suggest that if a manager gives a more positive presentation during the earnings conference call, the investments in the following four years are significantly higher. Vuong tests also document that sentiment produced by Finance-BERT can better predict a firm's future investment than that by Loughran and McDonald's financial dictionary.

5. Conclusion

This paper introduces state of the art unsupervised deep learning NLP algorithms that can take into account the contextual relation among words and examines their effectiveness in summarizing sentiments in financial texts. We first show that Google's Bidirectional Encoder Representations from Transformers (BERT) can significantly outperform other approaches popular in finance and accounting including the Loughran and McDonald financial dictionary, the naïve Bayes, and a shallow machine learning algorithm (Word2Vec), especially in sentences mislabeled by other algorithms as neutral, suggesting that incorporating contextual information helps uncover sentiments that bag-of-words approaches do not identify. We further document that pre-training BERT with financial texts (Finance-BERT), which allows it to learn the contextual relation of words in the finance domain, improves the algorithm's performance.

Next, we show that these improvements carry out of sample and result in a more accurate measure of how investors interpret financial text sentiments. Specifically, conference call texts are informative to investors and other less accurate approaches underestimate the economic significance of textual informativeness by at least 25%. Moreover, we show that BERT and

Finance-BERT are significantly more robust when the empirical analyses have lower power due to small sample size. Last, we show that textual sentiments summarized by Finance-BERT can better predict future earnings and capital expenditure, after controlling for financial statement based determinants commonly used in finance and accounting research.

In addition to demonstrating the effectiveness and potential of deep learning NLP algorithms in financial economic research, our results have implications for the choice of NLP algorithms. Specifically, we show that simple bag-of-words approaches that do not require substantial computing resources can produce satisfactory results when the empirical tests have sufficient power. However, when researchers are concerned with Type II errors, they should use newer and more accurate deep learning NLP algorithms.

References

ALSENTZER, E.; J. R. MURPHY; W. BOAG; W.-H. WENG; D. JIN; T. NAUMANN and M. MCDERMOTT. "Publicly available clinical BERT embeddings." *Working Paper* (2019).

BALDI, P.; P. SADOWSKI and D. WHITESON. "Searching for exotic particles in high-energy physics with deep learning." *Nature communications* **5** (2014): 4308.

BAMBER, L. S.; O. E. BARRON and T. L. STOBER. "Trading Volume and Different Aspects of Disagreement Coincident with Earnings Announcements." *The Accounting Review* **72** (1997): 575-597.

BELTAGY, I.; A. COHAN and K. LO. "Scibert: Pretrained contextualized embeddings for scientific text." *Working Paper* (2019).

BLEI, D. M.; A. Y. NG and M. I. JORDAN. "Latent dirichlet allocation." *Journal of machine Learning research* **3** (2003): 993-1022.

BODNARUK, A.; T. LOUGHRAN and B. MCDONALD. "Using 10-K text to gauge financial constraints." *Journal of Financial and Quantitative Analysis* **50** (2015): 623-646.

BONSALL IV, S. B.; A. J. LEONE; B. P. MILLER and K. RENNEKAMP. "A plain English measure of financial reporting readability." *Journal of Accounting and Economics* **63** (2018): 329-357.

BROWN, S. V. and J. W. TUCKER. "Large-sample evidence on firms' year-over-year MD&A modifications." *Journal of Accounting Research* **49** (2011): 309-346.

BUEHLMAIER, M. M. and T. M. WHITED. "Are financial constraints priced? Evidence from textual analysis." *The Review of Financial Studies* **31** (2018): 2693-2728.

CHEN, Y.; Y. LI; R. NARAYAN; A. SUBRAMANIAN and X. XIE. "Gene expression inference with deep learning." *Bioinformatics* **32** (2016): 1832-1839.

DAS, S. R. "Text and context: Language analytics in finance." *Foundations and Trends® in Finance* **8** (2014): 145-261.

DEVLIN, J.; M.-W. CHANG; K. LEE and K. TOUTANOVA. "Bert: Pre-training of deep bidirectional transformers for language understanding." *Working Paper* (2018).

EL-HAJ, M.; P. RAYSON; M. WALKER; S. YOUNG and V. SIMAKI. "In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse." *Journal of Business Finance & Accounting* **46** (2019): 265-306.

ELLIOTT, W. B.; K. M. RENNEKAMP and B. J. WHITE. "Does concrete language in disclosures increase willingness to invest?" *Review of Accounting Studies* **20** (2015): 839-865.

FRIEDMAN, J.; T. HASTIE and R. TIBSHIRANI. *The elements of statistical learning*. Volume 1: Springer series in statistics New York, 2001.

HENRY, E. "Market reaction to verbal components of earnings press releases: Event study using a predictive algorithm." *Journal of Emerging Technologies in Accounting* **3** (2006): 1-19.

HENRY, E. "Are investors influenced by how earnings press releases are written?" *The Journal of Business Communication (1973)* **45** (2008): 363-407.

HENRY, E. and A. J. LEONE. "Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone." *The Accounting Review* **91** (2015): 153-178.

HINTON, G.; L. DENG; D. YU; G. DAHL; A.-R. MOHAMED; N. JAITLY; A. SENIOR; V. VANHOUCKE; P. NGUYEN and B. KINGSBURY. "Deep neural networks for acoustic modeling in speech recognition." *IEEE Signal processing magazine* **29** (2012).

HOBERG, G. and G. M. PHILLIPS. "Industry choice and product language." *Working Paper* (2015).

HUANG, A. H.; R. LEHAVY; A. Y. ZANG and R. ZHENG. "Analyst information discovery and interpretation roles: A topic modeling approach." *Management Science* **64** (2018): 2833-2855.

HUANG, A. H.; A. Y. ZANG and R. ZHENG. "Evidence on the information content of text in analyst reports." *The Accounting Review* **89** (2014): 2151-2180.

HUANG, X.; A. NEKRASOV and S. H. TEOH. "Headline salience, managerial opportunism, and over-and underreactions to Earnings." *The Accounting Review* **93** (2018): 231-255.

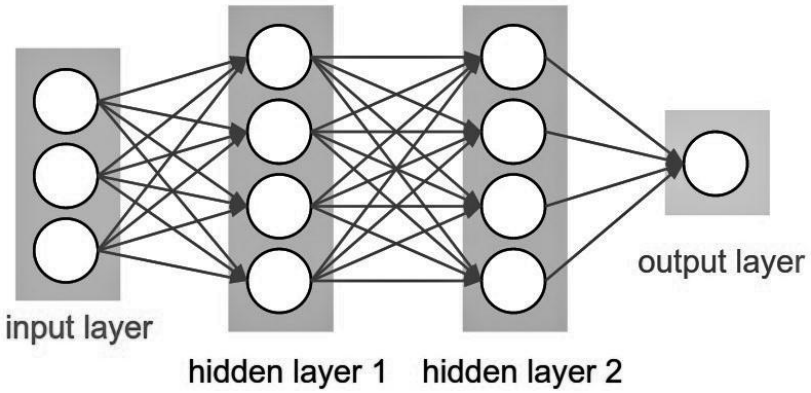
JOULIN, A.; E. GRAVE; P. BOJANOWSKI; M. DOUZE; H. JÉGOU and T. MIKOLOV. "Fasttext. zip: Compressing text classification models." *Working Paper* (2017).

KEARNEY, C. and S. LIU. "Textual sentiment in finance: A survey of methods and models." *International Review of Financial Analysis* **33** (2014): 171-185.

KRIZHEVSKY, A.; I. SUTSKEVER and G. E. HINTON. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* (2012): 1097-1105.

LANG, M. and L. STICE-LAWRENCE. “Textual analysis and international financial reporting: Large sample evidence.” *Journal of Accounting and Economics* **60** (2015): 110-135.
LARCKER, D. F. and A. A. ZAKOLYUKINA. “Detecting deceptive discussions in conference calls.” *Journal of Accounting Research* **50** (2012): 495-540.
LECUN, Y.; Y. BENGIO and G. HINTON. “Deep learning.” *Nature* **521** (2015): 436.
LEE, J.; W. YOON; S. KIM; D. KIM; S. KIM; C. H. SO and J. KANG. “Biobert: pre-trained biomedical language representation model for biomedical text mining.” *Working Paper* (2019).
LEWIS, D. D. “Naive (Bayes) at forty: The independence assumption in information retrieval.” *European conference on machine learning* (1998): 4-15.
LI, F. “Annual report readability, current earnings, and earnings persistence.” *Journal of Accounting and economics* **45** (2008): 221-247.
LI, F. “Survey of the Literature.” *Journal of accounting literature* **29** (2010): 143-165.
LI, F. “The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach.” *Journal of Accounting Research* **48** (2010): 1049-1102.
LI, F.; R. LUNDHOLM and M. MINNIS. “A Measure of Competition Based on 10-K Filings.” *Journal of Accounting Research* **51** (2013): 399-436.
LOUGHRAN, T. and B. MCDONALD. “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks.” *The Journal of Finance* **66** (2011): 35-65.
LOUGHRAN, T. and B. MCDONALD. “Measuring readability in financial disclosures.” *The Journal of Finance* **69** (2014): 1643-1671.
LOUGHRAN, T.; B. MCDONALD and H. YUN. “A wolf in sheep’s clothing: The use of ethics-related terms in 10-K reports.” *Journal of Business Ethics* **89** (2009): 39-49.
LOUGHRAN, T. I. M. and B. MCDONALD. “Textual Analysis in Accounting and Finance: A Survey.” *Journal of Accounting Research* **54** (2016): 1187-1230.
LUNDHOLM, R. J.; R. ROGO and J. L. ZHANG. “Restoring the tower of Babel: How foreign firms communicate with US investors.” *The Accounting Review* **89** (2014): 1453-1485.
MANNING, C. D. and H. SCHÜTZE 'Foundations of statistical natural language processing, vol. 999,' in *Book Foundations of statistical natural language processing, vol. 999*, edited by Editor. City: MIT Press, 1999.
MATSUMOTO, D.; M. PRONK and E. ROELOFSEN. “What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions.” *The Accounting Review* **86** (2011): 1383-1414.
MIKOLOV, T.; K. CHEN; G. CORRADO and J. DEAN. “Efficient estimation of word representations in vector space.” *Working Paper* (2013).
MUSLU, V.; S. RADHAKRISHNAN; K. R. SUBRAMANYAM and D. LIM. “Forward-Looking MD&A Disclosures and the Information Environment.” *Management Science* **61** (2015): 931-948.
PENNINGTON, J.; R. SOCHER and C. MANNING. 'Glove: Global vectors for word representation.' Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Year.
PETERS, M. E.; M. NEUMANN; M. IYYER; M. GARDNER; C. CLARK; K. LEE and L. ZETTLEMOYER. “Deep contextualized word representations.” *Working Paper* (2018).
RADFORD, A.; K. NARASIMHAN; T. SALIMANS and I. SUTSKEVER. “Improving language understanding by generative pre-training.” *Working Paper* (2018).
SOCHER, R.; Y. BENGIO and C. D. MANNING. “Deep learning for NLP (without magic).” *Tutorial Abstracts of ACL 2012* (2012): 5-5.
STRUBELL, E.; A. GANESH and A. MCCALLUM. “Energy and Policy Considerations for Deep Learning in NLP.” *Working Paper* (2019).
VUONG, Q. H. “Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses.” *Econometrica* **57** (1989).
WALLACH, H. M. 'Topic modeling: beyond bag-of-words.' Proceedings of the 23rd international conference on Machine learning.

Figure 1 A graphic demonstration of 3-layer neural network algorithm, including one input layer with three inputs, two hidden layers of 4 neurons each and one output layer.



Appendix: Variable Definition

Name	Definition
CAR	Cumulative abnormal returns from day -1 to day +1 around announcement date, multiplied by 100. Abnormal returns are defined as raw returns minus value-weighted market return.
AbVOL	Cumulative abnormal volume from day -1 to day +1 around announcement date, where abnormal volumes are defined as $\frac{V_t - M}{S}$. V is the trading volume in a stock on day t. M is the mean, and S is the standard deviation of its trading volume during 60-day period that end five days prior to the announcement date.
Earn	The quarterly/yearly earnings scaled by the book value of assets.
CAPEX	Capital expenditure scaled by total asset.
Tone	(#Positive sentences - #Negative sentence)/# Total sentences. We define tone on the sentence level. In regressions, tone variables are standardized (rescaled to have a mean of 0 and a standard deviation of 1). Under BERT/ Finance BERT and Word2Vec, each word in a sentence is represented as a vector and is positive (negative) if the vector is closer to the positive (negative) word vector. Then, aggregate the sentiment to a sentence level with weight in some fashion. In Word2Vec, each word is represented by a single static vector regardless of its context, while BERT/ Finance BERT considers the context of the word which has dynamic vectors. In Loughran-McDonald (LM), if a sentence with #negative words>1, then this sentence is negative; if a sentence with #positive words >1 and without negative words, then this sentence is positive; otherwise, the sentence is neutral. In naïve Bayes (NB), the sentence is reduced into a list of words with each word weighted in some fashion (e.g., the frequency of the word in the sentence) and is classified into a specific category (positive, negative and neutral) which has the highest aggregate probability (multiply each word's probability under a specific category and the probability of a specific category).
Neg	The number of negative sentences scaled by the total number of sentence in the file.
Pos	The number of positive sentences scaled by the total number of sentence in the file.
UE	Actual EPS minus analyst estimated EPS, scaled by share price and multiplied by 100. The analyst estimated EPS is obtained from the I/B/E/S consensus file.
Size	The logarithm of market value of equity.
Loss	A dummy variable equals 1 if actual earnings per share < 0, otherwise 0.
Return	The contemporaneous stock returns in the fiscal quarter calculated using CRSP monthly return data.
Accruals	The accruals (earnings minus cash flow from operations) scaled by the book value of assets.
MTB	The market value of equity plus the book value of total liabilities scaled by The book value of total assets.
RetVol	The stock return volatility calculated using 12 months of monthly return data before the fiscal quarter ending date.
EarnVol	The standard deviation of ROA, calculated using data from the last five years.
#BusSeg	The logarithm of 1 plus the number of business segment.
#GeoSeg	The logarithm of 1 plus the number of geographic segment.
Age	The number of years since a firm appears in CRSP monthly file.
M&A	A dummy variable that equals 1 if a firm makes a merger or acquisition in a given fiscal quarter and 0 otherwise.
SEO	A dummy that equals 1 if a firm has seasoned equity offering in a fiscal quarter and 0 otherwise.
SI	The amount of special items scaled by the book value of assets.

Table 1 Comparison of NLP Algorithms' Performance in Sentiment Classification in Training Sample

This table reports the classification accuracy in NLP algorithms including Finance-BERT, BERT, Loughran and McDonald word list (LM), the naïve Bayes and Word2Vec. In the bracket, the first (second) number represents the proportion of the cell as a percentage of the total number of sentences in the row (column). Panel A compares the classification outcome of Finance-BERT with human. Panel B compares the classification outcome of BERT with human and Finance-BERT. Panel C compares the classification outcome of LM with human and Finance-BERT. Panel D compares the classification outcome of the naïve Bayes with human and Finance-BERT. Panel E compares the classification outcome of Word2Vec with human and Finance-BERT.

Panel A: Comparison of Finance-BERT and human labeled sentiments

	Human Labeled			
	Positive	Neutral	Negative	Total
Finance-BERT Labeled				
Positive	3,205	360	117	3,682
	(87%, 89.6%)	(9.8%, 7.8%)	(3.2%, 6.4%)	(100%, 36.8%)
Neutral	258	4,074	160	4,492
	(5.7%, 7.2%)	(90.7%, 88.8%)	(3.6%, 8.7%)	(100%, 44.9%)
Negative	114	152	1,560	1,826
	(6.2%, 3.2%)	(8.3%, 3.3%)	(85.4%, 84.9%)	(100%, 18.3%)
Total	3,577	4,586	1,837	10,000
	(35.8%, 100%)	(45.9%, 100%)	(18.4%, 100%)	(100%, 100%)

Panel B: Comparison of BERT and human labeled/ Finance BERT.

	Human Labeled				Finance-BERT Labeled			
	Positive	Neutral	Negative	Total	Positive	Neutral	Negative	Total
BERT Labeled								
Positive	3,049	391	164	3,604	3,252	209	143	3,604
	(84.6%, 85.2%)	(10.8%, 8.5%)	(4.6%, 8.9%)	(100%, 36%)	(90.2%, 88.3%)	(5.8%, 4.7%)	(4%, 7.8%)	(100%, 36%)
Neutral	404	4,056	227	4,687	338	4,173	176	4,687
	(8.6%, 11.3%)	(86.5%, 88.4%)	(4.8%, 12.4%)	(100%, 46.9%)	(7.2%, 9.2%)	(89%, 92.9%)	(3.8%, 9.6%)	(100%, 46.9%)
Negative	124	139	1,446	1,709	92	110	1,507	1,709
	(7.3%, 3.5%)	(8.1%, 3%)	(84.6%, 78.7%)	(100%, 17.1%)	(5.4%, 2.5%)	(6.4%, 2.4%)	(88.2%, 82.5%)	(100%, 17.1%)
Total	3,577	4,586	1,837	10,000	3,682	4,492	1,826	10,000
	(35.8%, 100%)	(45.9%, 100%)	(18.4%, 100%)	(100%, 100%)	(36.8%, 100%)	(44.9%, 100%)	(18.3%, 100%)	(100%, 100%)

Table 1 (Cont'd) Comparison of NLP algorithms' Performance in Sentiment Classification in Training Sample

Panel C: Comparison of Loughran and McDonald word list and human labeled/Finance-BERT.

	Human Labeled				Finance-BERT Labeled			
	Positive	Neutral	Negative	Total	Positive	Neutral	Negative	Total
LM Labeled								
Positive	1,404	294	105	1,803	1,460	245	98	1,803
	(77.9%, 39.3%)	(16.3%, 6.4%)	(5.8%, 5.7%)	(100%, 18%)	(81%, 39.7%)	(13.6%, 5.5%)	(5.4%, 5.4%)	(100%, 18%)
Neutral	1,703	3,765	731	6,199	1,764	3,726	709	6,199
	(27.5%, 47.6%)	(60.7%, 82.1%)	(11.8%, 39.8%)	(100%, 62%)	(28.5%, 47.9%)	(60.1%, 82.9%)	(11.4%, 38.8%)	(100%, 62%)
Negative	470	527	1,001	1,998	458	521	1,019	1,998
	(23.5%, 13.1%)	(26.4%, 11.5%)	(50.1%, 54.5%)	(100%, 20%)	(22.9%, 12.4%)	(26.1%, 11.6%)	(51%, 55.8%)	(100%, 20%)
Total	3,577	4,586	1,837	10,000	3,682	4,492	1,826	10,000
	(35.8%, 100%)	(45.9%, 100%)	(18.4%, 100%)	(100%, 100%)	(36.8%, 100%)	(44.9%, 100%)	(18.3%, 100%)	(100%, 100%)

Panel D: Comparison of the naïve Bayes and human labeled/Finance-BERT.

	Human Labeled				Finance-BERT Labeled			
	Positive	Neutral	Negative	Total	Positive	Neutral	Negative	Total
NB Labeled								
Positive	2,976	360	328	3,664	2,949	354	361	3,664
	(81.2%, 83.2%)	(9.8%, 7.8%)	(9%, 17.9%)	(100%, 36.6%)	(80.5%, 80.1%)	(9.7%, 7.9%)	(9.9%, 19.8%)	(100%, 36.6%)
Neutral	577	4,185	404	5,166	668	4,052	446	5,166
	(11.2%, 16.1%)	(81%, 91.3%)	(7.8%, 22%)	(100%, 51.7%)	(12.9%, 18.1%)	(78.4%, 90.2%)	(8.6%, 24.4%)	(100%, 51.7%)
Negative	24	41	1,105	1,170	65	86	1,019	1,170
	(2.1%, 0.7%)	(3.5%, 0.9%)	(94.4%, 60.2%)	(100%, 11.7%)	(5.6%, 1.8%)	(7.4%, 1.9%)	(87.1%, 55.8%)	(100%, 11.7%)
Total	3,577	4,586	1,837	10,000	3,682	4,492	1,826	10,000
	(83.2%, 100%)	(7.8%, 100%)	(17.9%, 100%)	(36.6%, 100%)	(36.8%, 100%)	(44.9%, 100%)	(18.3%, 100%)	(100%, 100%)

Table 1 (Cont'd) Comparison of NLP algorithms' Performance in Sentiment Classification in Training Sample

Panel E: Comparison of Word2Vec and human labeled/Finance-BERT.

	Human Labeled				Finance-BERT Labeled			
	Positive	Neutral	Negative	Total	Positive	Neutral	Negative	Total
Word2Vec Labeled								
Positive	1,211	704	473	2,388	1,274	653	461	2,388
	(50.7%, 33.9%)	(29.5%, 15.4%)	(19.8%, 25.7%)	(100%, 23.9%)	(53.4%, 34.6%)	(27.3%, 14.5%)	(19.3%, 25.2%)	(100%, 23.9%)
Neutral	2,339	3,825	1,307	7,471	2,375	3,790	1,306	7,471
	(31.3%, 65.4%)	(51.2%, 83.4%)	(17.5%, 71.1%)	(100%, 74.7%)	(31.8%, 64.5%)	(50.7%, 84.4%)	(17.5%, 71.5%)	(100%, 74.7%)
Negative	27	57	57	141	33	49	59	141
	(19.1%, 0.8%)	(40.4%, 1.2%)	(40.4%, 3.1%)	(100%, 1.4%)	(23.4%, 0.9%)	(34.8%, 1.1%)	(41.8%, 3.2%)	(100%, 1.4%)
Total	3,577	4,586	1,837	10,000	3,682	4,492	1,826	10,000
	(35.8%, 100%)	(45.9%, 100%)	(18.4%, 100%)	(100%, 100%)	(36.8%, 100%)	(44.9%, 100%)	(18.3%, 100%)	(100%, 100%)

Table 2 Comparison of Loughran and McDonald and Finance-BERT Performance Conditional on Human Labeled Tone in Training Sample.

This table tabulate the sentiment classification of LM and Finance-BERT. Panels A, B and C report the sample of sentences labeled by researchers as positive, neutral and negative respectively. In the bracket, the first (second) number represents the proportion of the cell as a percentage of the total number of sentences in the row (column).

Panel A: Comparison of LM and Finance-BERT in sentences manually labeled as Positive.

	Finance BERT Labeled			
	Positive	Neutral	Negative	Total
LM Labeled				
Positive	1,359	30	15	1,404
	(96.8%, 42.4%)	(2.1%, 11.6%)	(1.1%, 13.2%)	(100%, 39.3%)
Neutral	1,458	198	47	1,703
	(85.6%, 45.5%)	(11.6%, 76.7%)	(2.8%, 41.2%)	(100%, 47.6%)
Negative	388	30	52	470
	(82.6%, 12.1%)	(6.4%, 11.6%)	(11.1%, 45.6%)	(100%, 13.1%)
Total	3,205	258	114	3,577
	(89.6%, 100%)	(7.2%, 100%)	(3.2%, 100%)	(100%, 100%)

Panel B: Comparison of LM and Finance-BERT in sentences manually labeled as Neutral.

	Finance BERT Labeled			
	Positive	Neutral	Negative	Total
LM Labeled				
Positive	83	206	5	294
	(28.2%, 23.1%)	(70.1%, 5.1%)	(1.7%, 3.3%)	(100%, 6.4%)
Neutral	241	3,436	88	3,765
	(6.4%, 66.9%)	(91.3%, 84.3%)	(2.3%, 57.9%)	(100%, 82.1%)
Negative	36	432	59	527
	(6.8%, 10%)	(82%, 10.6%)	(11.2%, 38.8%)	(100%, 11.5%)
Total	360	4,074	152	4,586
	(7.8%, 100%)	(88.8%, 100%)	(3.3%, 100%)	(100%, 100%)

Panel C: Comparison of LM and Finance-BERT in sentences manually labeled as Negative.

	Finance BERT Labeled			
	Positive	Neutral	Negative	Total
LM Labeled				
Positive	18	9	78	105
	(17.1%, 15.4%)	(8.6%, 5.6%)	(74.3%, 5%)	(100%, 5.7%)
Neutral	65	92	574	731
	(8.9%, 55.6%)	(12.6%, 57.5%)	(78.5%, 36.8%)	(100%, 39.8%)
Negative	34	59	908	1,001
	(3.4%, 29.1%)	(5.9%, 36.9%)	(90.7%, 58.2%)	(100%, 54.5%)
Total	117	160	1,560	1,837
	(6.4%, 100%)	(8.7%, 100%)	(84.9%, 100%)	(100%, 100%)

Table 3 Sample Selection

This table presents the sample selection procedure for the sample used regressions.

Sample selection criteria	# of firms	# of conference calls
Conference call transcripts, 2003-2012	690	18,607
Retain: firms with GVKEY	635	17,170
Retain: firms have financial information to calculate control variables	632	16,840
Final sample	632	16,840

Table 4 Descriptive Statistics

Panel A: Summary statistics

This panel presents the descriptive statistics for the sample used in regressions. Variable definitions are in Appendix.

Variable	N	Mean	Stdev	Q1	Median	Q3
CAR	16,840	0.08	5.95	-3	0.05	3.33
$AbVOL$	16,840	4.99	5.71	1.41	3.67	6.91
$Tone_{FinBERT}$	16,840	0.39	0.17	0.27	0.4	0.51
$Tone_{BERT}$	16,840	0.38	0.17	0.26	0.38	0.5
$Tone_{LM}$	16,840	0.09	0.15	0	0.1	0.2
$Tone_{NB}$	16,840	0.43	0.14	0.33	0.43	0.53
$Tone_{W2V}$	16,840	0.26	0.09	0.19	0.25	0.31
$Pos_{FinBERT}$	16,840	0.51	0.13	0.42	0.51	0.6
Pos_{BERT}	16,840	0.51	0.12	0.42	0.51	0.6
Pos_{LM}	16,840	0.29	0.09	0.22	0.28	0.35
Pos_{NB}	16,840	0.48	0.13	0.39	0.48	0.57
Pos_{W2V}	16,840	0.28	0.09	0.21	0.27	0.34
$Neg_{FinBERT}$	16,840	0.12	0.07	0.06	0.1	0.16
Neg_{BERT}	16,840	0.13	0.08	0.08	0.12	0.18
Neg_{LM}	16,840	0.19	0.09	0.13	0.18	0.24
Neg_{NB}	16,840	0.05	0.04	0.02	0.04	0.07
Neg_{W2V}	16,840	0.02	0.02	0.01	0.02	0.03
UE	16,840	0.06	0.69	0	0.04	0.16
$Size$	16,840	9.34	1.09	8.57	9.24	9.98
$Loss$	16,840	0.1	0.3	0	0	0
$Earn$	16,840	0.02	0.02	0	0.01	0.02
$Return$	16,840	0.01	0.13	-0.07	0	0.08
$Accruals$	16,838	-0.01	0.03	-0.02	-0.01	0
MTB	16,840	1.91	1.08	1.16	1.54	2.27
$RetVol$	16,840	0.09	0.05	0.05	0.07	0.1
$EarnVol$	16,840	0.02	0.02	0	0.01	0.02
$\#BusSeg$	16,840	2.65	2.32	1	3	4
$\#GeoSeg$	16,840	2.98	2.74	1	2	4
Age	16,840	35.09	22.73	16	33	47
$M\&A$	16,840	0.24	0.42	0	0	0
SEO	16,840	0.01	0.12	0	0	0
SI	16,840	0	0.01	0	0	0

Table 4 (Cont'd) Descriptive Statistics

Panel B: Pearson correlation table

This panel presents the Pearson correlation table. Correlation coefficients in bold indicate the significance at 0.01 or better. Variable definitions are in Appendix.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
(1)	CAR																					
(2)	$AbVOL$	**-0.15**																				
(3)	$Tone_{FinBERT}$	**0.14**	0.03																			
(4)	$Tone_{BERT}$	**0.14**	0.03	**0.97**																		
(5)	$Tone_{LM}$	0.11	0.03	**0.77**	**0.77**																	
(6)	$Tone_{NB}$	0.08	**0.07**	**0.76**	**0.75**	0.63																
(7)	$Tone_{W2V}$	0.04	**0.06**	**0.47**	**0.46**	0.34	0.58															
(8)	$Earn$	**0.07**	**0.05**	**0.20**	**0.20**	**0.25**	**0.19**	**0.10**														
(9)	UE	**0.19**	0.00	**0.13**	**0.13**	**0.14**	**0.10**	**0.04**	**0.16**													
(10)	$Size$	-0.01	**-0.04**	**0.17**	**0.17**	**0.13**	0.13	**0.12**	**0.23**	**0.04**												
(11)	$Loss$	**-0.08**	**0.00**	**-0.16**	**-0.17**	**-0.19**	**-0.15**	-0.09	**-0.56**	**-0.23**	**-0.22**											
(12)	$Return$	-0.04	-0.01	**0.13**	**0.12**	**0.12**	**0.08**	0.02	**0.05**	**0.11**	**0.06**	**-0.05**										
(13)	$Accruals$	-0.04	**-0.02**	**0.03**	**0.03**	**0.00**	0.06	**0.04**	**0.19**	**0.08**	**0.07**	**-0.24**	**0.03**									
(14)	MTB	0.00	0.12	**0.22**	**0.22**	**0.24**	**0.17**	**0.13**	**0.60**	**0.01**	**0.18**	**-0.15**	**0.11**	**-0.05**								
(15)	$RetVol$	**0.00**	**0.01**	**-0.08**	**-0.08**	**-0.07**	**-0.05**	-0.04	**-0.22**	**-0.01**	**-0.33**	**0.33**	**0.10**	**-0.13**	-0.09							
(16)	$EarnVol$	**-0.02**	**0.02**	**-0.03**	**-0.03**	**0.02**	**-0.04**	**0.00**	**-0.08**	**0.01**	**-0.17**	**0.27**	0.01	**-0.14**	**0.13**	**0.33**						
(17)	$\#BusSeg$	0.00	**-0.02**	0.03	0.02	0.02	**0.05**	**0.08**	**-0.03**	**0.04**	**0.04**	**0.00**	0.01	0.02	**-0.11**	**-0.05**	**-0.01**					
(18)	$\#GeoSeg$	0.00	0.02	**0.01**	**0.01**	0.08	**0.08**	**0.04**	**0.13**	**0.03**	**0.09**	**0.02**	0.02	**-0.02**	**0.13**	**0.09**	**0.16**	**0.15**				
(19)	Age	-0.02	**-0.09**	**-0.07**	**-0.07**	-0.06	**-0.01**	-0.04	**-0.02**	**0.00**	**0.20**	**-0.04**	**-0.01**	**0.04**	**-0.15**	**-0.20**	**-0.16**	**0.23**	**0.13**			
(20)	$M\&A$	0.00	0.00	**0.09**	**0.09**	0.07	**0.08**	**0.09**	**0.05**	**0.00**	**0.16**	**-0.04**	-0.02	**0.02**	**0.05**	**-0.08**	**-0.03**	**0.05**	**0.02**	**-0.03**		
(21)	SEO	0.00	-0.06	-0.02	-0.02	**-0.05**	**-0.05**	**-0.05**	**-0.07**	0.00	**-0.01**	**0.05**	**0.02**	0.01	**-0.06**	**0.11**	**-0.01**	**-0.01**	**-0.03**	**-0.01**	0.00	
(22)	SI	**0.04**	**-0.02**	**0.07**	**0.07**	**0.08**	**0.06**	**0.03**	**0.44**	**0.04**	**0.10**	**-0.45**	**0.05**	**0.38**	0.01	**-0.13**	**-0.19**	**-0.01**	**-0.05**	**0.02**	**0.01**	0.01

Table 5 Market Reaction to Conference Calls and Textual Sentiments

This table reports the relation between textual sentiments in conference call transcripts and market reactions. Market reactions are measured with cumulative abnormal returns and cumulative abnormal trading volume in a three-day window around conference call hosting dates. Panel A shows results of OLS regressions: $CAR = \alpha + \beta_1 Tone_j + \beta_2 Controls + \varepsilon$. Tone_j is the sentiment of Conference calls captured by approach j. Panel B compares the explanation power of $\text{Tone}_{\text{FinBERT}}$ ($\text{Tone}_{\text{BERT}}$) with tone measures from other NLP approaches in regressions. Panel C shows results of OLS regressions: $AbVOL = \alpha + \beta_1 Pos_j + \beta_2 Neg_j + \beta_3 Controls + \varepsilon$. Pos_j and Neg_j is the sentiment of conference calls predicted by approach j. Panel D compares the explanation power of $Pos_{FinBERT}$ (Pos_{BERT}) as well as $Neg_{FinBERT}$ (Neg_{BERT}) with tone measures from other NLP approaches in regressions. All regressions include year fixed effects. t-stats based on standard errors clustered by firm are reported in parentheses below the coefficients. ***, **, * indicate significance at the 0.01, 0.05, and 0.10 level, respectively. All variables are defined in Appendix.

Panel A: Regression of cumulative abnormal return on textual sentiments

	(1)	(2)	(3)	(4)	(5)
Dependent Variable			CAR		
$Tone_{FinBERT}$	**0.734*** **				
	(15.01)				
$Tone_{BERT}$		**0.709*** **			
		(15.08)			
$Tone_{LM}$			**0.464*** **		
			(9.77)		
$Tone_{NB}$				**0.369*** **	
				(8.10)	
$Tone_{W2V}$					**0.175*** **
					(3.86)
$Earn$	8.096**	8.493***	8.886***	10.976***	13.098***
	(2.46)	(2.59)	(2.69)	(3.31)	(3.99)
UE	1.468***	1.472***	1.499***	1.528***	1.560***
	(11.63)	(11.67)	(11.80)	(11.97)	(12.14)
$Size$	-0.271***	-0.268***	-0.216***	-0.214***	-0.201***
	(-5.55)	(-5.50)	(-4.41)	(-4.35)	(-4.09)
$Loss$	-0.365	-0.347	-0.386	-0.405	-0.429*
	(-1.45)	(-1.38)	(-1.52)	(-1.60)	(-1.71)
Intercept	2.343***	2.313***	1.854***	1.810***	1.652***
	(4.27)	(4.21)	(3.35)	(3.24)	(2.97)
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	16,840	16,840	16,840	16,840	16,840
Adj. R^2	0.054	0.053	0.045	0.043	0.041

Table 5 (Cont'd) Market Reaction to Conference Calls and Textual Sentiments

Panel B: Comparison of explanation power in regressions of abnormal return

	Z-statistic	p-value
$Tone_{FinBERT}$ *vs* $Tone_{BERT}$	**2.131****	**0.033**
$Tone_{FinBERT}$ *vs* $Tone_{LM}$	**7.422*****	**<0.001**
$Tone_{FinBERT}$ *vs* $Tone_{NB}$	**8.288*****	**<0.001**
$Tone_{FinBERT}$ *vs* $Tone_{W2V}$	**8.260*****	**<0.001**
$Tone_{BERT}$ *vs* $Tone_{LM}$	**6.878*****	**<0.001**
$Tone_{BERT}$ *vs* $Tone_{NB}$	**7.789*****	**<0.001**
$Tone_{BERT}$ *vs* $Tone_{W2V}$	**7.935*****	**<0.001**

Table 5 (Cont'd) Market Reaction to Conference Calls and Textual Sentiments

Panel C: Regression of cumulative abnormal trading volume on textual sentiments

	(1)	(2)	(3)	(4)	(5)
Dependent Variable			$AbVOL$		
$Pos_{FinBERT}$	**0.448*** **				
	(7.35)				
$Neg_{FinBERT}$	**0.422*** **				
	(7.07)				
Pos_{BERT}		**0.429*** **			
		(6.88)			
Neg_{BERT}		**0.369*** **			
		(5.92)			
Pos_{LM}			**0.261*** **		
			(3.94)		
Neg_{LM}			**0.142** **		
			(2.25)		
Pos_{NB}				**0.468*** **	
				(7.77)	
Neg_{NB}				**0.206*** **	
				(4.17)	
Pos_{W2V}					**0.348*** **
					(5.66)
Neg_{W2V}					0.018
					(0.39)
$\|UE\|$	0.275***	0.264***	0.238***	0.304***	0.281***
	(3.54)	(3.40)	(3.11)	(3.95)	(3.62)
$Size$	-0.373***	-0.368***	-0.350***	-0.361***	-0.371***
	(-5.90)	(-5.83)	(-5.48)	(-5.89)	(-6.15)
MTB	0.700***	0.702***	0.703***	0.664***	0.655***
	(10.56)	(10.49)	(10.18)	(10.26)	(10.33)
Intercept	6.302***	6.257***	6.058***	6.298***	6.375***
	(10.06)	(10.00)	(9.67)	(10.33)	(10.70)
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	16,840	16,840	16,840	16,840	16,840
Adj. R^2	0.034	0.032	0.027	0.035	0.030

Table 5 (Cont'd) Market Reaction to Conference Calls and Textual Sentiments

Panel D: Comparison of explanation power in regressions of trading volume

	Z-statistics	p-value
$Tone_{FinBERT}$ *vs* $Tone_{BERT}$	**3.943*** **	**<0.001**
$Tone_{FinBERT}$ *vs* $Tone_{LM}$	**5.322*** **	**<0.001**
$Tone_{FinBERT}$ *vs* $Tone_{NB}$	-1.006	0.314
$Tone_{FinBERT}$ *vs* $Tone_{W2V}$	**3.016*** **	**0.003**
$Tone_{BERT}$ *vs* $Tone_{LM}$	**4.403*** **	**<0.001**
$Tone_{BERT}$ *vs* $Tone_{NB}$	**-2.928*** **	**0.003**
$Tone_{BERT}$ *vs* $Tone_{W2V}$	**1.857***	**0.063**

Table 6 Comparison of Finance BERT, BERT, Loughran and McDonald, the Naïve Bayes and Word2Vec Performance in Small Sample

This table reports the summary statistics of bootstrapping regression of market reactions on textual sentiments in conference call transcripts. This table replicates regressions in Table 5 based on bootstrapped sample with 400 iterations. Market reactions are measured with abnormal returns and abnormal trading volume. Panel A estimates the OLS regressions: $CAR = \alpha + \beta_1 Tone_j + \beta_2 Controls + \varepsilon$, while Panel B estimates the OLS regressions: $AbVOL = \alpha + \beta_1 Pos_j + \beta_2 Neg_j + \beta_3 Controls + \varepsilon$. Tone$_j$, Pos_j and Neg_j are the sentiment of conference calls captured by approach j. # Positive sign is the frequency of sentiment effects that are in the positive direction, and # Positive significant is the frequency of sentiment effects that are in the positive direction and significant at either p < 10%, p < 5%, or p < 1% (two-tailed). All variables are defined in Appendix.

Panel A: Regression of cumulative abnormal returns on tone measures

	Mean	# Positive sign	# Positively significant		
			10%	5%	1%
$Tone_{FinBERT}$	0.743	400	397	392	369
$Tone_{BERT}$	0.722	400	395	390	357
$Tone_{LM}$	0.478	400	319	285	182
$Tone_{NB}$	0.377	391	267	223	148
$Tone_{W2V}$	0.174	323	112	71	23

Panel B: Regression of cumulative abnormal trading volume on tone measures

	Mean	# Positive sign	# Positively significant		
			10%	5%	1%
$Pos_{FinBERT}$	0.482	399	347	323	228
Pos_{BERT}	0.464	398	340	297	215
Pos_{LM}	0.297	381	209	158	78
Pos_{NB}	0.497	400	374	355	274
Pos_{W2V}	0.366	394	307	267	157
$Neg_{FinBERT}$	0.335	390	232	194	86
Neg_{BERT}	0.291	380	199	134	62
Neg_{LM}	0.084	276	43	20	6
Neg_{NB}	0.130	316	73	45	10
Neg_{W2V}	0.008	210	19	5	2

Table 7 Comparison of Loughran and McDonald and Finance BERT across Industries

Panel A: Comparison of LM and Finance BERT

This panel tabulates the statistics of discrepancy in the LM word list labeled tone and the Finance BERT labeled tone across industries. The statistics are the number of sentences labeled by the two algorithms as a percentage of total number of sentences in presentation sections of all firms in that industry. Industry classifications are based on two-digit GICS codes.

	FinBERT: Positive		FinBERT: Neutral		FinBERT: Negative		
GICS Industry	LM: Negative	LM: Neutral	LM: Positive	LM: Negative	LM: Positive	LM: Neutral	Total
10 Energy	4.619	16.510	3.058	6.314	0.520	4.497	35.518
15 Materials	7.087	18.264	2.672	5.654	0.679	4.790	39.146
20 Industrials	6.886	20.361	2.397	5.001	0.639	4.977	40.261
25 Consumer Discretionary	6.594	19.102	3.023	4.800	0.677	4.200	38.396
30 Consumer Staples	6.736	21.124	2.776	4.324	0.666	4.191	39.817
35 Health Care	6.248	19.537	3.520	6.418	0.530	3.265	39.518
40 Financials	8.750	16.699	2.920	8.383	0.665	4.109	41.526
45 Information Technology	5.193	20.688	3.232	5.328	0.506	4.129	39.076
50 Communication Services	5.951	19.872	2.722	4.166	0.390	3.216	36.317
55 Utilities	5.426	13.234	4.040	8.705	0.549	4.094	36.048
Observations	144,498	414,571	67,210	132,740	13,324	92,547	

Table 7 (Cont'd) Comparison of Loughran and McDonald and Finance BERT across Industries

Panel B: Regression of cumulative abnormal return on tone measures across industries

This panel reports the relation between textual sentiments in conference call transcripts and three-day cumulative abnormal returns across industries. Industry classifications are based on two-digit GICS codes. We estimate the OLS regressions $CAR = \alpha + \beta_1 Tone_j + \beta_2 Controls + \varepsilon$. Tone_j is the sentiment of Conference calls captured by approach j. All regressions include year fixed effects. t-stats based on standard errors clustered by firm are reported in parentheses below the coefficients. ***, **, * indicate significance at the 0.01, 0.05, and 0.10 level, respectively. All variables are defined in Appendix.

Sample	GIC 10	GIC 15	GIC 20	GIC 25	GIC 30	GIC 35	GIC 40	GIC 45	GIC 50	GIC 55
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Dependent Variable						CAR				
$Tone_{FinBERT}$	**0.596*** **	**0.722*** **	**0.764*** **	**0.928*** **	**0.906*** **	**1.070*** **	**0.621*** **	**1.144*** **	0.549	**0.399** **
	(3.47)	**(3.94)**	**(5.87)**	**(7.03)**	**(6.17)**	**(5.79)**	**(4.43)**	**(7.00)**	(1.09)	**(2.40)**
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,107	1,077	1,979	2,743	1,471	1,856	2,324	2,860	305	1,118
Adj. R2	0.060	0.090	0.070	0.077	0.099	0.059	0.044	0.044	0.058	0.093
$Tone_{LM}$	**0.563** **	**0.490*** **	**0.452*** **	**0.652*** **	**0.754*** **	**0.745*** **	**0.244** **	**0.814*** **	0.618	0.202
	(2.68)	**(2.94)**	**(3.51)**	**(4.33)**	**(6.15)**	**(4.28)**	**(2.63)**	**(5.71)**	(1.32)	(1.45)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,107	1,077	1,979	2,743	1,471	1,856	2,324	2,860	305	1,118
Adj. R2	0.057	0.078	0.057	0.069	0.091	0.046	0.036	0.034	0.059	0.088
Vuong's test: $Tone_{FinBERT}$ versus $Tone_{LM}$										
Z-statistic	0.828	**1.989** **	**2.984*** **	**2.820*** **	**1.692***	**2.528** **	**2.678*** **	**3.048*** **	-0.221	1.521
p-value	0.408	**0.047**	**0.003**	**0.005**	**0.091**	**0.011**	**0.007**	**0.002**	0.825	0.128

Table 7 (Cont'd) Comparison of Loughran and McDonald and Finance BERT across Industries

Panel C: Regression of cumulative abnormal trading volume on Tone measures across industries

This panel reports the relation between textual sentiments in conference call transcripts and three-day cumulative abnormal trading volume around conference call dates across industries. Industry classifications are based on two-digit GICS codes. We estimate the OLS regression: Ab$VOL = \alpha + \beta_1 Pos_j + \beta_2 Neg_J + \beta_3 Controls + \varepsilon$. Pos_j and Neg_j are the sentiment of conference calls captured by approach j. All regressions include year fixed effects. t-stats based on standard errors clustered by firm are reported in parentheses below the coefficients. ***, **, * indicate significance at the 0.01, 0.05, and 0.10 level, respectively. All variables are defined in Appendix.

Sample	GIC 10	GIC 15	GIC 20	GIC 25	GIC 30	GIC 35	GIC 40	GIC 45	GIC 50	GIC 55
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Dependent Variable						AbVOL				
$Pos_{FinBERT}$	0.186	0.111	-0.079	**0.461*** **	0.172	0.203	0.100	**0.324***	0.410	-0.007
	(1.09)	(0.40)	(-0.41)	**(2.68)**	(0.94)	(1.25)	(0.79)	**(1.70)**	(0.86)	(-0.05)
$Neg_{FinBERT}$	**0.148**	**0.453****	**0.675*****	**0.370****	0.243	**0.577*****	**0.343*****	0.235	0.436	0.198
	(1.04)	**(2.69)**	**(4.50)**	**(2.54)**	(0.99)	**(3.02)**	**(2.85)**	(1.28)	(1.23)	(1.42)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,107	1,077	1,979	2,743	1,471	1,856	2,324	2,860	305	1,118
Adj. R2	0.047	0.015	0.047	0.045	0.033	0.021	0.014	0.013	0.009	0.014
Pos_{LM}	0.270	-0.051	-0.232	0.272	-0.108	0.099	-0.206	0.092	0.738	-0.088
	(1.55)	(-0.19)	(-1.39)	(1.62)	(-0.54)	(0.70)	(-1.54)	(0.49)	(1.72)	(-0.51)
Neg_{LM}	**0.417****	**0.333***	**0.247***	0.184	-0.328	0.063	0.096	-0.080	-0.047	0.163
	(2.48)	**(1.75)**	**(1.69)**	(1.29)	(-1.17)	(0.37)	(0.80)	(-0.39)	(-0.09)	(0.94)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,107	1,077	1,979	2,743	1,471	1,856	2,324	2,860	305	1,118
Adj. R2	0.053	0.012	0.033	0.042	0.033	0.015	0.013	0.012	0.024	0.014
Vuong's test: FinBERT versus LM										
Z-statistic	**-1.643***	0.713	**2.788*****	1.180	-0.181	1.578	0.590	0.848	-0.938	0.016
p-value	**0.100**	0.476	**0.005**	0.238	0.856	0.115	0.555	0.396	0.348	0.987

Table 8 Comparison of Loughran and McDonald and Finance BERT across Firm Characteristics

This table reports the relation between textual sentiments in conference call transcripts and market reactions across firm characteristics. Market reactions are measured with cumulative abnormal returns and cumulative abnormal trading volume in a three-day window around conference call hosting dates. Panel A shows results of OLS regression: $CAR = \alpha + \beta_1 Tone_j + \beta_2 Controls + \varepsilon$, while Panel B shows results of OLS regression: $AbVOL = \alpha + \beta_1 Pos_j + \beta_2 Neg_j + \beta_3 Controls + \varepsilon$. Tone_j, Pos_j, and Neg_j represent sentiments of conference calls captured by approach j. All regressions include year fixed effects. t-stats based on standard errors clustered by firm are reported in parentheses below the coefficients. ***, **, * indicate significance at the 0.01, 0.05, and 0.10 level, respectively. All variables are defined in Appendix.

Panel A: Regression of cumulative abnormal return on tone measures

Sample	Low_Earn	High_Earn	Low_Size	High_Size	Low_ToneDiff	High_ToneDiff
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable				CAR		
$Tone_{FinBERT}$	**0.709*** **	**0.751*** **	**0.886*** **	**0.595*** **	**0.742*** **	**0.673*** **
	(9.78)	**(11.21)**	**(11.66)**	**(10.15)**	**(9.78)**	**(8.75)**
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8,420	8,420	8,420	8,420	8,420	8,420
Adj. R^2	0.055	0.059	0.061	0.045	0.051	0.052
$Tone_{LM}$	**0.338*** **	**0.590*** **	**0.583*** **	**0.361*** **	**0.602*** **	**0.424*** **
	(5.03)	**(9.07)**	**(7.89)**	**(6.16)**	**(8.30)**	**(6.59)**
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8,420	8,420	8,420	8,420	8,420	8,420
Adj. R^2	0.046	0.052	0.053	0.036	0.047	0.048
Vuong's test: $Tone_{FinBERT}$ versus $Tone_{LM}$						
Z-statistic	**5.564*** **	**4.141*** **	**5.098*** **	**5.450*** **	**3.961*** **	**4.133*** **
p-value	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

Table 8 (Cont'd) Comparison of Loughran and McDonald and Finance BERT across Firm Characteristics

Panel B: Regression of cumulative abnormal trading volume on tone measures

Sample	Low_Earn	High_Earn	Low_Size	High_Size	Low_ToneDiff	High_ToneDiff
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable			AbVOL			
$Pos_{FinBERT}$	**0.484*** **	**0.418*** **	**0.427*** **	**0.464*** **	**0.624*** **	**0.357*** **
	(6.70)	**(4.58)**	**(5.06)**	**(6.03)**	**(7.64)**	**(4.19)**
$Neg_{FinBERT}$	**0.532*** **	**0.321*** **	**0.357*** **	**0.464*** **	**0.385*** **	**0.463*** **
	(7.09)	**(3.78)**	**(4.82)**	**(5.14)**	**(5.52)**	**(5.05)**
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8,420	8,420	8,420	8,420	8,420	8,420
Adj. R^2	0.027	0.036	0.023	0.043	0.042	0.027
Pos_{LM}	**0.280*** **	**0.245*** **	**0.275*** **	**0.255*** **	**0.430*** **	0.066
	(3.42)	**(2.78)**	**(3.14)**	**(2.86)**	**(5.63)**	(0.75)
Neg_{LM}	**0.230*** **	0.065	0.096	**0.172***	**0.293*** **	0.024
	(3.05)	(0.67)	(1.14)	**(1.90)**	**(3.57)**	(0.28)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8,420	8,420	8,420	8,420	8,420	8,420
Adj. R^2	0.017	0.032	0.019	0.035	0.036	0.021
Vuong's test: $Tone_{FinBERT}$ versus $Tone_{LM}$						
Z-statistic	**4.649*** **	**2.818*** **	**2.941*** **	**4.237*** **	**3.448*** **	**3.699*** **
p-value	**<0.001**	**0.005**	**0.003**	**<0.001**	**0.001**	**<0.001**

Table 9 Future Earnings and Tone Measures

This table reports the relation between textual sentiments in conference call transcripts and future earnings. We estimate the OLS regression: $Earn = \alpha + \beta_1 Tone_j + \beta_2 Controls + \varepsilon$. Tone_j is the sentiment of conference calls captured by approach j. All regressions include year and firm fixed effects. t-stats based on standard errors clustered by firm are reported in parentheses below the coefficients. ***, **, * indicate significance at the 0.01, 0.05, and 0.10 level, respectively. All variables are defined in Appendix.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variables	$Earn_{y+1}$	$Earn_{y+1}$	$Earn_{y+2}$	$Earn_{y+2}$	$Earn_{y+3}$	$Earn_{y+3}$	$Earn_{y+4}$	$Earn_{y+4}$
$Tone_{FinBERT}$	**0.006*** **		**0.004*** **		0.001		-0.000	
	(6.47)		**(5.21)**		(1.16)		(-0.06)	
$Tone_{LM}$		**0.004*** **		**0.003*** **		**0.002***		0.001
		(4.29)		**(3.71)**		**(1.96)**		(1.19)
$Earn$	0.628***	0.646***	-0.039	-0.028	-0.023	-0.028	-0.137	-0.146
	(6.21)	(6.39)	(-0.39)	(-0.28)	(-0.26)	(-0.32)	(-1.46)	(-1.54)
$Return$	0.021***	0.023***	0.019***	0.020***	0.017***	0.017***	0.005	0.004
	(4.27)	(4.62)	(3.86)	(4.04)	(3.58)	(3.52)	(1.01)	(0.87)
$Accruals$	-0.036	-0.037	-0.032	-0.033	-0.020	-0.020	-0.049*	-0.049*
	(-1.49)	(-1.51)	(-1.57)	(-1.60)	(-1.12)	(-1.11)	(-1.85)	(-1.84)
$Size$	-0.018***	-0.017***	-0.027***	-0.027***	-0.033***	-0.033***	-0.028***	-0.028***
	(-3.71)	(-3.59)	(-5.89)	(-5.84)	(-7.20)	(-7.23)	(-5.99)	(-6.07)
MTB	0.024***	0.024***	0.015***	0.015***	0.013***	0.013***	0.013***	0.013***
	(7.39)	(7.48)	(4.15)	(4.21)	(3.59)	(3.58)	(3.57)	(3.56)
$RetVol$	-0.014	-0.012	-0.096***	-0.094***	-0.110***	-0.108***	-0.059*	-0.058*
	(-0.54)	(-0.45)	(-3.24)	(-3.17)	(-2.96)	(-2.93)	(-1.76)	(-1.73)
$EarnVol$	0.313*	0.314*	0.447**	0.447**	0.130	0.127	0.055	0.053
	(1.91)	(1.91)	(2.24)	(2.24)	(0.67)	(0.66)	(0.34)	(0.32)
$\#BusSeg$	0.001	0.001	0.000	0.001	0.000	0.000	0.001	0.001
	(1.02)	(1.10)	(0.53)	(0.60)	(0.13)	(0.16)	(0.62)	(0.63)
$\#GeoSeg$	0.001	0.001	0.002	0.002	0.000	0.000	-0.001	-0.001
	(0.87)	(0.85)	(1.08)	(1.07)	(0.30)	(0.29)	(-0.87)	(-0.86)
Age	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
	(-0.19)	(-0.15)	(-0.72)	(-0.68)	(-0.95)	(-0.92)	(-0.18)	(-0.16)
$M\&A$	0.001	0.001	0.000	0.000	0.001	0.001	-0.001	-0.001
	(1.00)	(1.04)	(0.20)	(0.23)	(0.48)	(0.47)	(-0.63)	(-0.64)

SEO	-0.014***	-0.014***	0.000	0.001	0.003	0.003	0.006*	0.006*
	(-2.77)	(-2.67)	(0.06)	(0.18)	(0.70)	(0.71)	(1.78)	(1.76)
SI	-0.689***	-0.705***	-0.020	-0.032	-0.047	-0.054	0.179	0.175
	(-4.72)	(-4.79)	(-0.14)	(-0.22)	(-0.36)	(-0.40)	(1.22)	(1.18)
Intercept	0.003	0.003	0.004	0.004	0.006**	0.006**	0.002	0.002
	(1.12)	(1.15)	(1.54)	(1.56)	(1.99)	(1.99)	(0.69)	(0.69)
Year, firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	15,394	15,394	15,394	15,394	15,394	15,394	15,394	15,394
Adj. R^2	0.121	0.117	0.065	0.063	0.057	0.057	0.041	0.041
Vuong's test: $Tone_{FinBERT}$ versus $Tone_{LM}$								
Z-statistic	**5.090*** **		**3.073*** **		-1.331		-0.901	
p-value	**<0.001**		**0.002**		0.183		0.367	

Table 10 Future Investments and Tone measures

This table reports the relation between textual sentiments in conference call transcripts and future capital expenditures. We estimate the OLS regression: $CAPEX = \alpha + \beta_1 Tone_j + \beta_2 Controls + \varepsilon$. Tone_j is the sentiment of conference calls captured by approach j. All regressions include year and firm fixed effects. t-stats based on standard errors clustered by firm are reported in parentheses below the coefficients. ***, **, * indicate significance at the 0.01, 0.05, and 0.10 level, respectively. All variables are defined in Appendix.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variables	$CAPEX_{y+1}$	$CAPEX_{y+1}$	$CAPEX_{y+2}$	$CAPEX_{y+2}$	$CAPEX_{y+3}$	$CAPEX_{y+3}$	$CAPEX_{y+4}$	$CAPEX_{y+4}$
$Tone_{FinBERT}$	0.014		**0.053****		**0.071*****		**0.042***	
	(0.62)		**(2.57)**		**(3.11)**		**(1.92)**	
$Tone_{LM}$		-0.017		0.027		**0.049****		0.019
		(-0.85)		(1.37)		**(2.55)**		(1.03)
$Earn$	5.075**	5.304**	2.591	2.808	-1.285	-1.100	-1.855*	-1.665
	(2.24)	(2.34)	(1.39)	(1.51)	(-1.00)	(-0.86)	(-1.65)	(-1.50)
$Return$	-0.324***	-0.306***	-0.122	-0.102	-0.030	-0.012	0.090	0.108
	(-3.68)	(-3.40)	(-1.35)	(-1.12)	(-0.40)	(-0.15)	(1.08)	(1.25)
$Accruals$	-0.328	-0.335	-0.667	-0.675	-0.625*	-0.634**	-0.614**	-0.621**
	(-0.72)	(-0.73)	(-1.55)	(-1.56)	(-1.96)	(-1.98)	(-2.12)	(-2.14)
$Size$	0.192*	0.199*	0.118	0.125	0.060	0.067	-0.040	-0.034
	(1.86)	(1.92)	(1.34)	(1.41)	(0.77)	(0.86)	(-0.50)	(-0.43)
MTB	0.260***	0.262***	0.235***	0.238***	0.188***	0.191***	0.129**	0.132**
	(4.24)	(4.28)	(3.92)	(3.97)	(3.33)	(3.37)	(2.18)	(2.22)
$RetVol$	0.817	0.796	1.363**	1.375**	0.938	0.968	0.538	0.545
	(1.22)	(1.18)	(2.13)	(2.13)	(1.49)	(1.53)	(0.83)	(0.84)
$EarnVol$	2.528	2.590	2.353	2.379	3.888*	3.891*	2.636	2.662
	(0.97)	(0.99)	(0.96)	(0.97)	(1.68)	(1.69)	(1.41)	(1.42)
$\#BusSeg$	-0.011	-0.011	-0.002	-0.002	0.003	0.004	0.021	0.021
	(-0.42)	(-0.43)	(-0.09)	(-0.07)	(0.18)	(0.23)	(0.97)	(0.99)
$\#GeoSeg$	-0.009	-0.009	0.026	0.026	0.028	0.027	0.010	0.010
	(-0.32)	(-0.32)	(0.91)	(0.90)	(1.08)	(1.07)	(0.44)	(0.43)
Age	-0.023*	-0.023*	-0.030**	-0.030**	-0.032***	-0.032***	-0.037***	-0.037***
	(-1.79)	(-1.81)	(-2.22)	(-2.22)	(-2.63)	(-2.62)	(-3.30)	(-3.29)
$M\&A$	0.015	0.016	-0.003	-0.002	0.016	0.016	0.008	0.008
	(0.52)	(0.54)	(-0.12)	(-0.09)	(0.80)	(0.83)	(0.42)	(0.44)
SEO	-0.033	-0.031	-0.026	-0.021	-0.071	-0.066	-0.056	-0.052

	(-0.55)	(-0.51)	(-0.35)	(-0.29)	(-0.93)	(-0.87)	(-0.93)	(-0.87)
SI	-5.624**	-5.558**	-3.297	-3.399	-0.287	-0.476	1.423	1.351
	(-2.03)	(-2.02)	(-1.26)	(-1.30)	(-0.15)	(-0.25)	(0.80)	(0.76)
Intercept	-0.062	-0.062	-0.005	-0.004	0.103	0.104	0.120**	0.121**
	(-1.00)	(-1.00)	(-0.08)	(-0.07)	(1.52)	(1.53)	(2.16)	(2.16)
Year, firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	15,394	15,394	15,394	15,394	15,394	15,394	15,394	15,394
Adj. R^2	0.059	0.059	0.055	0.054	0.054	0.053	0.048	0.048
Vuong's test: $Tone_{FinBERT}$ versus $Tone_{LM}$								
Z-statistic		-0.143		**2.504****		**2.421****		**2.130****
p-value		0.886		**0.012**		**0.015**		**0.033**