Forthcoming, *Review of Asset Pricing Studies*, 2021

Is Positive Sentiment in Corporate Annual Reports Informative? Evidence from Deep Learning*

Mehran Azimi
University of Massachusetts Boston

Anup Agrawal
University of Alabama

* We thank an anonymous referee, Hui Chen (the editor), Jonathan Brogaard, Stephen V. Brown (discussant), Mark Chen, Doug Cook, Mike Cooper, Marco Enriquez, Jerry Hoberg, Ravi Jagannathan, Erik Johnson, Anzhela Knyazeva, Diana Knyazeva, Lei Kong, Kelvin Liu, Kevin Mullally, Yahui Pan, Sugata Ray, Ken Rosen, Majeed Simaan (discussant), Andy Wu (discussant), Feng Zhang and conference and seminar participants at the AFA Poster Session, CFEA-NYU, FMA, MFA, SEC, University of Alabama, University of North Carolina Wilmington, University of Massachusetts Boston, University of Utah, Christopher Newport University, and Loyola Marymount University for helpful comments. Send correspondence to Anup Agrawal, University of Alabama, Culverhouse College of Business, Tuscaloosa, AL 35487-0224. Telephone: (205) 348-8970, E-mail: aagrawal@cba.ua.edu. The authors acknowledge support from a summer research grant from the Culverhouse College of Business, University of Alabama (Azimi) and the William A. Powell, Jr. Chair in Finance and Banking (Agrawal). All errors are our own.

Is Positive Sentiment in Corporate Annual Reports Informative? Evidence from Deep Learning

Abstract

We use a novel text classification approach from deep learning to more accurately measure sentiment in a large sample of 10-Ks. In contrast to most prior literature, we find that positive, and negative, sentiment predicts abnormal return and abnormal trading volume around 10-K filing date and future firm fundamentals and policies. Our results suggest that the qualitative information contained in corporate annual reports is richer than previously found. Both positive and negative sentiments are informative when measured accurately, but they do not have symmetric implications, suggesting that a net sentiment measure advocated by prior studies would be less informative. (*JEL* C81, D83, G10, G14, G30, M41)

Is Positive Sentiment in Corporate Annual Reports Informative? Evidence from Deep Learning

Introduction

Text has become an important source of data in economics and finance (see, e.g., Gentzkow, Kelly, and Taddy 2019 for a review of methods and applications). The sentiment or tone in text has been widely analyzed in finance (for excellent reviews of the literature, see Kearney and Liu 2014 and Loughran and McDonald 2016). Despite their widespread use, extant methods for measuring sentiment have low accuracy, which likely results in low power and incorrect inferences. For instance, implicit and explicit negation makes measuring positive sentiment challenging. Consequently, the literature is inconclusive regarding the information content of positive sentiment in financial text. In other words, it is unclear whether positive sentiment has information content and whether the market reacts to it (see the review by Loughran and McDonald 2016). In this paper, we introduce a state-of-the-art textual classification method for measuring the sentiment in financial text that is accurate, intuitive, and interpretable. We then use the method to address the unresolved issue regarding the information content of positive sentiment and re-evaluate previously established results on negative sentiment in corporate annual reports, filed with the SEC as 10-Ks. The method we introduce has broad applications because it can accurately mimic humans in eliciting what a text is about and its stance on the subject. More importantly, it can perform this task on large data sets. We illustrate the benefits of using this classification approach in the context of sentiment analysis.

Our approach to measuring sentiment is to read a text document and determine what percentage of its sentences are positive, negative and neutral. Though intuitive and interpretable, this approach is not feasible manually given that we have more than 200 million sentences in our sample. We employ recent technological advances in Natural Language Processing (NLP) and train a machine to perform this task with high accuracy. Our method achieves a leap in classification accuracy from 45% - 77% under existing methods to about 90%. We demonstrate the benefits of using our approach by comparing it with the two

most common methods in the literature and briefly describe how our method works. (Section 2 and Appendix A provide more details.)

By far, the most common method to measure sentiment in the finance literature is based on word dictionaries. The most influential study in this strand, Loughran and McDonald (2011; henceforth, LM), provides a list of words that are positive, negative, uncertain, etc. in finance texts. Measuring sentiment based on the frequency of the appearance of positive and negative words is simple but has several drawbacks. First, it ignores the context in which words appear. Second, the negation of positive words is hard to detect, especially implicit negation.[1] Third, there is no feasible external validation of the measure unless the method is applied to sentences instead of a full document.

A variant of this method assigns a weight to each word in a document to calculate a weighted sum of words. Jegadeesh and Wu (2013) is a notable study that finds a term-weighting scheme based on stock returns. The general drawback of this method, in addition to the drawbacks of the word list method, is that there is no theoretical framework that guides researchers as to which weighting scheme is appropriate. So, researchers face too many weighting schemes to choose from (see Loughran and McDonald 2016). Moreover, this method is less interpretable compared to a regular word-based method. In addition, this approach usually needs a word list to begin with, due to a degrees of freedom problem. Lastly and most importantly, using variables such as stock returns outside of a text document to find a weighing scheme assumes that the appearance and frequency of the words are related to those outside variables - an assumption that is itself often the question to be answered.

The second common method in this literature is the Naïve Bayesian Classification (NBC) method. It is a statistical method that, similar to our method, classifies sentences into the desired classes. The

[1] For instance, the tone of the following sentence from a 10-K is negative while the words in italic are positive. "For these and other reasons, these competitors may *achieve greater* acceptance in the marketplace than our company, limiting our ability to *gain* market share and customer loyalty and increase our revenues."

difference with our approach is in the underlying method and hence its accuracy. Under NBC, a sentence (or a document) is represented by a vector that shows how often each word appears in the sentence. Using a sample of labeled sentences, the model estimates the parameters which are then used to classify the "unseen" sentences into the categories. NBC ignores the relation between words and the sequential nature of the text[2]. Though intuitive and interpretable, this method has significantly lower accuracy than our method. In addition, the problem of negation seems to persist.

Our approach is based on classifying sentences into classes. As in a typical classification problem, a function operates on features and provides the probability that an observation belongs to each class. In our study, an observation is a sentence and classes are positive, negative, and neutral sentiments. In what follows, we describe the method we use to calculate features, i.e., word-embedding. We then explain our choice of the function, i.e., neural networks.

We start by mapping each word into a vector of low dimension. This process is called word-embedding. The goal is to reduce the dimension while preserving the semantic and syntactic aspects of words. We implement word-embedding with a structure suggested by Mikolov et al. (2013a) using more than 7 billion words and 220 million sentences from the full text of all 10-K filings by U.S. public companies made during 1994-2017. The output of word-embedding represents each word with a low-dimension vector. Similar words have close vector representations measured by cosine similarity (Table A1 shows several examples).

We then use recurrent neural networks (RNN), which takes the sequence of word vectors in a sentence and classifies the sentiment expressed in the sentence into one of three classes: negative, positive, and neutral. Using RNN allows us to capture complex non-linear dependencies between words, while taking

[2] NBC can add sequences of two or more words (Bi-grams and N-grams) as standalone features of the document. However, the number of parameters explodes as the sequence gets larger. Moreover, this variant of NBC is expected to work well in cases where negation is explicit and occurs in close proximity of a positive word, *e.g.,* 'The movie was not good,' which is not common in financial text.

into account the sequential nature of the words in a sentence. Taken together, the two steps result in a sentiment classifier that takes into account the relation between words and the sequential nature of text.[3] We train our RNN classifier using 8,000 manually labeled sentences that are randomly selected from 10-K filings. We use two criteria, namely accuracy and *F1-score* (defined in section 2), to select the best measure among LM, NBC, and our deep learning approach.

The accuracy of existing methods is 45% for LM[4] and 78% for NBC. Our method results in a substantial increase in accuracy to 91%. The 78% accuracy of the NBC method is likely an overestimate because our random sample of sentences contains only 10,600 unique words, which is substantially less than the 45,191 total words in our dictionary. As a result, all the information in the words that are not represented in our training sample is lost and NBC is more likely to misclassify out-of-sample sentences. Our method significantly mitigates this issue because word-embedding allows the classifier to learn about unseen words since our sample contains words with similar connotations.

Our second criterion, F1-score, takes into account both Type I and Type II errors in classification (see, e.g., Loughran and McDonald 2016). Our method has an F1-score of 84.8%, while it is 66.9% for NBC and 46.1% for LM. Thus, the improvement in accuracy and F1-score of our approach over the two prior approaches is quite substantial. In addition, we use a regularization method to mitigate overfitting

[3] Since word-embedding is performed before sentiment classification, the output of word-embedding does not contain the tonal aspect of words, thus precluding a look-ahead bias in subsequent predictive regressions.

[4] LM method computes the sentiment of a document, rather than a sentence. In this section, to compare the accuracy of different methods, we classify the sentiment of a sentence under LM method as positive (negative) if it has more (fewer) positive than negative words. In the rest of the paper, consistent with the prior literature, we calculate positive (negative) sentiment under LM method as the ratio of the number of positive (negative) words to the total number of words in a document.

when training the model. As a result, the performance of our classifier in an out-of-sample set of 1,500 randomly selected sentences, with 90% accuracy and 84.5% F1-score, is very close to the in-sample performance.

Based on these results, we select our method as the appropriate method to perform sentiment classification and to measure sentiment. Armed with an accurate and reliable measure of sentiment, we next delve into the empirical questions regarding sentiment. We first examine whether the market reacts to 10-K sentiment. We then examine whether the sentiment is informative, i.e., whether it has predictive power regarding future firm fundamentals and policies. We interpret our results and briefly discuss plausible economic mechanisms that could explain the results but leave their thorough investigation for future research. Throughout, we also perform the analysis using the two commonly used sentiment measures, i.e. NBC and LM, to identify situations where the previous methods provide inferences that are correct and those where they are not. The choice of a sentiment measure is thus independent of our subsequent analysis.

We start by examining the relation between our sentiment measures and the reaction of stock prices and trading volumes to the 10-K filing. We find that negative (positive) sentiment significantly predicts lower (higher) abnormal return over days (0, +3) around the 10-K filing date, i.e., the filing period. After controlling for quantitative information in the filing and other relevant variables, a one standard deviation increase in negative (positive) sentiment predicts a change in cumulative abnormal return of -0.13% (0.07%) during the filing period. Under LM method, positive sentiment is unrelated to the filing abnormal return. Under NBC method, neither negative nor positive sentiment measure is significantly related to the abnormal return at the 10-K filing.

We also find that both positive and negative sentiment are related to higher abnormal return over event windows of up to one month after the filing period. This finding suggests that during the filing period the market underreacts to positive sentiment and overreacts to negative sentiment in the 10-K filing. LM sentiment measures fail to capture this dynamic. NBC positive sentiment exhibits weaker relations and only for longer periods after the filing date. In addition, negative (positive) sentiment predicts significantly higher (lower) abnormal trading volume around the filing date, suggesting that it reflects more (less)

concerns and uncertainty about the future, which increases (decrease) the divergence of opinion across investors. In multivariate analysis, a one standard deviation increase in negative (positive) sentiment predicts a 0.13 (0.04) standard deviation increase (decrease) in abnormal trading volume. The differential magnitudes suggest that investors are more responsive to negative sentiment than to positive sentiment. Overall, these results show that positive and negative sentiment measures do not have symmetric relations with abnormal return and trading volume. This asymmetric relation generally holds in the rest of our empirical results. Our finding that positive textual sentiment in 10-K filings sensibly and reliably predicts investor reactions to the filing is new to the literature, which has largely been unable to find significant results with positive sentiment, mainly because of the inability of existing methods to measure positive sentiment reliably. This a key advantage of our deep learning approach over existing methods of textual sentiment analysis.

We next examine the relation between sentiment and future firm fundamentals. We find that positive sentiment predicts higher return on assets, higher operating cash flow, and higher net income over the next year, while negative sentiment predicts lower values of these performance measures. Positive LM sentiment predicts lower future profitability, which is counterintuitive, but consistent with the measure being inaccurate. While NBC sentiment measures have the same signs as our deep learning method, the former have up to 60% lower economic significance, particularly for positive sentiment.

We next evaluate the informativeness of the sentiment in the 10-K filing regarding future firm policies. The sentiment in corporate annual reports reflects general business environment, outlook, and investment opportunities which are related to the need for holding cash. We empirically examine the relation between sentiment and future cash holdings. We find that negative sentiment predicts higher future cash holdings, which suggests that firms increase cash holdings when expecting more uncertainty and an unfavorable business environment. Consistent with this interpretation, positive sentiment predicts lower future cash holdings. The estimated effect of negative sentiment is three times larger in magnitude than positive sentiment. Comparing with other methods, LM estimates the effect of positive sentiment with a wrong sign, while NBC positive sentiment has a smaller economic effect on future cash holdings.

Our finding that positive sentiment predicts higher future cash flow from operations triggers a natural question: what is the extra cash flow used for? To investigate this issue, we examine the relationship between sentiment and future use of leverage. Using book leverage to remove the effect of change in market value, we find that a one standard deviation increase in positive sentiment predicts a 0.13 standard deviation decrease in leverage in the next period, suggesting that the extra cash generated in the future is used to reduce leverage. On the other hand, negative sentiment predicts higher leverage, but the magnitude of this relation is much smaller than that of positive sentiment. The results using LM sentiment and NBC positive measures are consistent with our deep learning measures, but NBC negative sentiment has no predictive power. Overall, the fact that our approach yields results on future firm fundamentals and policies that are more sensible is another major advantage of our approach over the existing methods.

Finally, motivated by Cohen, Malloy, and Nguyen (2020), we examine whether changes in sentiment are informative. We repeat our analyses using changes, instead of levels, of sentiment as independent variables. We find that an increase in positive sentiment predicts higher abnormal return at the 10-K filing date. While the coefficient of change in negative sentiment is negative, it is statistically insignificant. Moreover, changes in sentiment predict future profitability, cash holdings, and leverage. The results for changes in positive sentiment are much stronger than for changes in negative sentiment, both statistically and economically. In contrast, changes in LM and NBC sentiment measures largely fail to predict filing abnormal returns, future profitability and leverage.

Overall, we find persuasive empirical evidence that, in contrast to prior studies, positive sentiment in 10-K filings is informative and that the market reacts to it. The effects of positive sentiment and negative sentiment in corporate filings are often asymmetric, which implies that using a net sentiment measure advocated by prior studies would result in loss of information. More importantly, our findings suggest that employing this state-of-the-art technique for textual analysis can provide more reliable measures of sentiment. The word-embedding matrix and the NN classifier can be shared and used easily, and researchers can improve the accuracy of the classifier by using their own labelled sentences, which would substantially reduce the cost of using this approach. Finally, in addition to measuring general sentiment in other sources

of textual data in finance, this method can be used for tasks such as topic-specific content analysis, e.g., classifying text into topics such as competition, innovation, financial constraints, supply chain disruptions or foreign demand shocks, and to measure the tone within each topic.

The cost of using our approach is learning this new technology and the manual work needed to classify the sentences in the training set. However, NBC shares these features. LM method doesn't require this manual work if word lists are already developed in the language of study and source of textual data, e.g., news media, social media, etc. If not, researchers need to develop their own word lists which requires a significant amount of manual work. In terms of computational power, performing word-embedding, training the classifier, and running the classifier on the full sample takes about one to two weeks on an average desktop computer. The benefits of using our approach are significant improvements in accuracy and F1-score of sentiment measures, which mitigate concerns about low power and incorrect inferences under previous methods. Moreover, our approach can be modified and extended to measure the source of tone-induced return predictability. Our approach can also be used to measure the stance of a text on any subject. In sum, this method allows us to extract and quantify significant amount of information from textual data.

The paper contributes to the literature on textual content analysis (see, e.g., Huang et al. 2017; Li, Lundholm and Minnis 2013) and sentiment analysis (see, e.g., Henry 2008; Tetlock, Saar-Tsechansky and Macskassy 2008) by introducing a novel text classification approach. Our approach to measure sentiment is sentence-based, rather than word-based, and circumvents the need to develop word lists or to choose a term-weighting scheme. Our approach also makes use of the relationship between words in context and considers a sentence as a sequence of words rather than a bag-of-words in which order does not matter. These two properties are the main advantages of this approach compared to the NBC approach (see, e.g., Li 2010; Huang, Zang and Zheng 2014), resulting in higher accuracy of sentiment classification. More specifically, the paper contributes to the literature on sentiment analysis of 10-Ks (see, e.g., Loughran and McDonald 2011), finds new evidence on its information content, and addresses the unresolved issue regarding positive sentiment. More broadly, the paper contributes to the literature on qualitative information

in accounting and finance (see, e.g., Mayew and Venkatachalam 2012; Coval and Shumway 2001). Finally, the paper contributes to the literature on corporate disclosures (see, e.g., Dyer, Lang and Stice-Lawrence 2017; Li 2010) by providing evidence on the information content of 10-K filings.

1. Related Literature

Textual content analysis is a growing literature in finance. In this section, we briefly discuss the literature on content analysis based on the most popular methods, followed by the papers on sentiment analysis relevant to this study. Kearney and Liu (2014) and Loughran and McDonald (2016) provide detailed reviews of the finance literature on textual sentiment and textual analysis, respectively. Gentzkow, Kelly and Taddy (2019) survey statistical methods for analyzing textual data and its applications in economics and related social sciences.

One strand of this literature relies on word-based sentiment measures and field-specific dictionaries. Earlier sentiment studies use DICTION, Harvard General Inquirer, and Henry (2008) word lists to measure the tone or sentiment of a financial document. Most recent studies use Loughran and McDonald's (2011) word lists, especially their lists of negative and uncertain words, because they have been found to be more relevant for financial documents.

Other studies develop and use topic-specific word lists. Hoberg and Maksimovic (2015) use a word list to identify financially-constrained firms. Li, Lundholm and Minnis (2013) measure competition by counting the number of occurrences of the word *compete* and its variants in 10-K filings. Qiu and Wang (2017) use a word list to measure skilled labor risk that firms face. Loughran, McDonald and Yun (2009) find a relation between ethics-related word count in a stock's 10-K filing and the probability of it being a 'sin' stock.

Another strand of the content analysis literature applies techniques from NLP and machine learning. Several studies employ NBC for sentiment analysis. Huang, Zang and Zheng (2014) and Li (2010) use this method to measure the sentiment in analyst reports and forward-looking statements in 10-K filings,

respectively. Ji, Talavera and Yin (2018), Antweiler and Frank (2004), Ryans (Forthcoming), and Buehlmaier and Whited (2017) have also applied NBC in different settings.

Finally, several studies use a topic modeling approach called Latent Dirichlet Allocation (LDA) that is most suitable for assigning interpretable topics to a document. Huang et al. (2017) use LDA to show that analysts discuss topics beyond what firms disclose. Dyer, Lang and Stice-Lawrence (2017) employ LDA to explore changes in 10-K disclosures over time. Bellstam, Bhagat and Cookson (forthcoming) apply LDA, together with LM word lists, to analyst reports to construct a measure of innovation. Hanley and Hoberg (2019) use LDA, together with word-embedding that we employ in this paper, to identify interpretable emerging risks in the financial sector. While LDA has not been used for sentiment analysis in finance, it can be. Similar to word-embedding techniques, LDA outputs a vector representation of words, which can be fed to a NN to build a classifier.

Sentiment analysis in finance has established that sentiment is informative for stock prices, firm fundamentals, and the overall stock market performance. This literature uses several sources of textual data such as corporate disclosures, analyst reports, news articles, earnings conference calls, and social media. Most of the literature has focused on negative and uncertain words to measure sentiment. Tetlock, Saar-Tsechansky and Macskassy (2008) show that negative words in news stories predict earnings and that the market reacts to that information. Huang, Zang and Zheng (2014) find that negative and positive sentiment in analyst reports are related to abnormal return and future earnings growth. Feldman et al. (2010) find that changes in the tone of the management discussion and analysis (MD&A) sections of 10-K filings are related to the filing period excess return. Li (2010), using NBC to construct a single tone measure, finds that the tone of forward-looking statements in MD&A predicts future profitability and liquidity. Cohen, Malloy, and Nguyen (2020) find that at the time of a 10-K or 10-Q filing, investors don't react to changes in the language used from the previous filing. But these changes, identified using document similarity measures, predict future stock returns and profitability.

Loughran and McDonald (2011) find that negative, but not positive, words in 10-K filings are related to abnormal returns around the filings. Our study comes closest to this paper in that both examine the information content of the sentiment in 10-K filings. LM establish new word lists and show that negative and uncertain words are related to variables such as abnormal return, trading volume, and fraud. Loughran and McDonald (2016) caution that researchers need to deal with the negation of positive words to examine positive sentiment. Our paper uses deep learning to measure sentiment more accurately and intuitively, re-examines several previously established results on negative sentiment, and finds new evidence on the information content of positive sentiment.

2. Sentiment Classification

In this section we briefly discuss the method we use for sentiment classification. A more detailed discussion is in Appendix A. Our approach is sentence-based, *i.e.* it assigns sentiment to each sentence. This approach classifies the sentiment in sentences similar to the way a human being (i.e., an intelligent agent) would do it. Since we use a large textual dataset, manually performing sentiment classification is nearly impossible. We borrow from the artificial intelligence literature to perform this task.

Our approach to sentiment classification is a two-step process. First, we use a dimensionality reduction technique, *i.e.* word-embedding, and find vector representation of words, in which each word is represented by a vector of low dimension. The idea behind the method is to maximize the probability of choosing the *current* word, given a set of words surrounding it in a sentence. The algorithm finds close vector representation for words that surround the *current* word in different sentences. The parameters associated with each word in this set up construct the vector representation. The results of word-embedding depend on the textual data that is used, among other factors. Generally, it is desirable to use as much relevant textual data as possible. To perform word-embedding, we use the full text of all 10-K filings by U.S. public companies over 1994-2017. The choice of vector size, i.e., the word-embedding dimension, is somewhat arbitrary, but the recommended range is between 20 and 500. We choose 200 for this dimension in an attempt to get high accuracy in sentiment classification (which uses the output of word-embedding), while

keeping the computational cost reasonable.[5] Word-embedding is known to preserve semantic and syntactic features of words. Similar words have a similar representation measured by cosine similarity. In a recent study, Li et al. (forthcoming) use word-embedding to find words that are relevant to corporate culture. We then represent each sentence as a sequence of vectors corresponding to the words in the sentence.

In the second step, we train a neural network (NN) to classify a sentence into three categories: negative, positive and neutral. We use recurrent NN (RNN) as it is better suited to sequential data such as text (see, e.g., LeCun, Bengio, and Hinton 2015). More specifically, we employ long short-term memory (LSTM) network, introduced by Hochreiter and Schmidhuber (1997), that enables the network to retain information from observations that are far from the end of the sequence.[6] To train our NN, we manually classify 8,000 randomly selected sentences (train-set) into the three categories[7]. Our first criterion in measuring the performance of the classifier is accuracy, which is defined as the percentage of all sentences

[5] As discussed below, our procedure yields an accuracy of 91% in-sample and 90% out-of-sample.

[6] Our choice of the structure of the sentiment classifier, *i.e.* word-embedding followed by LSTM network, is a natural choice in NLP. Wang et al. (2015) employ a similar structure to perform sentiment classification on Twitter posts. They achieve comparable accuracy to the best available data-driven approaches at the time, and higher accuracy than several feature-engineering approaches. We use the same structure but perform word-embedding independently of RNN.

[7] Can 'the benefit of hindsight' affect how we label the sentiment of some sentences, which could then affect our subsequent predictive results? Well, for labeling sentiment, we only observe the sentences and do not need any other information related to the firm, date, context, returns, etc. While it is possible to take that information into account when manually labeling the sentences to perform a possibly more accurate classification, it is impossible to tell how labeling a sentence differently would affect the ultimate classifier we train, the results of millions of sentences to be classified by the classifier, and the eventual empirical results.

whose sentiment is correctly classified. The in-sample accuracy of the trained NN is 91%. We then examine the out-of-sample performance of the classifier. We use an additional 1,500 manually labelled sentences (test-set) and find an out-of-sample accuracy of 90%.

Panels A and B of Table 1 show the distribution of categories for the train-set and the test-set, respectively. Note that negative sentences that are classified as positive and vice versa are rare. Panel C shows the accuracy if we use LM word lists to classify sentences. This part is for comparison with other studies (e.g., Huang, Zang and Zheng 2014) as the method to calculate the sentiment in a 10-K is based on the number of words, not the number of sentences. However, it illustrates that LM positive and negative words often appear in neutral contexts. Panel D presents the same analysis using NBC.

To quantify this analysis, we use *F1-score* as our second criterion to measure the performance of our classifier. It is defined as the harmonic mean of *Precision* and *Recall*. Precision for class C is # of sentences correctly classified as C / total # of sentences classified as C. Recall for class C is # of sentences correctly classified as C / (# of sentences correctly classified as C + # of sentences incorrectly not classified as C). For a multiway classification problem, F1-score is the average of the F1-scores across classes. Precision, recall, and F1-score for each class can be calculated using the accuracy matrix in Table 1. Notably, precision and recall for the positive class using our deep learning method are 80% and 69% respectively. Precision and recall for the positive class under the LM method are 25% and 68%, while they are 43% and 47% under the NBC method. Consistently across all classes, our deep learning sentiment classifier achieves higher precision and recall compared to LM and NBC methods.

We use the trained NN to label all the sentences in a 10-K filing to calculate the overall sentiment of the filing. Table A2 provides some examples of sentences we classify as negative, positive and neutral to train the NN. We also report negative (positive) words based on LM word lists in sentences in which the sentiment is not negative (positive) to illustrate that the meaning of words depends on the context in which they are used.

Our approach to sentiment classification uses the relation between words and considers a sentence as a sequence of words. The former is achieved by using word-embedding and the latter is achieved by using RNN for sentiment classification. Word-embedding enables the classifier to accurately classify sentences in out-of-sample data even if some words do not exist in the train-set. The classifier can relate the 'unseen' words to similar 'have seen' words in the train-set. This is one of the main advantages of this method compared to NBC. Overall, our approach is sentence-based, which is by its nature more accurate and intuitive than word-based measures. It also achieves high accuracy compared to the extant sentence-based methods used in finance and accounting.

3. Data

We obtain data on firm fundamentals from Compustat, and stock prices and trading volumes from CRSP. We compute cumulative abnormal returns using Eventus. We use the GVKEY-CIK Link table from the SEC Analytics Suite to link each 10-K filing with a Compustat firm. We obtain all 10-K and 10-K405[8] filings by U.S. public companies during 1994 to 2017 from the Software Repository for Accounting and Finance (SRAF) website, maintained by Professor Bill McDonald.[9] SRAF has parsed EDGAR filings to remove encodings unrelated to the textual content of the filings. We start our matching process by downloading 193,692 10-K filings, excluding duplicates and firms that file multiple filings on the same date. We then find a matching GVKEY, using the GVKEY-CIK Link table which results in 156,288 filings. Next, we find Permno match and only include share codes equal to 10 and 11 (i.e., equity securities issued by companies incorporated in the U.S.), resulting in 98,602 filings. We then exclude utility and financial

[8] Form 10-K405 is a Form 10-K that indicates that an officer or director of the company failed to file their insider trading disclosures (Forms 3, 4 and 5) on time. Form 10-K405 was discontinued after 2002. We follow Loughran and McDonald (2011) and do not include 10-KSB and 10-KSB405 filings, mostly by penny stock firms, that existed until 2009.

[9] Available at: http://sraf.nd.edu/

firms and all filings with less than 200 sentences. For each firm, we only include the first filing for each reporting period in case of multiple reports. The final sample consists of 62,726[10] firm-year observations with non-missing cumulative abnormal returns to estimate equation (1).

To perform word-embedding, 10-K filings need to be preprocessed. Inputs to the algorithm are sentences, therefore we tokenize each 10-K filing into sentences. Next, each sentence needs to be tokenized into words. We convert all words into lowercase, exclude words that appear in less than 100 filings, and exclude words that appear less than 500 times in all of the filings combined. This procedure results in a dictionary of 45,191 words. While the choices of 100 and 500 are arbitrary, the idea is to produce a dictionary that is not too large, so as to save computational cost when performing word-embedding. The pre-processing results in 220 million sentences and 7.5 billion words in more than 190,000 10-K filings[11].

After pre-processing, all the sentences are fed to an algorithm to compute the word-embedding matrix. One popular, efficient, and scalable choice for implementing word-embedding is the Gensim software. Specifically, we use the Word2vec[12] module that implements Mikolov's (2013a and 2013b) proposed structure. This module takes as hyper-parameters the number of surrounding words, the dimension of the word vectors, and several other parameters that determine the sampling frequency, hardware configuration, training algorithms, etc. We set the dimension of word-embedding to 200 for this study.

To construct measures of positive and negative sentiment, we use the trained NN to classify all the sentences in each 10-K filing into positive, negative and neutral. The total number of negative (positive)

[10] For comparison, Jegadeesh and Wu (2013) report 45,860 filings during 1995-2010, without excluding utility firms.

[11] For word-embedding, it is desirable to use as much relevant text as available. So, we use all filings, instead of trying to find a GVKEY or Permno match.

[12] Available at: https://radimrehurek.com/gensim/models/word2vec.html

sentences divided by the total number of sentences in each filing is our measure of negative (positive) sentiment. We also calculate the sentiment based on LM word lists for each filing, as defined in Appendix B. Panel A of Table 2 shows Pearson correlations between our sentiment measures and those of LM. It is interesting to note that the correlation between our and LM's negative (positive) sentiment measures is 0.56 (0.51), i.e., roughly mid-way between 0 and 1. Panel B of Table 2 shows summary statistics of our sentiment measures and firm-level variables.

4. Empirical Results

In the previous section, we describe the process of calculating the sentiment in 10-K filings based on the sentiment of all the sentences in each filing. We choose to analyze the full text of 10-Ks, instead of its sections such as Risk Factors or MD&A, for two reasons. First, prior studies (e.g., Loughran and McDonald 2011) find that the MD&A section is not informative. Second, the Risk Factors section generally has negative sentiment which can be measured relatively accurately using negative words. The full text of 10-K is more suitable for investigation since there are comparable studies (*e.g.*, Loughran and McDonald 2011; Jegadeesh and Wu 2013) on it, and both negative and positive sentiment is prevalent in it.

Sentiment is a general concept that is quantified. Sentences can have positive or negative sentiment, but they can be about different topics. Managers express facts and opinions on a variety of topics in 10-K filings. A negative sentence can be about competition a firm faces, regulations that affect its operations and profitability, lawsuits against the firm, its inability to raise funds, the loss of key personnel, and many other issues. Each of these cases can affect firm fundamentals to different extents, but they are all expected to affect profitability negatively. In sentiment analysis, we aggregate all these topics and provide a unified measure of negative and positive sentiments.

The sentiment in a 10-K filing reflects managers' opinions of the firm's operating results over the past year and their view of what the future holds for the firm. To the extent that these opinions and views are informative beyond the quantitative information in 10-K filings, the market should respond to them and they should be reflected in future fundamentals of the firm, on average. To test the former prediction, we

examine the response of stock prices and trading volumes to the sentiment in 10-K filings. To test the latter, we examine whether the sentiment in 10-K filings predicts future firm fundamentals.

4.1 Does sentiment predict abnormal returns?

The first question we address after computing an intuitive and accurate measure of sentiment is: Is the sentiment in 10-K filings associated with abnormal stock returns around the 10-K filing date? Previous studies find that negative sentiment predicts negative abnormal returns. Jegadeesh and Wu (2013) find that both negative and positive sentiments are associated with abnormal returns. We start by re-examining these central results and estimate the following equation:

$$CAR = \alpha + \beta_1 \,.\, Negative + \beta_2 \,.\, Positive + \gamma \,.\, Controls \qquad (1)$$

where *CAR* is the cumulative abnormal return (based on Fama-French three factor model plus momentum) over days 0 to +3 around the filing date[13], *Negative* and *Positive* are our measures of negative and positive sentiment respectively, and *Controls* is a set of control variables that captures quantitative information included in the 10-K filing, namely *Total Assets, Tobin's Q, Market cap, Cash, Leverage* and *ROA*. All the variables are defined in Appendix B. Following Jegadeesh and Wu (2013), we also include the abnormal return over days [-1, +1] around the earnings announcement (*EARet*) in our set of control variables in equation (1). We also estimate the same set of regressions using sentiment measures computed using word lists similar to Loughran and McDonald (2011) and NBC. For comparison, all sentiment measures are normalized to have a mean of zero and a standard deviation of one.

The results are shown in Table 3. Column 1 shows a regression that includes just our negative and positive sentiment measures and control variables. Columns 2 and 3 replace our sentiment measures with LM and NBC sentiment measures. Columns 4 to 6 add year-quarter fixed effects and industry fixed

[13] Our choice of this time window to measure the abnormal return to 10-K filings follows prior studies (see, e.g., Loughran and McDonald 2011; Jegadeesh and Wu 2013).

effects.[14] In columns 7 to 9 we exclude observations for which there is an earnings announcement within 2 days prior to the 10-K filing date. In all the specifications, higher negative sentiment predicts lower cumulative abnormal return around the filing date, which is consistent with previous studies. The coefficient of *LM Neg*, the negative sentiment calculated using LM negative word list, is also negative and statistically significant, consistent with the results of Loughran and McDonald (2011).

Notably, our positive sentiment measure predicts higher cumulative abnormal return. In line with most previous findings, the positive sentiment measured by positive words, *LM Pos,* is unrelated to the abnormal return in any specification. NBC sentiment measures are not related to abnormal return in any of the specifications. As shown in column 1, after including control variables, a one standard deviation increase in negative (positive) sentiment predicts a change in cumulative abnormal return of -0.13% (0.07%). Not only is positive sentiment related to abnormal return, its estimated coefficient is non-trivial. In sum, both negative and positive sentiments are significantly related to abnormal return in opposite directions. Our finding that positive sentiment in a 10-K filing predicts the abnormal return to the filing is new compared to most of the prior literature, except for Jegadeesh and Wu (2013).

We next examine whether these relationships in a short time-window after the 10-K filing date continue or reverse over longer windows after the filing period. Consistent with Jegadeesh and Wu (2013), we re-estimate equation (1) after replacing the dependent variable with the cumulative abnormal return calculated over three different windows after the first trading week following the 10-K filing. The lengths of these windows are one week (5 trading days), two weeks (10 trading days), and one month (22 trading days). Table 4 shows the results. Negative sentiment, which predicts lower abnormal return during the filing

[14] We do not include firm fixed effect in our analysis because we don't have enough degrees of freedom. Our sample is limited by electronic filings of 10-Ks, which only began widely since 1996. (Only a few firms filed electronically with the SEC during the transition period of 1994-1995.) Nevertheless, our results are qualitatively similar if we include firm fixed effects.

period, predicts higher abnormal return after the filing period, which suggests that the market overreacts to negative sentiment during the filing period. But positive sentiment predicts higher abnormal return both during and after the filing period, suggesting that the market underreacts to positive sentiment during the filing period[15]. Table 4 also shows the corresponding analysis using LM word lists and NBC. Word-based sentiment measures are unrelated to abnormal returns after the filing period. Both positive and negative NBC sentiment measures, which are unrelated to filing abnormal returns, predict higher abnormal returns after the filing period, although positive sentiment becomes significant only over longer time windows.

The asymmetric reaction of the market to positive and negative sentiment during the filing period is related to the literature on reversal, drift and information transmission. While many studies find underreaction to the hard information in news such as announcements of earnings or M&A and to changes in analyst recommendations, many others focus on soft information. For instance, Tetlock, Saar-Tsechansky and Macskassy (2008), Feldman, et al. (2010), and Jegadeesh and Wu (2013) find that the market doesn't respond fully and immediately to the qualitative information contained in media news and corporate public reports. The evidence in this literature is mixed (see, e.g., Tetlock 2014) and tends to find overreaction to media news and underreaction to the more sophisticated soft information in corporate reports. The evidence on the direction of the response to positive and negative news is also mixed. Frank and Sanati (2018) propose a unified framework to explain price response to news shocks and focus on investor type and market conditions rather than the information itself. We believe that our result is best viewed in the context of lazy prices (see Cohen, Malloy, and Nguyen 2020) in the sense that the market seems to be inattentive to the information contained in corporate annual reports. The reaction to the sentiment in reports over the filing period is comparable in magnitude to that of the post-filing period. This result differs from studies that find that the post-disclosure effect is significantly smaller than the disclosure-

[15] Jegadeesh and Wu (2013) find that the market underreacts to both sentiment measures during the filing period.

period effect. Perhaps this result is not surprising given that 10-K filings tend to be complex and lengthy reports that appear to be overlooked by even sophisticated investors. On the other hand, news reports tend to be short, easy to interpret, and catch a lot of attention from investors, especially retail investors. Therefore, the market response to the information differs depending on information attributes as well as market conditions and investor type. Our analysis of the market response based on firms' information environment further supports this idea.

We also examine the performance of a trading strategy based on the sentiment measures. We rank firms with December fiscal year end at the end of March of each year based on their negative and positive sentiment. We then construct a portfolio that longs stocks in the highest (lowest) quintile of positive (negative) sentiment and short sells stocks in the lowest (highest) quintile of positive (negative) sentiment. The portfolio is rebalanced once a year at the end of March.[16] We regress the return of the portfolio on Fama-French three factors and calculate alpha. In untabulated results, we find that the alpha is statistically insignificant using either our positive or negative sentiment measures. This result is consistent with Loughran and McDonald (2011).

In addition, we test whether the information environment of firms affects the market reaction at the time of 10-K filings. One would expect that firms with low analyst coverage will have greater information asymmetry between managers and investors. Therefore, the market response to the information in 10-K filings should be stronger for such firms. On the other hand, these firms are usually smaller with less diversified operations, making them less complex with lower information asymmetry. These two effects are in an opposite direction and we cannot predict *ex ante* whether the market reacts more strongly to the sentiment in 10-K filings for firms with low analyst coverage or for firms with high analyst coverage. To examine this issue, we partition firms at the median based on analyst coverage into high and low coverage groups and estimate equation (1) separately for each group. We then compare the estimated coefficients. In

[16] The results are similar if we hold the portfolio for three months, instead of one year.

untabulated results, the estimated coefficients of our sentiment measures are not statistically different between the two groups. We also partition firms based on the dispersion of analyst forecasts as an alternate measure of information asymmetry, and repeat the previous analysis. Again, we find no statistically significant difference between the estimated coefficients of the sentiment measure between the two groups.

Overall, we find that our sentiment measures predict abnormal return during and after the 10-K filing period up to one month. LM positive sentiment is unrelated to abnormal return and LM negative sentiment only predicts abnormal return during the filing period but not after that. NBC sentiment does not predict abnormal return during the filing period and predicts return after the filing period in some specifications.

4.2 Does sentiment predict abnormal volume?

We next examine the relation between the sentiment measures and abnormal trading volume over days 0 to +3 around the 10-K filing date. We estimate the same equation as in equation (1), with abnormal trading volume as the dependent variable. We calculate abnormal trading volume following Loughran and McDonald (2011) using the mean (M) and standard deviation (S) of trading volume during the 60-day period that ends 5 days prior to the filing date. Thus, abnormal volume for a firm over day t is computed as $AV_t = (V_t - M) / S$, where V_t is its trading volume on day t. The mean of AV_t over days t = 0 to +3 is our measure of abnormal trading volume for a firm. The results are shown in Table 5.

In all specifications, higher negative sentiment predicts higher abnormal trading volume, and higher positive sentiment predicts lower abnormal trading volume. Higher negative sentiment potentially reflects more uncertainty, raises investor concerns about the firm's future and increases asymmetric information among investors, resulting in higher divergence of investors' opinion and higher abnormal trading volume. On the other hand, higher positive sentiment signals that managers expect less uncertainty about the future and reflects more resolved concerns that firms might have faced, resulting in lower abnormal trading volume. The results are similar when using NBC, but LM word lists provide mixed results. In column (1), a one standard deviation increase in negative (positive) sentiment predicts 0.65/4.94 = 0.13 (0.18/4.94 =

0.04) standard deviation increase (decrease) in abnormal trading volume. The absolute values of the estimated coefficients of negative and positive sentiment are statistically different at the 1% level of significance. This asymmetric result suggests that investors are more responsive to negative sentiment than to positive sentiment.

These results are also consistent with our results on the market reaction during and after the filing period. Negative 10-K sentiment predicts higher trading volume that leads to prices exceeding their intrinsic values, leading to a reversal, consistent with our finding that negative 10-K sentiment predicts a reversal in stock prices after the filing period. The negative relation between positive sentiment and abnormal trading volume is consistent with prices not fully adjusting to positive 10-K sentiment over the filing period.

Overall, we find in section 4 so far that positive sentiment, as well as negative sentiment, predicts filing period abnormal return and abnormal trading volume. In addition, the results on abnormal return after the filing period and the asymmetric results on trading volume suggest that positive sentiment is by nature different from negative sentiment. When manually labeling 9,500 sentences, we observe that positive and negative sentences tend to discuss different topics. Aggregating these two measures to construct a net sentiment measure would likely result in loss of information embedded in them. Our results in the next sub-section further support this idea.

4.3 Does sentiment predict future firm fundamentals?

In their annual reports, firms usually discuss their outlook on the economy, industry, and firm, disclose risk factors, explain the firm's future directions, and report key factors affecting revenues and expenses. Whether this textual information, and the sentiment expressed in it, contains information regarding future firm fundamentals that is not captured by the quantitative information in the report is an empirical question. Most prior studies find that only negative sentiment has information content about firm fundamentals. In this section, we re-examine these findings and also investigate whether positive sentiment is informative.

We start by estimating the following regression:

$$ROA_{(t+1)} = \alpha + \beta_1 \,.\, Negative_{(t)} + \beta_2 \,.\, Positive_{(t)} + \gamma \,.\, Controls_{(t)} \qquad (2)$$

where *ROA* is the return on assets, *Negative* and *Positive* are normalized measures of negative and positive sentiment, and *Controls* is a set of control variables found by the prior literature to affect profitability. The coefficients of interest are β_1 and β_2. In a series of specifications, we successively add year-quarter and industry fixed effects. The results in panel A of Table 6 support the idea that the sentiment conveyed by managers in the 10-K filing is informative about future firm profitability. Positive sentiment predicts higher future *ROA* and negative sentiment predicts lower future *ROA*. In column (1), a one standard deviation increase in positive (negative) sentiment predicts 1.7 (2.8) percentage point increase (decrease) in *ROA* the next year. When we repeat this analysis using sentiment measures based on word lists, while the results are similar for negative sentiment, positive sentiment predicts *lower* future profitability. These results suggest that our deep learning approach adds considerable value, especially for measuring positive sentiment. The NBC sentiment measures predict future *ROA* similar to our measures, but its positive sentiment is economically less significant than the deep learning approach in all three specifications. In untabulated results, we find qualitatively similar results when using net income as the left-hand side variable.

Next, we estimate the regression in equation 2 using $Op.\ CFlow_{(t+1)}$ as the dependent variable. *Op. CFlow* is net operating cash flow divided by total assets. The results in panel B of Table 6 show that positive (negative) 10-K sentiment predicts higher (lower) cash flow the next year. In column (1), a one standard deviation increase in positive (negative) sentiment predicts a +1.4 (-1.9) percentage point change in future operating cash flow. Here too, positive sentiment is informative and its effect is roughly of the same order of magnitude as the negative sentiment. When we repeat this analysis with sentiment measures using word lists, negative sentiment significantly predicts lower future *Op. CFlow.* But the coefficient of positive sentiment is also negative, consistent with the conclusion of previous studies that find that positive sentiment based on positive word lists provides an inaccurate measure of sentiment (see, e.g., the review by Loughran and McDonald 2016). Using NBC sentiment measures provides qualitatively similar results

to our deep learning approach. In sum, the results in Table 6 suggest that both measures of sentiment using the deep learning method are informative with respect to future profitability in an intuitive manner, and their relationship with future profitability is not symmetric.

4.4 Does sentiment predict future firm policies?

As numerous prior studies (see, e.g., Bates, Kahle, and Stulz 2009; Acharya, Davydenko, and Strebulaev 2012) find, managers use cash holding as a precautionary measure against risk, which should be reflected in the sentiment in annual reports. Negative sentiment generally reflects poor past performance or increased uncertainty and concern about the future, which implies higher future cash holding. Positive sentiment, on the other hand, generally reflects performance above expectations or a favorable business environment, which suggest lower future cash holding because managers are less concerned about risks. But if firms are financially constrained, growth opportunities and positive sentiment could be positively related to future cash holding (see e.g., Bolton, Chen, and Wang 2011). To investigate this issue, we estimate equation (2) after replacing the dependent variable with $Cash_{t+1}$, defined as cash plus cash equivalents divided by total assets. In Table 7, the estimated coefficients of our sentiment measures are consistently significant across all specifications and have opposite signs, i.e., negative sentiment predicts higher future cash holding, while positive sentiment predicts lower future cash holding. The absolute value of the estimated coefficient of negative sentiment is about three times that of positive sentiment and they are statistically different from each other at the 1% level. This asymmetric result suggests that managers respond in the face of uncertainty and negative outlook by raising cash holdings more than they reduce them when the outlook is favorable. When measured using word lists, both negative and positive sentiments predict higher future cash holdings, which is counterintuitive. This result supports previous studies about the unreliability of positive sentiment measure using word lists and is in line with the results in Tables 3, 4 and 6. The results using NBC sentiment measures are qualitatively similar to our deep learning measures, though the economic significance of NBC positive sentiment is somewhat weaker.

Our results so far show that positive sentiment predicts higher future operating cash flow, higher profitability, but lower cash holding. What is the extra cash generated from operations used for? One possibility is that it is used to pay off debt. To find out if this is the case, we examine the relation between sentiment and future leverage. We use book leverage because market leverage is mechanically related to market capitalization and our sentiment measures. We estimate the regression in equation (2) with $Leverage_{t+1}$ as the dependent variable. Table 8 shows that positive sentiment predicts lower future leverage ratio, suggesting that the extra cash generated from operations is used to reduce leverage. On the other hand, negative sentiment is marginally associated with higher future leverage. The magnitude of the estimated coefficient of the positive sentiment is about 4 to 9 times larger than that of the negative sentiment and they are statistically different at the 1% level. This asymmetric result is consistent with the hypothesis that firms that express high negative sentiment have less flexibility to change their leverage ratio than firms with high positive sentiment. The results using LM sentiment and NBC positive measures are consistent with our deep learning measures, but NBC negative sentiment has no predictive power.

In untabulated results, positive (negative) sentiment predicts higher (lower) valuation, measured by Tobin's Q the next year. We measure Q as (the market value of common stock + book values of preferred stock, long-term debt and debt in current liabilities) divided by the book value of total assets. We also examine whether our sentiment measures predict investment activities in the future. We find that neither negative nor positive sentiment predicts investments (measured by capital expenditures, R&D expenses, or changes in net or gross property, plant and equipment (PP&E), each scaled by total assets at the beginning of the fiscal year) during the next year. There are two potential explanations of this result. First, investment activities are determined by long-term considerations and are not affected by temporary business environments, which are reflected in the sentiment in annual reports. Second, the overall sentiment in annual reports is a noisy measure of investment plans and outlook discussed in 10-Ks. We leave a fuller investigation of this issue to future research.

4.5 Information content of changes in sentiment

Our final set of analyses examines whether the change in sentiment in 10-Ks relative to last year is informative. Cohen, Malloy, and Nguyen (2020) find that firms that change the language in their 10-K filings experience negative future stock returns that reflect changes in firm fundamentals, but investors are inattentive to these changes. Motivated by their findings, we next examine whether changes in the level of sentiment predict abnormal stock returns at the 10-K filing, and future fundamentals and firm policies. Accordingly, we repeat our analyses in prior sections after replacing sentiment levels by their first differences as our main explanatory variables[17]. We start by examining the stock price reaction around the 10-K filing. In different specifications, we exclude observations with an earnings announcement close to the filing date, as in section 4.1, and include year-quarter and industry fixed effects. Table 9 presents the results. Change in positive sentiment predicts positive filing abnormal returns, but change is negative sentiment does not. Changes in LM and NBC sentiment measures do not predict filing abnormal returns.

Table 10 examines the predictive power of sentiment changes on future profitability and cash flow. In Panel A, higher positive (negative) sentiment predicts higher (lower) future profitability. For changes in LM and NBC measures, negative sentiment does not matter, while higher positive sentiment predicts higher future profitability in most specifications. In Panel B, only the change in our positive sentiment matters for cash flow. Higher positive sentiment predicts higher future operating cash flow. LM and NBC sentiment measures are insignificant.

[17] The correlation between changes in positive sentiment and changes in negative sentiment is 0.51. To explore whether the lower power of our results in this section is due to multicollinearity, we include only the change in one sentiment measure. The results are qualitatively very similar, suggesting that multicollinearity is not a big concern here.

Finally, Table 11 shows this analysis on future cash holdings and leverage. In Panel A, changes in both our sentiment measures significantly predict future cash holdings. Higher negative (positive) sentiment predicts higher (lower) cash holdings. Changes in NBC sentiment measures yield similar results. For LM measures, only positive sentiment changes significantly predict (higher) cash holdings. In Panel B, only our positive sentiment measure significantly predicts (lower) future leverage. Coefficients of changes in LM and NBC sentiment measures are insignificant.

In sum, we find that changes in sentiment measures, especially positive sentiment, contain information about future firm fundamentals and that the market reacts to that information. This information also leads to changes in future firm policies.

5. Conclusion

This paper brings state-of-the-art techniques from natural language processing and deep learning to finance for content analysis and sentiment classification. We apply word-embedding to find vector representation of words that preserves semantic and syntactic features of words, and apply deep learning to train a sentiment-classifier. The trained sentiment-classifier achieves an out-of-sample accuracy of 90%. We then examine the information content of positive and negative sentiment measures based on our NN classifier. Unlike prior studies based on word-based classifiers, we find that both negative *and positive* sentiments are informative. Positive (negative) sentiment predicts higher (lower) abnormal return and lower (higher) abnormal trading volume around the 10-K filing date. The market overreacts to negative sentiment and underreacts to positive sentiment during the filing period. All of these effects are larger for negative sentiment than for positive sentiment. Positive sentiment also predicts higher future profitability, higher operating cash flow, lower cash holding, and lower financial leverage. Negative sentiment predicts these variables in the opposite direction. Except for cash holding, the magnitudes of these effects are greater for *positive* sentiment than for negative sentiment. We find generally similar results when we examine the change in sentiment instead of its level. We conclude that (1) the text of corporate annual reports has richer

information content than previously found, (2) positive sentiment is also informative besides negative sentiment, and (3) calculating a net sentiment measure would likely result in loss of information.

The deep learning method used in this paper provides an intuitive, interpretable, and verifiable sentiment measure, and circumvents the need to develop word lists and term-weighting schemes. Moreover, researchers using textual data in non-English languages with no established finance word lists can also use this method. In addition to general sentiment analysis, this method can be applied to content analysis in specific areas. Examples of topics that firms discuss in annual reports are innovation, competition, access to external financing and the risk posed by large customers and suppliers. Researchers can extract information on such topics in a way similar to a classification task. Exploring the economic mechanisms that explain the predictive power of sentiment and investigating managers' strategic disclosure behavior are other promising pathways for future research. Considering the vast amount of textual data (e.g., various corporate disclosures, analyst reports, conference calls, news articles, and social media) and new textual analysis techniques such as the deep learning technique introduced in this paper, this is an exciting research area that holds much promise.

References

Acharya, V., S. A. Davydenko, and I. A. Strebulaev. 2012. Cash holdings and credit risk. *Review of Financial Studies* 25:3572-3609.

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

Bates, T. W., K. M. Kahle, and R. M. Stulz. 2009. Why do US firms hold so much more cash than they used to? *Journal of Finance* 64:1985-2021.

Bellstam, G., S. Bhagat, and J. A. Cookson. Forthcoming. A text-based analysis of corporate innovation. *Management Science*.

Bolton, P., H. Chen, and N. Wang. 2011. A unified theory of Tobin's q, corporate investment, financing, and risk management. *Journal of Finance* 66:1545-1578.

Buehlmaier, M. M. M., and T. M. Whited. 2018. Are financial constraints priced? Evidence from textual analysis. *Review of Financial Studies* 31:2693-2728.

Chollet, F. 2015. Keras: The Python Deep Learning Library. https://keras.io

Cohen, L., C. Malloy, and Q. Nguyen. 2020. Lazy prices. *Journal of Finance* 75:1371-1415.

Coval, J. D., and T. Shumway. 2001. Is sound just noise? *Journal of Finance* 56:1887-1910.

Dyer, T., M. Lang, and L. Stice-Lawrence. 2017. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics* 64:221-245.

Fama, E. F., and K. R. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33:3-56.

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

Frank, M. Z., and A. Sanati. 2018. How does the stock market absorb shocks? *Journal of Financial Economics* 129:136-153.

Gentzkow, M., B. T. Kelly, and M. Taddy. 2019. Text as data. *Journal of Economic Literature* 57:535-74.

Hanley, K. W., and G. Hoberg. 2019. Dynamic interpretation of emerging risks in the financial sector. *Review of Financial Studies* 32:4543-4603.

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

Hoberg, G., and V. Maksimovic. 2014. Redefining financial constraints: A text-based analysis. *Review of Financial Studies* 28:1312-1352.

Hochreiter, S., and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9:1735-1780.

Huang, A. H., R. Lehavy, A. Y. Zang, and R. Zheng. 2017. Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science* 64:2833–2855.

Huang, A. H., A. Y. Zang, and R. Zheng. 2014. Evidence on the information content of text in analyst reports. *Accounting Review* 89:2151-2180.

Jegadeesh, N., and D. Wu. 2013. Word power: A new approach for content analysis. *Journal of Financial Economics* 110:712-729.

Ji, J., O. Talavera, and S. Yin. 2018. The Hidden Information Content: Evidence from the Tone of Independent Director Reports. Working paper, University of Sheffield.

Kearney, C., and S. Liu. 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis* 33:171-185.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521:436–444.

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

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

Li, K., F. Mai, R. Shen, and X. Yan. Forthcoming. Measuring corporate culture using machine learning. *Review of Financial Studies*.

Loughran, T., B. McDonald, and H. Yun. 2009. A wolf in sheep's clothing: The use of ethics-related terms in 10-K reports. *Journal of Business Ethics* 89:39-49.

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

Loughran, T., and B. McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research* 54:1187-1230.

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

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013a. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*: 3111-3119.

Qiu, Y., and T. Y. Wang. 2017. Skilled Labor Risk and Compensation Policies. Working paper, Temple University.

Rehurek, R., and P. Sojka. 2010. Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Ryans, J. Forthcoming. Textual classification of SEC comment letters. *Review of Accounting Studies*.

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

Tetlock, P. C. 2014. Information transmission in finance. *Annual Review of Financial Economics* 6:365–384.

Wang, X., Y. Liu, S. U. N. Chengjie, B. Wang, and X. Wang. 2015. Predicting polarities of tweets by composing word-embeddings with long short-term memory. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* 1:1343-1353.

Table 1

Accuracy of alternative classification methods

Panel A: Train-Set (8,000 Sentences)

		Manually Labeled		
		Negative	*Neutral*	*Positive*
Neural Network Classification	*Negative*	20.3%	2.2%	0.4%
	Neutral	3.5%	64.8%	2.0%
	Positive	0.2%	1.2%	5.4%

Panel B: Test-Set (1,500 Sentences)

		Manually Labeled		
		Negative	*Neutral*	*Positive*
Neural Network Classification	*Negative*	20.2%	2.3%	0.3%
	Neutral	4.0%	63.5%	2.2%
	Positive	0.1%	1.5%	5.9%

Panel C: Classification Using LM word list (9,500 Sentences)

		Manually Labeled		
		Negative	*Neutral*	*Positive*
Classification Based on LM words	*Negative*	17.1%	28.0%	0.9%
	Neutral	4.2%	26.6%	1.6%
	Positive	2.6%	13.6%	5.4%

Panel D: NBC Classification (Average 10-fold out-of-sample)

		Manually Labeled		
		Negative	*Neutral*	*Positive*
Naïve Bayes Classification	*Negative*	19.1%	8.8%	2.0%
	Neutral	4.3%	54.9%	2.1%
	Positive	0.4%	4.6%	3.7%

This table reports the distribution of sentences into three sentiment categories: *negative, positive,* and *neutral.* Panel A (B) shows the train-set (test-set), which consists of 8,000 (1,500) sentences. The sum of the percentages on the main diagonal in each panel measures the accuracy of the NN classification. We use stratified random sampling to select 9,500 sentences to assure that the data is balanced, *i.e.* the neutral category does not dominate the sample. Stratifies are based on Loughran and McDonald's (2011) word lists. 2,000 sentences are completely random; 5,000 sentences include at least one word from LM's negative or positive word lists; 2,000 sentences include at least one word from their list of uncertain words, and 500 sentences include at least on word from their list of constraint words. Panel C shows the classification based on LM word lists. A sentence is positive (negative, neutral) if the number of positive words minus the number of negative words in the sentence is positive (negative, zero). Panel D shows the classification based on NBC classifier. Numbers are the average of 10-fold out-of-sample accuracy. Sentences are randomly partitioned into 10 groups. 10 NBC classifiers are trained each time on 90% of the data. The accuracy is calculated on the 10% out-of-sample data each time.

Table 2
Correlations and summary statistics

Panel A

	Negative	Positive	LM Neg	LM Pos	NBC Neg	NBC Pos
Negative	1					
Positive	0.23	1				
LM Neg	0.56	-0.15	1			
LM Pos	0.27	0.51	0.06	1		
NBC Neg	0.93	0.33	0.42	0.31	1	
NBC Pos	0.15	0.79	-0.25	0.43	0.26	1

Panel B

	Count	Mean	Sd
Negative	62726	0.12	0.06
Positive	62726	0.05	0.03
LM Neg	62726	0.016	0.004
LM Pos	62726	0.006	0.002
NBC Neg	62726	0.18	0.08
NBC Pos	62726	0.08	0.04
Assets ($million)	62726	2983	18206
Market cap ($million)	62683	3304	17407
Leverage	62456	0.22	0.22
Cash	62711	0.23	0.25
ROA	62453	0.03	0.36
R&D	62726	0.08	0.17
Tobin's Q	62382	1.93	2.00
Op. CFlow	62539	0.01	0.30
Tangibility	62650	0.24	0.22
B/M	62643	0.57	0.62
EARet	61134	0.05%	9.5%
Abn. Trading volume	62726	1.42	4.94
CAR(0, +3)	62726	-0.35%	8.3%

Panel A shows Pearson correlations among the sentiment measures. Panel B shows summary statistics of sentiment measures, firm fundamentals, cumulative abnormal returns, and abnormal trading volume. Variables are defined in Appendix B.

Table 3
Filing abnormal return and sentiment

	Dependent variable: CAR(0, +3)								
Independent variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Negative	-0.13***			-0.14***			-0.19***		
	(0.038)			(0.051)			(0.056)		
Positive	0.07**			0.09**			0.09**		
	(0.034)			(0.036)			(0.037)		
LM Neg		-0.09**			-0.08*			-0.15***	
		(0.035)			(0.041)			(0.042)	
LM Pos		0.01			0.01			-0.01	
		(0.034)			(0.036)			(0.034)	
NBC Neg			-0.06			-0.06			-0.08
			(0.037)			(0.051)			(0.056)
NBC Pos			0.01			0.04			0.03
			(0.035)			(0.039)			(0.039)
Obs.	60,536	60,536	60,536	60,103	60,103	60,103	44,514	44,514	44,514
Adj. R-sq.	0.062	0.062	0.062	0.063	0.063	0.062	0.005	0.005	0.005
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
YQ FE				Yes	Yes	Yes	Yes	Yes	Yes
Ind. FE				Yes	Yes	Yes	Yes	Yes	Yes

The table presents estimates of the OLS regressions of *CAR(0, +3)*, the cumulative abnormal return in percentages over days 0 to +3 around the 10-K filing date. Abnormal return is computed using the three Fama and French factors and momentum. The main explanatory variables of interest are *Negative* and *Positive, LM Neg* and *LM Pos*, and *NBC Neg* and *NBC Pos. Negative (Positive)* is the ratio of the number of negative (positive) sentences based on our deep learning approach to the total number of sentences in a 10-K filing. *LM Neg* (*LM Pos*) is the ratio of the number of negative (positive) words based on Loughran and McDonald's (2011) word lists to the total number of words in a filing. Positive words that are preceded within the last three words by {no, not, none, neither, never, nobody} are considered negative. *NBC Neg (NBC Pos)* is the ratio of the number of negative (positive) sentences based on Naïve Bayes classifier to the total number of sentences in a 10-K filing. Columns 7, 8, and 9 exclude filings for which there is an earnings announcement within 2 days before the 10-K filing date. All sentiment measures are normalized to have a mean of 0 and a standard deviation of 1. Control variables are *Total Assets, Tobin's Q, Market cap, Cash, Leverage, ROA,* and *EARet*, as defined in Appendix B. Year_Quarter fixed effect is based on the year and quarter of the filing date. Industry fixed effect is based on Fama and French (1993) 48-industry classification. The coefficients of the constant, control variables, and fixed effects are omitted for brevity. Standard errors are in parentheses and are clustered by firm. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Table 4

Post-filing abnormal return and sentiment

	Dependent variable								
	CAR (+5, +9)			CAR (+5, +14)			CAR (+5, +26)		
Ind. Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Negative	0.11**			0.25***			0.32***		
	(0.051)			(0.073)			(0.107)		
Positive	0.08**			0.18***			0.36***		
	(0.037)			(0.052)			(0.077)		
LM Neg		0.01			0.07			0.08	
		(0.040)			(0.059)			(0.085)	
LM Pos		0.01			0.06			0.10	
		(0.035)			(0.050)			(0.077)	
NBC Neg			0.14***			0.29***			0.31***
			(0.052)			(0.074)			(0.108)
NBC Pos			0.05			0.09*			0.25***
			(0.040)			(0.056)			(0.082)
Obs.	60,031	60,031	60,031	60,031	60,031	60,031	60,033	60,033	60,033
Adj. R-sq.	0.009	0.008	0.009	0.016	0.015	0.016	0.036	0.036	0.036

The table presents estimates of OLS regressions of *CAR(+5 +T)*, the cumulative abnormal return, in percentages over days +5 to +T following the 10-K filing date. Abnormal return is computed using the three Fama and French factors and momentum. The main explanatory variables of interest are *Negative* and *Positive, LM Neg* and *LM Pos*, and *NBC Neg* and *NBC Pos. Negative (Positive)* is the ratio of the number of negative (positive) sentences based on our deep learning approach to the total number of sentences for each filing. *LM Neg* (*LM Pos*) is the ratio of the number of negative (positive) words based on Loughran and McDonald's (2011) word lists to the total number of words. Positive words that are preceded within the last three words by {no, not, none, neither, never, nobody} are considered negative. *NBC Neg (NBC Pos)* is the ratio of the number of negative (positive) sentences based on Naïve Bayes classifier to the total number of sentences in a 10-K filing. All sentiment measures are normalized to have a mean of 0 and a standard deviation of 1. All the columns include control variables and Year_Quarter and Industry fixed effects. Control variables are *Total Assets, Tobin's Q, Market cap, Cash, Leverage, ROA,* and *EARet*, as defined in Appendix B. Year_Quarter fixed effect is based on the year and quarter of filing date. Industry fixed effect is based on Fama and French (1993) 48-industry classification. The coefficients of the constant, control variables, and fixed effects are omitted for brevity. Standard errors are in parentheses and are clustered by firm. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Table 5

Abnormal trading volume and sentiment

Dependent variable: Abnormal Volume

Ind. variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Negative	0.65***			0.16***			0.06**		
	(0.03)			(0.04)			(0.03)		
Positive	-0.18***			-0.14***			-0.06***		
	(0.03)			(0.03)			(0.02)		
LM Neg		0.39***			0.09***			0.02	
		(0.03)			(0.03)			(0.02)	
LM Pos		-0.02			-0.08***			-0.02	
		(0.03)			(0.03)			(0.02)	
NBC Neg			0.67***			0.18***			0.07**
			(0.03)			(0.04)			(0.03)
NBC Pos			-0.33***			-0.15***			-0.05**
			(0.02)			(0.03)			(0.02)
Obs.	62,107	62,107	62,107	61,660	61,660	61,660	44,507	44,507	44,507
Adj. R-sq.	0.015	0.007	0.017	0.043	0.042	0.043	0.010	0.010	0.010
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
YQ FE				Yes	Yes	Yes	Yes	Yes	Yes
Ind. FE				Yes	Yes	Yes	Yes	Yes	Yes

The table presents estimates of OLS regressions of the average abnormal trading volume, *Abnormal Volume (AV)*, in a stock over days t = 0 to +3 around the 10-K filing date. AV equals the mean of AV_t over days t = 0 to +3. $AV_t = (V_t – M) / S$, where V_t 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 the 60-day period that ends five days prior to the filing date. *Negative (Positive)* is the ratio of the number of negative (positive) sentences based on our deep learning approach to the total number of sentences in a 10-K filing. *LM Neg* (*LM Pos*) is the ratio of the number of negative (positive) words based on Loughran and McDonald's (2011) word lists to the total number of words. Positive words that are preceded within the last three words, by {no, not, none, neither, never, nobody} are considered negative. *NBC Neg (NBC Pos)* is the ratio of the number of negative (positive) sentences based on Naïve Bayes classifier to the total number of sentences in a 10-K filing. Columns 7,8, and 9 exclude filings for which there is an earnings announcement within 2 days prior to the 10-K filing date. All sentiment measures are normalized to have a mean of 0 and a standard deviation of 1. The standard deviation of the dependent variable is 4.94. Control variables are *Total Assets, Tobin's Q, Market cap, Cash, Leverage,* and *ROA,* as defined in Appendix B. Year_Quarter fixed effect is based on the year and quarter of the filing date. Industry fixed effect is based on Fama and French (1993) 48-industry classification. The coefficients of the constant, control variables, and fixed effects are omitted for brevity. Standard errors are in parentheses and are clustered by firm. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Table 6

Future profitability and sentiment

Panel A	Dependent variable: ROA_{t+1}								
Ind. Var.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Negative	-0.028***			-0.021***			-0.013***		
	(0.001)			(0.002)			(0.002)		
Positive	0.017***			0.016***			0.012***		
	(0.001)			(0.001)			(0.001)		
LM Neg		-0.017***			-0.010***			-0.007***	
		(0.001)			(0.001)			(0.001)	
LM Pos		-0.016***			-0.015***			-0.007***	
		(0.001)			(0.001)			(0.001)	
NBC Neg			-0.026***			-0.020***			-0.008***
			(0.001)			(0.002)			(0.002)
NBC Pos			0.012***			0.009***			0.005***
			(0.001)			(0.001)			(0.001)
ROA	0.508***	0.514***	0.513***	0.506***	0.509***	0.511***	0.480***	0.482***	0.484***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
B/M	0.026***	0.022***	0.023***	0.026***	0.022***	0.023***	0.014***	0.012***	0.012***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
Market cap	0.018***	0.018***	0.018***	0.019***	0.019***	0.018***	0.018***	0.019***	0.018***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ROA Vol.	-0.153***	-0.168***	-0.156***	-0.150***	-0.163***	-0.154***	-0.129***	-0.135***	-0.133***
	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.027)	(0.027)	(0.027)
Ret. Vol.	-0.189***	-0.193***	-0.205***	-0.235***	-0.250***	-0.248***	-0.217***	-0.224***	-0.230***
	(0.015)	(0.015)	(0.015)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
Obs.	53,830	53,830	53,830	53,830	53,830	53,830	53,488	53,488	53,488
Adj. R-sq.	0.562	0.559	0.560	0.565	0.564	0.563	0.586	0.585	0.584
YQ FE				Yes	Yes	Yes	Yes	Yes	Yes
Ind. FE							Yes	Yes	Yes

Panel B	Dependent variable: Op. $CFlow_{t+1}$								
Ind. Var.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Negative	-0.019***			-0.014***			-0.008***		
	(0.001)			(0.001)			(0.002)		
Positive	0.014***			0.013***			0.010***		
	(0.001)			(0.001)			(0.001)		
LM Neg		-0.013***			-0.009***			-0.007***	
		(0.001)			(0.001)			(0.001)	
LM Pos		-0.014***			-0.014***			-0.006***	
		(0.001)			(0.001)			(0.001)	
NBC Neg			-0.018***			-0.013***			-0.004**
			(0.001)			(0.002)			(0.002)
NBC Pos			0.010***			0.008***			0.005***
			(0.001)			(0.001)			(0.001)
Op. CFlow	0.483***	0.483***	0.486***	0.480***	0.480***	0.484***	0.448***	0.450***	0.451***
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
B/M	0.033***	0.029***	0.031***	0.031***	0.027***	0.029***	0.021***	0.020***	0.020***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Market cap	0.015***	0.016***	0.015***	0.016***	0.017***	0.016***	0.015***	0.016***	0.015***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ROA Vol.	-0.161***	-0.173***	-0.165***	-0.159***	-0.169***	-0.163***	-0.144***	-0.148***	-0.147***
	(0.024)	(0.024)	(0.024)	(0.023)	(0.023)	(0.024)	(0.022)	(0.022)	(0.022)
Ret. Vol.	-0.158***	-0.158***	-0.171***	-0.196***	-0.201***	-0.206***	-0.196***	-0.195***	-0.205***
	(0.013)	(0.013)	(0.013)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Obs.	53,845	53,845	53,845	53,845	53,845	53,845	53,504	53,504	53,504
Adj. R-sq.	0.507	0.506	0.505	0.509	0.509	0.507	0.532	0.532	0.531
YQ FE				Yes	Yes	Yes	Yes	Yes	Yes
Ind. FE							Yes	Yes	Yes

The table presents estimates of OLS regressions of a profitability measure. In panel A, the dependent variable is $ROA_{(t+1)}$, with a standard deviation of 0.36. In panel B, the dependent variable is $Op.\ CFlow_{(t+1)}$, the net operating cash flow from the Cash Flow Statement divided by total assets, with a standard deviation of 0.3. All independent variables have subscript *t*, which denotes the year of the 10-K reporting period. *Negative (Positive)* is the ratio of the number of negative (positive) sentences based on our deep learning approach to the total number of sentences in a filing. *LM Neg* (*LM Pos*) is the ratio of the number of negative (positive) words based on Loughran and McDonald's (2011) word lists to the total number of words in a filing. Positive words that are preceded within the last three words by {no, not, none, neither, never, nobody} are considered negative. *NBC Neg (NBC Pos)* is the ratio of the number of negative (positive) sentences based on Naïve Bayes classifier to the total number of sentences in a 10-K filing. All sentiment measures are normalized to have a mean of 0 and a standard deviation of 1. Control variables are defined in Appendix B. Year_Quarter fixed effect is based on the year and quarter of the 10-K reporting period. Industry fixed effect is based on Fama and French (1993) 48-industry classification. The coefficients of the constant and fixed effects are omitted for brevity. Standard errors are in parentheses and are clustered by firm. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Table 7
Future cash holdings and sentiment

	Dependent variable: $Cash_{t+1}$								
Ind. Var.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Negative	0.010***			0.010***			0.009***		
	(0.001)			(0.001)			(0.001)		
Positive	-0.003***			-0.003***			-0.003***		
	(0.000)			(0.000)			(0.000)		
LM Neg		0.008***			0.006***			0.005***	
		(0.000)			(0.001)			(0.001)	
LM Pos		0.006***			0.006***			0.004***	
		(0.001)			(0.001)			(0.001)	
NBC Neg			0.009***			0.009***			0.008***
			(0.001)			(0.001)			(0.001)
NBC Pos			-0.002***			-0.001***			-0.002***
			(0.000)			(0.000)			(0.001)
Cash	0.840***	0.838***	0.842***	0.837***	0.835***	0.839***	0.812***	0.814***	0.814***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
B/M	-0.005***	-0.004***	-0.004***	-0.008***	-0.007***	-0.007***	-0.005***	-0.004***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ROA	-0.002	-0.001	-0.004	-0.002	-0.001	-0.004	0.001	0.002	-0.000
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Log(Sale)	-0.005***	-0.005***	-0.004***	-0.005***	-0.006***	-0.005***	-0.004***	-0.005***	-0.004***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Sales Growth	-0.016***	-0.016***	-0.016***	-0.016***	-0.016***	-0.016***	-0.015***	-0.015***	-0.015***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ROA. Vol.	0.003	0.006	0.004	0.004	0.007	0.005	0.006	0.008	0.006
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Ret. Vol.	0.041***	0.039***	0.046***	0.024***	0.029***	0.031***	0.024***	0.027***	0.029***
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Obs.	52,948	52,948	52,948	52,948	52,948	52,948	52,662	52,662	52,662
Adj. R-sq.	0.815	0.815	0.815	0.817	0.817	0.817	0.820	0.819	0.819
YQ FE				Yes	Yes	Yes	Yes	Yes	Yes
Ind. FE							Yes	Yes	Yes

The table presents estimates of OLS regressions of $Cash_{t+1}$, which equals (cash plus cash equivalents) divided by *Total Assets.* All independent variables have subscript *t*, which denotes the year of the 10-K reporting period. *Negative (Positive), LM Neg* (*LM Pos*), and *NBC Neg (NBC Pos)* are sentiment measures and defined in Table 6. All sentiment measures are normalized to have a mean of 0 and a standard deviation of 1. The standard deviation of the dependent variable is 0.25. Control variables are defined in Appendix B. Year_Quarter fixed effect and industry fixed effect are defined in Table 6. The coefficients of the constant and fixed effects are omitted for brevity. Standard errors are in parentheses and are clustered by firm. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively

Table 8
Future leverage and sentiment

	Dependent variable: Leverage$_{t+1}$								
Ind. Var.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Negative	0.003			0.004*			0.005**		
	(0.002)			(0.002)			(0.002)		
Positive	-0.028***			-0.027***			-0.020***		
	(0.002)			(0.002)			(0.002)		
LM Neg		0.007***			0.009***			0.010***	
		(0.002)			(0.002)			(0.002)	
LM Pos		-0.015***			-0.015***			-0.015***	
		(0.002)			(0.002)			(0.002)	
NBC Neg			-0.003			0.000			0.000
			(0.002)			(0.002)			(0.002)
NBC Pos			-0.027***			-0.029***			-0.022***
			(0.002)			(0.002)			(0.002)
Tobin's Q	0.003***	0.002**	0.002***	0.002*	0.001	0.001	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Cash	-0.279***	-0.268***	-0.267***	-0.280***	-0.266***	-0.267***	-0.288***	-0.280***	-0.277***
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
ROA	-0.090***	-0.096***	-0.090***	-0.087***	-0.094***	-0.087***	-0.082***	-0.085***	-0.082***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
R&D	-0.029***	-0.017	-0.016	-0.036***	-0.025**	-0.024**	-0.044***	-0.036***	-0.035***
	(0.011)	(0.012)	(0.011)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Total Assets	0.019***	0.019***	0.019***	0.019***	0.019***	0.019***	0.016***	0.017***	0.016***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Tangibility	0.147***	0.167***	0.155***	0.140***	0.159***	0.146***	0.146***	0.150***	0.148***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.014)	(0.014)	(0.014)
Obs.	59,146	59,146	59,146	59,146	59,146	59,146	58,770	58,770	58,770
Adj. R-sq.	0.217	0.208	0.218	0.229	0.221	0.231	0.270	0.269	0.272
YQ FE				Yes	Yes	Yes	Yes	Yes	Yes
Ind. FE							Yes	Yes	Yes

The table presents estimates of OLS regressions of *Leverage$_{t+1}$*, defined as (long term debt plus debt in current liabilities) divided by *Total Assets.* All independent variables have subscript *t*, which denotes the year of the 10-K reporting period. *Negative (Positive), LM Neg* (*LM Pos*), and *NBC Neg (NBC Pos)* are sentiment measures and defined in Table 6. All sentiment measures are normalized to have a mean of 0 and a standard deviation of 1. The standard deviation of the dependent variable is 0.22. Control variables are defined in Appendix B. Year_Quarter fixed effect is based on year and quarter of the 10-K reporting period. Industry fixed effect is based on Fama and French (1993) 48-industry classification. The coefficients of the constant and fixed effects are omitted for brevity. Standard errors are in parentheses and are clustered by firm. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Table 9

Change in sentiment and filing abnormal return

	Dependent variable: CAR(0, +3)								
Ind. Var.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Δ Negative	-0.03			0.01			-0.04		
	(0.044)			(0.047)			(0.045)		
Δ Positive	0.07*			0.07*			0.08*		
	(0.042)			(0.041)			(0.042)		
Δ LM Neg		-0.01			-0.04			-0.01	
		(0.038)			(0.037)			(0.038)	
Δ LM Pos		0.03			0.04			0.03	
		(0.033)			(0.032)			(0.033)	
Δ NBC Neg			-0.02			0.06			-0.02
			(0.049)			(0.053)			(0.050)
Δ NBC Pos			0.05			0.03			0.05
			(0.049)			(0.049)			(0.050)
Observations	52,306	52,306	52,306	38,361	38,361	38,361	51,955	51,955	51,955
Adj. R-sq.	0.064	0.064	0.064	0.003	0.003	0.003	0.065	0.065	0.065
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
YQ FE							Yes	Yes	Yes
Ind. FE							Yes	Yes	Yes

The table presents estimates of the OLS regressions of *CAR(0, +3)*, the cumulative abnormal return in percentages over days 0 to +3 around the 10-K filing date. Abnormal return is computed using the three Fama and French factors and momentum. The main explanatory variables of interest are first difference (Δ) in *Negative* and *Positive, LM Neg* and *LM Pos*, and *NBC Neg* and *NBC Pos. Negative (Positive)* is the ratio of the number of negative (positive) sentences based on our deep learning approach to the total number of sentences in a 10-K filing. *LM Neg* (*LM Pos*) is the ratio of the number of negative (positive) words based on Loughran and McDonald's (2011) word lists to the total number of words in a filing. Positive words that are preceded within the last three words by {no, not, none, neither, never, nobody} are considered negative. *NBC Neg (NBC Pos)* is the ratio of the number of negative (positive) sentences based on Naïve Bayes classifier to the total number of sentences in a 10-K filing. Columns 4,5, and 6 exclude filings for which there is an earnings announcement within 2 days prior to the 10-K filing date. All independent variables are normalized to have a mean of 0 and a standard deviation of 1. Control variables are *Total Assets, Tobin's Q, Market cap, Cash, Leverage, ROA,* and *EARet*, as defined in Appendix B. Year_Quarter fixed effect is based on the year and quarter of the filing date. Industry fixed effect is based on Fama and French (1993) 48-industry classification. The coefficients of the constant, control variables, and fixed effects are omitted for brevity. Standard errors are in parentheses and are clustered by firm. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Table 10

Change in sentiment and future profitability

Panel A	Dependent variable: ROA_{t+1}					
Ind. Var.	(1)	(2)	(3)	(4)	(5)	(6)
Δ Negative	-0.003**			-0.003***		
	(0.001)			(0.001)		
Δ Positive	0.004***			0.004***		
	(0.001)			(0.001)		
Δ LM Neg		-0.001			-0.001	
		(0.001)			(0.001)	
Δ LM Pos		0.002**			0.002**	
		(0.001)			(0.001)	
Δ NBC Neg			-0.001			-0.002
			(0.001)			(0.001)
Δ NBC Pos			0.002			0.002*
			(0.001)			(0.001)
Obs.	46,078	46,078	46,078	45,792	45,792	45,792
Adj. R-sq.	0.627	0.627	0.627	0.640	0.640	0.640
YQ and Ind. FE				Yes	Yes	Yes

Panel B	Dependent variable: Op. $CFlow_{t+1}$					
Ind. Var.	(1)	(2)	(3)	(4)	(5)	(6)
Δ Negative	0.001			-0.000		
	(0.001)			(0.001)		
Δ Positive	0.002**			0.002**		
	(0.001)			(0.001)		
Δ LM Neg		-0.001			-0.001	
		(0.001)			(0.001)	
Δ LM Pos		0.001			0.001	
		(0.001)			(0.001)	
Δ NBC Neg			0.001			0.000
			(0.001)			(0.001)
Δ NBC Pos			0.001			0.001
			(0.001)			(0.001)
Obs.	46,090	46,090	46,090	45,804	45,804	45,804
Adj. R-sq.	0.559	0.559	0.559	0.573	0.573	0.573
YQ and Ind. FE				Yes	Yes	Yes

The table presents estimates of OLS regressions of a profitability measure. In panel A, the dependent variable is $ROA_{(t+1)}$. In panel B, the dependent variable is $Op.\ CFlow_{(t+1)}$. All columns include control variables similar to Table 3. Independent variables are first difference (Δ) of sentiment measures and are normalized to have mean of 0 and a standard deviation of 1. Fixed effects are defined similar to Table 3. The coefficients of the constant, controls, and fixed effects are omitted for brevity. Standard errors are in parentheses and are clustered by firm. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Table 11

Change in sentiment, future cash holdings, and future leverage

Panel A	Dependent variable: $Cash_{t+1}$					
Ind. Var.	(1)	(2)	(3)	(4)	(5)	(6)
Δ Negative	0.002**			0.002**		
	(0.001)			(0.001)		
Δ Positive	-0.002***			-0.002***		
	(0.001)			(0.001)		
Δ LM Neg		0.001**			0.001*	
		(0.000)			(0.000)	
Δ LM Pos		-0.001			-0.001	
		(0.000)			(0.000)	
Δ NBC Neg			0.002***			0.002***
			(0.001)			(0.001)
Δ NBC Pos			-0.003***			-0.002***
			(0.001)			(0.001)
Obs.	45,393	45,393	45,393	45,134	45,134	45,134
Adj. R-sq.	0.819	0.819	0.819	0.823	0.823	0.823
YQ and Ind. FE				Yes	Yes	Yes

Panel B	Dependent variable: $Leverage_{t+1}$					
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
Negative	0.001			0.001		
	(0.001)			(0.001)		
Positive	-0.002***			-0.002***		
	(0.001)			(0.001)		
LM Neg		-0.000			-0.000	
		(0.001)			(0.001)	
LM Pos		-0.000			-0.001	
		(0.001)			(0.001)	
NBC Neg			-0.001			-0.001
			(0.001)			(0.001)
NBC Pos			0.000			-0.000
			(0.001)			(0.001)
Obs.	49,228	49,228	49,228	48,924	48,924	48,924
Adj. R-sq.	0.208	0.208	0.208	0.268	0.268	0.268
YQ and Ind. FE				Yes	Yes	Yes

The table presents estimates of OLS regressions of $Cash_{t+1}$ (panel A) and $Leverage_{t+1}$ (Panel B). All columns include control variables similar to Tables 7 and 8. Independent variables are first difference (Δ) of sentiment measures and are normalized to have mean of 0 and a standard deviation of 1. Fixed effects are defined similar to Tables 7 and 8. The coefficients of the constant, controls, and fixed effects are omitted for brevity. Standard errors are in parentheses and are clustered by firm. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Appendix A

Sentiment Classification using Deep Learning

A.1. Neural networks

This appendix provides a brief introduction to neural networks and the method we use for sentiment classification. The left side of Figure A1 shows the basic building block of neural networks. Each input, x_i, is a real number that is multiplied by a weight, w_i, shown as a line connecting x_i to node n. The sum of the products of x_i and w_i, z_i, is the input to node n. The node applies a function to the input and provides a real number as the output. A logistic regression model can be represented using this structure with features as $x_1, x_2, \ldots, x_n$, coefficients as $w_1, w_2, \ldots, w_n$, and y as the output of the node with function $y = 1 / (1 + e^{-z})$. Nodes can be stacked up to build a *layer* as shown on the right side of Figure A1. The output of each node in a layer can be the input to the next layer which can be the output layer. The function that operates on the input to a node and generates the output of that node is called the activation function. Activation functions are determined before training the NN. Training neural networks refers to computing all the weights, w_i, in all the layers in order to minimize a pre-defined cost (or loss) function that depends on the outputs and the weights in the NN. All the layers between the input and the output layer are called hidden layers. Deep neural networks are NN that are built using many hidden layers. NN can perform complicated tasks due to their ability to capture complex nonlinearities.

Recurrent NN (RNN) have a different structure and data flow than the feed-forward NN described above, but they have the same building blocks. Figure A2 shows a diagram of a simple RNN. x_t is the input (which can be a vector) at time t to a NN presented as a rectangle. This NN creates an output, y_t, and a state variable, s_{t+1}, that is used together with x_{t+1} in the next time step. The NN in each time step is the same, *i.e.* it has the same structure with the same set of weights to be calculated during training. For the sentiment classification task in this paper, x_t represents a word in a sentence and y_T (where T is the length of the sentence) represents a three-dimension output that shows the probability that the sentence belongs to each

sentiment category. In the next section, we discuss word-embedding to find a vector representation of words, x_t, to be used in the RNN-based sentiment classifier.

A.2. Word-embedding

Words can be represented numerically by vectors with the dimension equal to the number of words in a dictionary - the collection of all different words in the corpus under study. All elements of such a vector are zero except one which equals to 1 and corresponds to a specific word - this vector is called a one-hot vector. In this representation, only the exact same words in a text would have the same vector. While preserving the true dimensionality of words, this method has several drawbacks in practice. It does not capture any similarity between words. 'Loan' and 'Debt' are as similar or different as 'Finance' and 'Zoology'. In addition, any analysis using this word representation method requires the algorithm to have seen all the significant words in the dictionary enough times during training. Word-embedding is an NLP technique that can mitigate both concerns by finding a low-dimension (20 to 500) vector representation of words.

There are many word-embedding techniques all of which result in a low-dimension representation of words. With word-embedding, each word is represented by a continuous vector of an arbitrary dimension (200 in this paper). Mikolov et al. (2013a) propose two novel structures using neural networks to estimate word-embedding at a low computational cost with high accuracy. In another study, Mikolov et al. (2013b), further suggest some modifications to improve the quality and efficiency of word-embedding that can be performed on very large data sets. Figure A3 shows an example of a simple structure proposed by Mikolov et al. (2013a). Input is the one-hot vector of a word right before the *current* word in a sentence. The matrix w_{dxN} (where N is the number of words in the dictionary and d is the word-embedding dimension) represents all the weights that connect the input vector to the hidden layer, which is the word-embedding matrix that we use once the NN is trained. The hidden layer is connected to the output layer which is a Softmax classifier. Each output shows the probability that the corresponding word in the dictionary is the *current* word. The output with the highest probability is the predicted *current* word. The model is trained to

maximize the probability of predicting the *current* word correctly given the input word. We use a structure proposed by Mikolov et al. (2013a), called continuous bag-of-words (CBOW).

In a CBOW structure, given a set of neighboring words in a sentence, the probability of occurrence of the *current* word is maximized. Since the order of neighboring words does not affect the results, CBOW is a bag-of-words method. The model takes as input the average of one-hot vectors of neighboring words, instead of a single one-hot vector shown in Figure A3. The word-embedding matrix and parameters of the Softmax classifier are estimated to maximize the likelihood of predicting the *current* word correctly. Each column of the word-embedding matrix represents a word in the dictionary. Results of word-embedding should not be evaluated on a standalone basis, rather based on a downstream task for which it is being used. The downstream task in our study is sentiment classification discussed in the next section. Nevertheless, for illustration, we show five most similar words to 12 different financial words based on the results of our word-embedding in Table A1. *Score* is calculated based on the cosine similarity of the vectors corresponding to each pair of words. In general, word-embedding is known to preserve semantic and syntactic aspects of words. In a recent finance study, Li et al. (forthcoming) use word-embedding to find a lexicon of words related to corporate culture.

A.3. Sentiment classifier

Next, we can represent each sentence as a sequence of vectors of the dimension chosen for word-embedding. We can then use NN and train a model to take a sentence as input and classify the sentiment in each sentence into negative, positive, and neutral. To do that, we need to have a train-set that includes manually labelled sentences and choose a NN structure and train it. We manually classify 9,500 randomly[18]

[18] We use stratified random sampling to select 9,500 sentences to assure that the data is not unbalanced, *i.e.* the occurrence of positive and negative sentences is not rare. Stratas are based on LM's (2011) word lists and include 2,000 sentences chosen completely at random; 5,000 sentences that include at least one word from LM negative or positive word lists; 2,000 sentences that include at least one word from LM uncertain

selected sentences into three categories: negative, positive, and neutral. Recurrent neural network is a structure that captures the dynamics of sequential data. A specific type of RNN, long short-term memory (LSTM), proposed by Hochreiter and Schmidhuber (1997), avoids the problems of vanishing and exploding gradients when training the model. LSTM network can also learn from observations far back in the sequence, implying that it can 'memorize' words in long sentences that occurred near the beginning. We train an LSTM network (with a Softmax output layer) on the train-set of 8,000 sentences[19], known as the in-sample data set in the forecasting literature. The other 1,500[20] sentences are then used to evaluate the out-of-sample performance of the trained model. As shown in Table 1, the accuracy of this model for in-sample and out-of-sample sentiment classification is about 91% and 90%, respectively[21].

The choice of the type of NN and the hyper-parameters[22] of the model are discretionary and researchers can evaluate the performance of different models. While the level of accuracy we achieve can potentially be improved, it is quite high in the sentiment analysis literature and significantly higher than the accuracy of the word list and NBC methods used in finance. Regarding implementation, researchers have several choices to train a NN. Tensorflow by Google, which is now open source, has a strong active

words; and 500 sentences that include at least one word from LM constraint words. The accuracy of the classifier across the stratas is very similar.

[19] More precisely, we use 8,000 sentences as our train and development set to fine tune the classifier and to ensure that the classifier is not over-fitting the train-set.

[20] For the purpose of evaluation, the appropriate size of the out-of-sample set is 10% to 20% of the size of in-sample train-set.

[21] Note that in Table 2, the percentage of positive sentences is relatively small. This is due to the nature of the textual data we use, *i.e.* 10-K filings.

[22] Some examples of hyper-parameters are the number of hidden layers, the number of nodes in each layer, the dimension of word-embedding, the method of training and its parameters.

community and many sample codes for machine learning tasks are available on GitHub and many weblogs. Theano is another popular choice. This paper uses Keras[23], also an open source library, which requires less coding than many other choices. It is modular and user-friendly and is tailored to standard machine learning tasks that researchers in other disciplines may also find helpful.

Finally, we address some questions that researchers might encounter when using our approach. First, note that the look ahead bias in performing word-embedding and training the classifier doesn't apply in our setting for at least three reasons. One, word-embedding only learns the semantic and syntactic features of words. Unless the meaning of a word changes over time,[24] using text data from different time periods should provide similar results on tone or sentiment, as long as the corpus is large enough. Two, since we perform word-embedding independently of sentiment classification, the tonal information of words doesn't affect the outcome of word-embedding. Three, regarding manual classification, when we label each sentence, we have no knowledge of the firm that the sentence belongs to, the date of the disclosure, the market response to the filing, and most importantly, how our classification affects the ultimate classifier that we use on 200 million sentences and its effect on our empirical results. This last point makes it almost impossible for researchers to see the outcome of the empirical results when performing the manual classification. When classifying sentences, we don't find situations where we need more information about the context or past events.

Second, there are several points to note regarding our choice of the training sample and its size. One, the training sample should come from the same text data, here 10-Ks. However, if labelled sentences from other sources are available, one can use them together with 10-K sentences to train a classifier. The

[23] We use Python in all steps, *i.e.* pre-processing 10-K filings, performing word-embedding, and training the sentiment classifier. All the packages mentioned in the paper can be imported and used in Python.

[24] E.g., "apple" and "amazon" had quite different meanings prior to 2000s, referring to a fruit and a forest back then and now to two tech-giants.

potential benefit is the need for a smaller set of labelled sentences from 10-Ks to achieve desired accuracy, hence reducing the manual work. Alternatively, one can train a classifier using sentences from other sources and then use 10-K labelled sentences to improve the classifier. Presumably, the more similar the other source is to 10-Ks, the higher the potential benefit of using it. Two, regarding the sample size, generally when improving the accuracy of a classifier is not possible by changing the structure of the classifier or fine-tuning hyperparameters of the model, the last resort is to increase the sample size. Using 1,000 and 3,000 sentences in our training set, we find accuracy of 79% and 85%, respectively. We choose a sample size of 8,000 to improve the accuracy of our classifier to 91%.[25]

Finally, how about measuring sentiment by performing classification on paragraphs rather than sentences? This method has several drawbacks. First, paragraphs can be nuanced, containing both positive and negative sentences, so classifying a paragraph into one category can be misleading. Second, manually labelling paragraphs requires significantly more work. Third, the size of the training sample probably needs to be larger, requiring even more manual work, since a paragraph likely has more information than a sentence. Classifying at the document level shares these problems. There are also technical issues as parsing 10-Ks into paragraphs is more prone to error than parsing into sentences. With the current NLP technology, performing sentiment analysis on sentences seems to be a better choice.

[25] We expect similar results as long as the set of 8,000 sentences in the training sample is similarly chosen randomly from 10-Ks. However, it is possible to use manually labelled sentences from other sources to augment 10-K sentences. The potential benefit would be to reduce the manual work and use already labelled sentences by other researchers in other contexts. But the out-of-sample performance of the classifier should be evaluated using only 10-K sentences.

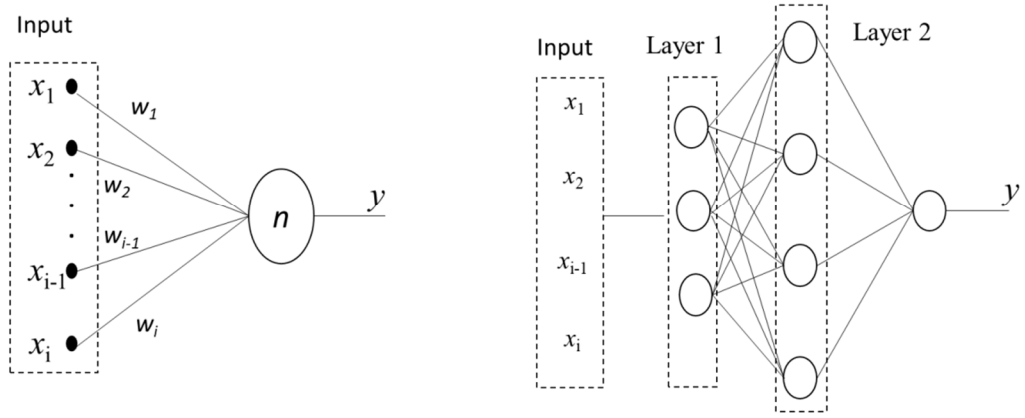

Figure A1

The figure on the left shows the building block of neural networks (NN). The inputs are x_1, x_2, …, x_i, which are real numbers. Solid lines represent weights, and y is the output of node n which is a function of $\sum x_i \cdot w_i$. The figure on the right shows a simple NN with 2 hidden layers. All inputs are connected to all nodes in layer 1; y is the output of the NN.

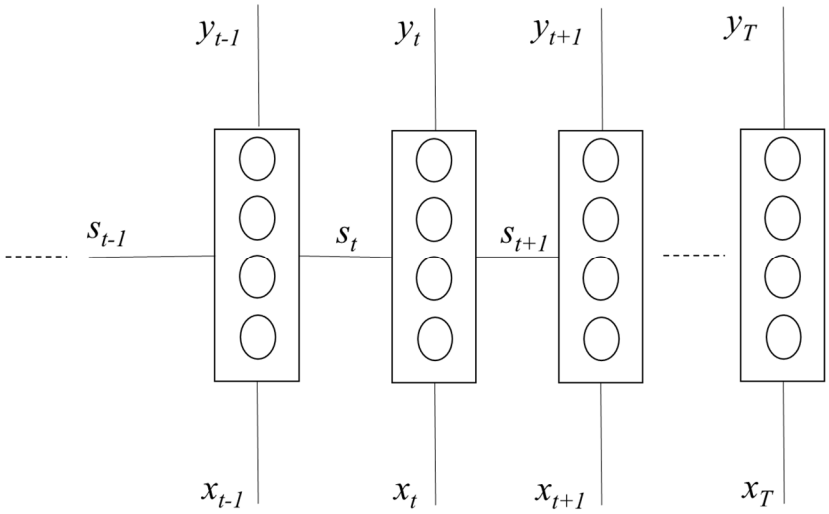

Figure A2

This figure shows the structure and data flow of a simple recurrent neural network (RNN). The input is x_t which has a time stamp, and the output is y_t. The building blocks are the same at all time steps. The state variable s_t carries forward the information from time *t-1* to time *t*.

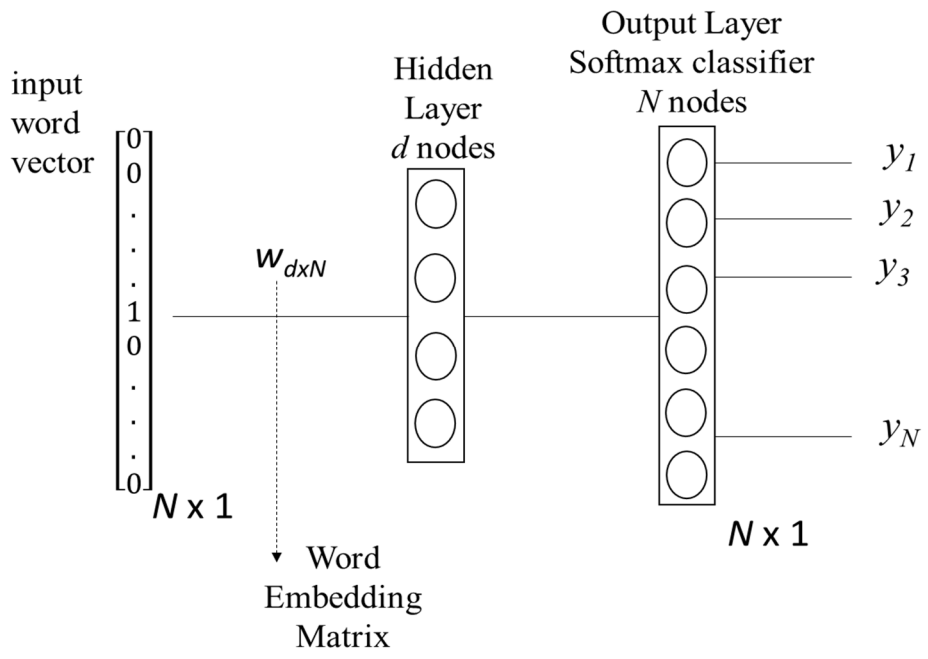

Figure A3

A simple structure to perform word-embedding using neural networks (NN) proposed by Mikolov et al. (2013a). The input is the one-hot vector associated with a neighboring word to the current word. Each output represents the probability that the NN assigns to that word being the target word based on the input word. The word-embedding matrix is associated with the weights that connect the input vector to the hidden layer, *d* is the dimension of word-embedding, and *N* is the number of words in the dictionary.

Table A1

Illustration of word similarity using the word-embedding output.

Word	*penalties*	Score	*competition*	Score	*operations*	Score
Most Similar	*fines*	0.72	*intense*	0.80	*results*	0.70
	penalty	0.68	*competitive*	0.75	*operating*	0.64
	criminal	0.64	*compete*	0.73	*business*	0.58
	civil	0.61	*competing*	0.72	*condition*	0.58
	underpayment	0.55	*competitors*	0.66	*profitability*	0.57

Word	*skilled*	Score	*profit*	Score	*mercedes*	Score
Most Similar	*talented*	0.68	*margins*	0.70	*volvo*	0.70
	nurses	0.67	*gross*	0.70	*chevrolet*	0.69
	personnel	0.66	*margin*	0.63	*toyota*	0.68
	trained	0.66	*profits*	0.63	*mazda*	0.67
	professionals	0.65	*revenues*	0.62	*lexus*	0.67

Word	*risk*	Score	*loss*	Score	*loan*	Score
Most Similar	*risks*	0.74	*losses*	0.72	*loans*	0.81
	exposure	0.64	*gain*	0.62	*mortgage*	0.71
	exposed	0.63	*net*	0.57	*credit*	0.68
	exposures	0.63	*income*	0.57	*lender*	0.61
	sensitivity	0.58	*earnings*	0.56	*lending*	0.60

Word	*innovation*	Score	*patent*	Score	*research*	Score
Most Similar	*innovative*	0.72	*patents*	0.91	*development*	0.76
	excellence	0.70	*uspto*	0.76	*collaborative*	0.60
	innovations	0.66	*trademark*	0.74	*commercialization*	0.60
	innovate	0.61	*intellectual*	0.74	*crada*	0.59
	creativity	0.61	*infringement*	0.67	*preclinical*	0.59

The table shows the five most similar words to 12 selected words based on the results of word-embedding. Score is cosine similarity. Each word is associated with a vector of dimension 200 calculated in the word-embedding stage. Score is calculated using the cosine similarity function. (If v_1 and v_2 are two word vectors, cosine similarity is calculated as $(v_1 \,.\, v_2) \,/\, \|v_1\| \,.\, \|v_2\|$, where the numerator is the inner product of the two vectors and $\|\,.\,\|$ represents geometric magnitude.)

Table A2

Examples of sentences and their sentiment.

Positive Words	Negative Sentence
achieve, greater, gain	For these and other reasons, these competitors may achieve greater acceptance in the marketplace than our company, limiting our ability to gain market share and customer loyalty and increase our revenues.
greater, better, able	Furthermore, competitors who have greater financial resources may be better able to provide a broader range of financing alternatives to their customers in connection with sales of their products.
enjoy, advantages, greater	Many of these potential competitors are likely to enjoy substantial competitive advantages, including greater resources that can be devoted to the development, promotion and sale of their products.
successful, alliances, able	There can be no assurance that we will be successful in our ongoing strategic alliances or that we will be able to find further suitable business relationships as we develop new products and strategies.
successful, able, achieve, profitability	There can be no assurance that any of the Company's business strategies will be successful or that the Company will be able to achieve profitability on a quarterly or annual basis.
able, opportunities, opportunities, favorable	We cannot assure you that we will be able to identify suitable acquisition or joint venture opportunities in the future or that any such opportunities, if identified, will be consummated on favorable terms, if at all.
successfully, enhance, advantage, opportunities	If additional financing is not available when required or is not available on acceptable terms, we may be unable to fund our expansion, successfully promote our brand name, develop or enhance our products and services, take advantage of business opportunities, or respond to competitive pressures, any of which could have a material adverse effect on our business.
collaborative, achieve, profitability	Our long-term liquidity also depends upon our ability to attract and maintain collaborative relationships, to increase revenues from the sale of our products, to develop and market new products and ultimately, to achieve profitability.
able, success, able, achieve	Even if we are able to develop new products, the success of each new product depends on several factors including whether we selected the proper product and our ability to introduce it at the right time, whether the product is able to achieve acceptable production yields and whether the market accepts the new product.
efficiencies, benefit, achieved	Although Stratos expects that the elimination of duplicative costs, as well as the realization of other efficiencies related to the integration of the businesses, may offset incremental transaction, merger-related and restructuring costs over time, we cannot give any assurance that this net benefit will be achieved in the near term, or at all.

Table A2 (cont.)

Positive words	Neutral Sentence
gain, greater, gain	If a business combination results in a bargain purchase for us, the economic gain resulting from the fair value received being greater than the purchase price is recorded as a gain included in other income (expense), net, in the Consolidated Statements of Comprehensive Loss.
improvements, improvements, improvements	The estimated lives used in determining depreciation and amortization are: Buildings and improvements 12-40 years, Warehouse and office equipment 5-7 years, and Automobiles 3-5 years. Leasehold improvements are amortized over the lives of the respective leases or the service lives of the improvements, whichever is shorter.
superior, opportunity, superior	If the Company receives a Superior Proposal, Parent must be given the opportunity to match the Superior Proposal.
enables, exceptional, strength	Specialty steels are made with a high alloy content, which enables their use in environments that demand exceptional hardness, toughness, strength and resistance to heat, corrosion or abrasion, or combinations thereof.
greater, greater, advances	Majority Lenders means Lenders having greater than 50% of the total Commitments or, if the Commitments have been terminated in full, Lenders holding greater than 50% of the then aggregate unpaid principal amount of the Advances.

Negative Words	Positive Sentence
disputes, difficulty	We believe that we maintain a satisfactory working relationship with our employees, and we have not experienced any significant labor disputes or any difficulty in recruiting staff for our operations.
serious, adverse, unexpected, irreversible	No serious adverse events and no unexpected or irreversible side effects were reported in the Ceplene study.
Problems	We also maintain a separate technical support group dedicated to answering specific customer inquiries and assisting customers with the operation of products and finding low cost solutions to manufacturing problems.
Bad	In 2003, we reduced bad debt expense by $0.4 million versus 2002.
Unable	We believe the effect of this law will be to accelerate sales of our needleless systems, although we are unable to estimate the amount or timing of such sales.
claims, against	These agreements released all legal claims against us.
dismissing, claims, against	On November 28, 2012, the Federal Court in the MDL entered an order dismissing all claims against Nalco.
against, damage	Lower Lakes maintains insurance on its fleet for risks commonly insured against by vessel owners and operators, including hull and machinery insurance, war risks insurance and protection and indemnity insurance (which includes environmental damage and pollution insurance).
Susceptible	Management believes that the Company's container manufacturing capabilities makes the Company less susceptible than its competitors to ocean-going container price fluctuations, particularly since the cost of used containers is affected by many factors, only one of which is the cost of steel from which the Company can manufacture new containers.
damage, loss, interruption	We also maintain coverage for property damage or loss, general liability, business interruption, travel-accident, directors and officers liability and workers compensation.

Table A2 (cont.)

Negative Words	Neutral Sentences
loss, impairment, loss, loss	We consider the likelihood of loss or impairment of an asset or the incurrence of a liability, as well as our ability to reasonably estimate the amount of loss in determining loss contingencies.
critical, critical, doubtful, restructuring	Our critical accounting policies are as follows: revenue recognition; allowance for doubtful accounts; accounting for income taxes; and restructuring charge.
impairment, impairment, impairment, loss	If it is more likely than not that a goodwill impairment exists, the second step of the goodwill impairment test must be performed to measure the amount of the goodwill impairment loss, if any.
impairment, loss, impairment, impairment	Unproved oil and gas properties that are individually significant are periodically assessed for impairment of value, and a loss is recognized at the time of impairment by providing an impairment allowance.
disclose, loss, litigation, claims	We account for and disclose loss contingencies such as pending litigation and actual or possible claims and assessments in accordance with the FASB s authoritative guidance on accounting for contingencies.

This table presents several sentences classified under our approach as negative (positive) or neutral, and the positive (negative) words in them based on the Loughran and McDonald (2011) word lists.

Appendix B: Variable Definitions

Negative	Ratio of the number of negative sentences based on our deep learning approach to the total number of sentences in a 10-K filing
Positive	Ratio of the number of positive sentences based on our deep learning approach to the total number of sentences in a 10-K filing
LM Neg	Ratio of the number of negative words based on Loughran and McDonald's (2011) negative word list to the total number of words in a 10-K filing. Positive words preceded within the last three words by {no, not, none, neither, never, nobody} are considered negative
LM Pos	Ratio of the number of positive words based on Loughran and McDonald's (2011) positive word list to the total number of words in a 10-K filing. Positive words preceded within the last three words by {no, not, none, neither, never, nobody} are considered negative
NBC Neg	Ratio of the number of negative sentences based on Naïve Bayes classifier to the total number of sentences in a 10-K filing
NBC Pos	Ratio of the number of positive sentences based on Naïve Bayes classifier to the total number of sentences in a 10-K filing
Abnormal Volume	The average trading volume over the 4-day event window [0, +3], where volume is standardized based on its mean and standard deviation over days [-65, -6] before the 10-K filing date
B/M	Book value of common equity divided by market value of common equity
CAR(0, +3)	Cumulative abnormal return over days [0, +3] using the three Fama and French factors and momentum
Cash	Cash and cash equivalents divided by total assets, *che / at*
EARet	Cumulative abnormal return over days [-1, +1] surrounding earnings announcement date
Leverage	Leverage ratio, measured as (long-term debt plus debt in current liabilities) divided by total assets, *(ldtt + dlc) / at*
Log(Sale)	Natural log of total sales, ln(*sale)*
Market cap	Natural log of market value of common shares, ln(*prcc_f * csho)*
Op. CFlow	Cash flow from operating activities divided by lagged total assets, $oancf_t / at_{(t-1)}$
ROA	Operating income before depreciation divided by lagged total assets, $oibdp_t / at_{(t-1)}$
ROA Vol.	Standard deviation of ROA over the last 5 years

Ret. Vol.	Standard deviation of monthly returns over the last 12 months
R&D	Research and development expenses divided by lagged total assets, $xrd_t / at_{(t-1)}$
Sales Growth	Sales growth over the last year ($Sale_t$ - $Sale_{t-1}$) / $Sale_{t-1}$
Tangibility	Property, plant, and equipment divided by total assets *ppent*/*a*t
Tobin's Q	*((prcc_f * csho)+pstk+dltt+dlc) / at*
Total Assets	Natural log of total assets, ln(*at)*