Contextualized News in Corporate Disclosures: A Neural Language Approach

Abstract

I quantify and explain value-relevant news in textual disclosures using word context. I improve upon current methods by applying a new textual analysis approach—a *BERT-based* neural language model—to characterize disclosures as sequentially connected and interacting elements (rather than stand-alone words). I denote this enhanced measurement as contextualized, and I apply it to predicting the magnitude and direction of disclosure news. The contextualized text in earnings announcements (1) explains three times more variation in short-window stock returns than text measured using traditional narrative attributes or recent machine learning techniques, and (2) offers large incremental explanatory power relative to reported earnings modeled using traditional or machine learning methods. Contextualized disclosures also strongly predict future earnings, with most news arising from (a) word order (i.e., context), (b) text describing numbers, and (c) text at the beginning of disclosures. This study highlights the importance of contextualized disclosures for researchers, regulators, and practitioners.

Keywords: disclosure, earnings announcements, textual analysis, neural language models

Data Availability: SEC filings, and the Compustat, CRSP, and I/B/E/S/ databases

JEL Classifications: G14, G32, M21, M41

1. Introduction

This study applies a recent innovation in textual analysis, namely a *BERT-based* neural language model (i.e., Bidirectional Encoder Representations from Transformers), to better quantify and understand news in corporate disclosures.[1] *BERT-based* models are a new class of neural language models that show superior ability to capture the sequential connections and interactions of words and other elements within a textual document (i.e., context), and thus better reflect the totality of the information being conveyed in textual disclosures.[2] I train a new *BERT-based* language model to predict the magnitude and direction of value-relevant news (i.e., the predicted "out-of-sample" contemporaneous stock market reaction) for a large sample of quarterly earnings announcement texts. Thus, I derive a single summary information variable that synthesizes the news, based on word context (i.e., contextualized), within quarterly disclosures. Using this new approach, I find that the contextualized news extracted from quarterly earnings announcement text (1) explains at least *three* times more variation in short-window abnormal stock returns than disclosures modeled using traditional narrative attributes (e.g., "tone") or recent machine learning techniques (e.g., based on one/two-word phrases, also known as "N-Grams"), and (2) offers large incremental explanatory power relative to reported earnings modeled using

---

1 See Devlin et al. (2019) for the original *BERT* model, together with Siano and Wysocki (2021) and Lee and Zhong (2021) for applications within the accounting domain. Throughout this study, I refer to *BERT-based* (or neural language) models to indicate a class of transformer-based deep (i.e., large) neural networks that are (i) pre-trained to "learn" general linguistic structures from a broad set of nonaccounting texts (e.g., Wikipedia posts), and (ii) "fine-tuned" (or adapted) to the technical terminology of the financial accounting domain (see Section 3). Thus, the results and insights in this study do not depend upon the use of a single model, but rather a *class* of language models equipped with the aforementioned features. In this study, I distinguish *BERT*-based deep (i.e., large) neural networks from alternative machine learning methods to model textual disclosures.

2 I define *context* (and *word context*) as the modeling of (i) the text preceding and following each element in a disclosure, and (ii) the sequential semantic relations across each element in a disclosure (alternatively called "co-text" as outlined in Catford, 1978). In addition, I use (*textual*) *document* or *disclosure* to indicate texts that include ordered collections of words organized in sentences and paragraphs.

traditional or machine learning methods.[3] In addition, I show that contextualized disclosures strongly predict the level and volatility of future quarterly earnings, and that most news arises from (a) word order (i.e., context), (b) text describing numbers, and (c) text at the beginning of disclosures. Overall, these findings highlight the decision-relevance of contextualized disclosures relative to alternative measurements of text, based on traditional word counts or more recent machine learning techniques, and demonstrate the promise of new neural language models for understanding corporate disclosures.

The textual analysis literature in accounting and finance (see surveys by Li, 2010; Das, 2014; Loughran and McDonald, 2016) has traditionally (i) identified *independent* disclosure attributes within text (e.g., "tone"), (ii) quantified these attributes using empirical procedures (e.g., "tone" is usually calculated as the difference between "positive" and "negative" words to total words in a document), and (iii) analyzed the information content of accounting text by using individual narrative attributes to explain stock returns (e.g., Loughran and McDonald, 2011). Such traditional approaches have a major limitation: characterizing subsets of words and text attributes in isolation leads to measurement error since words in a document are strongly connected and interrelated. For example, the word "loss" is classified as a "negative" word by traditional approaches; however, if used in the (word) context of a one-time (i.e., transitory) charge, the same term might not convey "bad news". This limitation likely contributes to the modest explanatory power of traditional textual analysis methods in correlating disclosures with stock market outcomes (Lev, 2018).

[3] *Narrative attributes* refer to measurements of text based on word counts. *Traditional* [*machine learning*] *methods* or *techniques* indicate "linear" (i.e., through OLS) and "in-sample" ["non-linear" (i.e., using non-linear machine learnable functions) and "out-of-sample"] modeling of text or financial statement items.

More recent studies address such limitation by applying machine learning techniques to model disclosure text (see the survey by Bochkay et al., 2022). Recent approaches generally (i) extract one or two-word phrases (i.e., "unigrams" and "bigrams", or more in general, "N-Grams") from disclosure text, (ii) model their occurrence and co-occurrence through non-linear functions, and (iii) fit the text to a classification or regression outcome (e.g., Li, 2010; Frankel et al., 2016; Frankel et al., 2021). Such techniques offer two advantages relative to traditional methods: non-linear functions can capture some interactions across words; and text can be characterized based on a notion of "local word context" (e.g., two consecutive words). However, these latter approaches still suffer from two fundamental limitations: they are unable to characterize the semantic role of and similarities among words in a document (e.g., whether "reported" is a verb and whether it is similar to "recognized"); and they largely ignore word order (e.g., "did not report a loss related to" and "did report a loss, not related to" might be treated equally). Therefore, these methods cannot fully capture complex ordered relations across elements in a text (i.e., word context).

Recent innovations in language modeling and neural networks show promise in capturing complex word context for documents in a variety of settings.[4] In particular, *BERT-based* language models are neural networks with two key features enabling them to capture complex word context. First, they distill general semantic relations among thousands of words in a language through pre-

[4] *BERT-based* models represent a current best practice for natural language processing tasks due to their superior accuracy demonstrated outside of and within the accounting domain, and their accessibility (i.e., the pre-trained model and several software implementations are freely available). However, alternative neural network approaches exist. For instance, Meursault (2019) applies a convolutional network to predict absolute raw stock returns using text. A similar neural network approach is described in Li (2019). These latter approaches do not leverage large-scale pretraining and lacks two key features that make *BERT-based* models superior at characterizing *context*: bi-directionality and attention-based transformers. Overall, alternative models suffer from an incomplete (i.e., local) word-context characterization that affects modeling accuracy and associated inferences. Replicability is also harder to achieve as the software is usually not open source.

training on massive datasets. Second, they represent each word in a document in terms of its *ordered* relationship to any other words in the same text: this allows for a rich and contextualized characterization of linguistic features within that document. *BERT*-based models can be fine-tuned to a specific setting to directly explain, through language, investors' actions in stock markets. Notably, the direct training of a *BERT-based* model to explain stock market outcomes (e.g., abnormal price revisions) enhances our ability to understand the linkages between information contained in textual disclosures and how market participants incorporate this information into their decision-making such as price formation. Therefore, I train a *BERT-based* neural network to capture the contextualized news in earnings announcement disclosures by modeling the rich content and word context of text in earnings press releases and how it relates to contemporaneous abnormal stock returns (in further analyses, I also investigate trading volumes and return volatility). I then assess the ability of the model to explain ("out-of-sample") short-window market outcomes through the news contained in the text of quarterly earnings announcements.[5] Notably, this approach generates a *single* information variable that condenses the links between the contextualized news in the disclosure and the expected (per the *BERT-based* neural language model) stock market outcomes. I choose the setting of earnings announcements because (i) a large literature documents their decision-relevance (Ball and Brown, 1968; Beaver, 1968); (ii) they offer recurring disclosures for a broad cross-section of firms over time; and (iii) they are relatively shorter than regulatory filings (e.g., 10-K and 10-Q reports) and thus offer a more tractable computational task.

[5] My study is not meant to extract an ex-ante trading signal. I investigate how text maps into contemporaneous, rather than future, investors' actions.

I show that the contextualized representation of a textual disclosure allows for a richer characterization of textual news content and thus better explains stock market reactions to a disclosure compared to traditional and recent machine learning methods that treat disclosures as independent words or (only) pairs or words. I find that contextualized disclosures explain at least *three* times more variation in short-window market outcomes than disclosures modeled using traditional narrative attributes or recent machine learning techniques and offer large incremental explanatory power relative to reported earnings modeled using traditional or machine learning methods. To facilitate inferences, I link my findings to valuation theory (Miller and Modigliani, 1958; Ohlson, 1995) and demonstrate that contextualized disclosures are also strongly predictive of the level and uncertainty of future accounting earnings, above and beyond current period's earnings numbers. This is consistent with the contextualized news in textual disclosures mapping into firm value and investor trading decisions through a channel of revised expectations about future earnings.

This study makes three primary contributions, building on the capital markets literature, the textual analysis literature in accounting and finance, and the emerging literature on investors' use and processing of accounting information. First, I offer new evidence on the information content of contextualized disclosures for investor stock pricing decisions (Healy and Palepu, 2001; Holthausen and Watts, 2001; Kothari, 2001). Lev (1989, p. 173) describes accounting earnings as *"… an information variable that explains only about 5% of stock return variability and whose relation with returns is unstable."* My study documents a large *incremental* explanatory power of contextualized news in textual disclosures for (short-window) stock market outcomes relative to reported financial statement items. Notably, I characterize financial statement items not only through traditional (i.e., linear, "in-sample") approaches, but also using recent machine learning (i.e., non-linear, "out-of-sample") techniques. Therefore, my analyses better highlight the incremental information content of disclosures, relative to reported numbers, as opposed to differences in modeling approaches.[6] In addition, I offer a theory-based explanation for the power of contextualized news in explaining abnormal returns, demonstrating that contextualized disclosures are strong predictors of future earnings and map into investors' actions through a channel of future earnings numbers that are relevant for investor valuation decisions.

Second, I build upon the textual analysis literature in accounting and finance (Li, 2010; Loughran and McDonald, 2016; Bochkay et al., 2022) by introducing a framework and offering new applied insights about the use of neural language methodologies to extract information from narrative text. I use a *BERT-based* model to extract a *single* information variable, based on word context, from the text of corporate disclosures; such approach can be applied to a wide range of accounting research questions, and thus improve traditional and recent text analysis tools in accounting and finance (Frankel et al., 2021). Therefore, I extend economic-based and measurement insights into recent "big data" applications for textual accounting disclosures.

Third, I contribute to the recent debate surrounding investors' use and processing of accounting information (Blankespoor et al., 2020). My findings suggest that investors use, process, and interpret unstructured information signals—sourced from the disclosure system—in a contextualized form rather than as a collection of single or pairs of words. My collective results support that the contextual modeling of disclosures promises a richer understanding of disclosures and investors' processing of accounting information.

[6] The goal of these analyses is to document disclosures' *incremental* information content rather than comparing absolute explanatory power between disclosures and reported numbers. In fact, given the short-window nature of the study, most of reported numbers' information content for pricing revisions has likely been anticipated by market participants at the time of the earnings announcement (Collins et al., 1994).

2. Background and Related Work

Starting from Ball and Brown (1968), a broad stream of empirical literature investigates the fundamental role of accounting information in capital markets.[7] The traditional experimental variables synthesizing accounting informativeness are reported income numbers (Kothari, 2001). Despite being widely utilized by equity investors, research questions the relatively low association of earnings numbers with stock returns (Lev, 1989; Lev and Zarowin, 1999). In response, researchers have increasingly extended the set of examined accounting signals to recognize the richness of the reporting system and the importance and complementarity of the disclosure system. Several studies focus on alternative performance metrics, such as cash flows, accruals, and a wide array of financial statement variables (e.g., Dechow, 1994; Ou and Penman, 1989; Wilson, 1987); others investigate disclosed numbers such as management's forecasts (e.g., Waymire, 1984). Unsurprisingly, these analyses confirm the importance of other reported and disclosed numbers for valuation decisions. More recent valuation studies investigate disclosed textual information. Researchers traditionally identify attributes of narrative disclosures that may inform pricing decisions, proposing empirical procedures to characterize those attributes (see surveys by Li, 2010; Das, 2014; Loughran and McDonald, 2016). The related textual analysis literature in accounting and finance suggests that (at least some) narrative information signals are significantly related to stock returns and trading volumes (Loughran and McDonald, 2011). However, these statistical associations exhibit low explanatory power, thus offering only modest evidence about the economic role of the narrative disclosure system for stock market outcomes (Lev, 2018).

[7] As noted in Kothari (2001), prominent roles for accounting information include (i) valuation and security analysis, and (ii) contracting. In this study, I focus on the valuation role of accounting disclosures (with a clear emphasis on text-based information signals). In Section 6 (i.e., Conclusions), I discuss promising avenues to extend empirical contributions to several areas of capital markets research in accounting.

One limitation likely affects the ability of traditional textual analytics at explaining investors' reactions. Specifically, the measurement procedures introduce substantial noise due to the extensive use of "bag-of-words" approaches, and the stand-alone word focus of traditional textual metrics (Loughran and McDonald, 2016). "Bag-of-words" approaches treat words irrespective of their position in a document, thus losing contextual connections. This can lead to misclassified words (e.g., words that are used with "neutral" sentiment may be classified as "negative"). Moreover, the design of empirical textual attributes largely reflects researchers' pragmatism rather than theory or users' actions.[8] It should be noted that while a theory exists linking income numbers to the market valuation of firms' equity (Miller and Modigliani, 1958; Ohlson, 1995), the mapping of qualitative attributes of textual disclosures to market valuation remains underdeveloped.

To address these limitations, more recent studies apply machine learning techniques allowing some interactions among words in a disclosure, and "learn" the (likely non-linear) mapping of text features to financial accounting outcomes. For example, Frankel et al. (2016) applies Support Vector Machine (i.e., SVM) algorithms to characterize "Management Discussion and Analysis" (i.e., MD&A) disclosures and explain, through them, accounting accruals. Donovan et al. (2021) uses a combination of machine learning techniques—including, SVM, Latent Dirichlet Allocation (or LDA), and Random Forest (or RF)—to measure credit risk through MD&A and conference calls text. Frankel et al. (2021) applies Random Forest protocols to measure disclosure sentiment within 10-K and conference calls narratives and compare it with

[8] For instance, the theoretical justification is unclear for the choice of measuring "readability" through the "Fog Index" (i.e., [average words per sentence + percentage of complex words] * 0.4), or "tone" through a sentiment polarity index (i.e., [positive words – negative words] / total words). Moreover, the highly non-linear nature of linguistic features and contextual connections seems to be inconsistent with a linear and independent characterization of textual signals/attributes.

traditional dictionary-based sentiment metrics. Overall, this research collectively finds that machine learning methodologies offer considerable advantages (i.e., evaluated through predictive accuracy) over traditional "bag-of-words" approaches.

However, even the above machine learning approaches are unable to characterize complex word context and lack fundamental features desirable to understand disclosures and investor use of accounting text. First, these methods cannot identify the semantic role of and similarities among words in a document; therefore, machine learning techniques are generally unable to characterize the "meaning" (i.e., the connection among words related to a similar semantic object) of a textual token. Second, these machine learning techniques largely ignore word order; therefore, they cannot characterize the "conditional meaning" (i.e., conditional on surrounding words) of a textual token. Note that sensitivity to word order is an especially desirable feature for a language model because it allows to more closely mimic investors' processing of textual signals, and therefore better understand nuances of their decision process.

Advances in the computer science domain, related to deep (i.e., large) neural networks, offer an opportunity to mitigate the prior two limitations. In particular, recent developments of *BERT-based* language models (see Devlin et al., 2019; Siano and Wysocki, 2021; Lee and Zhong, 2021) now allow researchers to (i) characterize each word in a document in terms of its *sequential* relationship to any other words in the same text, thus allowing for a (richer) context-based characterization of narrative disclosures, and (ii) empirically model and optimize which features of a textual disclosure are related to valuation outcomes (e.g., stock returns and trading volumes).[9]

[9] Contemporaneous advances in language modeling promise performance improvements with respect to current *BERT-based* implementations. The goal of this work is (i) to describe a framework to characterize contextualized disclosure news and investigate its mapping into investors' actions, and (ii) to offer economic-based interpretations for such a mapping. The intent of this study is *not* to propose *BERT-based* language models as the definitive solution to language modeling tasks.

Notably, game-theoretical work at the intersection of information economics and linguistics suggests that contextual information affects economic decisions (Rubinstein, 2000; Glazer and Rubinstein, 2001). *BERT-based* language models are large neural networks that (a) learn the general semantic role of and similarity across parts of speech, (b) preserve contextual word order, and therefore (c) can better quantify and characterize the decision usefulness of any textual disclosures (including, for example, earnings), for stock pricing revisions, than alternative machine learning techniques.

3. Data and Methodology

3.1 Sample Selection

I source earnings announcement texts from 8-K filings downloaded from the SEC EDGAR database, and Business Wire U.S. press releases downloaded from Factiva. I gather daily stock returns and trading volumes from CRSP, firms' quarterly fundamentals from Compustat, and analysts' quarterly estimates from I/B/E/S. I scrape each EDGAR 8-K filing and Business Wire U.S. press release to extract the earnings announcement text, a company's identifier, and the announcement date. Each narrative disclosure is tagged with a company's identifier (i.e., CIK or Ticker) and an earnings announcement date. I then use the identifier and date to match the observation with daily and quarterly data from CRSP and Compustat (or I/B/E/S), respectively.

I retain unique firm-quarter earnings announcements, excluding disclosures with fewer than 5 sentences or 100 total words. I also exclude generic cautionary statements unlikely to affect investors' decision-making, as well as tables due to non-comparable variation in their layout across firms and quarter-years. I use the full remaining text of the selected disclosures.[10] Finally, I exclude low liquidity observations (suggesting a weaker expected connection between disclosures and market reactions), defined as "penny stocks" (i.e., stocks with a unit price lower than $1) and companies with less than $10 million in total assets.

The final sample includes 101,412 quarterly earnings disclosures with available stock returns, trading volumes, and financial statement variables over 1989–2019. Table 1 presents the sample and the selection criteria.

<See Table 1>

3.2 Measuring Contextualized News in Earnings Announcement Disclosures

I apply a new *BERT-based* language model to each of the quarterly earnings disclosures to extract a single information variable (solely from disclosure text and based on word context) predicting the magnitude and direction of the disclosure news. I then use the predicted contextualized news to explain "out-of-sample" capital market outcomes (e.g., 3-day cumulative abnormal returns around the earnings announcements). Following state-of-the-art protocols in machine learning, I implement specific procedures including random stratified sampling and cross-validation tests, which mitigate potential modeling biases.[11] Moreover, I design and perform a

---

[10] Tables are characterized by differences in formatting, sequencing of information, separation from the main earnings announcement text, types and formatting of numbers. These differences vary across and within-firm over time. *BERT-based* models are not explicitly pre-trained on tabular information and thus lack training to maximize signal extraction from tables. In untabulated tests, I use the entire text of the earnings announcements (inclusive of tables and generic cautionary statements) and find a 1.5% lower explanatory power on average; however, the tenor of my results does not change. Appendix B offers more details about the parsing strategy for tables and generic cautionary statements.

[11] To confirm that the results do not depend upon the random choice of the training and fine-tuning samples, I fit two models using the "out-of-sample" documents for fine-tuning and the fine-tuning sample for "out-of-sample" prediction (i.e., I apply a "two-fold protocol" for cross-validation). Cross-validation results (not reported) are largely consistent with the main reported results.

series of interpretability analyses to offer economic-based insights into contextualized news and *BERT-based* models' functioning.

To model earnings announcement disclosures, I exploit recent advances in *machine transfer learning*.[12] A language model (i.e., neural network) is first pre-trained on large-scale texts to learn generalizable domain-invariant features of language. Subsequently, the model's "knowledge" (i.e., the set of learned features and parameters) is transferred and "fine-tuned" (i.e., adapted) to a particular domain of interest (e.g., financial accounting disclosures). The pre-trained model is readily available to users, which is a major advantage given the high computational intensiveness of large-scale pre-training. Thus, researchers only need to fine-tune the model. Fine-tuning corresponds to a supervised learning task; the neural network is fit to narrative disclosures labeled with a chosen outcome (e.g., 3-day returns or trading volumes), and it learns a set of parameters mapping linguistic features to that outcome. This adapts the final layers of the pre-trained neural network's parameters to the linguistic nuances of the reference domain, a step particularly important for domains characterized by technical terminology such as financial accounting. The goal is to use the pre-trained and fine-tuned *BERT-based* language model to predict the outcome of interest (e.g., contemporaneous 3-day returns or trading volumes) for "out-of-sample" texts (i.e., observations not used in the fine-tuning step).

The *machine transfer learning* framework requires choosing a language model (i.e., neural network). I opt for *BERT-based* neural language models, which have achieved high modeling accuracy and relatively low cost in several natural language processing tasks, including within the

[12] The transfer learning framework I employ is similar to that proposed by Siano and Wysocki (2021) and Huang et al. (2021). However, the focal objective of these prior studies is to classify the "tone" or "sentiment" of a given text rather than directly modeling, through contextualized narrative disclosures, investment-relevant actions in equity markets. Moreover, their modeling choices and implementation are substantially different from what I propose. Sections 3.4 and footnote 15 will describe these differences in more detail.

accounting domain (Siano and Wysocki, 2021; Lee and Zhong, 2021). These models are especially effective at characterizing context-based disclosures, since they treat text as an *ordered* sequence of terms, and represent each term based on the preceding and following text (i.e., context). *BERT* (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019) was introduced by Google Research in 2019, with subsequent research teams sharing new versions of the model. However, the core features of *BERT-based* models' success remain unchanged. First, the neural network is pre-trained using a *"bi-directional"* approach, wherein for each linguistic token, both the preceding and following portions of text characterize its semantic role. Second, this language model is built around *"attention-based transformers"* that describe each word in a language considering its relationship with (i) any other words in a language, and (ii) any words in a particular disclosure (see Siano and Wysocki, 2021). Thus, *BERT-based* models learn two sets of key features in language (e.g., Jawahar, et al., 2019; Tenney et al., 2019): the syntactic role of parts of speech (e.g., whether a token is a noun, a pronoun, a verb, a company name or a number), and the sequential relation across parts of speech (e.g., subject-object agreement, subject-verb agreement, co-reference, cause-effect links).

3.3 Model Choice and Implementation

Among recent *BERT-based* language models, I choose the *RoBERTa* model (i.e., Robustly Optimized BERT Pretraining Approach, Liu et al., 2019) due to its high accuracy and low resource intensiveness documented in the financial accounting domain (Siano and Wysocki, 2021).[13] For modeling contemporaneous market reactions, I split the full sample of earnings announcement texts (meeting the above selection criteria) into two halves: 50% (i.e., 115,821 data points) is

[13] Appendix C reports the URL link to freely download and use *RoBERTa*.

employed for fine-tuning, and the remaining 50% for "out-of-sample" modeling. The observations are assigned to the two samples using a random approach stratified by date (i.e., quarter-year), leading to balanced samples in terms of observations belonging to each quarter-year. I finally retain only *BERT*-based predictions for 101,412 firm-quarters with available financial statement items and firm fundamentals. Of note, the split approach I implement is unlikely to suffer from overfitting bias as the prediction task involves contemporaneous (rather than future) market outcomes, and the one-period time-series correlation of 3-day cumulative abnormal stock returns is low (i.e., about 0.01).[14]

I "fine-tune" the *RoBERTa* language model on the text of quarterly earnings announcements labeled with 3-day cumulative abnormal stock returns. The neural networks' hyperparameters (i.e., learning rate and the number of training epochs), the software used, and the computational time are described in Appendix C. Once fine-tuned, I use *RoBERTa* to predict contemporaneous (i.e., same-quarter) short-window stock returns based solely on the "out-of-sample" earnings announcement texts (i.e., 101,412 disclosures never processed before by the model). Therefore, the fine-tuned *RoBERTa* model has "learned" how contextualized earnings disclosures likely map into contemporaneous investors' actions. I also fine-tune the *RoBERTa* model on the text of quarterly earnings announcements labeled with the firm's earnings for the

[14] The randomly stratified 50-50% split approach balances the trade-off between maximizing accuracy and minimizing training cost. One could also implement a split approach with a higher percentage of training/fine-tuning observations (e.g., 80% of data points training and 20% for "out-of-sample" modeling) as it is usually done in machine learning applications. This approach would likely offer higher modeling accuracy, but also lower generalizability due to the smaller "out-of-sample" set. An alternative method would be applying a (yearly) rolling window for fine-tuning and prediction. Such a methodology would imply fitting yearly models exploiting a fixed (e.g., 5-year) or variable-length (e.g., all prior years) window of yearly observations. Rolling fine-tuning is computationally intensive and is usually implemented to model future outcomes exhibiting high time-series correlation (i.e., to minimize potential overfitting biases). It would offer relatively higher accuracy in modeling non-stale language features (since business language evolves over time), but this may come at the cost of a lower number of per-year training/fine-tuning observations.

next quarter and the standard deviation of earnings for the next four quarters. Since the outcome involves future earnings, and future earnings are highly serially correlated (with a one-period time-series correlation of about 0.5), for these tasks I estimate a model each year using the prior 5 years of textual data to limit any overfitting biases.

With respect to prior uses of *BERT-based* models, I introduce two novel features.[15] First, I overcome the limit in the maximum sequence length that *BERT-based* models can process (i.e., 512 tokens). In particular, I slice each earnings announcement into 512-token "windows" of text and process all the sequences for a given document. For each earnings announcement, the predicted output (e.g., contemporaneous short-widow market reaction) is then averaged across all the sequences. Second, I add a linear layer to the neural network to obtain a continuous (rather than categorical) output. This is equivalent to letting the *BERT-based* models perform a regression rather than (the usual) classification task.

3.4 Benchmarks

I study the incremental information content of contextualized earnings disclosures relative to (i) largely a-contextual characterizations of the same text disclosure, and (ii) relevant financial statement items. Since contextualized news derives from the non-linear *BERT*-based approach, I model both alternative text measurements and relevant reported numbers using (a) traditional (i.e., linear, "in-sample"), and (b) recent machine learning (i.e., non-linear, "out-of-sample") methods to better isolate incremental information content as opposed to differences in modeling choices. Recent research documents substantial gains in predictive power when stock returns are modeled

[15] Siano and Wysocki (2021) limit their analyses to the first 15 sentences of the earnings announcement. Both Siano and Wysocki (2021) and Huang et al. (2020) apply *BERT-based* models to perform a two or multi-class classification task in which the output is a classification label rather than a continuous variable.

through non-linear machine learning techniques using either disclosure text (e.g., Frankel et al., 2021) or reported numbers (Barth et al., 2022) as inputs. Therefore, I apply a Gradient Boosting machine learning approach (e.g., Krupa and Minutti-Meza, 2021) to predict stock returns "out-of-sample" using largely a-contextual characterizations of earnings announcement disclosure text, and relevant financial statement items.[16] Appendix D provides details about the implementation of Gradient Boosting protocols.

3.5 Understanding Contextualized News

I offer direct evidence about the role of word context within contextualized news. First, I "mask" (i.e., modify) the text of earnings disclosures by alternatively (i) *deleting* connecting (i.e., "stop") words that are usually considered uninformative, thus ignored in traditional textual analysis (e.g., Frankel et al., 2021), and (ii) *randomizing* all words.[17] I then predict contemporaneous market reactions using the modified earnings disclosures as inputs to the *BERT-based* language model and finally test the associated "out-of-sample" explanatory power. If *BERT*-based neural language models find word-context informative, their modeling accuracy should exhibit substantial sensitivity to both connecting words and word order.

Second, I "fine-tune" (i.e., train) the *BERT-based* language model on text that *only* contains the 100 most frequent accounting performance words found within earnings announcement texts, or "tone" words commonly employed in the traditional textual analysis literature. I then use the

[16] I choose Gradient Boosting after comparing its predictive accuracy with that of Support Vector Regressions (e.g., Donovan et al., 2021) and Random Forest (e.g., Frankel et al., 2021) approaches and finding that, for the main prediction task assessed in this paper, it outperforms alternative methods with respect to accuracy and computational time. Of note, I train and test the Gradient Boosting algorithm on samples comparable to those used for *BERT*-based fine-tuning and prediction.

[17] The deletion of connecting words and the randomization of all words only involves "out-of-sample" observations. The *BERT-based* language model is *not* fine-tuned on the modified texts. Connecting words include "stop-words" from the "spaCy" library (https://spacy.io/).

fine-tuned model to predict "out-of-sample" capital market outcomes for observations that *only* include frequent performance or "tone" words, respectively. If *BERT*-based models characterize the full set of sequentially connected elements in a disclosure, they should not be able to achieve high predictive accuracy by exploiting only a subset of words.

I then demonstrate, with two examples, how the *BERT-based* word context modeling of news can offer insights into features of corporate disclosures. First, I investigate which *contents* affect contextualized news the most. In particular, following the traditional capital markets research and recent work about the role of numbers within text (Siano and Wysocki, 2018), I study contextualized news in text that *discusses* numbers, and that does not discuss numbers.[18] If numbers represent one of the primary elements of interest to investors, within disclosures, textual discussions about quantitative signals may associate with relatively higher contextualized news. Therefore, I predict contemporaneous short-window returns twice: first using text discussing numbers and then using other text. I also assess the role of numerical tokens (rather than text discussing numerical tokens) within earnings disclosures. The goal is to understand whether most contextualized news arises from numbers within the text rather than the contextualized verbal contents describing numbers. I therefore delete dollar and percentage amounts within text discussing numbers and use these modified sequences as inputs to the *BERT-based* model to predict short-window outcomes. Second, similar to Cheng et al. (2021), I examine the *position* within accounting earnings disclosures in which most contextualized news is found. Therefore, I select earnings announcements characterized by at least three sequences of 512 tokens (i.e., the maximum textual sequence length that *BERT-based* models can process; see also Section 3.3) and

[18] I identify sentences that include dollar amounts or percentages, and that are immediately adjacent to them as "sentences that discuss numbers". I characterize other sentences as "sentences that do not discuss numbers."

assess contextualized news individually for each sequence (i.e., the first, the second, and the third within each document in order of appearance).

I also examine contextualized news to understand properties of corporate disclosures. In particular, consistent with research on the asymmetric release of “good” versus “bad” news (e.g., Basu, 1997; Kothari et al., 2009), I study asymmetric responses to contextualized news within the different settings of “good” and “bad” news release (defined based on positive and negative cumulative abnormal returns in the earnings announcement period, respectively).

4. Research Design

I use OLS regressions to test explanatory power (i.e., “Adjusted” and “Within” R^2) of contextualized news for short-window capital market outcomes around earnings-announcement disclosure events as follows:

$$\textit{Capital Market Reaction}_{iq} = \alpha_1 + \beta_1 \textit{ Text News}_{\textit{(Word Context)}iq} + \textit{Fixed Effects}_i + \varepsilon_{iq} \qquad (1)$$

Capital Market Reaction is the (contemporaneous) 3-day cumulative abnormal stock return (i.e., *CAR*); alternatively, it is the 3-day abnormal trading volume and stock return volatility within the sensitivity analyses. The experimental variable is *Text News* $_{\textit{(Word Context)}}$ that captures the relevant capital market news attribute contained solely in the text, modeled using word context, of a firm’s earnings announcement disclosure—i.e., *CAR* modeled “out-of-sample” through *BERT-based* language models. The variables are measured each calendar quarter “*q*”, from 1989 to 2019, for each firm “*i*” in the sample. In the main analyses, I estimate the model across firms; in additional tests, I include fixed effects for each company and test the explanatory power within firm.

I focus my tests on the incremental explanatory power of *Text News* $_{(Word\ Context)}$ relative to (i) largely a-contextual characterizations of text, and (ii) relevant financial statement items. I model both of the latter benchmark measures using traditional (i.e., linear, "in-sample") as well as machine learning (i.e., non-linear, "out-of-sample") approaches. Therefore, I estimate and compare the following two models first including alternative disclosure text measurements and financial statement items modeled through traditional approaches, and then alternatively modeled using recent machine learning techniques.

$$\begin{aligned} \textit{Capital Market Reaction}_{iq} = \;& \alpha_1 + \gamma_1\, \textit{Text News}_{\,(No\ Context)iq} \\ & + \gamma_2\, \textit{Financial Statement Items}_{iq} \\ & + \gamma_3\, \textit{Controls}_{iq} + \textit{Fixed Effects}_i + \eta_{iq} \end{aligned} \tag{2}$$

$$\begin{aligned} \textit{Capital Market Reaction}_{iq} = \;& \alpha_2 + \delta_1\, \textit{Text News}_{\,(Word\ Context)iq} \\ & + \delta_2\, \textit{Text News}_{\,(No\ Context)iq} \\ & + \delta_3\, \textit{Financial Statement Items}_{iq} \\ & + \delta_4\, \textit{Controls}_{iq} + \textit{Fixed Effects}_i + \chi_{iq} \end{aligned} \tag{3}$$

Text News $_{(No\ Context)}$ is disclosure text modeled using methodologies that are unable to fully characterize word context. These include dictionary-based *Narrative Attributes* and *N-Grams* used as inputs to machine learning models. *Narrative Attributes* include five linguistic constructs for which prior literature documents an association with investors' actions: *Tone* (measured as ["positive" – "negative"]/total words using the Henry dictionary, 2008),[19] *Readability* (computed using the Gunning Fog 1952 index), *Length* (or the natural logarithm of total words in the earnings

[19] Henry (2008) develops a dictionary that is tailored to earnings announcement texts. Therefore, this choice maximizes the signal-to-noise ratio of the tone measure.

announcement), *Numbers* (the total number of dollar and percentage amounts disclosed within text), and *Future* (or the number of future tense verbs scaled by total words). *N-Grams* are the most relevant 3,000 one and two-word phrases extracted from earnings announcement texts (following Frankel et al., 2021).[20]

Consistent with the extensive prior capital markets research in accounting, I also assess the incremental explanatory power of *Text News* $_{(Word\ Context)}$ relative to that of reported accounting numbers. *Financial Statement Items* include *Earn* (the level of reported earnings before extraordinary items scaled by lagged market capitalization) and *Surprise* (the earnings surprise relative to the most recent consensus analyst forecast); *Div* and *Div Chg* (the level and quarter-to-quarter difference in dividends per share) that account for firms' distributed cash flows; and *Leverage* and *Leverage Chg* (the level and quarter-to-quarter difference in short-term and long-term liabilities to total assets) that measure the composition of the capital structure. Importantly, dividend distributions and debt-to-equity ratios represent powerful predictors of future income numbers (Ou and Penman, 1989). I further include *Restr* (an indicator that measures whether restructuring costs are reported) and *Spec Items* (the absolute value of special items scaled by quarterly assets) to account for special business circumstances with likely implications for valuation. *Financial Statement Items* are used both as "in-sample" predictors (i.e., traditional

[20] As previously discussed, *N-Grams* are the textual input to a Gradient Boosting machine learning model that is used to predict the contemporaneous short-window capital market reaction around earnings disclosures. The training and prediction protocols are comparable to those implemented for *BERT*-based modeling (i.e., the same observations used for *BERT*-based fine-tuning and prediction are also used with *N-Grams*). One and two-word phrases (i.e., "unigrams" and "bigrams") are selected using a TF-IDF (i.e., term frequency-inverse document frequency) approach to evaluate the most relevant words in a document. As a comparison, Frankel et al. (2021) uses all two-word phrases in a textual disclosure. In robustness analyses, I extend the number of TF-IDF unigrams to 6,000 and 9,000 finding negligible changes in predictive accuracy. Appendix D provide details about the N-Grams implementation protocol.

approach), and alternatively to train a Gradient Boosting non-linear algorithm (i.e., recent machine learning approach) to predict "out-of-sample" short-window market reactions.

Finally, I explicitly control for relevant concurrent news events that might confound the role of disclosures in affecting investment actions and the firm's general risk and information environment. Thus, the regressions include *AF* (an indicator variable for whether an analyst earnings forecast occurs in the 3-day period surrounding the earnings announcement event); *Size* (the natural logarithm of a firm's market capitalization), and *Follow* (the natural logarithm of analysts following a firm in a quarter).[21]

I formally test differences in explanatory power for models (2) and (3) by bootstrapping R^2 using 10,000 repetitions. The main regressions are run without fixed effects; additional tests include company-level fixed effects to assess within-firm explanatory power. Standard errors are clustered by firm and quarter-year (for regressions without firm fixed effects) or quarter-year (for regressions with firm fixed effects).

I also assess the predictive power of contextualized corporate disclosures for the level and volatility of future accounting earnings. The primary dependent variables are $Earn_{\ Q+1}$ (next quarter's accounting earnings) and $SD\ Earn_{\ Q+1\text{-}Q+4}$ (the standard deviation of the future four quarters of accounting earnings). The experimental variables are future earnings' level and volatility predicted "out-of-sample" solely through *BERT-based* language models and the contextualized earnings disclosure – i.e., $Pred\ Earn_{\ (Word\ Context)}$ and $Pred\ SD\ Earn_{\ (Word\ Context)}$.

---

[21] In a series of robustness analyses (not reported), I also include the release of management's forecasts as an additional control to account for relevant concurrent news events. The results are largely unaffected.

5. Empirical Results

5.1 Descriptive Statistics

Table 2 presents the descriptive statistics for the primary set of 101,412 "out-of-sample" firm-quarter observations. Panel A offers three splits: (i) by fiscal quarter, (ii) by fiscal year, and (iii) by industry. The four fiscal quarters are equally represented in the final sample, and most of the observations occur within 2000-2019. The three most (least) prevalent industries are Banking, Business Services, and Pharmaceuticals (Tobacco, Fabricated Products, and Shipbuilding). Panel B shows statistics for the main variables. The mean (median) market capitalization is $3.8 billion ($530 million); the mean (median) quarterly unscaled earnings amount to $52 million ($3.9 million); 28% of firms report quarterly losses; and the average firm is followed by 8 analysts. Interestingly, 60% of earnings announcement events reflect a concurrent release of analysts' forecasts. Finally, earnings press releases are characterized by an average of approximately 800 words, 37 disclosed numbers, and a generally positive tonal orientation.

<See Table 2>

Table 3 presents further univariate evidence. Of note, Columns (1) and (2) show a strong correlation between *Text News* *(Word Context)* and the *Capital Market Reaction* main variable (i.e., 0.39 for price revisions measured using *CAR*). The contextualized disclosure measures are also strongly associated with the level and volatility of future income numbers (i.e., the univariate correlation is 0.61 and 0.53, respectively). Further, traditional textual attributes (*Tone* and *Numbers*) and recent machine learning characterizations of text (i.e., *Pred CAR N-Grams*) also are significantly associated with the outcomes of interest, that with much weaker than the correlation observed for *Text News* *(Word Context)*.

<See Table 3>

5.2 *The Explanatory Power of Contextualized News for Investors' Actions*

Figure 1 summarizes the incremental explanatory power of the contextualized news measure in explaining 3-day *CAR* around earnings announcements. On average, contextualized disclosures increase Adjusted R^2 by 10% relative to alternative measurements of disclosure text, financial statement numbers, and both.

<See Figure 1>

Table 4 provides further details about the absolute and incremental explanatory power of contextualized news. Panel A compares contextualized disclosure news with news derived from alternative text measurements and relevant reported numbers modeled using a traditional (i.e., linear, "in-sample") approach. *Text News $_{(Word\ Context)}$* explains approximately 15% of the across-firm variation (i.e., Adjusted R^2) in short-window price reactions, representing a ten-fold increase in explanatory power relative to traditionally-evaluated narrative attributes (i.e., 1.4% Adjusted R^2).[22] *Text News $_{(Word\ Context)}$* also offers a large incremental explanatory power (12%) relative to financial statement items in explaining investors' actions (4.7%). Of note, Column (5) shows that the contextualized news proxy offers dimensions of informativeness that are distinct and incremental to both traditional text attributes and reported income numbers, adding 11% of Adjusted R^2 (versus 5.5% provided by narrative attributes and reported numbers taken together).

Panel B compares contextualized news with news derived from alternative text measurements and relevant reported numbers modeled using recent machine learning (i.e., non-

[22] It should be noted that an upper-bound limit for public disclosures' information content, measured using OLS R^2, is likely significantly lower than 100%. In fact, the finance literature has demonstrated that a large fraction of stock returns volatility is alternatively driven by private information revealed through rational trading (e.g., French and Roll, 1986; Roll, 1988) or irrational/noise trading (e.g., Hirshleifer, 2001).

linear, "out-of-sample") techniques. In this panel, text and financial reported numbers are used as input to non-linear machine learning protocols to predict short-window market reactions "out-of-sample". Consistent with Frankel et al. (2021), Column (2) shows that largely a-contextual N-Grams, modeled through a machine learning Gradient Boosting algorithm, offer more than twice as much explanatory power for short-window market reactions than traditional narrative attributes (i.e., 3.9% vs. 1.4%). Notably, *Text News* $_{(Word\ Context)}$ significantly outperforms N-Grams-based "out-of-sample" predicted market reactions (i.e., *Pred CAR N-Grams*); in fact, it generates an 11% incremental Adjusted R^2, representing a 300% improvement. Column (3) demonstrates that non-linear modeling also affects financial statement items; the predicted (through reported numbers) "out-of-sample" market reaction (i.e., *Pred CAR Fin Stat Items*) offers an Adjusted R^2 of 8.9% compared to 4.7% provided by financial statement items modeled with traditional approaches. Relative to both *Pred CAR N-Grams* and *Pred CAR Fin Stat Items*, contextualized news generates a large 7.6% incremental explanatory power (i.e., a 76% increase in Adjusted R^2). All of the above differences are formally confirmed using bootstrapping.

Overall, I document that contextualized disclosure news explains an economically and statistically relevant fraction of price revisions around earnings announcement dates. Results also reveal incremental explanatory power at least *three* times higher relative to that using a-contextual alternative (traditional and machine learning) characterizations of disclosure text, as well as large incremental explanatory power relative to financial reported items (measured either through linear or recent non-linear methods), improving the overall R^2 by approximately 100%.

<See Table 4>

Table 5 provides evidence on within-firm absolute and incremental explanatory power of contextualized disclosures for market reactions. In particular, I replicate the Table 4 analyses including firm fixed effects in all regressions, and test for differences in Within R^2 (see deHaan, 2020). I continue to document large incremental within-firm explanatory power of contextualized news relative to both alternative measurements of disclosures and financial statement items.

<See Table 5>

5.3 The Explanatory Power of Contextualized News for Future Earnings

Next, I use valuation theory to interpret the strong explanatory power of contextualized news by investigating whether the latter measure offers persistent information signals to estimate future earnings numbers. That is, I use contextualized disclosures to predict the level and the uncertainty of future accounting earnings.[23] I also study the incremental predictive power of contextualized disclosures relative to traditional narrative attributes and financial statement items.[24] Table 6 Panel A shows that contextualized earnings press releases explain 33% of cross-sectional variation in future reported earnings. As a benchmark, text attributes offer a 7.2% and current reported numbers a 31% Adjusted R^2. Of note, *Pred Earn* $_{(Word\ Context)}$ offers at least 10 percentage points of incremental Adjusted R^2, improving the overall predictive accuracy by 30%.

Table 6 Panel B examines future earnings uncertainty, measured using the standard deviation of future 4-quarter earnings. *Pred SD Earn* $_{(Word\ Context)}$ explains 28% of future earnings uncertainty and offers at least 12% incremental Adjusted R^2, relative to both traditional narrative

[23] Recall, the level of future earnings is measured as *Pred Earn* $_{(Word\ Context)}$ (defined as the predicted next-quarter reported earnings obtained from a *BERT*-based neural language model based solely on the word-context-based text of a firm's quarterly earnings press release), and the uncertainty of future earnings is measured as *Pred SD Earn* $_{(Word\ Context)}$ (defined as the predicted standard deviation of future 4-quarter earnings obtained from a *BERT*-based neural language model based solely on the word-context-based text of a firm's quarterly earnings press release).

[24] Tables 4 and 5 document strong incremental explanatory power of contextualized news relative to both traditional and recent machine learning measurements of text and reported numbers. Therefore, this analysis focuses only on traditional narrative attributes and reported numbers. Untabulated tests show a 2.5 percentage points lower incremental R^2 (on average) when text and reported numbers are measured using machine learning methods. However, the tenor of the results does not change.

attributes and financial statement items; such an incremental R^2 improves the overall model predictive accuracy by at least 50%.

<See Table 6>

I further analyze and compare the predictive power of *Pred Earn* $_{(Word\ Context)}$ for earnings measured at different future dates: 1, 2, 3, and 4 quarters following the earnings announcement event. Table 7 confirms that contextualized disclosures explain substantially more variation in future period's earnings than narrative attributes over all these alternative time frames. Importantly, the incremental predictive power of *Pred Earn* $_{(Word\ Context)}$, relative to both narrative attributes and reported numbers, increases through time in favor of the contextualized text proxy, improving the predictive accuracy from less than 30% in "Q+1" to more than 40% in "Q+4".[25]

<See Table 7>

Overall, the evidence of Tables 6 and 7 suggests that contextualized disclosures offer persistent information signals useful to estimating the level and uncertainty of future reported earnings. Further, this appears consistent with mapping into investor actions through a channel of revised expectations about firms' future cash flows.

5.4 Understanding Contextualized News

Next, I study *sequential connections* within text that should impact contextualized news to assess whether word context, as opposed to subsets of keywords, drive the observed results. In Table 8 Panel A, I alternatively *delete* connecting ("stop") words (i.e., usually considered uninformative thus ignored by traditional and recent machine learning methods) and *randomize* the content of the earnings disclosure texts. I then use these modified disclosure texts as inputs to

[25] One caveat for this analysis is that the sample changes slightly between "Q+1" and "Q+4" passing from 97,914 observations to 93,754 data points.

the fine-tuned *BERT-based* model and test whether contextualized news relies (and, if so, to what extent) on the ordered and sequential connections (i.e., word context) found in the text of the earnings announcement observations. Results show that when disclosure content is deprived of connecting words then contextualized news decreases by more than 40%; and when disclosure words are randomized, the disclosures' explanatory power for market reactions drops by more than 80%. This evidence suggests that (i) word order represents a key determinant of the information content for contextualized disclosures, and (ii) connecting words, usually considered uninformative thus ignored by traditional or recent machine learning approaches (e.g., Frankel et al., 2021) actually provide useful word context and therefore affect disclosure informativeness. Restated, the *BERT-based* model extracts substantial signal from both the relative positions and the syntactic connections across parts of speech in a disclosure.[26] For comparison, results show that the sensitivity of N-Grams methods to such effects is relatively low, as they retain 90% of their predictive accuracy when connecting words are deleted, or words are randomized.

I also fine-tune the *BERT-based* language model on earnings announcement disclosures that contain *only* (i) the 100 most recurring performance-related words within earnings press releases or (ii) the Henry (2008) "tone" words. If *BERT-based* models characterize sequential contextual connections across parts of speech, then this fine-tuning strategy should reduce the amount of contextualized news that can be extracted from quarterly earnings disclosures. On the other hand, if *BERT-based* neural networks represent a relatively sophisticated methodology to count subsets of keywords (i.e., the traditional approach found within the textual analysis

[26] As described in Sections 1 and 3, *BERT-based* models interpret context as the text that precedes and follows any element within a disclosure. The characterization of context, therefore, depends upon modeling: (i) the words preceding and following any textual token in a disclosure, and (ii) the reciprocal and sequential connections across elements in that same textual disclosure.

literature), the amount of contextualized news should be only mildly affected. Table 8 Panel B presents the results. When only a subset of keywords is used for fine-tuning and prediction, *BERT-based* models' ability to extract contextualized news from earnings disclosures declines significantly, confirming the fundamental role of sequential contextual connections and contents for *BERT-based* language models' prediction accuracy. Comparison using N-Grams methods reveals them to retain most of their predictive accuracy when trained only on subsets of keywords.

<See Table 8>

I next use contextualized disclosure measurement to examine features and properties of corporate disclosures. First, I offer initial evidence on the disclosure *contents* most strongly associated with contextualized news. I distinguish text discussing quantitative information from other text, assessing whether contextualized news differs between the two. Table 9 Panel A reveals most contextualized news arises from text related to quantitative accounting information. Further, consistent with word context driving contextualized news, the news arising from the entire disclosure is higher than the news arising from quantitative and non-quantitative texts taken individually. I also examine whether most contextualized news arises from discussions related to numbers or from the disclosed numbers themselves. I delete dollar amounts and percentages from sample documents and extract a new measure of contextualized news that excludes numbers within the text of corporate disclosures. The results document that numerical information within text accounts for about 20% of contextualized news total explanatory power.

Second, I study the position within a disclosure in which most contextualized news is found. I restrict this analysis to earnings announcements including at least three sequences of 512 tokens (i.e., the maximum textual sequence length that *BERT-based* models can process). I divide each disclosure into three 512-token parts, using each as input to the fine-tuned *BERT-based* model, and extracting a predicted market outcome (i.e., 3-day *CAR*) for each sequence. Results shows that most contextualized news is concentrated within the first part of the document with about 70% of explanatory power.

Further, I demonstrate that contextualized measurement can assess the properties of corporate text. I use contextualized measurement of earning announcement press releases to investigate asymmetric responses to disclosures. Table 9 Panel B shows that during “bad news” events, contextualized news explains a larger proportion of short-window stock returns variation, as well as a higher economic magnitude of the response to contextualized news. This initial evidence highlights opportunities to utilize contextualized disclosures to examine properties of corporate text such as conservatism (Basu, 1997) and news withholding (Kothari et al., 2009).

<See Table 9>

Finally, I show that *BERT-based* models’ ability to extract news is not confined to signed stock returns. Table 10 presents consistent results alternatively using abnormal trading volumes and stock return volatility as outcome variables. Contextualized earnings press releases explain at least 20% of abnormal trading volumes and return volatility in the 3-day announcement period, and offer at least 12 percentage points of incremental R^2 relative to traditional narrative attributes and reported numbers.

<See Table 10>

6. Conclusions

This study offers novel empirical evidence about the fundamental role of contextualized information contained in textual disclosures for investor decision-making. It exploits recent advances in natural language processing to examine the effect of accounting disclosures on investor behavior. In particular, I use a *BERT-based* neural language model to improve upon traditional "bag-of-words" and more recent text analysis methods by (i) modeling disclosures as sequentially connected and interacting elements (rather than isolated words), and (ii) directly predicting the magnitude and direction of disclosure news. This approach addresses a key limitation in the textual analysis literature of "moving beyond assuming words occur as independent units" (Loughran and McDonald, 2016, page 1190).

Applying this approach, I derive a characterization of narrative disclosures, based on word context, to better quantify the decision-relevance of the disclosure system in equity markets. Specifically, I extract a single information variable from earnings press releases over 31 years, and then use this information variable to explain ("out-of-sample") abnormal price revisions and trading volumes in the 3-day window around earnings announcements. Contextualized news explains 15% of the variation in *signed* short-window returns, representing *three* times the explanatory power of disclosures modeled using traditionally-evaluated text attributes (e.g., "tone") or recent machine learning techniques (e.g., based on one/two-word phrases, also known as "N-Grams"). Importantly, contextualized news offers large incremental explanatory power for stock returns relative to reported earnings modeled using both traditional and machine learning methods, generating at least a 100% improvement in terms of Adjusted R^2. Linking my finding to valuation theory, I show that contextualized news strongly predicts the level and the uncertainty of future accounting earnings. To better understand the role of word context, I demonstrate that most news arises from word order, showing that alternative text analysis methods are largely insensitive to sequential word arrangements. Finally, I find that large news arises from text discussing numbers and text at the beginning of disclosures.

Collectively, these innovations allow for a richer and contextualized characterization of textual disclosures and provide new ways to study the fundamental attributes, properties, and outcomes of corporate disclosures (Leuz and Wysocki, 2016). My findings demonstrate that the contextualized characterization of disclosed text conveys strong and timely valuation-relevant signals that affect actions in equity markets. Therefore, a contextualized measurement of disclosures can enhance the reliability of inferences about the decision-usefulness of accounting text and offer fundamental insights into investor processing of unstructured information signals in capital markets (Blankespoor et al., 2020). My findings have implications for researchers, regulators, and practitioners that attempt to understand the decision-usefulness of corporate disclosures and their use by investors.

Finally, the single information variable that I measure can be applied to a wide array of textual documents (e.g., regulatory filings, analyst reports, managerial conference calls, and CSR reports) to examine issues including: the informativeness of disclosures over longer event horizons to capture a richer set of accounting information within unstructured textual contents; the time-series of disclosures' decision-relevance; the conditional complementarity and substitutability between the reporting and disclosure systems in capital markets; and the provision of and response to asymmetric disclosures.

References

Ball, R., Brown, P., 1968. An empirical evaluation of accounting income numbers. Journal of Accounting Research 6, 159–178.

Barth, M. E., Li, K., McClure, C., 2022. Evolution in value relevance of accounting information. The Accounting Review, Forthcoming.

Basu, S., 1997. The conservatism principle and the asymmetric timeliness of earnings. Journal of Accounting and Economics 24, 3–37.

Beaver, W., 1968. The information content of annual earnings announcements. Journal of Accounting Research 6, 67–92.

Beaver, W., McNichols M. F., Wang, Z. Z., 2020. Increased market response to earnings announcements in the 21st century: An empirical investigation. Journal of Accounting and Economics 69, 101244.

Blankespoor, E., deHaan, E., Marinovic, I., 2020. Disclosure processing costs, investors' information choice, and equity market outcomes: A review. Journal of Accounting and Economics 70, 101344.

Bochkay, K., Brown, S. V., Leone, A. J., Tucker, J. W., 2022. Textual analysis in accounting: What's next? Unpublished Working Paper.

Catford, J C., 1978. A linguistic theory of translation: An essay in applied linguistics. Oxford University Press.

Cheng, L., Roulstone, D., Van Buskirk, A., 2021. Are investors influenced by the order of information in earnings press releases? The Accounting Review 96, 413–433.

Collins, D., S. P. Kothari, J. Shanken, Sloan, R., 1994. Lack of timeliness and noise as explanations for the low contemporaneous return-earnings association. Journal of Accounting and Economics 18, 289–324.

Das, S., 2014. Text and context: language analytics in finance. Foundations and Trends in Finance 8, 145–261.

Dechow, P., 1994. Accounting earnings and cash flows as measures of firm performance: the role of accounting accruals. Journal of Accounting and Economics 18, 3–42.

deHaan, E., 2021. Using and interpreting fixed effects models. Unpublished Working Paper.

Devlin, J., M. Chang, K. Lee, Toutanova, K., 2019. Pre-training of deep bi-directional transformers for language understanding. ArXiv Preprint arXiv:1810.04805.

Donovan, J., Jennings, J., Koharki, K., Lee, J., 2021. Measuring credit risk using qualitative disclosure. Review of Accounting Studies 26, 815–863.

Frankel, R., Jennings, J., Lee, J., 2016. Using unstructured and qualitative disclosures to explain accruals. Journal of Accounting and Economics 62, 209–227.

Frankel, R., Jennings, J., Lee, J., 2021. Disclosure sentiment: Machine learning vs. dictionary approaches. Management Science Preprint, 1–19.

French, K. R., Roll, R., 1986. Stock return variances: The arrival of information and the reaction of traders. Journal of Financial Economics 17, 5–26.

Glazer, J., Rubinstein, A., 2001. Debates and decisions: on a rational of argumentation rules. Games and Economic Behavior 36, 158–173.

Gunning, R., 1952. The technique of clear writing. McGraw-Hill International Book Co.

Healy, P. M., Palepu, K. G., 2001. Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. Journal of Accounting and Economics 31, 405–440.
Henry, E., 2008. Are investors influenced by how earnings press releases are written? Journal of Business Communication 45, 363–407.
Hirshleifer, D., 2001. Investor psychology and asset pricing. Journal of Finance 56, 1533-1597.
Holthausen, R. W., Watts, R. L., 2001. The relevance of the value-relevance literature for financial accounting standard setting. Journal of Accounting and Economics 31, 3–75.
Huang, A., Hui, W., Yang, Y., 2022. FinBERT—A large language model approach to extracting information from financial text. Unpublished Working Paper.
Jawahar, G., Sagot, B., Seddah, D., 2019. What does BERT learn about the structure of language? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.
Kothari, S. P., 2001. Capital markets research in accounting. Journal of Accounting and Economics 1, 105–231.
Kothari, S. P., Shu, S., Wysocki, P. D., 2009. Do managers withhold bad news? Journal of Accounting Research 47, 241–276.
Krupa, J., Minutti-Meza, M., 2022. Regression and machine learning methods to predict discrete outcomes in accounting research. Journal of Financial Reporting, Forthcoming.
Lee, M. C., Zhong, Q., 2021. Shall we talk? The role of interactive investor platforms in corporate communication. Journal of Accounting and Economics, Forthcoming.
Leuz, C., Wysocki, P. D., 2016. The economics of disclosure and financial reporting regulation: Evidence and suggestions for future research. Journal of Accounting Research 54, 525–622.
Lev, B., 1989. On the usefulness of earnings and earnings research: Lessons and directions from two decades of empirical research. Journal of Accounting Research 27, 153–192.
Lev, B., 2018. The deteriorating usefulness of financial report information and how to reverse it. Accounting and Business Research 48, 465–493.
Lev, B., Zarowin, P., 1999. The boundaries of financial reporting and how to extend them. Journal of Accounting Research 2, 353–385.
Li, F., 2010. Textual analysis of corporate disclosures: a survey of the literature. Journal of Accounting Literature 29, 143–165.
Li, K., 2019. Linguistic fundamentals in earnings press releases. Doctoral Dissertation, Stanford University.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: A robustly optimized BERT pretraining approach. ArXiv preprint arXiv:1907.11692I.
Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance 66, 35–65.
Loughran, T., McDonald, B., 2016. Textual analysis in accounting and finance: A survey. Journal of Accounting Research 54, 1187–1230.
Meursault, V., 2019. The language of earnings announcements. Unpublished Working Paper.
Miller, M. H., Modigliani, F., 1958. The cost of capital, corporation finance and the theory of investment. American Economic Review 48, 261–297.
Ohlson, J., 1995. Earnings, book values, and dividends in equity valuation. Contemporary Accounting Research 11, 661–687.

Ou, J., Penman, S., 1989. Financial statement analysis and the prediction of stock returns. Journal of Accounting and Economics 11, 295–329.

Roll, R., 2000. R-Squared. Journal of Finance 43, 541-566.

Rubinstein, A., 2000. Economics and language: Five essays. Cambridge University Press.

Siano, F., Wysocki, P. D., 2018. The primacy of numbers in financial and accounting disclosures: Implications for textual analysis research. Unpublished Working Paper.

Siano, F., Wysocki, P. D., 2021. Transfer learning and textual analysis of accounting disclosures: Applying big data methods to small(er) datasets. Accounting Horizons 35, 217–244.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., Pavlick, E., 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. ArXiv preprint arXiv:1905.06316.

Waymire, G., 1984. Additional evidence on the information content of management earnings forecasts. Journal of Accounting Research 22, 703–718.

Wilson, P., 1987. The incremental information content of the accrual and funds components of earnings after controlling for earnings. The Accounting Review 62, 293–322.

Appendix A

Variable Definitions

	Description	Source
Market Reaction		
CAR	Cumulative abnormal stock return computed as actual security returns minus the return on the overall market (including dividends), in the 3-day window surrounding an earnings announcement event (i.e., from day -1 to 1).	CRSP
Future Earnings		
Earn $_{Q+1}$	Next-quarter earnings before extraordinary items scaled by the market value of equity computed 4 quarters earlier.	Compustat
SD Earn $_{Q+1\text{-}Q+4}$	Standard deviation of future 4-quarter reported earnings. At least 2 quarterly earnings are required to compute the variable.	Compustat
Experimental Variables		
Text News $_{(Word\ Context)}$	The "news" contained within the text of a firm's quarterly earnings press release. The "news" is obtained from a *BERT*-based neural language model that is trained to explain 3-day abnormal stock returns based solely on the text of a firm's quarterly earnings press release. The *BERT*-based model uses word context—i.e., the ordered sequence of all elements in the disclosure—to explain abnormal stock returns.	8-K / Bus. Wire
Pred Earn $_{(Word\ Context)}$	The predicted next-quarter reported earnings obtained from a *BERT*-based neural language model based solely on the text of a firm's quarterly earnings press release. The *BERT*-based model uses word context—i.e., the ordered sequence of all elements in the disclosure—for its prediction.	8-K / Bus. Wire
Pred SD Earn $_{(Word\ Context)}$	The predicted standard deviation of future 4-quarter earnings obtained from a *BERT*-based neural language model based solely on the text of a firm's quarterly earnings press release. The *BERT*-based model uses word context—i.e., the ordered sequence of all elements in the disclosure—for its prediction.	8-K / Bus. Wire
Narrative Attributes		
Tone	Difference between Henry (2008) positive and negative words scaled by the total number of words in the document.	8-K / Bus. Wire
Fog	Gunning (1952) Fog Index computed as ([average words per sentence + percentage of complex words] * 0.4).	8-K / Bus. Wire
Length	Natural logarithm of the total number of words found in an earnings announcement text.	8-K / Bus. Wire
Numbers	Total number of numbers disclosed within the text of an earnings announcement.	8-K / Bus. Wire

Future	Total number of future-tense verbs scaled by the total number of words in the earnings announcement.	8-K / Bus. Wire
Financial Statement Items		
Earn	Quarterly earnings before extraordinary items scaled by the market value of equity computed 4 quarters earlier.	Compustat
Surprise	Earnings per share in quarter q less the mean quarter q earnings per share analyst forecast made prior to the earnings announcement date scaled by stock price at the end of quarter q. Missing values are replaced with the average by firm.	I/B/E/S
Div	Quarterly distributed dividends per share. Missing observations are replaced with a zero.	Compustat
Div Chg	Percentage difference in quarterly distributed dividends between consecutive quarters.	Compustat
Leverage	(Short-term liabilities + long term liabilities) / total assets all measured quarterly.	Compustat
Leverage Chg	Percentage difference in quarterly leverage between consecutive quarters.	Compustat
Restr	An indicator equal to "1" for quarterly restructuring expenses different from zero and "0" otherwise.	Compustat
Spec Items	The absolute value of special items scaled by quarterly assets.	Compustat
Controls		
Size	Natural logarithm of a firm's market capitalization.	CRSP
Following	Natural logarithm of the number of quarterly analyst earnings estimates. Missing observations are assigned a value of zero.	I/B/E/S
AF	An indicator equal to "1" if at least one analyst forecast is released within the earnings announcement event period (i.e., from day -1 to day 1).	I/B/E/S
Other *CAR* Predictions		
Pred CAR N-Grams	The short-window (i.e., 3-day) cumulative abnormal return modeled "out-of-sample" from top 3,000 "unigrams" and "bigrams" (i.e., "N-Grams"), by TF-IDF, extracted from quarterly earnings press releases. "N-Grams" are modeled non-linearly through a Gradient Boosting machine learning model.	8-K / Bus. Wire
Pred CAR Fin Stat Items	The short-window (i.e., 3-day) cumulative abnormal return modeled "out-of-sample" from all the "Financial Statement Items" variables (see above). The "Financial Statement Items" are modeled non-linearly through a Gradient Boosting machine learning model.	Compustat

Other Market Reactions		
AVOL	Abnormal trading volume computed as the difference between average share-outstanding-scaled volumes in the 3-day earnings announcement period (i.e., from day -1 to 1) minus average share-outstanding-scaled volumes in the non-announcement period (i.e., from day -130 to day -10 and from day 10 to day 130); the difference is then divided by the standard deviation of share-outstanding-scaled volumes in the non-announcement period (see Beaver et al., 2020).	CRSP
Vol	The standard deviation of daily abnormal returns for the 3-day period (i.e., from day -1 to 1) surrounding an earnings announcement event.	CRSP
Pred AVOL *(Word Context)*	The short-window (i.e., 3-day) abnormal trading volume modeled "out-of-sample" from a *BERT*-based neural language model based solely on the text of a firm's quarterly earnings press release. The *BERT*-based model uses word context—i.e., the ordered sequence of all elements in the disclosure—for its prediction.	8-K / Bus. Wire
Pred Vol *(Word Context)*	The short-window (i.e., 3-day) abnormal stock return volatility modeled "out-of-sample" from a *BERT*-based neural language model based solely on the text of a firm's quarterly earnings press release. The *BERT*-based model uses word context—i.e., the ordered sequence of all elements in the disclosure—for its prediction.	8-K / Bus. Wire

Appendix B

Textual Parsing

1. Downloading and Parsing of Earnings Announcement Texts

I download 8-K filings from SEC EDGAR using Python and the *sec-edgar-downloader* package (https://pypi.org/project/sec-edgar-downloader/). I download all 8-K filings for all available CIK identifiers within Compustat Fundamentals Quarterly within +/- 6 calendar days of the quarterly earnings announcement date.

I parse the HTML version of the downloaded 8-K filings using Python and the *Beautiful Soup* library (https://beautiful-soup-4.readthedocs.io/en/latest/). I extract the filing date using regular expressions searching for *"FILED AS OF DATE"* and then identify the earnings press release based on a (i) starting marker (e.g., "EX-99.1", "REPORTS", "EARNINGS") and an (ii) ending marker (e.g., "GRAPHIC"). I test the parsing algorithm on 30 disclosures and find that it accurately identifies the earnings announcement text in all instances. I discard documents for which starting and/or ending markers cannot be reliably identified.

I also download earnings press releases from Dow Jones Factiva. I collect textual documents from the Business Wire U.S. repository. I identify end-of-document markers (e.g., "DOCUMENT BWR0") and separate each earnings announcement. I also scrape each disclosure using regular expressions to extract the company's Ticker symbol (also verified by a firm's name) and the earnings announcement date.

I finally retain only non-duplicate SEC EDGAR and Business Wire U.S. textual disclosures and exclude documents with less than 5 sentences or 100 words – both identified using Python and the *NLTK* library (https://www.nltk.org/).

2. Identification of Tables and Generic Statements

I split each textual disclosure into sentences using Python and the NLTK library (https://www.nltk.org/). I test the sentence tokenization algorithm for 30 disclosures and find a

93% parsing accuracy. I then classify a sentence as “table” whenever it contains more than 20 non-breaking spaces (i.e., “\xa0”), dash symbols (i.e., “-“), or plus symbols (i.e., “+”). I exclude sentences classified as “tables” from the main analyses (see Section 3.1 of the paper for rationales and robustness checks).

I identify and exclude generic cautionary statements (see Section 3.1 of the paper for rationales and robustness checks) using regular expressions that match multiple tokens (e.g., “CERTAIN STATEMENTS”, “CAUTIONARY STATEMENTS”, “WORDS SUCH AS ‘BELIEVE’, ‘MAY’, ‘WILL’, ‘EXPECT’”). One example of cautionary statements for Ocean Bio-Chem Inc. is reported thereafter.

Ocean Bio-Chem Inc. (CIK: 350737) – Second Quarter 2017
(Announcement Date: August 14th, 2017)

“*Certain statements contained in this Press Release including without limitation, Company performance in the second half of 2017, the Company’s entry into the pet market and commencement of production in the expanded portion of the Company’s plant, constitute forward-looking statements. For this purpose, any statements contained in this report that are not statements of historical fact may be deemed forward-looking statements. Without limiting the generality of the foregoing, words such as "believe," "may," "will," "expect," "anticipate," "intend," "could" including the negative or other variations thereof or comparable terminology are intended to identify forward-looking statements.*”

Since long documents tend to display relatively more tables with non-standard formatting and relatively longer cautionary or generic discussions (that are more likely to go undetected using the prior algorithms), I exclude the last (and least informative) 512 tokens from disclosures with an above-average number of tokens. As a sensitivity check, I also run the main analyses without excluding the last 512 tokens for long documents; *Text News* $_{(Word\ Context)}$ (i.e., word-context-based news) exhibits a moderately lower correlation with *CAR* (i.e., 0.35 vs. 0.39) but the tenor of the results does not change. Note that 512 tokens represent the maximum sequence length that *BERT-based* models can process – see Section 3.3 of the paper and Appendix C for details about the method I implement to still be able to consider the entire earnings announcement text.

3. Identification of Numerical Tokens

I identify and count a number in the following cases: (i) a numerical substring is preceded by a dollar sign (“$”); (ii) a numerical substring is followed by the words million/billion/trillion; (iii) a numerical substring is followed by a percentage sign (“%”) or by the words “percent”/”pct”. I also identify numbers in parentheses (negative sign) and/or for which the previous markers (i) are preceded by one or two white spaces; (ii) are not preceded by any white spaces; (iii) are capitalized, fully or in part (applies to words).

4. Word Lists

I use the Henry (2008) word list to compute *Tone*. To perform tests involving performance-related words (see Sections 3.5 and 5.4), I tabulate and use the 100 most frequent accounting performance terms) found within earnings disclosures (note that some of these keywords overlap with the Henry 2008 *Tone* words):

Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
million	2.1%	months	0.3%	more	0.1%	record	0.1%
quarter	2.0%	loss	0.3%	services	0.1%	acquisition	0.1%
year	0.9%	growth	0.3%	stock	0.1%	continue	0.1%
net	0.8%	business	0.2%	strong	0.1%	development	0.1%
share	0.7%	revenues	0.2%	lower	0.1%	executive	0.1%
compared	0.7%	operations	0.2%	continued	0.1%	investment	0.1%
income	0.6%	adjusted	0.2%	nine	0.1%	comparable	0.1%
sales	0.5%	interest	0.2%	decrease	0.1%	reports	0.1%
company	0.5%	reported	0.2%	performance	0.1%	value	0.1%
second	0.4%	expenses	0.2%	full	0.1%	profit	0.1%
ended	0.4%	tax	0.2%	ebitda	0.1%	shares	0.1%
results	0.4%	related	0.2%	product	0.1%	customers	0.1%
increased	0.4%	costs	0.2%	segment	0.1%	debt	0.1%
increase	0.4%	up	0.2%	capital	0.1%	portfolio	0.1%
operating	0.4%	margin	0.2%	well	0.1%	markets	0.1%
diluted	0.4%	expense	0.1%	six	0.1%	charges	0.1%
revenue	0.4%	basis	0.1%	decreased	0.1%	guidance	0.1%
period	0.3%	three	0.1%	loans	0.1%	measures	0.1%
fourth	0.3%	billion	0.1%	exhibit	0.1%	service	0.1%
fiscal	0.3%	higher	0.1%	loan	0.1%	flow	0.1%
earnings	0.3%	products	0.1%	offset	0.1%	improved	0.1%
financial	0.3%	market	0.1%	cost	0.1%	production	0.1%
percent	0.3%	gross	0.1%	impact	0.1%	losses	0.1%
cash	0.3%	average	0.1%	information	0.1%	sale	0.1%
gaap	0.3%	release	0.1%	result	0.1%	equity	0.1%

To run the interpretability tests described within Section 3.5 of the paper, I delete from earnings disclosures the set of "stop-words" (i.e., connecting words) listed within the spaCy library (https://spacy.io/).

5. Other Textual Attributes

To compute the Gunning (1952) *Fog* (i.e., "readability") index, I tokenize each disclosure in words and sentences using Python and the *NLTK* library (https://www.nltk.org/). I also classify "complex" words (i.e., words with more than two syllables) using Python and the *Pyphen* library (https://pypi.org/project/pyphen/).

To compute *Future* and identify future-tense verbs, I use Python and the part-of-speech (POS) tagger tool provided by the *NLTK* library (https://www.nltk.org/).

Appendix C

BERT-based Modeling Details

I fine-tune a pre-trained *RoBERTa* neural language model that can be freely downloaded from *GitHub* (https://github.com/pytorch/fairseq/tree/main/examples/roberta). In particular, I use the "RoBERTa Large" model and the *Pytorch* framework (https://pytorch.org/) for fine-tuning.[27]

I fine-tune (*RoBERTa*) *BERT-based* neural language models using a "learning rate" of 1e-6 and 3 "training epochs".[28] The choice of this set of hyperparameters largely follows the procedure and implementation described in Siano and Wysocki (2021) for earnings announcement texts.

For the main prediction tasks related to contemporaneous abnormal stock returns (i.e., 3-day *CAR*) the average computational time for fine-tuning is 4 hours per training epoch (i.e., about 12 hours in total) using an NVIDIA Tesla A100 GPU accessible through the Google Cloud Platform and Google Colab (https://research.google.com/colaboratory/faq.html). The "out-of-sample" modeling/prediction time amounts to 3 hours for each outcome of interest.

For tasks involving the prediction of *future* outcomes (i.e., future-quarter earnings and the standard deviation of future 4-quarter earnings), which are highly serially correlated (see also Section 3.3), I fine-tune a *RoBERTa* model each year between 1994 and 2019 using the disclosure observations over the prior 5 years (e.g., 1989-1993 to fine-tune the model in 1994, and so on). The fine-tuning time for each outcome is about 24 hours. The "out-of-sample" modeling/prediction

[27] I opt for a RoBERTa "Large" (as opposed to a "Base") model to maximize the modeling of contextual connections across parts of speech, that benefit from a higher number of trainable parameters. Notably, current and widely accessible computational resources make RoBERTa "Large" fine-tuning feasible.

[28] The *learning rate* represents the size of the step through which a loss or cost function is minimized. Too high learning rates could cause the global minimum of the cost function to be missed in the optimization process, while too large ones may increase computational time and render the model's training unfeasible. *Training epochs* can be thought of as the number of times that the entire dataset (training) is passed through the neural network's artificial neurons. Epochs are needed to properly minimize the models' loss function through an iterative process of gradient descent. The choice of the number of epochs is critical: too few iterations do not allow the model to properly minimize the loss or cost function, while too many iterations generally lead to in-sample overfitting and low out-of-sample accuracy.

time amounts to about 5 hours, for each outcome of interest. For this particular set of tasks, involving future earnings, I do not independently optimize the models' primary hyperparameters (i.e., learning rate and number of training epochs); therefore, the resulting prediction accuracy might underestimate the information content of earnings press releases for future earnings prediction. I make this choice for two reasons. First, the objective of tests involving future earnings is to offer generalizable insights into a channel through which word-context-based disclosures affect investor decision-making. Second, identifying optimal hyperparameter for this task would represent a computationally intensive analysis (i.e., it would require designing validation sets and searching through permutations of hyperparameters for more than 25 model-years).

BERT-based language models (including *RoBERTa*) can process textual sequences of maximum 512 tokens. To overcome this limit and offer a more complete characterization of contextualized news within corporate disclosures, I divide earnings announcements longer than 512 tokens into "windows" or subsequences of (maximum) 512 tokens. Each "window" or subsequence overlaps with the prior "window" or subsequence for 20% of the tokens (i.e., I define a "stride" of 80%).[29] Each generated subsequence is used for fine-tuning purposes. During "out-of-sample" modeling, *RoBERTa* outputs a prediction for each "window" or subsequence. I consider them all by computing the arithmetic mean of the predictions for all "windows", or subsequences, within an earnings disclosure.

[29] The "stride" can be defined as the distance chosen to slide the window when generating textual subsequences. The choice of the stride, as usual, should balance computational time and the model's accuracy. Smaller strides allow the model to learn from a larger number of examples and textual sequences but also require higher fine-tuning time.

Appendix D

Alternative Gradient Boosting Modeling of Text and Reported Numbers

BERT-based models are not the only approach available to characterize the text of corporate disclosures. In fact, several alternative techniques can be applied to extract signal from text and predict ("out-of-sample") an outcome of interest. I follow Frankel et al. (2021) and characterize corporate disclosures through one and two-word phrases (also known as "N-Grams"). I pre-process text in 5 steps (again, following Frankel et al., 2021); I (i) convert text into lowercase characters, (ii) remove "stop-words" included in the "spaCy" library, (iii) remove numbers and punctuation, (iv) lemmatize words,[30] and (v) select the top 3,000 one and two-word "N-Grams" based on the TF-IDF protocol.[31] I finally use "N-Grams" as inputs to a non-linear Gradient Boosting model to predict, through disclosure text, 3-day cumulative abnormal stock returns.

Gradient Boosting is a recent machine learning algorithm, based on regression trees, already employed in financial accounting research (e.g., Krupa and Minutti-Meza, 2021). I choose Gradient Boosting after comparing its predictive accuracy and computational time with alternative techniques: Support Vector Regressions (e.g., Frankel et al., 2016) and Random Forest (e.g., Frankel et al., 2021; Barth et al., 2022). I find that Gradient Boosting offers, on average, a 0.8% higher Adjusted R^2 relative to Support Vector Regressions and Random Forest approaches (i.e., with an average improvement of approximately 30% in terms of model fit); moreover, it requires

[30] "Lemmatization" is one of the most common pre-processing techniques in natural language processing because it allows to link semantically and morphologically similar words to a common underlying construct. For example, the words "report" and "reports" would be lemmatized as "report". Frankel et al. (2021) uses "Stemming" a similar technique that removes words suffixes but ignores the morphological role of words. For example, the words "result", "resulting", and "resulted" would be stemmed as "result". In sensitivity analyses, I also use "stemming" and find largely consistent evidence.

[31] TF-IDF (i.e., term frequency-inverse document frequency) is a natural language processing technique to evaluate how relevant is a word in a document based on its occurrence in a larger collection of documents (see, within the financial accounting domain, Loughran and McDonald, 2011). More infrequent words are considered to be more relevant and are thus given more weight. Research shows that this approach generally produces a better regression fit in natural language processing tasks (e.g., Loughran and McDonald, 2011 and 2016). Since Frankel et al. (2021) uses all one and two-word phrase N-Grams, I test the sensitivity of my results to selecting also top 6,000 and 9,000 N-Grams by TF-IDF; I find that results are largely unaffected by this choice.

a generally lower computational time (i.e., approximately 20 minutes for the main prediction task as opposed to an average of about 30 minutes for Support Vector Regressions and Random Forest). I choose Support Vector Regressions and Random Forest hyperparameters following Frankel et al. (2016) and Frankel et al. (2021), respectively. In particular, for the Random Forest protocol, I choose 5,000 regressions trees and a maximum number of textual features for each tree equal to the square root of the total number of N-Grams. In the same vein, I implement Gradient Boosting using 5,000 regression trees, 1,000 training iterations, and a 0.01 learning rate.[32] In all instances, I use a training/prediction protocol and corresponding observations that are identical to those used for *BERT*-based fine-tuning and "out-of-sample" prediction. For all the algorithms, I use widely available *Python* software implementations: (i) LightGBM Gradient Boosting, (ii) Sklearn LinearSVR, and (iii) Sklearn RandomForestRegressor.[33]

I apply a Gradient Boosting protocol (with the same hyperparameters described above) to predict, through the financial statement items (see Appendix A), 3-day cumulative abnormal returns around earnings announcement dates. Again, I use a training/prediction protocol and corresponding observations that are comparable to those used for *BERT*-based fine-tuning and "out-of-sample" prediction.

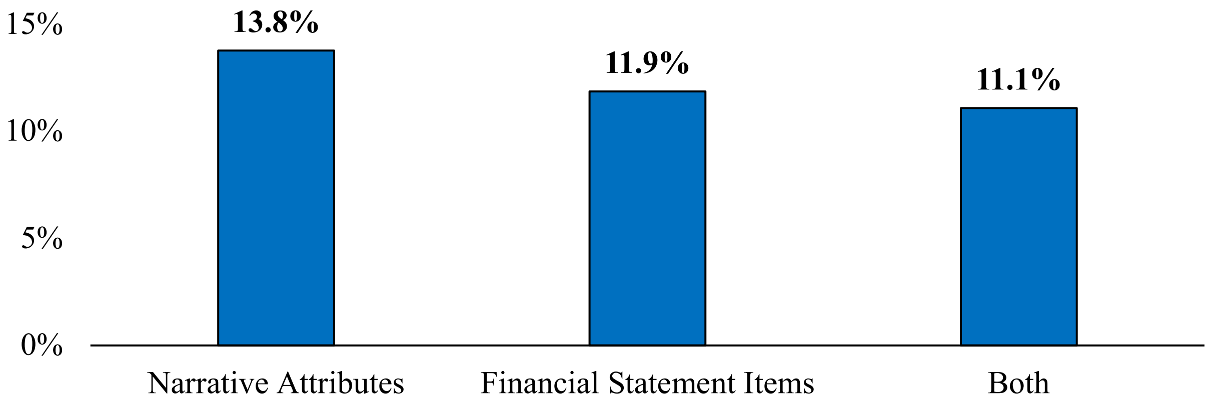
---

32 I use a "hold-out" sample of 500 earnings announcement press releases (i.e., a validation set never used neither for training/fine-tuning nor for "out-of-sample" prediction) to select the best combination of training iterations and learning rate using a grid search algorithm.

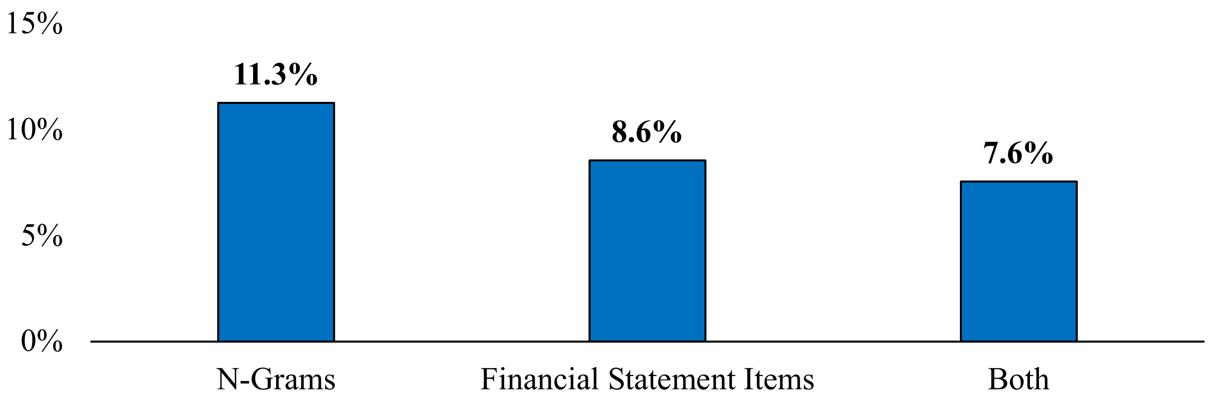
33 Available at:
(i) https://lightgbm.readthedocs.io/en/v3.3.2/#,
(ii) https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVR.html, and
(iii) https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

Figure 1
Incremental Explanatory Power of *BERT*-based Text News
for Short-Window Market Reactions (i.e., 3-day *CAR*)

Panel A: OLS R^2 incremental to text and financial statement items modeled linearly



Panel B: OLS R^2 incremental to text and financial statement items modeled non-linearly



This figure presents the incremental OLS explanatory power (i.e., incremental "Adjusted R^2") of *BERT*-based *Text News* $_{(Word\ Context)}$ for short-window market reactions (i.e., 3-day *CAR*) around earnings announcements. *Text News* $_{(Word\ Context)}$ is the "news" contained within the text of a firm's quarterly earnings press release and it is obtained from a *BERT*-based neural language model that is trained to explain 3-day abnormal stock returns based solely on the text of a firm's disclosure. The *BERT*-based model uses word context—i.e., the ordered sequence of all elements in the disclosure—to explain abnormal stock returns. Panels (A) [B] examine the incremental OLS "Adjusted R^2" of *Text News* $_{(Word\ Context)}$ relative to (i) alternative characterizations of disclosure text, (ii) reported numbers, and (iii) both (modeled "in-sample") [modeled "out-of-sample" using Gradient Boosting, a non-linear machine learning model]. In Panel (A) [B], disclosure text is characterized using (traditional narrative attributes, such as *Tone*) [one and two-word phrases, also known as "unigrams" and "bigrams" or, more in general, "N-Grams"]. The appendices describe the variables, *BERT*-based model, and Gradient Boosting protocol.

Table 1
Sample Selection

	Number
EDGAR 8-K filings and Business Wire quarterly earnings announcements (1989-2020)	261,124
Less:	
- observations with fewer than 5 sentences or 100 total words	(1,982)
- observations with missing return data	(18,216)
- "penny stocks" and firms with less than $10M in total assets	(9,284)
Total financial disclosures available for training and out-of-sample modeling	**231,642**
Stratified Training Sample	115,821
> Percentage of the total sample	*50%*
Out-of-sample observations	**115,821**
> Percentage of the total sample	*50%*
Less:	
- observations with missing financial items or firm characteristics	(14,409)
Total number of firm-quarters available for empirical tests	**101,412**

This table presents selection criteria for the final sample of 101,412 firm-quarter observations with available (i) SEC EDGAR 8-K filings or Business Wire U.S. earnings announcements texts, (ii) return data, and (iii) financial fundamentals. The final sample comprises textual observations characterized by a minimum of 5 sentences and 100 words. "Penny stocks" are defined as stocks with a unit price below $1. A stratified random selection procedure is used to construct the samples for *BERT-based* training (i.e., 50% of the total number of observations) and "out-of-sample" modeling (i.e., the remaining 50% of observations). Stratification involves quarter-years. Appendix C describes the *BERT*-based model choice and implementation details.

Table 2
Descriptive Statistics

Panel A – Observations by quarter, year, and industry

Fiscal Quarters	N	Firms	*Industries – Top/Bottom 5*	N	Firms
Quarter 1	24,265	6,907	Banking	13,024	1,129
Quarter 2	25,813	7,106	Business Services	11,284	1,123
Quarter 3	25,737	7,066	Pharmaceuticals	7,015	778
Quarter 4	25,597	7,229	Electronic Equip.	5,979	453
			Retail	4,694	368
Fiscal Years			Tobacco	105	7
			Fabricated Products	177	17
1989-2000	9,485	2,518	Shipbuilding	191	17
2001-2010	48,018	6,112	Defense	201	10
2011-2020	43,909	5,201	Precious Metals	221	29

Panel B – Variables

Primary Variables	N	Mean	Std. Dev.	p25	Median	p75
Market Reaction						
CAR	101,412	0.0006	0.0826	-0.0382	0.0002	0.0397
Future Earnings						
Earn $_{Q+1}$	97,914	0.0027	0.0349	-0.0030	0.0110	0.0203
SD Earn $_{Q+1\text{-}Q+4}$	96,674	0.0230	0.0360	0.0039	0.0090	0.0240
Experimental Variables						
Text News *(Word Context)*	101,412	0.0017	0.0242	-0.0080	0.0030	0.0130
Pred Earn *(Word Context)*	97,914	0.0030	0.0229	-0.0016	0.0115	0.0174
Pred SD Earn *(Word Context)*	96,674	0.0214	0.0216	0.0083	0.0150	0.0247
Narrative Attributes						
Tone	101,412	0.0199	0.0153	0.0090	0.0188	0.0298
Fog	101,412	18.2907	3.2368	16.1834	17.8409	19.7872
Length	101,412	6.6054	0.3803	6.4614	6.6732	6.8297
Numbers	101,412	36.3124	19.2139	22.0000	34.0000	48.0000
Future	101,412	0.0036	0.0036	0.0009	0.0027	0.0056
Financial Statement Items						
Earn	101,412	0.0015	0.0505	-0.0026	0.0112	0.0204
Surprise	101,412	0.0085	0.1303	-0.0114	0.0088	0.0367
Div	101,412	0.0902	0.1670	0.0000	0.0000	0.1200
Div Chg	101,412	-0.8339	19.6965	0.0000	0.0000	0.0110
Leverage	101,412	0.2197	0.2133	0.0336	0.1716	0.3404
Leverage Chg	101,412	2.4474	39.2841	-5.0186	0.0000	2.7171
Restr	101,412	0.2149	0.4107	0.0000	0.0000	0.0000
Spec Items	101,412	0.0052	0.0350	0.0000	0.0000	0.0014
Controls						
Size	101,412	6.3526	1.9630	4.8925	6.2699	7.6998
Following	101,412	1.4849	1.0646	0.6931	1.6094	2.3025
AF	101,412	0.5981	0.4902	0.0000	1.0000	1.0000
Other *CAR* Predictions						
Pred CAR N-Grams	101,412	0.0003	0.0090	-0.0048	0.0009	0.0062
Pred CAR Fin Stat Items	101,412	0.0006	0.0173	-0.0113	0.0012	0.0124

Panel B – Variables (Continued)						
Other Variables	N	Mean	Std. Dev.	p25	Median	p75
Other Market Reactions						
AVOL	101,412	0.8381	1.3168	-0.0874	0.4351	1.3434
Vol	101,412	0.0393	0.0349	0.0152	0.0289	0.0519
Pred AVOL *(Word Context)*	101,412	0.7958	0.6028	0.3214	0.6677	1.1719
Pred Vol *(Word Context)*	101,412	0.0395	0.0156	0.0275	0.0393	0.0500

This table reports the summary statistics for the primary sample of 101,412 firm-quarter observations with available disclosure and financial data. Panel A displays three partitions of the sample: (i) by fiscal quarter, (ii) by fiscal year, and (iii) by industry. Industries are defined following the Fama-French 48 classification (see https://mba.tuck.dartmouth.edu/pages/faculty/ken.french); the first 5 and the last 5 industries by number of observations are reported. Panel B includes sample statistics for the (a) primary and (b) other variables of interest. All variables are defined in Appendix A. Continuous variables are winsorized at the 1st and 99th percentiles.

Table 3
Correlations

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**	**(14)**	**(15)**	**(16)**	**(17)**	**(18)**	**(19)**	**(20)**	**(21)**	**(22)**	**(23)**
CAR	1																						
Earn $_{Q+1}$	**0.12**	1																					
SD Earn $_{Q+1\text{-}Q+4}$	**-0.04**	**-0.41**	1																				
Text News *(Word Context)*	**0.39**	**0.27**	**-0.12**	1																			
Pred Earn *(Word Context)*	**0.10**	**0.61**	**-0.34**	**0.28**	1																		
Pred SD Earn *(Word Context)*	**-0.05**	**-0.34**	**0.53**	**-0.14**	**-0.50**	1																	
Tone	**0.11**	**0.21**	**-0.20**	**0.32**	**0.29**	**-0.25**	1																
Fog	**-0.02**	**-0.09**	**0.07**	**-0.04**	**-0.12**	**0.06**	**-0.16**	1															
Length	**-0.01**	**-0.03**	**0.04**	**-0.04**	**-0.05**	**0.03**	**-0.05**	**0.46**	1														
Numbers	**0.03**	**0.13**	**-0.06**	**0.07**	**0.18**	**-0.09**	**0.05**	**0.16**	**0.60**	1													
Future	**-0.03**	**-0.09**	**0.05**	**-0.07**	**-0.12**	**0.06**	**-0.12**	**0.07**	**-0.07**	**-0.37**	1												
Earn	**0.14**	**0.53**	**-0.31**	**0.29**	**0.56**	**-0.35**	**0.20**	**-0.08**	**-0.04**	**0.10**	**-0.07**	1											
Surprise	**0.18**	**0.11**	**-0.08**	**0.21**	**0.09**	**-0.10**	**0.08**	**0.00**	**0.01**	**0.04**	**-0.03**	**0.19**	1										
Div	**-0.01**	**0.18**	**-0.14**	0.01	**0.26**	**-0.18**	**0.02**	**0.07**	**0.07**	**0.10**	**-0.08**	**0.15**	**0.05**	1									
Div Chg	0.00	**0.02**	**-0.03**	0.01	0.01	**-0.03**	**0.02**	**0.01**	0.01	**0.03**	-0.01	**0.02**	0.01	**0.19**	1								
Leverage	-0.01	**-0.04**	**0.17**	**-0.03**	**-0.04**	**0.15**	**-0.03**	**0.11**	**0.11**	**0.07**	**-0.03**	**-0.07**	**-0.03**	**0.16**	0.00	1							
Leverage Chg	**-0.02**	**-0.02**	**0.01**	**-0.04**	**-0.02**	0.00	-0.01	0.01	**0.01**	-0.01	0.01	**-0.06**	**-0.02**	**0.01**	**0.01**	**0.06**	1						
Restr	0.00	**-0.06**	**0.06**	0.01	**-0.08**	**0.05**	**-0.02**	**0.08**	**0.12**	**0.10**	0.00	**-0.07**	**0.03**	0.00	0.00	**0.10**	0.00	1					
Spec Items	**-0.02**	**-0.09**	**0.08**	**-0.07**	**-0.12**	**0.10**	**-0.06**	**0.02**	**0.02**	**-0.02**	**0.02**	**-0.28**	**-0.03**	**-0.04**	0.00	**0.01**	**0.04**	**0.08**	1				
Size	0.00	**0.23**	**-0.27**	**0.12**	**0.30**	**-0.30**	**0.19**	**0.08**	**0.19**	**0.18**	**-0.06**	**0.19**	**0.12**	**0.39**	**0.03**	**0.15**	**-0.01**	**0.20**	**-0.05**	1			
Following	**0.01**	**0.11**	**-0.18**	**0.08**	**0.14**	**-0.21**	**0.12**	**0.06**	**0.16**	**0.15**	**-0.02**	**0.09**	**0.08**	**0.17**	**0.03**	**0.08**	0.00	**0.16**	**-0.02**	**0.68**	1		
AF	**0.02**	**0.08**	**-0.15**	**0.07**	**0.10**	**-0.17**	**0.09**	**0.08**	**0.20**	**0.16**	0.01	**0.07**	**0.07**	**0.10**	**0.03**	**0.03**	0.00	**0.16**	**-0.02**	**0.48**	**0.73**	1	
Pred CAR N-Grams	**0.20**	**0.35**	**-0.23**	**0.54**	**0.44**	**-0.28**	**0.57**	**-0.10**	**-0.07**	**0.12**	**-0.11**	**0.35**	**0.15**	**0.06**	**0.02**	**-0.04**	**-0.03**	**-0.05**	**-0.09**	**0.20**	**0.12**	**0.10**	1
Pred CAR Fin Stat Items	**0.30**	**0.30**	**-0.13**	**0.41**	**0.28**	**-0.15**	**0.19**	**-0.04**	-0.01	**0.09**	**-0.06**	**0.49**	**0.60**	0.00	-0.01	**-0.05**	**-0.10**	0.01	**-0.11**	**0.14**	**0.10**	**0.09**	**0.33**

This table reports Pearson correlations for the primary sample of 101,412 firm-quarter observations with available disclosure and financial data. All variables are defined in Appendix A. Continuous variables are winsorized at the 1st and 99th percentiles. Bold coefficients indicate significance at <0.01 (two-tailed tests).

Table 4
Contextualized News and Price Revisions Around Earnings Announcements

Panel A: Linear modeling of narrative attributes and financial statement items (*N* = 101,412)

	Dependent Variable: *3-day CAR*				
	(1)	(2)	(3)	(4)	(5)
***Text News** (Word Context)*	**1.328** ***				**1.260** ***
Tone		0.594 ***		0.479 ***	-0.043 **
Fog		0.034 ***		0.039 ***	-0.000
Length		-0.009 ***		-0.005 ***	0.002 **
Numbers		0.020 ***		0.000 ***	-0.000
Future		0.005		-0.034	-0.045
Earn			0.208 ***	0.181 ***	0.057 ***
Surprise			0.102 ***	0.100 ***	0.066 ***
Div			-0.010 ***	-0.007 ***	0.002
Div Chg			-0.000	-0.000	-0.000 **
Leverage			0.007 ***	0.008 ***	0.007 ***
Leverage Chg			-0.003 ***	-0.003 ***	-0.000 *
Restr			0.001 *	0.002 **	0.001
Spec Items			0.035 ***	0.036 ***	0.031 ***
Size			-0.002 ***	-0.003 ***	-0.003 ***
Following			0.001 ***	0.002 ***	0.002 ***
AF			0.001	0.001	0.000
Fixed Effects	No	No	No	No	No
Adjusted R^2	15.15%	1.40%	4.70%	5.47%	16.56%
***Incremental Adjusted R^2 from Text News** (Word Context)*	-	**13.76%** ***	**11.85%** ***	**11.09%** ***	-

Panel B: Non-linear modeling of textual disclosures and financial statement items (*N* =101,412)

	Dependent Variable: *3-day CAR*				
	(1)	(2)	(3)	(4)	(5)
***Text News** (Word Context)*	**1.328** ***				**1.166** ***
Pred CAR N-Grams		1.800 ***		1.021 ***	-0.386 ***
Pred CAR Fin Stat Items			1.423 ***	1.249 ***	0.820 ***
Fixed Effects	No	No	No	No	No
Adjusted R^2	15.15%	3.90%	8.90%	10.01%	17.57%
***Incremental Adjusted R^2 from Text News** (Word Context)*	-	**11.27%** ***	**8.55%** ***	**7.56%** ***	-

This table presents the absolute and incremental OLS explanatory power (i.e., "Adjusted R^2") of *BERT*-based *Text News* *(Word Context)* for pricing revisions (i.e., 3-day *CAR*) around earnings announcement dates; the sample comprises N = 101,412 firm-quarter observations with available disclosure and financial data. *Text News* *(Word Context)* is the "news" contained within the text of a firm's quarterly earnings press release and it is obtained from a *BERT*-based neural language model that is trained to explain 3-day abnormal stock returns based solely on the text of a firm's disclosure. The *BERT*-based model uses word context—i.e., the ordered sequence of all elements in the disclosure—to explain abnormal stock returns. The dependent variable is 3-day *CAR*, the cumulative abnormal return for the 3-day period (day –1 to 1) surrounding an earnings announcement. Cumulative abnormal returns are computed as actual security returns minus the return on the overall market (including dividends).

Panels (A) [B] examine the incremental OLS "Adjusted R^2" of *Text News* *(Word Context)* relative to (i) alternative characterizations of disclosure text, (ii) reported numbers, and (iii) both (modeled "in-sample") [modeled "out-of-sample" using Gradient Boosting, a non-linear machine learning model]. In Panel (A) [B], disclosure text is characterized using (traditional narrative attributes, such as *Tone*) [one and two-word phrases, also known as "unigrams" and "bigrams" or, more in general, "N-Grams"].

Continuous variables are winsorized at the 1st and 99th percentiles. Standard errors are clustered by firm and quarter-year. ***, **, * indicate significance at <0.01, <0.05, <0.10. The statistical significance of the incremental "Adjusted R^2" is assessed using a bootstrapping protocol based on 10,000 iterations (with replacement). The *Fog* and *Numbers* coefficients are multiplied by 100 for better interpretability. The appendices describe the variables, *BERT*-based model implementation, and Gradient Boosting details. The training and prediction protocol (described in Section 3.3) is comparable for the *BERT*-based and Gradient Boosting models.

Table 5
Within-Firm Contextualized News

Panel A: Linear modeling of narrative attributes and financial statement items (*N* = 100,251)

	Dependent Variable: *3-day CAR*				
	(1)	(2)	(3)	(4)	(5)
Text News ***(Word Context)***	**1.383** ***				**1.289** ***
Tone		0.657 ***		0.603 ***	-0.043 *
Fog		0.022 *		0.001 ***	0.000
Length		-0.007 ***		-0.003 **	0.003 **
Numbers		0.017 ***		0.015 ***	0.000
Future		-0.067		0.002	0.021
Earn			0.200 ***	0.180 ***	0.066 ***
Surprise			0.133 ***	0.129 ***	0.086 ***
Div			-0.004	-0.002	0.007 **
Div Chg			-0.000	-0.000	-0.000 *
Leverage			-0.000	-0.001	-0.001
Leverage Chg			-0.002 ***	-0.002 ***	-0.000
Restr			0.002 **	0.004 ***	0.001 *
Spec Items			0.043 ***	0.043 ***	0.035 ***
Size			-0.012 ***	-0.015 ***	-0.013 ***
Following			0.001	0.002 **	0.003 ***
AF			0.002 **	0.002 **	0.002 **
Fixed Effects	Firm	Firm	Firm	Firm	Firm
Adjusted R^2	16.61%	3.25%	7.51%	8.39%	18.83%
Within R^2	14.82%	1.16%	5.52%	6.41%	17.14%
Incremental Within R^2 from Text News ***(Word Context)***	-	**13.65%** ***	**11.53%** ***	**10.73%** ***	-

Panel B: Non-linear modeling of textual disclosures and financial statement items (*N* = 100,251)

	Dependent Variable: *3-day CAR*				
	(1)	(2)	(3)	(4)	(5)
Text News ***(Word Context)***	**1.383** ***				**1.193** ***
Pred CAR N-Grams		1.867 ***		1.058 ***	-0.364 ***
Pred CAR Fin Stat Items			1.552 ***	1.395 ***	0.927 ***
Fixed Effects	Firm	Firm	Firm	Firm	Firm
Adjusted R^2	16.61%	5.35%	10.88%	11.83%	19.19%
Within R^2	14.82%	3.31%	8.96%	9.93%	17.45%
Incremental Within R^2 from Text News ***(Word Context)***	-	**11.48%** ***	**8.39%** ***	**7.52%** ***	-

This table presents the absolute and incremental "within-firm" OLS explanatory power (i.e., "Within R^2") of *BERT*-based *Text News $_{(Word\ Context)}$* for pricing revisions (i.e., 3-day *CAR*) around earnings announcement dates; the sample comprises N = 100,251 firm-quarter observations with available disclosure and financial data. N = 100,251 (rather than 101,412 as in the main analyses) due to the deletion of singletons. *Text News $_{(Word\ Context)}$* is the "news" contained within the text of a firm's quarterly earnings press release and it is obtained from a *BERT*-based neural language model that is trained to explain 3-day abnormal stock returns based solely on the text of a firm's disclosure. The *BERT*-based model uses word context—i.e., the ordered sequence of all elements in the disclosure—to explain abnormal stock returns. The dependent variable is 3-day *CAR*, the cumulative abnormal return for the 3-day period (day –1 to 1) surrounding an earnings announcement. Cumulative abnormal returns are computed as actual security returns minus the return on the overall market (including dividends).

Panels (A) [B] examine the incremental OLS "Within R^2" of *Text News $_{(Word\ Context)}$* relative to (i) alternative characterizations of disclosure text, (ii) reported numbers, and (iii) both (modeled "in-sample") [modeled "out-of-sample" using Gradient Boosting, a non-linear machine learning model]. In Panel (A) [B], disclosure text is characterized using (traditional narrative attributes, such as *Tone*) [one and two-word phrases, also known as "unigrams" and "bigrams" or, more in general, "N-Grams"]. In both panels, regression models include firm fixed effects; therefore, both "Adjusted R^2" and "Within R^2" are reported.

Continuous variables are winsorized at the 1st and 99th percentiles. Standard errors are clustered by quarter-year. ***, **, * indicate significance at <0.01, <0.05, <0.10. The statistical significance of the incremental "Within R^2" is assessed using a bootstrapping protocol based on 10,000 iterations (with replacement). The *Fog* and *Numbers* coefficients are multiplied by 100 for better interpretability. The appendices describe the variables, *BERT*-based model implementation, and Gradient Boosting details. The training and prediction protocol (described in Section 3.3) is comparable for the *BERT*-based and Gradient Boosting models.

Table 6
Contextualized Disclosures and Future Earnings

Panel A: Prediction of Future Earnings Through Contextualized Disclosures (*N* = 97,914)

	Dependent Variable: ***Earn*** $_{Q+1}$				
	(1)	(2)	(3)	(4)	(5)
Adjusted R^2	37.11%	7.15%	30.94%	32.64%	42.79%
Incremental Adjusted R^2 ***from Contextualized Text***	**-**	**30.14%** ***	**11.75%** ***	**10.15%** ***	**-**
Pred Earn (Word Context)	Yes	-	-	-	Yes
Narrative Attributes	-	Yes	-	Yes	Yes
Financial Statement Items	-	-	Yes	Yes	Yes
Controls	-	-	Yes	Yes	Yes
Fixed Effects	No	No	No	No	No

Panel B: Prediction of Future Earnings Volatility Through Contextualized Disclosures (*N* = 96,764)

	Dependent Variable: ***SD Earn*** $_{Q+1\text{-}Q+4}$				
	(1)	(2)	(3)	(4)	(5)
Adjusted R^2	27.79%	4.71%	18.68%	20.00%	32.52%
Incremental Adjusted R^2 ***from Contextualized Text***	**-**	**23.65%** ***	**13.58%** ***	**12.52%** ***	-
Pred SD Earn (Word Context)	Yes	-	-	-	Yes
Narrative Attributes	-	Yes	-	Yes	Yes
Financial Statement Items	-	-	Yes	Yes	Yes
Controls	-	-	Yes	Yes	Yes
Fixed Effects	No	No	No	No	No

This table presents the absolute and incremental OLS explanatory power (i.e., "Adjusted R^2") of *BERT*-based predictions for future quarterly earnings. Panels (A) [B] present results for the prediction of (next-quarter earnings, *Earn* $_{Q+1}$; *N* = 97,914 firm-quarter observations) [next-four-quarters standard deviation of earnings, *Earn* $_{Q+1\text{-}Q+4}$; *N* = 96,764 firm-quarter observations]. In Panel (A) [B], the experimental variable is (*Pred Earn (Word Context)*) [*Pred SD Earn (Word Context)*], which represents the predicted dependent variable, (next-quarter earnings) [next-four-quarters standard deviation of earnings], from a *BERT*-based neural language model based solely on the text of a firm's quarterly earnings press release. The *BERT*-based model uses word context—i.e., the ordered sequence of all elements in the disclosure—for its prediction.

Continuous variables are winsorized at the 1st and 99th percentiles. Standard errors are clustered by firm and quarter-year. ***, **, * indicate significance at <0.01, <0.05, <0.10. The statistical significance of the incremental "Adjusted R^2" is assessed using a bootstrapping protocol based on 10,000 iterations (with replacement). The appendices describe the variables and *BERT*-based model implementation. The training and prediction protocols are described in Section 3.3.

Table 7
Contextualized Disclosures and Future Earnings at Different Horizons

	Dependent Variable: *Earn* $_{Q+1}$ (N = 97,914)					Dependent Variable: *Earn* $_{Q+2}$ (N = 96,553)				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Adjusted R^2	37.11%	7.15%	30.94%	32.64%	42.79%	32.06%	6.24%	23.14%	24.90%	35.34%
Incremental Adjusted R^2 from Contextualized Text	-	**30.14%** ***	**11.75%** ***	**10.15%** ***	-	-	**26.09%** ***	**12.03%** ***	**10.44%** ***	-
Pred Earn $_{(Word\ Context)}$	Yes	-	-	-	Yes	Yes	-	-	-	Yes
Narrative Attributes	-	Yes	-	Yes	Yes	-	Yes	-	Yes	Yes
Financial Statement Items	-	-	Yes	Yes	Yes	-	-	Yes	Yes	Yes
Controls	-	-	Yes	Yes	Yes	-	-	Yes	Yes	Yes
Fixed Effects	No	No	No	No	No	No	No	No	No	No
	Dependent Variable: *Earn* $_{Q+3}$ (N = 95,109)					**Dependent Variable: *Earn* $_{Q+4}$ (N = 93,754)**				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Adjusted R^2	30.64%	5.69%	20.71%	22.43%	33.02%	29.93%	5.21%	21.61%	23.03%	32.51%
Incremental Adjusted R^2 from Contextualized Text	-	**25.19%** ***	**12.13%** ***	**10.60%** ***	-	-	**24.92%** ***	**10.73%** ***	**9.48%** ***	
Pred Earn $_{(Word\ Context)}$	Yes	-	-	-	Yes	Yes	-	-	-	Yes
Narrative Attributes	-	Yes	-	Yes	Yes	-	Yes	-	Yes	Yes
Financial Statement Items	-	-	Yes	Yes	Yes	-	-	Yes	Yes	Yes
Controls	-	-	Yes	Yes	Yes	-	-	Yes	Yes	Yes
Fixed Effects	No	No	No	No	No	No	No	No	No	No

This table presents the absolute and incremental OLS explanatory power (i.e., "Adjusted R^2") of *BERT*-based predictions for future quarterly earnings (from Q+1 to Q+4). The experimental variable is *Pred Earn* $_{(Word\ Context)}$ which represents the predicted dependent variable modeled "out-of-sample"—using a *BERT*-based model—from the contextualized text of firms' quarterly press releases. Continuous variables are winsorized at the 1st and 99th percentiles. Standard errors are clustered by firm and quarter-year. ***, **, * indicate significance at <0.01, <0.05, <0.10. The statistical significance of the incremental "Adjusted R^2" is assessed using a bootstrapping protocol based on 10,000 iterations (with replacement). The appendices describe the variables and *BERT*-based model implementation. The training and prediction protocol (described in Section 3.3).

Table 8
Explaining Contextualized News: The Role of Word Context

Panel A: "Masking" Tests

	Dependent Variable: *3-day CAR*			
	Text News (Word Context)		*Pred CAR N-Grams*	
	Adjusted R^2	Diff. in R^2	Adjusted R^2	Diff. in R^2
Baseline i.e., the text includes the full set of ordered words	15.15%	-	4.08%	-
Deletion of "Connecting" or "Stop" Words	7.95%	7.20% ***	3.90%	0.18% ***
Text Randomization	2.64%	12.51% ***	3.64%	0.44% ***

Panel B: Alternative Training

	Dependent Variable: *3-day CAR*			
	Text News (Word Context)		*Pred CAR N-Grams*	
	Adjusted R^2	Diff. in R^2	Adjusted R^2	Diff. in R^2
Baseline i.e., the text includes the full set of ordered words	15.15%	-	4.08%	-
On Performance Words Only	3.12%	12.03% ***	2.58%	1.50% ***
On Tone Words Only	4.10%	11.05% ***	2.46%	1.62% ***

This table compares the absolute OLS explanatory power (i.e., "Adjusted R^2") of "out-of-sample" (i) *Text News (Word Context)* (i.e., the "news" contained within the text of a firm's quarterly earnings press release obtained from a *BERT*-based neural language model that is trained to explain 3-day abnormal stock returns based solely on the text of a firm's disclosure) and (ii) *Pred CAR N-Grams* (i.e., Gradient Boosting predictions of 3-day abnormal stock returns from "unigrams" and "bigrams") for pricing revisions around earnings announcements (i.e., 3-day *CAR*). Panels (A) [B] compare OLS R^2 for permutations of (the disclosure text used for "out-of-sample" prediction, i.e., "masking tests") [the disclosure text used for *both* fine-tuning and "out-of-sample" prediction, i.e., "alternative training"]. In Panel (A) [B] the "Baseline" prediction, i.e., obtained from disclosures containing all ordered words, is compared with (predictions from text with deleted "connecting" or "stop" words, and predictions from text with randomized word order) [predictions from models trained on "performance" words only, and predictions from models trained on "tone" words only]. "Stop" words are extracted from the "spaCy" library (https://spacy.io/). "Performance" words are reported in Appendix B. "Tone" words include the Henry (2008) tokens list. ***, **, * indicate significance at <0.01, <0.05, <0.10. The statistical significance of the difference in "Adjusted R^2" is assessed using a bootstrapping protocol based on 10,000 iterations (with replacement). The appendices describe the variables and *BERT*-based model implementation.

Table 9
Applying Contextualized News to Understand Features and Properties of Corporate Disclosures

Panel A: Examples to Understand Features of Corporate Disclosures

	Dependent Variable: *3-day CAR*	
	Adjusted R^2	Diff. in R^2
Baseline i.e., the text includes the full set of ordered words	15.15%	-
Quantitative Sentences	12.42%	2.73% ***
Non-Quantitative Sentences	4.56%	10.59% ***
Quantitative Sentences (Numbers Deleted)	9.63%	5.52% ***
Section 1	10.42%	4.73% ***
Section 2	8.30%	6.85% ***
Section 3	4.24%	10.91% ***

Panel B: Example to Understand Properties of Corporate Disclosures

	Dependent Variable: *3-day CAR*		
	Full Sample	"Good News" Sample	"Bad News" Sample
Text News $_{(Word\ Context)}$	0.594 ***	0.615 ***	0.633 ***
D	-0.107 ***		
Text News $_{(Word\ Context)}$ *x D*	0.147 ***		
Controls	Yes	Yes	Yes
Fixed Effects	No	No	No
Adjusted R^2	54.03%	11.77%	14.25%
Observations	101,412	50,883	50,529

This table offers examples about using *BERT-based* models and the resulting *Text News* $_{(Word\ Context)}$ (i.e., the "news" contained within the text of a firm's quarterly earnings press release obtained from a *BERT*-based neural language model that is trained to explain 3-day abnormal stock returns based solely on the text of a firm's disclosure) to (i) understand features, and (ii) examine properties of corporate disclosures.

Panel A reports *BERT*-based predictions using alternative portions of each disclosure text. First, predictions are extracted for the sample of $N = 101{,}412$ firm-quarter observations from (i) quantitative sentences, (ii) non-quantitative sentences, and (iii) quantitative sentences in which numerical tokens have been deleted. Sentences are classified as describing numbers whenever they include a dollar amount, include a percentage, or precede/follow a sentence that includes numbers. Second, predictions are extracted for a sample of N = 82,688 firm-quarter observations from text in different positions within an earnings announcement. Each earnings announcement is analyzed three times. First, the initial 512 tokens (approximately 15 sentences) are analyzed (i.e., "Section 1"). Second, the next 512 tokens (from sentence 15 to 30, approximately) are analyzed (i.e., "Section 2"). Finally, the subsequent 512 tokens (i.e., "Section 3") are considered (from sentence 30 to 45, approximately). Only earnings disclosures that are long enough to include at least three sequences of 512 tokens are analyzed (this leads to 82,688 observations). Panel B presents asymmetries in the directional magnitude and explanatory power of $Text\ News_{(Word\ Context)}$ for pricing revisions (i.e., 3-day *CAR*) around earnings announcement dates. Asymmetries are assessed with respect to "Good News" (i.e., classified as such in the presence of *CAR* >= 0) and "Bad News" (i.e., classified as such in the presence of *CAR* < 0). *D* is defined as an indicator variable that equals "1" if *CAR* < 0 and "0" otherwise. The sample comprises 101,412 firm-quarter observations. Controls include "narrative attributes", "financial statement items", and other "controls" as defined in Appendix A.

***, **, * indicate significance at <0.01, <0.05, <0.10. The statistical significance of the difference in "Adjusted R^2" is assessed using a bootstrapping protocol based on 10,000 iterations (with replacement). The appendices describe the variables and *BERT*-based model implementation.

Table 10
Applying Contextualized Measurement to Alternative Dependent Variables

Panel A: Contextualized News and Abnormal Trading Volumes (*N* = 101,412)

	Dependent Variable: *AVOL*				
	(1)	(2)	(3)	(4)	(5)
Adjusted R^2	21.11%	1.47%	8.40%	8.76%	21.32%
Incremental Adjusted R^2 from Contextualized Text	-	**19.66%** ***	**12.89%** ***	**12.56%** ***	-
Pred AVOL (Word Context)	Yes	-	-	-	Yes
Narrative Attributes	-	Yes	-	Yes	Yes
Financial Statement Items	-	-	Yes	Yes	Yes
Other Controls	-	-	Yes	Yes	Yes
Fixed Effects	No	No	No	No	No

Panel B: Contextualized News and Abnormal Stock Return Volatility (*N* = 101,412)

	Dependent Variable: *Vol*				
	(1)	(2)	(3)	(4)	(5)
Adjusted R^2	21.48%	1.20%	7.87%	8.33%	21.99%
Incremental Adjusted R^2 from Contextualized Text	-	**20.32%** ***	**14.07%** ***	**13.66%** ***	-
Pred Vol (Word Context)	Yes	-	-	-	Yes
Narrative Attributes	-	Yes	-	Yes	Yes
Financial Statement Items	-	-	Yes	Yes	Yes
Other Controls	-	-	Yes	Yes	Yes
Fixed Effects	No	No	No	No	No

This table presents absolute and incremental OLS explanatory power (i.e., "Adjusted R^2") of *BERT*-based "out-of-sample" predictions for alternative market reaction variables, abnormal trading volumes (i.e., *AVOL*) and stock returns volatility (i.e., *Vol*), around earnings announcement dates. The sample comprises *N* = 101,412 firm-quarter observations with available disclosure and financial data. In Panel A, the dependent variable is the abnormal trading volume for the 3-day period (i.e., from day -1 to 1) surrounding an earnings announcement event. Abnormal trading volumes are computed as the difference between average share-outstanding-scaled volumes in the 3-day announcement period (i.e., from day -1 to 1) minus average share-outstanding-scaled volumes in the non-announcement period (i.e., from day -130 to day -10 and from day 10 to day 130); the difference is then divided by the standard deviation of share-outstanding-scaled volumes in the non-announcement period. In Panel B the dependent variable is the standard deviation of abnormal returns for the 3-day period (i.e., from day -1 to 1) surrounding an earnings announcement event. Abnormal returns are computed as actual security returns minus the return on the overall market (including dividends). The experimental variables are *Pred AVOL (Word Context)* (in Panel A) and *Pred Vol (Word Context)* (in Panel B); they represent abnormal volumes and the standard deviation of abnormal returns modeled "out-of-sample" through contextualized earnings disclosures, respectively.

Continuous variables are winsorized at the 1st and 99th percentiles. Standard errors are clustered by firm and quarter-year. ***, **, * indicate significance at <0.01, <0.05, <0.10. The statistical significance of the incremental "Adjusted R^2" is assessed using a bootstrapping protocol based on 10,000 iterations (with replacement). The appendices describe the variables, *BERT*-based model implementation, and Gradient Boosting details. The training and prediction protocol (described in Section 3.3) is comparable for the *BERT*-based and Gradient Boosting models.