12-2022

Understanding sentiment through context

Richard M.CROWLEY
Singapore Management University, rcrowley@smu.edu.sg

M.H. Franco WONG

Singapore Management University

School of Accountancy Research Paper Series Vol. 11, No. 1

(Paper No: 2023-160)

Understanding Sentiment Through Context

Richard M. Crowley

M.H. Franco Wong

May 2023

Understanding Sentiment Through Context[†]

Richard M. Crowley
School of Accountancy
Singapore Management University

M.H. Franco Wong
Rotman School of Management
University of Toronto

December 30, 2022

[†] We gratefully acknowledge the helpful comments and suggestions by Khrystyna Bochkay (discussant), Allen Huang, Hai Lu, the seminar participants of the Chinese University of Hong Kong, Hong Kong University of Science and Technology, Indian Institute of Management Bangalore, the Norwegian School of Economics, and University of Toronto, the PhD seminar participants at Singapore Management University, and the participants at the Singapore Management University SOAR Accounting Symposium. We appreciate financial support from the Singapore Management University, Rotman School of Management, and the Social Sciences and Humanities Research Council of Canada [Grant number 435-2022-0283].

Understanding Sentiment Through Context

Abstract

We examine whether empirical results using text-based sentiment of U.S. annual reports depend on the underlying context, within documents, from which sentiment is measured. We construct a clause-level measure of context, showing that sentiment is driven by many different contexts and that positive and negative sentiment are driven by different contexts. We then construct context-level sentiment measures and examine whether sentiment works as expected at the context-level across four prediction problems. Our results demonstrate that document-level sentiment exhibits significant noise in prediction and suggest that document-level aggregation of sentiment leads to missed empirical nuances. The contexts driving sentiment results vary substantially by outcome, suggesting lower empirical internal validity for document-level sentiment. Using three additional sentiment measures, we document the same inferences, concluding that document-level aggregation likely leads to lower internal validity. Sentiment is thus best applied at the level of specific contexts rather than across whole documents.

1. Introduction

We investigate whether text-based sentiment of financial documents is related to the underlying context within the documents from which sentiment is measured and how aggregation at the document level affects sentiment measurement. Context is defined as "the parts of a discourse that surround a word or passage and can throw light on its meaning (The Merriam-Webster dictionary)" or "the text or speech that comes immediately before and after a particular phrase or piece of text and helps to explain its meaning (The Cambridge dictionary)."

This study is motivated by the prevalence of applying sentiment classification methodologies in ways that ignore the context within a document, such as through assuming each word in a document is independent and that sentiment can be aggregated across a full document. Many studies in the literature do this using a dictionary of sentiment terms to calculate sentiment; according to Bochkay et al. [2022], 97.4% of all accounting publications from 2010 to 2021 that examine sentiment use a dictionary to do so. Despite the prevalence of such approaches, financial documents contain a variety of contexts embedded within. If the context in documents is not important for sentiment analysis, then the working assumptions and practices needed will be innocuous. Indeed, prior studies have found document-level sentiment useful in various prediction problems (see surveys by Li [2010a]; Loughran and McDonald [2016]; Gentzkow, Kelly and Taddy [2019]; El-Haj et al. [2019]).

Conceptually speaking, context is important for understanding sentiment in two ways. First, context helps directly with measuring the sentiment of a phrase. For example, the word "loss" takes on a different meaning in each of the following three phrases: "net *loss* increased $3.0 million," "net *loss* was due to $4.7 million goodwill write off," and "*loss* ratio decreased to 43% as result of segment's adherence to underwriting guidelines." While all three of these phrases are accounting related, only the first two have a negative sentiment. Furthermore, the reason behind the loss for the second phrase is possibly transient, while the reason for the first loss implies that it was repeated. The third phrase is accounting related but refers to a provision or policy and is thus neutral in sentiment. Hence, the sentiment and implications we ascribe to the word "loss" varies with the context behind these three phrases. As such, context can directly help with classifying a phrase's sentiment. While dictionary methods commonly used in the literature are unable to capture context in this way, some methods do leverage collections of words to help in classifying sentiment of phrases (see, e.g.: Antweiler and Frank [2004]; Li [2010b]; Huang, Zang and Zheng [2014]; Loughran and McDonald [2020a], Azimi and Agrawal [2021]; Huang, Wang and Yang [2022]).

Second, context helps to interpret how sentiment conceptually relates to economic phenomena. For instance, consider the occurrence of future material weaknesses. Prior literature documents that more negative sentiment is associated with more material weaknesses in the next year (Loughran and McDonald [2011]). Considering just negative sentiment tied to earnings discussion, the same result is likely to be found. However, if we consider negative sentiment about cautionary statements or risk disclosures, this may indicate a company with better control systems which is likely to have *fewer* future material weaknesses. As such, the context in which the sentiment arises can lead to different economic predictions. As prior research usually measures sentiment on documents that contain more than one context, these differential interpretations across contexts get masked out, potentially causing researchers to not identify their construct of interest and to find inconsistent results.

It is an empirical question whether context is crucial for understanding sentiment to the point that its omission will induce significant deviations in empirical inferences that rely on sentiment classification. We examine this question by separating a document into its distinct contexts and measuring the sentiment of each context using several well-known sentiment classification methods. We compare these context-based sentiment measures against their context-free sentiment counterpart (document-level sentiment) in four prediction problems (stock return, volume, volatility, and material weaknesses). We also examine the effect of aggregation across contexts on sentiment measurement and the corresponding implications for using sentiment analysis in accounting research.

We develop a four-step automated approach to identify contexts from a sample of documents. First, we process each document to extract the clauses contained in the document. Second, we filter clauses to remove redundant discussion. This is done at the sentence level and aims to keep a set of clauses that are each as short as possible, yet without dropping clauses that include any accounting/finance terms or words from the Loughran and McDonald [2011] (henceforth LM) positive and negative word lists. Third, we abstract away from the language of clauses to get a representation of each clause's meaning. Lastly, we cluster across all clauses to group them by meaning. The resultant clusters represent the distinct contexts of the documents.

We conduct our analysis using 35,362 MD&A sections of 10-K filings for the period from 1994 through 2018. Using our context construction process, we identify 137 contexts, with each context containing an average of 165,469 clauses (ranging from 35,387 to 459,303 clauses). We validate the coherence of the contexts using an intrusion task: We randomly take three clauses from one context and one clause from another (the intruder). Four testers can pick out the intruder 86% of the time, on average. Moreover, we conduct an empirical validation of the context measures and find that they capture a large portion of the sentiment of the MD&As. To facilitate the interpretation of our findings, we manually label the contexts according to our reading of a random sample of clauses from each context.

We first examine whether the Loughran and McDonald [2011] sentiment measures are related to the contexts in the direction we predicted according to the observed sentiment of the clauses in the contexts. We regress the LM negative and LM positive measures of the MD&A section on the 137 contexts, using the least absolute shrinkage and selection operator (LASSO) regression method with 10-fold cross validation. We find that 92 and 79 contexts exhibit significant explanatory power for explaining LM negative and LM positive sentiment, respectively, suggesting that negative sentiment depends more on context than positive sentiment. Furthermore, we find a much higher adjusted R^2 for LM negative sentiment, which suggests that negative sentiment depends more on context than positive sentiment. This may help to explain the findings of prior studies that the LM dictionary is better at capturing negative sentiment than positive sentiment. Finally, we divide the contexts into four groups based on their association with the document-level LM sentiment measures: high sentiment, skewed toward negative sentiment, skewed toward positive sentiment, and low sentiment.

Next, we examine the ability of the LM sentiment measures to capture contextual meaning by regressing filing-period excess return, filing-period abnormal volume, post-filing stock volatility, and future material weaknesses on a set of 137 context-level sentiment measures. Context-level sentiment variables are created to measure the percentage of clauses within each context classified as having a negative, positive, and neutral sentiment. The filing-period excess return regressions show that six contexts have a negative effect and six have a positive impact on excess return in the negative sentiment regression, while six and two contexts exhibit statistically negative and positive coefficient estimates in the positive sentiment regression. In the neutral sentiment regression, 13 contexts are significantly associated with filing-period excess returns. Taken together, these findings suggest that sentiment from only a handful of contexts drives the document-level sentiment result, and that document-level sentiment misses some of the nuance in how sentiment relates to filing-period excess return.

The abnormal volume and stock volatility regressions tell a similar story. While we find that most statistically significant context-level sentiment measures have the expected positive effect on filing-period abnormal trading volume, the effect is concentrated in only ten contexts. For post-filing stock volatility, we find a handful of context-level sentiment measures pointing in both positive and negative directions. These findings are different from those of Loughran and McDonald [2011], which documents that a higher percentage of positive or negative words is associated with larger trading volume and stock volatility. Lastly, we document similar but more disparate relationship between future material weaknesses and sentiment. For negative sentiment, 12 contexts predict more material weaknesses, while 13 contexts predict fewer. Similarly, 12 (9) contexts positively (negatively) predict material weaknesses for positive sentiment. This latter finding is opposite to our expectation and inconsistent with the results documented in Loughran and McDonald [2011].

If our context approach randomly assigns clauses into arbitrary groups, we would expect the estimated coefficients on these contexts to have the same sign and similar magnitude as the document-level LM sentiment measures, though with some (possibly substantial) added noise. Even if the groupings were non-random, the same results should attain unless the effect of sentiment in the regressions is influenced by the source of the non-random grouping. Our context construction process is one such non-random grouping which we developed to preserve the syntax and grammar of each clause. Hence, our finding that most estimated coefficients on context-level sentiment variables are different in sign and magnitude than those on the document-level sentiment measures is consistent with sentiment being affected by context.

As a falsification test, we construct a set of context-like measures that are based on randomly assigning each clause in each document to one of 137 groups under a uniform distribution. We replicate our primary findings and find that the randomized context-like measures result in fewer statistically significant coefficients and substantially lower adjusted R^2. This finding suggests that the contexts in our primary results are empirically useful in linking sentiment to various outcomes.

In additional analyses we extend our findings to other commonly used dictionaries (the Henry [2008] and General Inquirer dictionaries), indicating that our findings are not unique to the LM sentiment dictionaries. This suggests that our results are more likely a due to document-level aggregation of sentiment or application of bag-of-words methods. To rule out bag-of-words as the primary culprit, we re-run all our tests using FinBERT in place of the LM dictionary approach. FinBERT is a neural-network approach that utilizes sentence-level context internally in its calculation of sentiment to minimize measurement error of sentiment (Huang, Wang, and Yang [2022]). Unlike our approach, however, FinBERT does not provide a sentiment measure to researchers that can be directly used in empirical tests. Using FinBERT, we continue to find that sentiment depends on context, implying that document-level aggregation of sentiment is the source of our findings.

Finally, we formally examine the implication of context aggregation for sentiment measurement. Given our finding that sentiment varies with context, we expect that measuring sentiment at the document-level will be problematic on documents that have multiple contexts embedded in them. When aggregating to the document level, sentiment measures capture variations in both sentiment and context, failing to identify the underlying construct of interest. We conduct a simulation by estimating our main analyses on samples constructed from a random selection of *n* contexts, for *n* from 2 to 136. We run this random process 1,000 times for each *n* and separately for negative and positive sentiments. The results suggest that aggregating across contexts can suppress information that is captured at the context level. By aggregating across contexts we lose some of the nuance that suggests different contexts drive different sentiment results.

This study contributes to the textual analysis literature in accounting and finance by extending the prior literature through unravelling textual sentiment to understand what it is that researchers capture when using such methods. To do this, we evaluate common sentiment classification methods by examining whether the sentiment measures work as intended regardless of the contexts in financial documents. We find evidence consistently showing that context matters in sentiment analysis. Specifically, we show that sentiment does not seem to be consistent across contexts, and positive and negative sentiments are driven by different contexts. More importantly, only a limited number of contexts exhibit predictive power for the four outcome variables we examine, and the predictive contexts vary by the outcomes being predicted. Taken together, these results suggest that sentiment captures many constructs, perhaps dependent upon the underlying contexts of the texts. We then show that the importance of context is not restricted to dictionary-based sentiment methods, indicating that sentiment as a construct must be studied more granularly. Lastly, we show that document-level aggregation suppresses context-level information and reduces the explanatory power of the sentiment measures.

Taken together, we conclude that sentiment (in the context of financial documents) does not appear to be one specific construct, but a conflation of many constructs. As such, we recommend future researchers avoid broad, document-level measurement of sentiment. Instead, we suggest researchers focus on sentiment within a specific context where the theoretical linkage between the measured construct and the extant research question is clearer. An example is Hassan, Hollander, Van Lent, and Tahoun [2019], which measures political sentiment conditioned on political risk statements in quarterly earnings conference calls.[1]

Section 2 describes the existing textual analysis methods and the methodology behind our context construction approach. Section 3 presents the research design. Section 4 reports the empirical findings. Section 5 concludes.

2. Methodology

2.1. Prior Literature: Bag-of-words, Topic Modeling, Naïve Bayes, and FinBERT

The bag-of-words method has been pervasive in the textual analysis of financial disclosures (e.g., see surveys by Li [2010a]; Loughran and McDonald [2016]; Gentzkow, Kelly and Taddy [2019]; El-Haj et al. [2019]; Loughran and McDonald [2020a]; Bochkay et al. [2022]). It involves parsing a document into its individual words (tokens) and counting the frequency of these words against attribute-specific word lists (e.g., positive and negative) to extract sentiments from the document. The word lists used in the literature vary from a few specific keywords to dictionaries with over 100 tokens.[2] The popularity of textual analysis in accounting and finance

[1] The automated process we developed to extract distinct contexts from a sample of documents provides an approach for parsing contextual meaning when a broad set of contexts is needed. It is similar in spirit to topic modelling, in that it can be used to agnostically assign text to groups based on some definition of meaning. However, our context construction approach is finer-grained, able to accurately classify short snippets of text (parts of sentences), whereas topic modelling excels at labelling large sections of text (paragraphs to full documents). As such, our approach can aid future researchers in classifying more precise context-dependent measures at the clause or sentence level.

[2] Papers focused on only a few specific keywords include Li [2006], Loughran et al. [2009], and Hassan et al. [2021a; 2021b]. Li [2006] captures the risk sentiment of 10-K annual reports using words related to risk or uncertainty, while Loughran et al. [2009] measures “sin” using ethics-related terms. Hassan et al. [2021a] and

research has led to the development of finance-specific word lists by Henry [2008] and Loughran and McDonald [2011].[3] Besides general-purpose word lists, various studies have created custom word lists to capture context-specific attributes.[4] The key assumption of the bag-of-words method is that each word is independent and, hence, it ignores word order, sentence structure, and grammar when measuring the sentiment of sentences. This assumption does not reflect how language works, but it reduces the complexity of working with text as data. Two alternative methods have been used in the literature to partially overcome this shortcoming within a bag of words framework: topic modeling and naïve Bayes.

Topic modeling is an unsupervised machine learning technique which looks for patterns in how words covary within and across documents. Topic modeling is usually implemented using Latent Dirichlet Allocation (LDA), which is bag-of-words algorithm. Dyer, Lang, and Stice-Lawrence [2017] uses LDA to identify the major topics that led to an increase in the length of 10-K reports over time. Huang et al. [2018] quantifies the information intermediary role of analysts by applying LDA to extract the common topics being discussed in both earnings conference calls and analyst reports. Brown, Crowley, and Elliott [2020] uses LDA to obtain a set of semantically meaningful topics for predicting intentional misreporting. While LDA provides some measure of the overall content in a document, a drawback of the method is that it provides only a document-level measure. Furthermore, it does not measure context at all due to

---

Hassan et al. [2021b] measure exposure to epidemics and Brexit at the firm level from quarterly earnings conference calls. Papers utilizing larger dictionaries include Tetlock [2007] and Kothari et al. [2009], which use the Harvard General Inquirer word lists that include over 1,000 tokens.

[3] Loughran and McDonald [2011] and Loughran and McDonald [2015] show that the LM dictionaries are better for analyzing the sentiment of financial documents than the general-purpose Harvard IV/General Inquirer and DICTION lists, respectively.

[4] Managerial deception or extreme negative and positive words (Larcker and Zakolyukina [2012]), competition (Li, Lundholm and Minnis [2013]), financial constraints (Bodnaruk, Loughran and McDonald [2015]), corporate culture (Audi, Loughran and McDonald [2016]), firm complexity (Loughran and McDonald [2020b]), and extreme language (Bochkay et al. 2020).

its bag-of-words implementation. As such, the algorithm is not intended for labeling contexts of short snippets of text (e.g., sentences), making it difficult to examine the context of discussion in a more fine-grained manner using LDA.[5]

Naïve Bayes is a supervised machine learning technique, in which a training dataset is used to estimate the parameters of a Naïve Bayes model to classify out-of-sample data. Antweiler and Frank [2004] manually labels 1,000 stock message board postings and then use them to train a naïve Bayes algorithm to classify posting tone. Similarly, Li [2010b] and Huang, Zang and Zheng [2014], among others, use pre-labelled training data to "teach" naïve Bayes models to interpret the content of 10-K filings and analyst research reports, respectively. While naïve Bayes uses context internally for training supervised classification, it does not provide any measure of this context to the researcher.

More recent studies have attempted to factor in word order in measurement. Azimi and Agrawal [2021] uses neural networks, another supervised learning technique, to capture the sequences and dependencies between words, and estimate the model using a training dataset with 8,000 manually labelled sentences. Shanthikumar, Wang, and Wu [2021] analyzes social media comments using Amazon Web Services' Comprehend tool that accounts for lexical and semantic information from text. Siano and Wysocki [2021] applies the BERT language model, developed by Google and pre-trained on unlabeled data, to capture context rather than words for predicting changes in sales. Siano [2022] trains a similar model to predict earnings announcement event return and future earnings. Huang, Wang and Yang [2022] train a BERT language model

[5] There are some exceptions to this, which generally rely on using tweaked LDA-like models and/or short documents for training. For example, individual social media posts on Twitter (tweets) can be labeled using the Twitter-LDA model of Zhao et al. [2011]. Since tweets are generally sentence length, this algorithm can have a similar usage to the method developed in this study on that specific source of data. However, the assumptions of the Twitter-LDA model make it difficult to apply to any other data sources. The Twitter-LDA model is validated for use in classifying firm disclosure in Crowley, Huang and Lu [2022a].

specifically for financial contexts, termed FinBERT, and use it to classify sentiment, ESG content, and forward-looking statements. While these papers do leverage context in a fine-grained manner (typically at the sentence level), the sentence-level context is only used internally by the algorithm to solve a classification problem, and only this output of that classification is human interpretable.[6] As such, these methods do not directly help researchers to understand the context of the text.

2.2. Context Construction Methodology

Our primary goal, methodologically, is to extract all clauses from our documents and assign each of them to an appropriate context in an unsupervised manner. This will then allow us to examine the impact of sentiment in a context-specific manner across all contexts contained in our documents. Using an unsupervised approach provides two distinct benefits over a manual or supervised approach. First, an unsupervised approach does not require the researchers to have complete information on what is contained in the documents – the approach will extract all relevant contexts of a sufficient prevalence in a less biased fashion.[7] As such, an unsupervised approach should provide a more comprehensive collection of contexts. Second, an unsupervised approach does not require manual classification to execute. While we do validate our measure with a manual task, future researchers are able to apply our methodology without any manual classification. As such, the method is easy to replicate and extend to related domains.

[6] BERT and neural network-based methods do capture context internally, but do so in a high dimensional and non-human intelligible manner. As such, it is difficult for researchers to use such methods to examine context jointly with other measures.

[7] Unbiased in this case refers to researcher bias. Researcher bias is prevalent in supervised methodologies as researchers naturally include what they know in the classification, and anything they are not aware of cannot be included. There are still some potential biases in the methodology we outline in this section, stemming from 1) training sets used in constructing the open-source machine learning methods we implement and 2) statistical limitations where contexts with minimal representation in the data may not be included in the final output.

2.2.1. Constructing Contexts

We develop a four-stage automated approach to classify the text of the documents into distinct contexts. The first stage is clause segmentation using Stanford NLP's Open Information Extraction algorithm (Angeli, Premkumar and Manning [2015]; henceforth, Open IE), used to isolate self-contained parts of sentences. Open IE is a natural language processing method that summarises a sentence into relation triples in the form of (subject; relation verb; object), which we use to construct the clauses in the documents. The second stage is to filter out irrelevant or duplicate clauses to remove artifacts from the clause segmentation process as well as to decrease the dimensionality of the data. The third stage is to abstract away from language to the meaning of the clauses, to allow for classifying contexts in a language-agnostic manner. We do this by processing all clauses using the Universal Sentence Encoder algorithm to get a 512-dimensional representation of the clauses' meanings (Cer et al. [2018]). The final stage is then clustering clauses into contexts based on their underlying meaning – the resulting clusters are our contexts. We accomplish this using Mini-Batch K-Means (Sculley [2010]) optimized using the Gap statistic (Tibshirani, Walther and Hastie [2001]). The resulting contexts are then used throughout our analysis to examine how sentiment behaves in a context-dependent manner.

We apply our context construction method to the Management Discussion and Analysis (MD&A) sections of 10-K reports to minimize computation time.[8] We collect all annual 10-K reports from 1994 to 2018 and process each annual report using the python parser developed in Brown, Crowley and Elliott [2020], including using the same MD&A regex-based extraction

[8] Clause segmentation using Open IE applied to MD&A sections of 10-K filings takes ~6.5 days on a 6-core processor (using 11 threads) with 128GB of RAM. Processing all clauses using Universal Sentence Encoder was efficiently done on a GPU (GTX 1060) in around 2 hours. The Mini-Batch K-Means procedure takes around 1 day to run across 24 threads with 256GB of RAM (but takes longer than 1 week if using less RAM). The remaining operations described in this section take only minutes to run. All three parts of this process scale somewhat linearly (or worse) with the number of sentences processed; as such, what takes around 1 week of computation time on MD&As would take around 2 months of computation time on full text 10-K filings on the same hardware.

method. As shown in the top panel in Table 1, we have processed 208,169 annual reports (188,030 10-K filings and 20,139 10-K405 filings) to extract a total of 107,596 MD&A sections. In our empirical analysis we will further restrict to the 35,362 MD&As that match all the requirements listed in the bottom panel in Table 1.

We parse the 107,596 MD&As and obtain 179,703,756 clauses. After filtering out similar and overlapping clauses, the number of clauses drops to 48,576,229, a 73% reduction. Finally, we cluster these clauses based on their meaning to arrive at 137 clusters. We refer to these 137 clusters as the *contexts* within the 10-K MD&A filings. Appendix A describes the four stages of the context construction process in greater detail, and Appendix B details the optimization used to determine the number of contexts in the data.

2.2.2. Labelling Contexts

To interpret the contexts, we start by hand-labeling each context. To do this, we randomly pull 20 clauses from each context, interpreting them to determine a label.[9] The output of this process is presented in brief in Appendix C, showing two of the 20 clauses used for labeling each context. At the same time as the labeling exercise, we also hand-code a broader classification (presented in Appendix C as well). The labels for contexts were chosen to reflect the underlying meaning of the clauses included in each context. Intuitively, some contexts may overlap with positive or negative sentiment, but this overlap reflects natural variations in how certain types of information are discussed.[10]

---

[9] Two of the authors of this study independently labeled each context based on the 20 randomly selected clauses. The authors manually coded: 1) a name for the context, 2) a broader category for the context, and 3) any observed sentiment of the context. Any disagreement was settled through discussion by the author team.

[10] Of the 137 contexts we labeled, only seven labels contain words from the LM sentiment dictionaries. Of these, four reflect the information from the clauses directly ("*accounting losses,*" "cautionary statements," "*continuation or going concern*,' and "*losses*") and two reflect the grammatical structure of the clauses ("*modal weak statements*" and "*modal strong statements*"). One final context, "*negative accounting outcomes*" is a catchall for a variety of outcomes: impairment, inventory write-downs, AUM decreases, revenue declines, and working capital deficits, among others.

Based on our hand-coding, we find there to be four overarching types of contexts: *accounting*, *business operations*, *changes*, and *ungrouped* text. A particularly relevant set of contexts focused on *accounting* makes up 24 of the 137 contexts. Within the *accounting* contexts we find a variety of discussion on accounting policies (8 contexts), including "*accounting assumptions*,", "*fair value measurement*," and "*tax*." We also find discussion of accounting standards (2 contexts), general and balance sheet discussion (6 contexts) including discussion of "*deferred tax*" and "*cash flows*," and income statement discussion (8 contexts). Another set of highly relevant contexts is discussion of *business operations*, including 44 contexts covering everything from debt, equity and investment (8 contexts), expectations and future (6 contexts), macroeconomics (5 contexts), operations (16 contexts), outcomes (4 contexts), and structure (5 contexts). Examples of specific contexts within *business operations* include "*share transactions*," "*market risk*," "*sales of goods or assets*," "*growth*," and "*contracting*." The last relevant set of contexts includes *changes* in financial or performance figures, representing 14 contexts. In total, these business-relevant contexts comprise 82 of our 137 contexts.

The remaining contexts tend to be non-business related and focus more on grammar or usage of certain words. We find that 11 contexts are focused on specific grammar patterns such as modal weak statements, and another ten contexts relate to timeframes. Another 23 contexts relate to usage of specific words but generally lack specific useful information, and the remaining 11 contexts absorb various unrelated text. We refer to all these non-business-relevant contexts as *ungrouped* contexts – that is, they are not grouped based on the information contained in the statements, but on other aspects. These contexts are largely devoid of useful information, and thus should be of less interest. At least some of these contexts are unavoidable as natural language itself is not naturally clustered, and thus it is likely that there will be some clauses that do not match the rest. Furthermore, a fair number of the clauses within the ungrouped contexts arise from Open IE having extracted clauses that are too short to have useful information, and thus attenuation toward a specific word or pattern is inevitable. In other cases, these contexts may arise from the lack of power K-Means exhibits in clustering – it requires all clusters to be of the same shape (hyper-spherical), and thus some clauses on the boundaries of clusters may end up getting clustered together.[11]

Table 2, Panel A (Panel B) presents the most and least frequent business-relevant contexts based on the number of clauses in the context (number of documents including the context). While an ungrouped context is the most common individual context (not tabulated), purely unrelated contexts only account for 11.8% of all clauses. Business-relevant contexts account for 55.7% of all clauses, while specific grammar patterns (including timeframes and word mentions) accounts for the remaining 32.5% of all clauses. Of the business-relevant contexts, the most prevalent discussion is on "*increases in accounts,*" "*loans issued,*" "*mixed business activities,*" "*revenue recognition*," and "*sales of goods or assets*." The least frequent contexts focus on fine-grained issues like "*partnerships*," "*cautionary statements*," "*new accounting standards*," "*deferred tax,"* and "*depreciation and amortization*." Based on the number of documents represented, we find that almost all MD&A sections discuss "*increases in accounts*" and *"increases in performance,*" "*decreases in performance*", "*cash flows*," and "*financing and investment*." The least widespread topics again include "*partnerships,*" and "*deferred tax,*" though "*accounting standards*," '*energy*" and "*leases*" are also in fewer MD&As.

[11] While a more precise clustering methodology could be used, such methodologies would also impose significantly higher computation costs. As such, we view K-Means as a fairly optimal choice, as it is quick enough to run the full Gap-statistic simulation while simultaneously accurate enough to identify 82 business-relevant contexts including some granular discussion, such as "*deferred tax*," "*leases*," and "*partnerships*."

Panels C and D introduce the primary adjustment we apply to our context measure: subsetting contexts based on the LM sentiment dictionaries. Panel C shows which contexts are most and least likely to contain negative discussion, as defined by the LM dictionary. We see that the "*losses*" context has by far the most common negative context, with over 70% more clauses than the second highest context, "*decreases in value or performance."* The top two contexts are negative in over 90% of all clauses. Other commonly discussed contexts under the LM negative dictionary are about uncertainty, performance outcomes, and broader economics discussion. The least negative contexts are mostly ungrouped contexts, with only one business-relevant context in the bottom ten: "*income statement items*."

Panel D presents similar statistics for positive clauses, where a positive clause is defined as a clause with more tokens in the LM positive dictionary than it has in the LM negative dictionary. We see, rather unexpectedly, that two of the top contexts are "*tax*" and "*accounting standards*." Both do not appear to be positive topics, but they are easily explained by the inclusion of one word in the LM positive dictionary: "effective." The phrase "effective tax rate" is included in over 36,000 clauses within the "*tax*" context, and it changes the sentiment of the clauses to positive 34,000 times (94% of clauses). Similarly, the phrase "is effective" appears in nearly 13,000 clauses in the "*accounting standards*" context, changing the sentiment for all but 54 of the clauses (99.6% of clauses in the context). In both cases, the word effective is almost entirely misclassifying the sentiment of these clauses as positive. This may in part explain the lack of effectiveness of positive LM sentiment, as such misclassification adds a significant amount of noise to the measurement of positive sentiment. Looking at the other contexts, however, we see that the LM positive dictionary is picking up some useful contexts as well, such as "*changes in operating measures*" and "*growth*." In terms of the least common contexts, "*decreases in value* *or performance*" and "*losses*" are both generally negative and thus have minimal overlap with positive discussion.

2.2.3. Validation

To validate our contexts, we conduct an intrusion task following Brown, Crowley and Elliott [2020]. The task is presented as a series of multiple-choice questions, asking "Which statement does not belong?" along with presenting four clauses. Within each question, three of these clauses are from the same context, while a fourth, the intruder, is from a different context. If the contexts are intelligible, then the intruder should be identifiable at a rate better than chance. The questions include a balanced sample across the 82 business-relevant contexts (equally weighted), and the clauses presented from each context are randomly selected from a uniform distribution across the context. Four research assistants examined a set of 500 intrusion questions each, correctly classifying the intruding clause correctly 86.0% of the time, on average (ranging from 82.8% to 88.0%). Compared to the experimental results from Brown, Crowley and Elliott [2020] for topics derived from LDA applied to 10-K filings, this is a remarkably high level of accuracy, indicating that the contexts our method assigns are easily intelligible.

3. Research Design

3.1. Sample and Data

Table 1 describes the sample selection process. The sample covers the period from 1994 through 2018. The top panel shows that we start with 188,030 10-K and 20,139 10-K405 filings, of which 107,596 have a Management Discussion and Analysis (MD&A) section that we can identify. As reported in the bottom panel, the sample drops to 101,877 after removing filings that cannot be parsed by Open IE, that do not match to the Loughran-McDonald data library, or that are released too close together by the same firm. The sample further decreases to 49,812 after excluding observations without a CIK in CRSP/Compustat Merged and without data in Compustat. The final sample has 35,362 firm-years of MD&As, after we impose additional filters on market capitalization, stock price, stock return, trading volume, stock exchanges, book-to-market ratio, and word counts.

Stock return, price, trading volume, market capitalization, and trading exchange data are retrieved from CRSP, while accounting data are from Compustat. We retrieve full-text sentiment measures from the Loughran McDonald Master file, and we also construct equivalent measures for full-text and MD&As based on the Brown, Crowley and Elliott [2020] parser. We obtain material weakness counts from Audit Analytics.

Table 3 reports summary statistics of various sentiment measures, our clauses, and the dependent variables and controls used in our regressions. These variables are defined in Appendix D. As the Loughran McDonald data library only provides full-text sentiment scores, we present our calculated full-text scores alongside the full-text scores from Loughran and McDonald's data to allay any concerns that differences in parsing methodologies may drastically affect the sentiment measures. While the mean and median for our calculated negative sentiment measure is slightly lower, the mean and median for positive sentiment are quite similar. Untabulated correlations show that the negative sentiment measure we calculate is 80.3% correlated with Loughran and McDonald's, while our positive sentiment measure is 81.7% correlated with their measure. The MD&A sentiment measures are less correlated, at 44.3% and 43.9% for negative and positive sentiments, respectively, and likewise have univariate statistics that deviate a bit more for negative sentiment. This is expected because the MD&A talks about a potentially different set of issues and contexts as compared to full-text 10-K filings.

We find an average of 641.1 clauses per MD&A after completing the procedure described in Section 2.2.1. As such, there is a large amount of text, and thus context, in the average MD&A. When filtering clauses based on the LM sentiment dictionaries, we find that there are 36.6 negative and 20.1 positive clauses per MD&A, on average. Since our context construction process removes many non-sentiment bearing words, the ratio of sentiment-containing clauses to all clauses is high at 8.9% of clauses.

3.2. Empirical Models

In the empirical analysis, we use a consistent framework in constructing our regressions. For our first set of tests, which investigate the relationship between LM sentiment and context, we use regressions of the following form:

$$Sentiment_{f,t} = \alpha + \sum_{i=1}^{137} \beta_i \, Context_{i,f,t} + \gamma \cdot Controls_{f,t} + \delta \cdot Industry\ FE + \varepsilon. \qquad (1)$$

In equation (1), *Sentiment* refers to the document-level sentiment of the MD&A section of the 10-K filing for firm *f* in year *t*. The 137 *Context* measures are defined as the number of clauses in each context in each MD&A divided by the total number of clauses in that same MD&A. Our interest in these tests is to observe to what extent sentiment is driven by context, as well as to see whether contexts that generally skew negative (e.g., accounting losses) or positive (e.g., increases in performance) are empirically related to sentiment. As controls, we include the log of market value, log of the book-to-market ratio, log of share turnover, pre-event Fama-French 3-factor model alpha over the trading day window [-252, -6], where day 0 is the 10-K filing date, and an indicator for the firm being listed on the NASDAQ exchange. We also include Fama and French [1997] 48 industry fixed effects. The control variables and fixed effects are implemented following Loughran and McDonald [2011].

In the second set of tests, we examine the ability of the LM sentiment measures and contexts in predicting four outcome variables from Loughran and McDonald [2011]. As our intention is to use prior results as a setting to benchmark the importance of using context in sentiment analysis, we use a regression structure that follows from Loughran and McDonald [2011]:

$$DV_{f,t} = \alpha + \beta Sentiment_{f,t} + \gamma \cdot Controls_{f,t} + \delta \cdot Industry\ FE + \varepsilon. \quad (2)$$

The dependent variables in this specification are one of the following (where day 0 is the 10-K filing date): filing-period excess return over days [0, +3]; filing-period abnormal volume over days [0, +3], normalized as a z-score by the volume in days [-60, -6]; post-event return volatility, calculated as the root mean squared error of the Fama-French 3-factor model over days [+6, +252]; or future material weaknesses, which is the number of material weaknesses flagged by Audit Analytics in the next fiscal year. *Sentiment*, *Controls*, and *Industry FE* are the same as defined under equation (1).

To examine the LM sentiment measures across different contexts, we modify equation (2) by replacing the overall sentiment measure with our context-level sentiment measures:

$$DV_{f,t} = \alpha + \sum_{i=1}^{137} \beta_i Sentiment_{Context,i,f,t} + \cdot Controls_{f,t} + \delta \cdot Industry\ FE + \varepsilon \quad (3)$$

The dependent variables, controls, and fixed effects are the same as in equation (2). The context-level sentiment variables, $Sentiment_{Context,i,f,t}$, are measured as the number of clauses within a context classified as the given sentiment (negative or positive), divided by the number of clauses in the document. Our expectations for these measures are discussed alongside the results, as they vary by dependent variable. We also estimate equation (3) on neutral sentiment, and we define the context measure to be neutral for any clause that is not labelled as positive or negative.[12] We

[12] If we instead define neutral sentiment as the absence of any sentiment in a clause, all results are consistent.

use the LM negative and LM positive word lists in the main analysis and supplement them with other sentiment classification methods as discussed in Sections 4.3 and 4.4.

3.3. Empirical Methods

We estimate equations (1) and (3) using the least absolute shrinkage and selection operator (i.e., LASSO regression) (Tibshirani [1996]) to control for potential issues stemming from multicollinearity. Note that LASSO regression is equivalent to applying L1 regularization, which is a standard approach to reducing multicollinearity when VIFs are high. In every regression where we implement LASSO, we do so using 10-fold cross validation, and we select the model that minimizes the root mean squared error (RMSE) of the predictions on the validation samples. LASSO regression is equivalent to adding an additional penalty to the minimization operation in the regression. In other words, instead of minimizing:

$$\min_{\beta,\gamma,\delta\in\mathbb{R}} \frac{1}{N}\ |\varepsilon|_2^2\,, \tag{4}$$

we are instead minimizing the following:

$$\min_{\beta,\gamma,\delta\in\mathbb{R}} \frac{1}{N}\ |\varepsilon|_2^2 + \lambda\left[\sum|\beta|_1 + \sum|\gamma|_1 + \sum|\delta|_1\right]. \tag{5}$$

The additional term in equation (5) as compared to equation (4) represents the L1 penalty and is essentially the sum of absolute values of each coefficient in the model, scaled by λ. The penalty term λ is determined via 10-fold cross validation. To derive p-values and the adjusted R^2, we reimplement the resulting LASSO model as a linear model.[13]

One potential drawback of using LASSO in this context is that it may drop individual contexts that, while causally linked to the dependent variable, may be econometrically redundant with other measures in the regression, as demonstrated in Mullainathan and Spiess [2017]. As

[13] This approach is the same as the Post-LASSO estimator introduced by Belloni and Chernozhukov [2013].

such, LASSO may understate the number of significant contexts in the regressions. To address this, we check the robustness of our estimation results for equation (3) using the Double LASSO procedure of Belloni, Chernozhukov, and Hansen [2014]. Double LASSO is a multi-stage variant of LASSO that guarantees that causally linked independent variables will not be dropped by the LASSO procedure. We take the variables of the LASSO as the first stage regression and use the Double LASSO procedure to add back in any contexts that may be causally linked to the dependent variable either directly or through any of the measures originally selected by LASSO.

3.4. Randomized Baseline

We construct an additional set of context-like measures that are based on random assignment. This allows for comparison against a baseline model with the same structure as our primary tests but without the additional context our measures provide. We construct this by randomly assigning each clause in each document to one of 137 groups under a uniform distribution. We then aggregate all clauses within each group in a document to mimic the *Context* and $Sentiment_{Context}$ measures from regressions (1) and (3).

4. Empirical Findings

4.1. Association between Sentiments and Contexts

In the first set of the analysis, we examine how LM sentiment measures are associated with contexts discussed in the MD&A sections of 10-K filings. Using equation (1), we regress LM sentiment measures on the 137 context-level variables, with control variables and fixed effects included. If sentiment does not depend on context, then we would expect each coefficient on the context measures to be zero.[14] Table 4 column (1) summarizes the results of the LM negative

[14] An alternative specification is to regress normalized document-level sentiment on all normalized context-level sentiment measures. Under this specification, if sentiment does not depend on context, then the sum of all coefficients should be one, and as such each individual coefficient should be 1/137. We find that 85 (68) coefficients

sentiment regression following equation (1). The estimated coefficients on 92 of the 137 contexts are significantly different from zero at 5% or less (54 negatives and 38 positives). Column (2) summarizes the results of the LM positive sentiment regression, finding 79 significant contexts (40 negatives and 39 positives).[15] As such, both positive and negative sentiment depend on many contexts.

Overall, our context regressions capture 52.6% of the variation in negative sentiment and 20.4% of the variation in positive sentiment in the documents.[16] At the bottom of Table 4 we also present results using the randomized baseline discussed in Section 3.4. Using this baseline, we see that only 21.1% of the variation in negative sentiment and 9.6% of the variation in positive sentiment are captured under this specification. As such, we can attribute the difference in R^2 between our specification and the randomized baseline to the context captured by our context measures. As such, it appears that context drives 31.5% of negative sentiment and 10.8% of positive sentiment within the MD&A section of 10-K filings. Thus, negative sentiment is much more strongly linked to context than positive sentiment.

To better explain how contexts are related to sentiment, we group the contexts together based on the pattern of the estimated coefficients presented in columns (1) and (2). Our first grouping is high sentiment: contexts that include both positive and negative sentiments (i.e., exhibit a statistical positive coefficient in both the negative and positive sentiment regressions). Such

are significantly different from 1/137 at $p<0.05$ for negative (positive) sentiment, indicating that sentiment is strongly tied to context.

[15] In Section 2.2.2, we confirm the validity of the 137 contexts using an intrusion task. We also use a modified version of equation (1) to validate the contexts. We regress the MD&A's negative and positive sentiment on the amount of each MD&A dedicated to each context with that sentiment. Untabulated results indicate that our context-level negative sentiment measures capture 82.3% of the MD&A-based LM negative sentiment score, as indicated by adjusted R^2, while our context-level positive sentiment measures capture 68.6% of the MD&A-based LM positive sentiment score.

[16] When using full text document sentiment as the dependent variable, our context-level sentiment on MD&As captures 45.6% (32.4%) of the variation in negative sentiment and 46.3% (35.4%) of the variation in positive sentiment using data taken from our parser (the Loughran-McDonald Master file).

contexts may contain a higher level of sentiment, but potentially non-directionally on average. This group includes seven business-relevant contexts. Counterintuitively, both "*reduction in accounts*" and "*decreases and offsets in performance*" are included in this group despite having a more negative sentiment overall. We note that the primary driver of positive sentiment for "*reduction in accounts*" is the word "effective," accounting for nearly one-quarter of all positive sentiment in that context.

Our second grouping includes contexts that skew towards negative sentiment (i.e., exhibit a statistical positive coefficient in the negative sentiment regression and a negative coefficient in the positive sentiment regression), including 16 business-relevant contexts. For instance, this group includes discussion such as "*accounting losses*," "*negative accounting outcomes*," "*risk factor disclosures*," and "*declines in different measures*," all of which frequently include negative discussions based on our observations The included contexts in this group frequently include negative discussion, and cover matters such as loss, risks, debt, and declines, all of which can point toward negative outcomes.

Our third grouping includes contexts that skew towards positive sentiment, including 20 business-relevant contexts. Included in these contexts is "*increases in performance*," which, based on our reading, frequently includes positive sentiment discussion. The inclusion of "*accounting standards*" and "*tax*" in this group is surprising, but as discussed in Section 2.2.2, positive sentiment from both contexts is erroneously driven by the word "effective."

Lastly, we find 11 contexts we term as low sentiment contexts, as they predict lower levels of sentiment overall (i.e., the estimated coefficients in both regressions are negative). These include topics such as "*depreciation and amortization*," "*credit facilities*," and "*subsidiaries*." Perhaps counterintuitively, we also find that "*large expenses*" does not lead to more negative sentiment, though it discusses restructuring and impairment (both of which are in the LM negative dictionary).

Taken together, the results reported in Table 4 suggest that the LM sentiment measure does a better job capturing negative sentiment than positive sentiment, which is consistent with the findings of Loughran and McDonald [2011].

4.2. Predictive Power of Sentiment and Context-level Sentiment

To shed light on the ability of the LM sentiment measures to capture contextual meaning, we follow Loughran and McDonald [2011] to examine the predictive power of these measures for four outcome variables: filing period excess returns, filing period abnormal volume, post-filing return volatility, and future material weaknesses. In addition to the original LM sentiment measures for the MD&A subsection, we create three sentiment-based variables for each context. The first variable is equal to the percentage of clauses in the MD&A with positive LM sentiment in each context. The second and third variables are equal to the percentages of clauses that have LM negative and LM neutral sentiment in a context, respectively. We run regressions of each dependent variable on each set of these 137 context-level sentiment measures (i.e., positive, negative, or neutral).

4.2.1. Filing Period Excess Returns

We first examine the stock market reaction to context-level sentiment separately for negative, positive, and neutral sentiments. The filing period covers day 0 to day 3, inclusive, where day 0 is the 10-K filing date. Excess return is computed as the difference between a firm's buy-and-hold stock return and the CRSP value-weighted buy-and-hold market index return over the filing period. Following from Loughran and McDonald [2011], we expect negative sentiment to be negatively associated with the excess stock return around the 3-day filing period, while positive sentiment should exhibit a positive association with excess stock return.

Table 5 reports the estimation results. Columns (1) and (3) summarize the results from two linear regressions following equation (2) in which filing-period excess return is regressed on the LM negative and LM positive sentiment measures, respectively (i.e., a replication of Loughran and McDonald [2011]). The other columns show the results from three LASSO regressions following equation (3) in which the document-level LM sentiment measures are replaced by 137 variables capturing the percentage of clauses in the document with LM negative sentiment (column 2), LM positive sentiment (column 4), and LM neutral sentiment (column 5) in the corresponding contexts.

Column (1) shows that, on average, negative sentiment predicts a negative return over the filing period. Column (2), however, indicates that negative sentiment is only statistically significant in 12 (six negative and six positive) of the 137 contexts at the 5% level. Given that the variables are capturing the negative sentiment present in the contexts, the six positive coefficient estimates are inconsistent with the intended purpose of the negative LM sentiment measure. Similarly, column (3) finds that positive sentiment is not associated with the filing-period excess return, while column (4) reports that only two of the estimated coefficients on the 137 context-level positive sentiment measures are statistically positive (versus six negative). In other words, for two contexts a higher percentage of positive clauses is associated with higher excess returns in the filing date event window as expected, but for six of them a higher percentage of positive clauses is associated with lower excess returns. Thus, more contexts exhibit a result that is opposite to our expectation than those that support it.

Moreover, the results in column (5) show that 13 of the estimated coefficients are statistically significant (six negative and seven positive), suggesting that the market reacts to these 13 contexts, even though the LM measures consider the sentiment of these contexts to be neutral. If the primary driver of information from the text was sentiment, then we would have expected to find fewer significant coefficients (or none) on the context-level neutral sentiment measures for this regression. However, we find the neutral sentiment specification to have more significant contexts than the regressions with negative- and positive sentiment measures.

Further disconcerting for the use of sentiment in predicting filing period excess returns is its explanatory power. Of the five regressions, the neutral sentiment measures (column 5) have the highest adjusted R^2, indicating that context, rather than sentiment, is the more useful predictor of filing-period excess return.

Lastly, at the bottom panel of Table 5 we summarize the results of the Double LASSO procedure. We find that this alternative estimation method minimally affects the results – sentiment results are still mixed, few negative and positive sentiment measures exhibit significant explanatory power in the expected direction, and neutral sentiment measures continue to be the strongest predictors of filing period excess return.[17]

4.2.2. Filing Period Abnormal Volume

The second outcome variable is the abnormal volume over the 10-K filing period between day 0 and day 3, where day 0 is the 10-K filing date. Abnormal volume is computed as the average volume over the 4-day filing period and is standardized using its mean and standard deviation over the period from day -60 to day -6. We expect that both negative and positive

[17] We note that in some regressions, applying Double LASSO *decreases* the number of significant coefficients. While the Double LASSO procedure increases the number of included coefficients, to the extent that these coefficients increase multicollinearity across our independent variables, it can lead to inflated standard errors and thus less significant results.

sentiments in the MD&A are positively associated with a large abnormal trading volume over the 10-K filing period.

Table 6 summarizes the estimation results. Column (1) reports that negative sentiment has no impact on filing-period abnormal volume, but column (3) shows that positive sentiment has a negative effect. In comparison, seven contexts exhibit significant coefficient estimates for negative sentiment (column 2), four for positive sentiment (column 4), and eight for neutral sentiment (column 5) at the 5% level. The signs of the significant estimated coefficients for both negative and positive sentiment are predominantly positive (nine versus two negative). In other words, a higher percentage of negative and positive clauses in specific contexts is associated with a larger abnormal trading volume around the filings of the 10-K reports, which is consistent with the extant literature. However, this effect is concentrated in a small number of contexts, which is still inconsistent with sentiment being a consistent construct regardless of context. Furthermore, context-level neutral sentiment is the best predictor for filing-period abnormal volume, indicating again that sentiment may not serve as a strong predictor in this setting. The results using Double LASSO, summarized in the bottom panel, are consistent with the results using LASSO regression.

4.2.3. Post-Filing Return Volatility

The third dependent variable is the return volatility over the post 10-K filing period. Post-filing return volatility is the standard deviation of the errors from a Fama-French [1993] regression on daily returns on days -252 to -6 applied to data from day +6 to day +252 following the 10-K filing date. We expect that both negative and positive sentiments in the MD&A are associated with higher return volatility in the period after 10-K filing.

Table 7 column (1) shows the expected effect of the sentiment measure—more negative sentiment leads to higher volatility, but column (3) reports an insignificant relation for positive sentiment. Focusing on negative sentiment, column (2) indicates that the estimated coefficients on 21 of the 137 contexts are significant at the 5% level (eight negative and 13 positive). Similarly, column (4) shows that 21 positive sentiment measures are statistically significant (eight have negative coefficients versus 13 positive). The negative estimated coefficients are opposite to our expectation. In addition, column (5) finds that 52 of the estimated coefficients on contexts conditional on neutral sentiment are statistically significant (46 negative and six positive). This suggests that the post-filing return volatility is reacting to those 52 contexts, even though the LM sentiment measures considered the sentiment of these contexts to be neutral.

Loughran and McDonald [2011] document that a higher percentage of positive or negative words in 10-K reports is associated with a larger stock volatility in the post 10-K filing period [+6, +252], which we replicated for negative sentiment in column (1) but not for positive sentiment in column (3). In contrast, our results based on contexts-level on sentiment are mixed. While we do have a fair number of contexts for negative- and positive sentiment (13 each) that exhibit a positive association with future stock volatility, we also have eight contexts for each sentiment that present the opposite result. Furthermore, context-level neutral sentiment is the strongest predictor for stock volatility in the post-filing period. The results estimated using Double LASSO, summarized in the bottom panel, are again robust.

4.2.4. Future Material Weaknesses

The final dependent variable we examine is the number of material weaknesses in the next fiscal year, as obtained from Audit Analytics. We expect that negative sentiment in the MD&A

is associated with more material weaknesses, while positive sentiment with less material weaknesses.

Table 8 column (1) finds that negative sentiment has no power predicting material weaknesses. In contrast, column (2) indicates that 25 of the estimated coefficients on the contexts are significant at the 5% level (13 negative and 12 positive) for negative sentiment. As such, the results on negative sentiment are mixed. On the other hand, column (3) shows that positive sentiment exhibits the expected moderating effect on material weaknesses. Column (4) indicates that 21 of the coefficients are statistically significant (nine negative and 12 positive). Hence, we find more positive than negative coefficients for contexts conditional on positive sentiment, counter to both our expectations as well as the negative coefficient on positive sentiment documented in column (3). As such, the results on positive sentiment are also mixed. Column (5) presents the results of regressing future material weaknesses on context-level neutral sentiment, where we find 27 significant contexts (eight negative and 19 positive). Furthermore, neutral contexts are the strongest predictor of material weaknesses. The qualitative results using Double LASSO are again consistent with the results using LASSO regression.

4.2.5. Variation of Context Loading across Outcome Variables

If the effect of sentiment is driven by a consistent reason, then we expect that the same set of contexts will be significant in predicting the four outcome variables across Tables 5 through 8, within sentiment. On the other hand, it is plausible that different contexts have differential explanatory power for different dependent variables, i.e., that not only the context around the sentiment dictionary words matters, but also the context of the problem being examined. We will start by looking at a couple examples. First, consider the "*cautionary statements*" context—this is not necessarily a topic that most investors would have an interest in, but it may inform us about the risks the firm faces. In fact, for negative sentiment, "*cautionary statements*" is only relevant for material weaknesses, where risk is a factor. Second, the "*discussion of accounting procedures*" context is significant in explaining excess return, abnormal volume, and post-event return volatility for negative sentiment. As such, this context is consistent—this is ideally how sentiment would work on all contexts if sentiment's interpretation was not dependent on the economic context being examined.

For negative sentiment, not a single context is significant across all four dependent variables in our regressions. Overall, however, 50 different contexts are statistically significant at a 5% or lower level across the regressions, yet 35 of these are significant for only one dependent variable. Another 13 are significant in two regressions, while the contexts "*discussion of accounting procedures*" and "*decreases in expenses or performance*" are significant for three dependent variables. Consequently, there appears to be little commonality in the context-level reasoning as to why negative sentiment is related to different dependent variables. Instead, negative sentiment appears to proxy for a completely separate construct under each dependent variable.

For positive sentiment, the results are much worse. Out of the 48 different significant contexts, 43 of them are only significant for one dependent variable, while four are significant for two dependent variables. Only one context is significant across three dependent variables, and it is an ungrouped context (*"decrease"* + *unrelated statements*). As such, positive sentiment behaves very differently across regressions, and likely represents different constructs when it is significant across these regressions with different dependent variables.

4.2.6. Falsification Test

To ensure that our main results are driven by context rather than by any empirical noise that arises from our regression and data structures, we replicate our findings from Table 5 through

Table 8 using the randomized baseline discussed in Section 3.4. We find that, across all regressions, for all sentiments, and for all dependent variables, using the randomized baseline results in fewer statistically significant coefficients for contexts as well as a substantially lower adjusted R^2 as compared to our primary results (not tabulated). This finding suggests that the contexts in our primary results are capturing something empirically useful, and that number of significant coefficients on contexts in Table 5 through Table 8 is greater than random chance would suggest. As such, this provides empirical comfort that our primary results are not driven by disaggregation in general, but instead by the specific way we disaggregate sentiment into context-level measures. In other words, this falsification test suggests that the identified contexts are the driver of our results.

4.3. Other Sentiment Dictionaries

To further examine whether dictionaries and sentiment are consistent measures in the context of 10-K filings, we replicate all our findings with two additional sentiment dictionaries. First, we replicate using the Henry [2008] dictionary, which is focused more on the reporting of earnings. The replication of Table 4 (not tabulated) suggests that this dictionary has a stronger relationship between positive sentiment and context than the LM dictionary, with 42.8% of the variation in positive sentiment being explainable by contexts. Next, we replicate Table 5 through Table 8 using the Henry [2008] dictionary. The results are presented in column (1) of each panel of Table 9, while the original results using the LM dictionary are presented in column (4) of each panel. We continue to find consistent evidence that both negative and positive sentiments predict each outcome in both directions when measuring sentiment at the context level (except for abnormal volume), thereby leading to largely inconsistent results again. Furthermore, neutral sentiment continues to be better than negative and positive sentiments in predicting each outcome variable.

We further continue to document that the contexts driving the regression results for each dependent variable are different.

Second, we examine the Harvard General Inquirer (GI) dictionary that has been used in many studies in accounting and finance. Like the Henry [2008] dictionary, contexts explain more of the variation in positive sentiment using the GI dictionary than when using the LM dictionary. However, much less of the variation in negative sentiment under the GI dictionary is explainable by context – only 23.0%. When replicating the results in Table 5 through Table 8 (presented in column (2) of each panel of Table 9), we again document inconsistent results for both negative and positive sentiments and that neutral sentiment is the strongest predictor for all four LM outcome variables. Yet again, the contexts driving these results differ across dependent variables. Taken together, our findings are not specific to the LM dictionary but apply more broadly to dictionaries used in the literature.

4.4. FinBERT: Breaking Away from Bag-of-words

While our prior results all speak to drawbacks of dictionary-based sentiment measures, which mechanically ignore context when assigning sentiment, it is *ex ante* unclear if the weaknesses are due to the bag-of-word nature of dictionaries or the nature of sentiment in financial documents. To examine whether this is specifically an issue with dictionary-based sentiment or sentiment measurement more broadly, we replicate our analysis using the sentiment approach of FinBERT (Huang, Wang and Yang [2022]). FinBERT is a pretrained sentiment classification neural network based on BERT, a language model developed by Google. FinBERT assigns sentiment at the sentence level, based on the words in the sentence as well as the word order of the sentence, breaking away from the bag-of-words approach. Furthermore, FinBERT effectively uses the context within the clause to inform its sentiment measurement. As such, if context simply removes measurement error present in dictionaries, we should find results that are much more consistent with prior expectations when replicating our main results using FinBERT. Alternatively, if we continue to find mixed signs across context-level sentiment measures when using FinBERT, this will strongly suggest that our results are not due to measurement error in bag of words dictionaries, but instead due to a more fundamental link between context and sentiment.

We apply the Huggingface implementation of FinBERT to obtain the sentiment of each clause in the MD&A section of the 10-K filing.[18] Using FinBERT to replicate Table 4 (not tabulated), we find that the resulting sentiment has the highest adjusted R^2 of all tested sentiment measures from the last two subsections, at 52.6% and 47.5% for negative and positive sentiment, respectively. However, in our replication of Table 5 through Table 8 (presented in column (3) of each panel of Table 9), we find the same mixed patterns as those for the dictionary-based sentiment measures. For all regressions, only a small number of contexts is significant and in the expected direction. For all tests except the abnormal volume test, we continue to observe many results that are counter to expectations. Furthermore, neutral sentiment continues to be the strongest predictor across all tests, indicating that context still outperforms sentiment even when using a more powerful neural network-based sentiment classification method. Lastly, the significant contexts across the regression results continue to be highly variable across dependent variables, suggesting that even FinBERT-based sentiment captures more than a single construct.

Overall, the FinBERT results shows that our mixed findings are not restricted to the bag-of-words method for sentiment measurement in a financial context. Since sentiment captures different constructs when different dependent variables are regressed on it even when context is

---

[18] Available at: https://huggingface.co/yiyanghkust/finbert-tone.

used during sentiment classification, this suggests that the mixed results we document are a feature of sentiment itself. As such, measuring sentiment at the context-level is important for deriving inferences from the sentiment of financial text. Examining context-level sentiment by applying sentiment to a specific context of interest can facilitate making a more economics-driven argument for the association between sentiment and an outcome of interest.

4.5. Effects of Aggregation across Contexts

The last subsection shows that even the latest FinBERT model exhibits mixed findings across the four prediction problems. We show that sentiment captures different constructs and, hence, should be applied to a specific context in order to derive a consistent construct.

To examine how a lack of granularity impacts our results reported in Tables 5 through 8, we conduct simulations to varying the level of aggregation of a sentiment measure. Specifically, we randomly select and combine *n* contexts at various levels of *n*, from two up to 136 (which is one less than the number of contexts) and rerun the four sets of regressions. We conduct this random process 1,000 times for each *n* to get an approximate distribution of the coefficients and their significance. We run the simulation separately for positive and negative sentiments and across three groupings: all contexts; only business-relevant contexts in the Accounting, Business operations, or Changes clusters; and Ungrouped contexts.

The results of our simulations are presented in Figure 2. Each graph plots the number of times the aggregated coefficient is statistically significant across 1,000 iterations (y-axis) against the percent of contexts aggregated (x-axis). Panel A examines aggregation of negative sentiment across contexts (corresponding to the results under columns (1) and (2) in Tables 5-8). For the event return regressions (first row of Panel A), we observe the negative coefficient (solid red line) on the aggregated sentiment measure as in Table 5 column (1). However, we note that this appears to be driven by ungrouped contexts, rather than by business-relevant contexts. Furthermore, the positive coefficients (dotted blue line) we observed in our context-level sentiment regression (Table 5 column (2)) appear to be driven by the business-relevant contexts. For abnormal volume (second row) we see that aggregation swamps out any statistically significant result, consistent with the insignificant coefficient on negative sentiment in Table 6. At low level of aggregation, we have coefficients that are significant in both directions. For return volatility (third row), we see that the positive coefficient quickly attains with aggregation and is consistent across all three context groupings. Lastly, for material weakness (fourth row) we see that aggregation again swamps out any result. A negative coefficient is attained in over one quarter of simulated aggregations (250/1,000) for the all and business-relevant sets of contexts at aggregation levels between 15% and 66%.

Panel B examines aggregation of positive sentiment across contexts (corresponding to the results under columns (3) and (4) in Tables 5-8). For the event return regression, we see the insignificant coefficient on positive sentiment in Table 5 column (3) was likely due to aggregation with ungrouped contexts, as the aggregated result for business-relevant contexts becomes positive nearly 100% of the time. This positive sign is what was theoretically predicted. However, we do note that there are many contexts with negative coefficients (the red solid line) as well, especially at lower levels of aggregation. Hence, more aggregation does adversely affect the result. For abnormal volume, we find that the negative and significant coefficient documented in Table 6 is again likely driven by ungrouped contexts, and that aggregation again hides the mixed nature of the result. For return volatility, we find that the result is consistent across all sets of contexts – at low levels of aggregation there is a clearly mixed result, while higher levels of aggregation lead to only a negative and significant coefficient. Lastly, for material weaknesses we see that the negative and significant coefficient in Table 8 column (3) is consistent with the findings only by aggregating over negative contexts. Here, we find that the result is not mixed unless ungrouped contexts are considered.

Overall, the simulation provides two insights. First, aggregating across contexts, even just those that are more business-relevant, can suppress information that is captured at the context level. By aggregating contexts, we lose some of the nuance that suggests that different contexts drive different sentiment results. Second, aggregation across all contexts, including the ungrouped contexts, can lead to less significant results and even results with opposite signs to what was theoretically predicted. These results suggest examining sentiment at a more granular level – e.g., at the level of a specific context of interest. Hassan et al. [2019] provides an example of focusing on a narrow context (political discussion) when measuring (political) sentiment. They measure this using a sentiment dictionary, applying it only to words that are within a ten-word radius from a political bigram. As such, sentiment in their study is measured on specific context within a larger set of discussion.

5. Conclusion

We examine how text-based sentiment of the MD&A section of 10-K filings is related to the underlying context within the documents under different contexts. To measure context in our setting, we construct a methodology to automatically classify all contexts within a set of documents at the clause level. Using this methodology, we obtain identify 137 contexts across all MD&As in our sample. We find that the Loughran and McDonald [2011] sentiment measures are significantly driven by context, with negative sentiment being driven more by context than positive sentiment. Second, context-level positive and negative sentiment do not always relate to four outcome variables (filing-period excess return, filing-period abnormal volume, post-period stock volatility, and post-period material weaknesses) in the same direction as document-level sentiment measures. Finally, we show that different sets of contexts exhibit predictive power for different outcome variables. Taken together, this study indicates that document-level sentiment is not a consistent measure empirically. Instead, context is needed to understand the empirical implications of sentiment from financial documents.

In practice, context does matter, and different contexts empirically lead to entirely different (and sometimes opposite) results. We show that this pattern holds across two other dictionaries (Henry [2008] and the Harvard General Inquirer dictionaries), indicating a persistent issue for sentiment classification. Furthermore, we show that using a method that internally uses context in its measurement of sentiment (FinBERT) does not alleviate the inconsistencies of sentiment as a construct. As such, it appears that regardless of the approach used (bag-of-words or word order-based) to measure sentiment, researchers must carefully apply sentiment measures to text in ways that are grounded well in theory or their research question. Measuring sentiment on large sets of text combining many different contexts leads to inconsistent results that can effectively show empirical results in any direction. To avoid such issues, future research should measure sentiment on more fine-grained contexts that match to the research question that is being examined.

We suggest two ways researchers can follow to implement context empirically. For research focused on a specific type of discussion, disclosure, or economic phenomenon, restricting to discussion explicitly about the outcome of interest can serve as a fine-grained context. Such an approach is used to measure the sentiment of political discussion in Hassan et al. [2019] through using a dictionary of political terms to identify political discussion and applying sentiment to text in a narrow range around the terms. A simpler approach is taken to measure Brexit sentiment in

Hassan et al. [2021b], just using the term "Brexit" to identify relevant discussion (as it has both low Type I and II error in their sample) and applying sentiment in a narrow range around the term. For research needing to examine a broader set of discussion, the methodology we develop in this study is a feasible approach. The approach can be applied to any set of documents without needing any hand coding for model training – only for hand labeling of the output for interpretation. With this approach, researchers get a customized set of contexts specific to their documents without needing to supervise the process or construct a dictionary. In either case, the methodology should be matched to the research question of interest.

References

ANGELI, G., M. PREMKUMAR, and C. MANNING. 'Leveraging linguistic structure for open domain information extraction.' In proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (2015): 344–354.

ANTWEILER, W., and M. FRANK. 'Is all that talk just noise? Information content of internet stock message boards.' Journal of Finance 59 (2004): 1259–1294.

AUDI, R., T. LOUGHRAN, and B. McDONALD. 'Trust, but verify: MD&A language and the role of trust in corporate culture.' Journal of Business Ethics 139 (2016): 551–561.

AZIMI, M., and A. AGRAWAL. 'Is positive sentiment in corporate annual reports informative? Evidence from deep learning.' *Review of Asset Pricing Studies* (2021): Forthcoming.

BELLONI, A., and V. CHERNOZHUKOV. 'Least squares after model selection in high-dimensional sparse models.' *Bernoulli* 19 (2013): 521–547.

BELLONI, A., V. CHERNOZHUKOV, and C. HANSEN. 'Inference on treatment effects after selection among high-dimensional controls.' *The Review of Economic Studies* 81 (2014): 608–650.

BOCHKAY, K., S. BROWN, A. LEONE, and J. TUCKER. 'Textual analysis in accounting: What's next?' *Contemporary Accounting Research*, (2022): Forthcoming.

BOCHKAY, K., J. HALES, and S. CHAVA. 'Hyperbole or Reality? Investor Response to Extreme Language in Earnings Conference Calls.' *The Accounting Review* 95 (2020): 31–60.

BROWN, N. C., R. M. CROWLEY, and W. B. ELLIOTT. 'What are you saying? Using topic to detect financial misreporting.' *Journal of Accounting Research* 58 (2020): 237–291.

BODNARUK, A., T. LOUGHRAN, and B. McDONALD. 'Using 10-K text to gauge financial constraints.' *Journal of Financial and Quantitative Analysis*, 50 (2015): 623–646.

CER, D., Y. YANG, S.Y. KONG, N. HUA, N. LIMTIACO, R. St JOHN, N. CONSTANT, M GUAJARDO-CESPEDES, S. YUAN, C. TAR, and Y.H. Sung. 'Universal sentence encoder.' Unpublished paper, Google, 2018.

CROWLEY, R. M., W. HUANG, and H. LU. 'Discretionary Dissemination on Twitter.' Unpublished paper, Rotman School of Management, 2022a.

CROWLEY, R. M., W. HUANG, and H. LU. 'Executive Tweets.' Unpublished paper, Rotman School of Management, 2022b.

DYER, T., M. LANG, and L. STICE-LAWRENCE. 'The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation.' *Journal of Accounting and Economics* 64 (2017): 221–245.

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

FAMA, E. F., and K. R. FRENCH. 'Common risk factors in the returns of stocks and bonds.' *Journal of Financial Economics* 33 (1993): 3–56.

FAMA, E. F., and K. R. FRENCH. 'Industry costs of equity.' *Journal of Financial Economics* 43 (1997): 153–193.

Gentzkow, M., B. T. Kelly, and M. Taddy. 'Text as data.' *Journal of Economic Literature*, 57 (2019): 535–574.

HASSAN, T. A., S. HOLLANDER, L. VAN LENT, and A. TAHOUN. 'Firm-level political risk: Measurement and effects.' *The Quarterly Journal of Economics* 134 (2019): 2135–2202.

HASSAN, T. A., S. HOLLANDER, L. VAN LENT, M. SCHWEDELER, and A. TAHOUN. 'Firm-level exposure to epidemic diseases: Covid-19, SARS, and H1N1.' Unpublished paper, NBER, 2021a.

HASSAN, T. A., S. HOLLANDER, L. VAN LENT, and A. TAHOUN. 'The global impact of Brexit uncertainty.' Unpublished paper, NBER, 2021b.

HENRY, E. 'Are Investors Influenced by How Earnings Press Releases Are Written?' *Journal of Business Communication* 45 (2008): pp. 363–407.

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

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

HUANG, A., H. WANG, and Y. YANG. 'FinBERT—A Large Language Model for Extracting Information from Financial Text.' *Contemporary Accounting Research* (2022): Forthcoming.

KOTHARI, S. P., X. LI, and J. E. SHORT. 'The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis.' *The Accounting Review* 84 (2009): 1639–70.

LARCKER, D. F., and A. A. ZAKOLYUKINA. 'Detecting Deceptive Discussions in Conference Calls.' *Journal of Accounting Research* 50 (2012): 495–540.

LI, F. 'The Implications of Annual Report's Risk Sentiment for Future Earnings and Stock Returns.' Unpublished paper, Shanghai Jiaotong University, 2006.

LI, F. 'Textual Analysis of Corporate Disclosures: A Survey of the Literature.' *Journal of Accounting Literature* 29 (2010a): 143–65.

LI, F. 'The Information Content of Forward-Looking Statements in Corporate Filings—A naïve Bayesian Machine Learning Approach.' *Journal of Accounting Research* 48 (2010b): 1049–102.

LI, F., R. LUNDHOLM, and M. MINNIS. 'A measure of competition based on 10-K filings.' *Journal of Accounting Research*, 51 (2013): 399–436.

LOUGHRAN, T., and B. McDONALD. 'When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks' *Journal of Finance* 66 (2011): 35–65.

LOUGHRAN, T., and B. McDONALD. 'The use of word lists in textual analysis.' *Journal of Behavioral Finance* 16 (2015): 1–11.

LOUGHRAN, T., and B. McDONALD. 'Textual analysis in accounting and finance: A survey.' *Journal of Accounting Research* 54 (2016): 1187–1230.

LOUGHRAN, T., and B. McDONALD. 'Textual Analysis in Finance.' *Annual Review of Financial Economics* 12 (2020a): 357–375.

LOUGHRAN, T., and B. McDONALD. 'Measuring firm complexity.' Unpublished paper, University of Notre Dame, 2020b.

LOUGHRAN, T., B. McDONALD, and H. YUN. 'A Wolf in Sheep's Clothing: The Use of Ethics-Related Terms in 10-K Reports.' *Journal of Business Ethics* 89 (2009): 39–49.

MULLAINATHAN, S., and J. SPIESS. 'Machine learning: an applied econometric approach.' Journal of Economic Perspectives 31 (2017): 87–106.

SCULLEY, D. 'Web-scale k-means clustering.' In Proceedings of *The 19th International Conference on World Wide Web*, (2010): 1177–1178.

SHANTHIKUMAR, D., Q. A. WANG, and S. WU. 'Does Interaction on Social Media Increase or Moderate Extremeness?' Unpublished paper, University of California, Irvine, 2021.

SIANO, F. 'Contextualized News in Corporate Disclosures: A Neural Language Approach.' Dissertation, Boston University, 2022.

SIANO, F., and P. D. WYSOCKI. 'Transfer Learning and Textual Analysis of Accounting Disclosures: Applying Big Data Methods to Small(er) Data Sets.' *Accounting Horizons* (2021): 217–244.

TETLOCK, P. C. ‘Giving content to investor sentiment: The role of media in the stock market.’ *Journal of Finance* 62 (2007): 1139–1168.

TIBSHIRANI, R. ‘Regression shrinkage and selection via the lasso.’ *Journal of the Royal Statistical Society: Series B* 58 (1996): 267–288.

TIBSHIRANI, R., G. WALTHER, and T. HASTIE. ‘Estimating the number of clusters in a data set via the gap statistic.’ *Journal of the Royal Statistical Society: Series B* 63 (2001): 411–423.

ZHAO, W. X., J. JIANG, J. WENG, J. HE, E.P. LIM, H. YAN, and X. LI. ‘Comparing twitter and traditional media using topic models.’ In proceedings of *The European Conference on Information Retrieval* (2011): 338–349.

Appendix A: Constructing Contexts

We apply our context construction method to the Management Discussion and Analysis (MD&A) sections of 10-K reports. We describe the four stages of the context construction process in details below.

Stage 1: Clause segmentation

We parse all 107,596 MD&As using Stanford NLP's Open Information Extraction algorithm, or Open IE (Angeli, Premkumar and Manning [2015]). Open IE is a method used to extract "relation triples;" i.e., snippets of text from sentences of the form (subject; relation; object). Open IE accomplishes this using a series of three steps. First, it uses a dependency parser to build a parse tree of the sentence. A parse tree is a tree of the grammatical structure of a sentence, which helps in parsing the sentence from a natural language perspective. This parse tree, along with a named entity recognition (NER) system, is also used to resolve any "co-references" (i.e., replacing ambiguous words like "it" or "her" with the entity that is logically being discussed). This ensures that individual clauses we construct later are self-contained. The second step is to extract clauses from each sentence. This is done using a multinomial logistic approach applied to features obtained from the dependency parser (such as subject/object relations and part of speech tags). This produces a list of distinct, self-contained clauses which can stand on their own as sentences. The final step is then to segment the clauses into triples of the form (subject; relation verb; object). This is done entirely using the dependency tree through a set of eleven predefined linguistic patterns and three regular expression patterns.[19]

As an example, consider the following phrase: "The company's earnings increased by 5% due to an improvement in operating efficiency." This sentence has three key takeaways: 1) it is discussing earnings, 2) earnings increased by 5%, and 3) the 5% increase is due to operating efficiency. The Open IE extractions for this sentence, as shown in Figure 1, are (company; has; earnings), (company's earnings; increased by; 5%), (company's earnings; increased due; improved operating efficiency), and (company's earnings; increased due; operating efficiency). It is clear to see that the first three extractions match to the three key takeaways from the sentence. As such, we can see that Open IE is effective at extracting the key context from this sentence. The fourth extraction is a repeat of the third, but slightly more concise, which demonstrates a drawback of the Open IE method: it frequently generates excess extractions with slight differences in wording. In the case of this sentence, we would prefer to keep the third extraction, as the fourth extraction drops a word from the LM positive sentiment dictionary. We handle this issue in the next step of our methodology.

Stage 2: Reducing clause duplication

Applying Open IE as described above yields a total of 179,703,756 extractions across all MD&A filings—an average of 1,670 extractions per MD&A. To combat the issue of near-duplicate overlapping extractions, as well as to reduce the dimensionality of the data, we filter the extractions. The filtering process is designed to keep the fewest extractions possible, each of the shortest length possible, such that they 1) cover as much of the sentence as possible while not

[19] Of the 14 patterns, six relate to word order between the nominal subject (main subject of the clause, a noun), the direct object (a noun or noun phrase within the predicate, and the verb linking the nominal subject and direct object together within the predicate of the clause. An additional five patterns relate to word order within noun phrases that implicitly provide a sufficient information to form a full clause, as do the three regular expression patterns. An example of such a noun phrase is "Meta's AI research team, FAIR," which tells us that Meta has an AI research team named FAIR.

being nested within one another, 2) retain all words in the LM positive and negative sentiment dictionaries, and 3) retain all accounting and finance related terms from Campbell Harvey's hypertextual finance glossary and NYSSCPA's Accounting Terminology Guide.[20]

The first criterion ensures that we keep as much of the information in the sentence as possible. The second criterion avoids dropping words from the LM dictionaries, as dropping the words could potentially hurt the performance of sentiment measures in our empirical tests. As such, we err on the side of caution and keep all LM words. The final criterion ensures that we do not drop any words that could be useful in an accounting or finance setting. While both the accounting and finance glossaries predominantly contain terms that are 1 word long, they also contain terms that are phrases (i.e., two or more words). For phrases, we discard any which would already be flagged based on the individual word terms within each dictionary. For any phrases that would not be flagged by the previous procedure, we manually examine the words contained in the phrase and add only words that are unambiguously accounting- or finance-related.

After isolating all relevant individual words, we then transform these dictionaries into text analysis dictionaries by inflecting all words to obtain their conjugations, adjective forms, adverb forms, plural forms, and singular forms using the *word_forms* python library. This is important, as words can be used in many ways to discuss the same concept; for instance, for the word "collateral," we would be just as interested in the words "collaterals," "collateralize," and "collateralized." Since these dictionaries were not constructed with text analytics use in mind, they do not generally contain more than one inflection of a word originally. We do not inflect the words in the LM dictionaries as these dictionaries are already inflected to some extent, e.g., both "procrastinate" and "procrastination" are in the negative sentiment dictionary, and these dictionaries were already designed with text processing in mind.

The words in the four dictionaries are commonly found in the filings. Of the 179,703,756 extractions, 21,362,577 contain a word from the LM negative dictionary, 12,144,144 contain a word from the LM positive dictionary, 171,098,180 contain a word from our dictionary based on Campbell Harvey's hypertextual finance glossary, and 152,337,061 contain a word from our dictionary based on the NYSSCPA's Accounting Terminology Guide. That there is such high overlap between the accounting and finance dictionaries and our extractions provides some initial empirical comfort that Open IE is extracting relevant information from the MD&As. Filtering based on our length, coverage, and dictionary criteria drops the number of extractions from 179,703,756 to 48,576,229, a 73 percent reduction. After this stage, we concatenate the three components of the triples together with spaces to form clauses. We use these clauses throughout the rest of our methodology.

Stage 3: Abstracting from language to meaning

At this point we keep all remaining clauses throughout, but we still need to reduce the dimensionality of these clauses in order to be able to make sense of them more broadly. To accomplish this, we use Universal Sentence Encoder (USE), specifically the Transformer variant of USE by Cer et al. [2018].[21] This model takes a sentence-length snippet of text and maps it to a 512-dimensional vector space, based on both word order and the words themselves. Each dimension of each vector is bounded between -1 and +1. While the dimensions themselves are

[20] See https://people.duke.edu/~charvey/Classes/wpg/glossary.htm and https://www.nysscpa.org/professional-resources/accounting-terminology-guide#sthash.4Fay4z8I.dpbs.

[21] This encoder is freely available online at https://tfhub.dev/google/universal-sentence-encoder/4 on TensorFlow Hub.

not human-intelligible, USE maps snippets with similar meanings more closely together under a Euclidean distance metric. As such, it can be used to determine which snippets are more similar, and USE is significantly more robust to variations in writing styles and word choice than other algorithms like cosine similarity. E.g., if given "how are you," "how old are you," and "what is your age," USE correctly maps the second and third to be close together, while the first is quite far away within the vector space. Cosine similarity, on the other hand, would say the first two are nearly identical, while the second and third have no similarity at all. Since the effect of word choice is particularly pronounced on smaller snippets of text like our clauses, USE is a natural choice. This method has been used in the accounting literature by Crowley, Huang, and Lu [2022b] to show the similarity in meaning between tweets from executives and their CEOs.

Before we apply USE to the clauses, we mask out certain language that we have found to be overly sensitive to text of the following types: percentages, dates, times, dollar amounts, quantities, and ordinal numbers. To mask these out, we use Named Entity Recognition as provided by the Python spaCy library to identify all such instances, replacing each problematic text type with a single word representing it.[22] While our results are largely unchanged if we do not mask these texts out, we find a variety of contexts include ranges over the unmasked dates and amounts, such as events getting clustered into short ranges of years rather than by event types. By implementing masking, we can effectively solve this issue. After masking, we apply USE to get the vectorized representation of each masked clause.

In practice, USE can group together text that is quite domain specific. For instance, we find that USE correctly maps clauses discussing liquidation (e.g., "entity's liquidation becomes evident") with discussion of going concern (e.g., "Company's auditors have expressed substantial doubt about our ability to continue as going concern"). We also find it can correctly match up various discussions of energy production, such as mapping discussion of all the following close together: 1) oil pipelines, 2) crude oil volumes, 3) oil companies (BP, Mobil), 4) nuclear energy generation, 5) natural gas, and 6) energy generation by utilities. We do caveat that it is not a perfect algorithm, of course, as we noted it had some confusion on two relevant words: "note" (as in note payable or financial statement note) and "interest" (interest rate versus financial interest in an entity). However, such confusion did not cover all uses of those words; it only occurred within two specific contexts.

Stage 4: Clustering into contexts

After mapping all 48,576,229 masked clauses to USE's 512-dimensional vector space, we then apply a clustering method to gather clauses that are similar in meaning. Since USE relies on Euclidean distance to measure similarity, we use a variant of K-Means, as it likewise clusters based on Euclidean distance. The K-means variant we use is the Mini-Batch K-Means by Sculley [2010]. While a traditional K-Means algorithm requires processing all data at once (which is a problem in our case, as the USE vectors total around 230GB), Mini-Batch K-Means allows for processing the vectors in batches of any size. We implement the algorithm with a batch size of one million. To determine the number of clusters, k, we run Mini-Batch K-Means for all k from 2 through 200. We then optimize the number of clusters using a simulated bootstrapping technique based on Tibshirani, Walther and Hastie [2001] to construct the Gap statistic. The

[22] We mask all dates with DATE, all times with TIME, all percentages with PERCENT, all dollar amounts with MONEY, all quantities with QUANTITY, and all ordinal numbers with ORDINAL. For percentages we additionally tried masking with x%, xx%, X%, XX%, x.x%, and X.X%, but found PERCENT to be most stably related to actual percentages in our documents when implementing USE. We similarly tried $X, $XX, and $XXX for dollar amounts, but again MONEY was the most stable mask to use.

criterion for the Gap statistic is intuitive – an optimal number of clusters, k, is the lowest k such that the error at k clusters is within a certain bound from the error at k+1 clusters, adjusted for the variation in error at k+1 clusters. The variation is derived from a bootstrapped standard error using synthetic data of the same shape as the original data. For more details about this process, see Appendix B. After iterating, we determined that 137 was the optimal number of clusters under the Gap statistic.[23] We refer to these 137 clusters as the contexts within the 10-K MD&A filings.

Lastly, using the output of the Mini-Batch K-Means algorithm at 137 contexts, we assign each clause to a context based on the closeness of the clause to the cluster centers within the vector space created by USE. These context assignments constitute the final measure that is used throughout our tests.

[23] As detailed in Appendix B, 137 is the optimal number of clusters to use when subclusters are likely to matter. We replicate our primary results using 13 as the optimal number of clusters (not tabulated). All primary results of Tables 5 through 8 continue to hold. However, we note that consistent with subclusters being particularly important in the context of 10-K filings, we find that the explanatory power of the regressions using only 13 clusters is much lower than our main results using 137 clusters.

Appendix B: Cluster Number Optimization

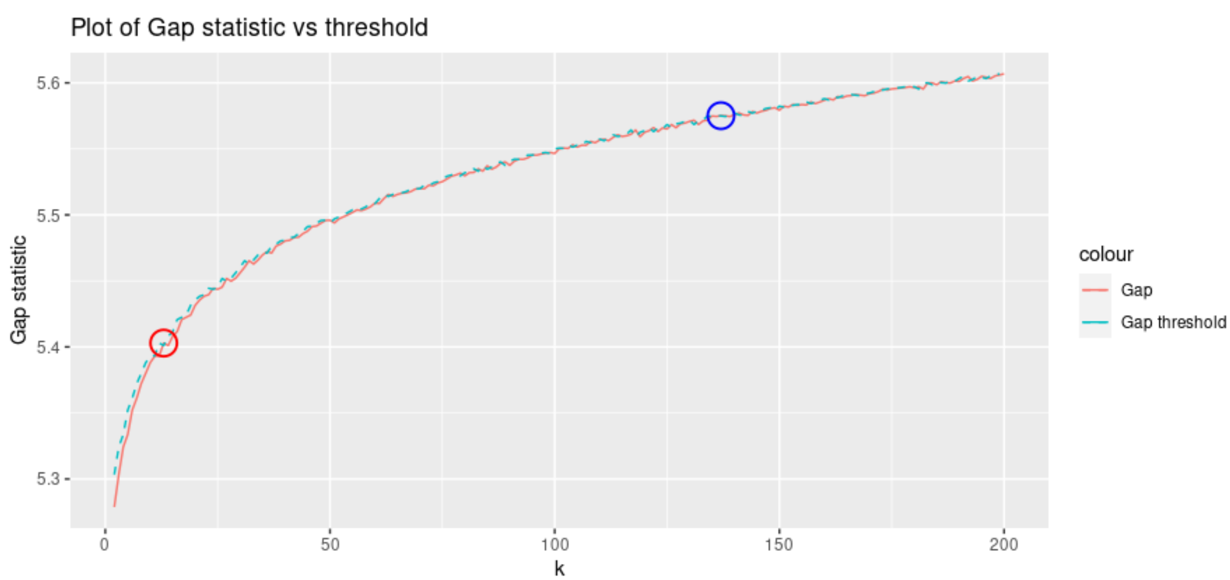
To optimize the number of clusters for the Mini-Batch K-Means algorithm, we use the Gap statistic of Tibshirani, Walther and Hastie [2001]. The Gap statistic is calculated at k clusters with B simulated samples, defining with W_k as the K-Means inertia score for the actual data at k clusters, $W^*_{k,r}$ as the K-Means inertia score for iteration r of the simulated samples at k clusters, and $\bar{l}$ as the average inertia across the B iterations at k clusters. It is calculated as follows:

$$Gap(k) = \left(\frac{1}{B}\right)\sum_{r=1}^{B} log\left(W^*_{k,r}\right) - log(W_k), and$$

$$s_k = sd_k\sqrt{1+\frac{1}{B}}\ , where\ sd_k = \sqrt{\left(\frac{1}{B}\right)\sum_{r=1}^{B}\left\{log\left(W^*_{k,r}\right) - \bar{l}\right\}^2}\ ,$$

To choose the optimal k based on the Gap statistic, we first follow Tibshirani, Walther and Hastie [2001] and select the lowest k such that $Gap(k) \geq Gap(k+1) - s_{k+1}$.

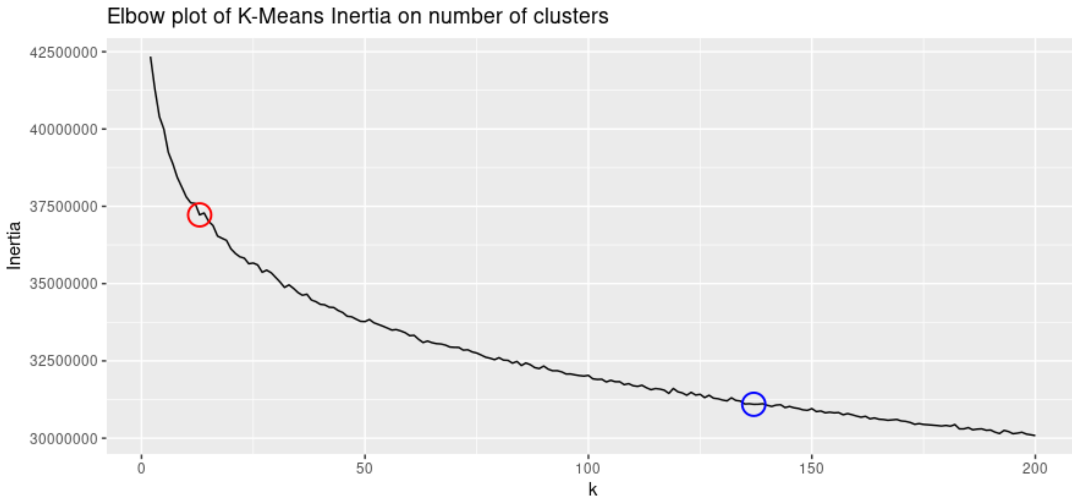
As each iteration of Mini-Batch K-Means is computationally expensive to run, we bootstrap ten samples for each k from two through 200. Each iteration uses data matching the same shape as ours: 48,576,229 observations, where each observation consisted of a 512-dimensional vector with each dimension bounded on [-1, +1] (to match the output of Universal Sentence Encoder). Based on this simulation, the first initial crossing is at a k of 13, with $Gap(13) = 5.402929$ and $Gap(14) = 5.401078$. The drop in Gap from k=13 to k=14 confirms that this is a relevant value, and it is illustrated in the plot below in the red circled area.



(Continued…)

Appendix B (Continued)

However, in Tibshirani, Walther and Hastie [2001] they caution that, in cases where there may be overarching clusters as well as subclusters (e.g., general accounting discussion versus individual matters such as taxation, accounting policies, income statement amounts, etc.), the initial k proposed by Gap may be too small. We confirm this with an elbow plot presented below, which shows that the spike upward in inertia at k=14 (in the red circle) is indeed temporary, and that there is still a steep downward slope after this point. This implies that more clusters may be efficient, up until inertia begins to plateau.



To find such a plateau point, we look for the first value of k at which Gap is optimal over two subsequent values of k rather than just one. Under this optimization, a k of 137 appears to be optimal, with $Gap(137) = 5.575112$, $Gap(138) = 5.574997$, and $Gap(139) = 5.574471$. As shown in the above plots, after k=137 (circled in blue in both plots), both Inertia and the Gap statistic begin to level off, indicating little need of a higher value of k. As such, we use k=137 for constructing our clusters.

Appendix C: Context Example Clauses

Appendix Table C Example extractions

Group: Accounting

Accounting policies

Accounting assumptions
- different assumptions are used in future periods
- assumptions affect reported amounts of commitments

Accounting policies
- Companys accounting policies are more fully described to audited consolidated financial statements
- accounting practices differ from U.S. GAAP

Accounting processes
- audit examining on test basis
- goodwill be tested for impairment

Cautionary statements
- results could differ materially from those currently anticipated
- results may differ from estimates

Discussion of accounting procedures
- we establish valuation allowance
- we establish objective evidence of fair value

Fair value measurement
- fair value was determined by employment of discounted cash flow model
- carrying value exceeds fair value

Revenue recognition
- Results recognized revenues through December 31 2001
- Maintenance revenues are recognized

Tax
- effective income tax rate was 34.8 %
- U.S. tax credits reflected in 2008 resulting from Brazilian tax rate increase

Accounting standards

Accounting standards
- Company has adoption of SFAS No. 153
- SFAS 162 amends certain of ARB 51s consolidation procedures for consistency with requirements of SFAS 141R

New accounting standard
- FASB issued Statement of Financial Accounting Standards No. 159
- Financial Accounting Standards Board issued Statement of Financial Accounting Standards No. 156

General and Balance sheet discussion

Account details
- Depreciation is provided on straight line basis over lives of assets
- amount payable is dependent upon level of unpaid claims

Cash flows
- Financing cash flows consist of repurchase of Control4 stock in open market
- Companys cash flows have have impacted by increase

Deferred tax
- net deferred tax asset was recorded as reduction to extent available
- we deferred tax liabilities

Negative accounting outcomes
- we recorded goodwill impairment loss
- We recorded 2,792,900 extinguishment loss

Noncurrent assets
- Real estate inventory is classified as non-current
- Long lived assets are evaluated

Securities and securities filings
- marketable securities are carried at fair value
- We have marketable securities derivatives

Income statement discussion

Accounting losses
- Company has loss from discontinued operations
- Company reported hurricane losses

Costs or expenses
- regulations would require expenditures
- Developed technologies are amortized

Depreciation and amortization
- Depreciation expense increased $ 4.2 million for year ended December 31 1997
- amortization using straight line method

Income statement items
- $ 4.8 million revenue in 1996
- annual revenue result return to profitability

Interest income or expense
- Net interest income decreased 15 %
- Interest expense decreased $ 57.6 million as result of $ 104.0 million growth in average interest bearing liability balances offset

Appendix Table C (Continued): Example extractions

Expenses and provisions
- Professional expenses increased due 178,000 increase in legal fees primarily associated with trademark protection
- Provision is made for expected difference between ambulance services fees collected

Large expenses
- cost are included as costs of acquisition
- restructuring charges include write down of

Losses
- impairment loss is recognized based excess
- Loan losses are charged for loan losses

Group: Business operations

Debt, Equity, and investment

Credit facilities
- Credit Facility restricts annual capital expenditures
- Corporation credit facility is secured by first priority security interest

Debt transactions
- SEC authorize issuance of short term debt
- standby letters of credit are until no longer required

Financing
- Company issued $ 152.1 million of convertible subordinated notes due with interest payable
- Company is presently evaluating different financing alternatives including project financing

Financing and investment
- Capital Resources repurchased stock before spin off
- PPMC invested over one million dollars

Funds and financing activities
- We fund operations through debt
- we repurchased at price in we open market share repurchase program

Loans
- Revolving Facility is with Lender
- compliance is with loan covenants

Obligations and covenants
- Commitments require payment of fee
- compliance is with each of our debt covenants

Share transactions
- Company fund stock repurchases
- stockholder approval increase stock to 150,000,000 shares

Expectations and future

Company expectations
- Company estimates that
- Company believes meet Company ongoing cash needs for foreseeable future

Continuation or going concern
- Company will continue monitor
- Companys auditors have expressed substantial doubt about our ability to continue as going concern

Expected outcomes
- trend will continue into future periods
- economic benefits are expected realized

Management expectations
- Management believes Company
- management expected maintain

Risk factor disclosures
- Disruption could have adverse effect on our business
- Delays could have adverse effect on demand for us products

Risk factor disclosures 2
- tender offer may may amended
- Changes could cause fluctuations in earnings

Macroeconomics

Economic and business conditions
- U.S. economic conditions worsened significantly in last quarter
- volatile trading conditions is in most markets

Interest rates
- interest rate can vary over terms of certain loans
- interest rate is 6.4 %

Market condition and competition
- competitive pressures is in industry
- there may greater competitive pressures is in markets

Market risk
- Market risk is potential loss arising from adverse changes in market rates
- We are exposed to market risk in course of We business

US Regulatory
- U.S. Food and Drug Administration (FDA) conducting studies
- FCC overstepped authority

Appendix Table C (Continued): Example extractions

Operations

Customers
- Speech Design 's main customers include leading PABX manufacturers
- new clients is in existing geographic areas

Employee matters
- Employee bonuses are accrued
- employee terminations are expected by end of second quarter

Energy
- Cameron Highway Oil Pipeline delivery of volumes
- Driving was demand for our duty diesel products

Expenses
- Companys management operating expenses
- costs related to program

General business description
- Software programs our cookies marketed as adware detectors
- Integrated Delivery Networks providing points by specializing in diagnosis of disease states

Investments
- investments is in commodity interest contracts
- significant investment is in capital assets

Investments and horizons
- proceeds are invested in short term securities
- short term advances are term funds

Leases
- decreased rental rates is in renewed leases
- operating leases is in normal course of business

Loans issued
- receivables held for sale to financial institution
- Lenders agreed loan

Mixed business activities
- we completed two additional closings of 2011 Equity PIPE
- strategic relationships is with our key suppliers

Options and ESOs
- options to purchase have have granted with exercise price per share
- anticipated decline is in tax benefits resulting from stock option exercises

Prices
- quoted prices is in active markets
- selling prices are determined based prices

Pricing
- economic activity prices factors
- indices had impact on demand for pricing

Products
- parties depend on on manufacturing our pharmaceutical products
- MRAM manufacture high performance spintronic products

Sales of goods or assets
- agreement expansion of Universal Music Group sales opportunities
- sales accounted for 13 % of total revenues

US-centric statements
- Revenues is in Americas
- our U.S. markets have had utilization in rig fleets

Outcomes

Economic impact on company
- expanded business activities is with combined impact of recently enacted Tax Cuts of 2017
- stronger Euro currency had favorable impact

Growth
- growth is in loan portfolio
- we saw marginal growth

Interest income or expense operations
- interest income depends upon amounts of interest earning assets earned
- we amortized $ 1.4 million to interest expense

Operating performance
- Company 's operations generate revenues from product sales
- Toy Segment has operating profit of

Structure

Contracting
- compliance is with applicable financial covenants
- contract provides indemnification

Contracting with other entities
- our agreement is with our third party logistics provider
- our contract is with UAE

Partners in partnership
- Investment Partnerships have incurred substantial legal costs based upon Mr. Calhouns fraud
- KPP issued 3,000,000 limited partnership units

Partnerships
- Partnership suffered losses in agricultural sector
- Partnership 's sole purpose is to trade in futures

Subsidiaries
- Company is wholly owned subsidiary of Golden State Holdings
- its owned subsidiary is in United Kingdom

Appendix Table C (Continued): Example extractions

Group: Changes

Changes

Change in sales
- Net Sales increased approximately 4 %
- Contract Packaging net sales increased 15 %

Changes in expenses
- Banks operating expenses decreased for fiscal 2012
- Research expense decreased for year

Changes in interest and forex rates
- changes is in interest rates
- Fluctuations is in value of local currencies

Changes in operating measures
- gross profit increased to 22.1 %
- Operating margins were negatively impacted by costs

Changes in revenue and expenses
- Selling increased from 16.0 %
- SG&A expenses declined from 15.4 %

Decreases in value or performance
- decline is in value of particular investment
- decline is in operating cash flows

Decreases in different measures
- revenues decreased over prior year
- short term borrowings decreased In 2014

Decreases in expenses or performance
- decrease is in bad debt expense
- decrease is in return on equity

Increase in expenses
- significant increases is in legal fees
- $ 1.1 million increase is in corporate administration expenses

Increases in accounts
- substantial increases is in premiums
- increase is in accrued compensation

Increases in income or revenue
- Company 's total revenues increased 54.2 %
- Operating income increased 3.7 million from $ 17.0 million

Increases in performance
- ETG 's gross profit margin increase was from principally improved product mix including higher margin product mix contributed by acquisitions
- increase was due aggressive marketing of segment's services to pharmaceutical industries while controlling direct costs

Reduction in accounts
- reduction is in interest income of 3,884
- reduction is in oil revenues

Decreases and offsets in performance
- unfavorable adjustments were result of pricing discounts
- decrease was due reduction in SoftPEG royalties

Ungrouped Text

Grammatical patterns

Cash headings
- Net cash used in activities
- Cash flow used in operations

Company information with name
- Delta CompuTec Inc. business is evolving is in stages
- CAPITAL RESOURCES Philadelphia Consolidated Holding Corp. is holding company

Dollar amounts
- $ 4,681,573 common stock issued for cash
- $ 56,983 services performed on contracts

Headers
- Management with Discussion of Financial Condition
- MANAGEMENT 'S DISCUSSION OF RESULTS OF OPERATIONS

Increases with time reference
- increase had occurred on August 31 2001
- increased borrowings is in 1997

Mentions of management
- Managements strategy includes soliciting
- full collectibility is doubtful in opinion of management

Modal strong statements
- management is required
- mandatory redemption is in whole

Modal weak statements
- We may raise capital To extent
- it is possible

Percents in year
- 19.1 % is in 2015
- 31.9 % is in 2002

Timing
- various points is in time
- related debt is in period incurred

"We" + verb + date
- We anticipate In 2005
- we had At October 31 2005

Appendix Table C (Continued): Example extractions

Timeframes

Date references with accounting content
- mortgage notes balance as February 28 , 2001
- management fee arrangement was cancelled upon completion of June 2006 public offering

Dates
- years ended December 31 2000
- CCI is from September 25 , 1997

Dates of events
- first clinic opened in July 2004
- Mark G. Mahowald joined US in March 2000

Dates with events
- Trust announced On March 5 1999
- FASB issued In September 2006

Dates with unrelated statements
- periods beginning after December 15 2016
- natural gas transported during 2002

Reporting periods
- corresponding period is in 2009
- comparable period is in 1998

Sales over a period
- sales is in particular period
- sales is in fiscal 1998

Time references + "company"
- Company repaid In February 2014
- Company recorded In 1999

Time references + "our"
- our realigned structure in 2004
- our average cost decreased In 2010

Unrelated events in time
- write off is in 2010
- portfolio is in 2002

Unrelated statements

Location names
- our are headquartered in San Jose
- U.S. service center is in Oklahoma City

Mixed accounting issues
- hedged transaction affected earnings
- allowance totaled $ 37.7 million

Mixed accounting terms
- liability is included
- EBITDA is included

Unrelated statements 1
- ASU No. 2013 Taxes Disclosures about Reclassification Adjustments
- currency collected is different in Europe

Unrelated statements 2
- they participate in USDA Biorefinery Assistance Program
- goodwill is primarily included in adhesives

Unrelated statements 3
- Additions are capitalized
- restaurants is in operation

Unrelated statements 4
- TGP recorded liability
- systems provided by NetGravity

Unrelated statements 5
- Included was non-recurring gain
- William W. Grant was Chairman

Unrelated statements 6
- restructuring charges totaled $ 1.4 million
- AstraZeneca expand commercialization

Unrelated statements about banks
- Federal Home Loan Bank remains in compliance with regulatory capital requirements
- Banks investment totaled $ 3,100,000

Unrelated transactions
- Merger was consummated
- transaction was deemed financing transaction under guidance

Unrelated statements with specific words

"Change" + unrelated statements
- change is with lower effective tax rates
- change is in inventories

"Changes" or "differences"
- Changes are recorded
- differences are expected reverse

"Company" + mixed accounts
- Company has interest expense
- Company received net cash

"Increase" + unrelated statements
- increase was result
- increase was higher

"Increase" or "decrease" + unrelated statements
- increase reflects impact of improved highway
- decrease related expenses

"Interest" + unrelated statements
- Company has net interest margin
- interest is in affiliate

Appendix Table C (Continued): Example extractions

"Company" + unrelated statements 1
- Company has total risk
- Company approves plan

"Company" + unrelated statements 2
- Company has ability to execute successfully
- Company will evaluate support

"Company" + unrelated statements 3
- company estimates initiatives
- Company entered into agreement to sell

"Company" + unrelated statements 4
- Company purchase number of trucks
- Company seek recovery against other potentially responsible parties

"Connection is with" + unrelated statements
- connection is with product development
- connection is with restructuring plan

"Decrease" + unrelated statements
- decrease was due to termination of production contract
- decrease is attributable

"Estimates" + unrelated statements
- estimates are revised
- AMP require estimation Of inputs

"Full" or "Total" + unrelated statements
- full valuation allowance is needed at December 31 2003
- Total capital is sum

"Net" + unrelated statements
- Our reported net income
- Globe Net develop audit Globe Nets website

"Notes" + unrelated statements
- See Note for discussion regarding goodwill
- Notes require annual principal payments

"Our" + unrelated statements
- our terms of our credit facilities
- our absent receipt

References to "consolidated statements"
- Consolidated Statements of Operations is for the years
- equity is in consolidated financial statements

"Statements" + unrelated statements
- amounts is in financial statements
- statements are based on expectations

"We" + unrelated statements 1
- we recorded expense
- we announced plans During quarter

"We" + unrelated statements 2
- we minimize variances
- We obtain debt

"We" or "our" + change statements
- our operating expenses increase in future
- our increase generated installment contracts

"We/our" + operations statements
- connection is with our operations in Brazil
- we commenced operations through December 31 2010

Appendix D: Variable Definitions

Variable	Definition
Sentiment measures	
Negative, Full 10-K, LM parser	Negative 10-K sentiment from the Loughran McDonald data files. Calculated as the number of individual words in the 10-K filing contained in the LM negative sentiment dictionary divided by the number total words in the 10-K filing.
Negative, Full 10-K, Our parser	After applying our 10-K parsing methodology to the raw text SEC files, it is calculated as the number of individual words in the parsed 10-K filing contained in the LM negative sentiment dictionary, divided by the number total words in the parsed 10-K filing.
Negative, MD&A, Our parser	After extracting the MD&A section from a 10-K using our parser, it is calculated as the number of individual words in the parsed MD&A contained in the LM negative sentiment dictionary, divided by the number total words in the parsed MD&A.
Positive, Full 10-K, LM parser	Positive 10-K sentiment from the Loughran McDonald data files. Calculated as the number of individual words in the 10-K filing contained in the LM negative sentiment dictionary divided by the number total words in the 10-K filing.
Positive, Full 10-K, Our parser	After applying our 10-K parsing methodology to the raw text SEC files, it is calculated as the number of individual words in the parsed 10-K filing contained in the LM positive sentiment dictionary, divided by the number total words in the parsed 10-K filing.
Positive, MD&A, Our parser	After extracting the MD&A section from a 10-K using our parser, it is calculated as the number of individual words in the parsed MD&A contained in the LM positive sentiment dictionary, divided by the number total words in the parsed MD&A.
Dependent variables	
Event period excess return	Holding period return from day 0 (filing date) to trading day +3, minus the CRSP Value Weighted Index return over the same interval.
Event period abnormal volume	Average trading volume of the stock over the period from day 0 (filing date) to trading day +3, standardized as a z-score using the mean and standard deviation of volume over days [-60, -6].
Post-event return volatility	The RMSE of an FF 3-factor model applied to trading days [+6, +252]. The coefficients of the model are determined based on trading days [-252, -6].
Material weakness count, t+1	The number of material weaknesses tied to the companies' next 10-K filing, per Audit Analytics.

(Continued…)

Appendix D (Continued)

Variable	**Definition**
Independent variables	
Context	The number of clauses in a filing that belong to the given context, divided by the total number of clauses extracted from the filing.
$\text{Sentiment}_{\text{Context}}$	The number of clauses in a filing with the specified sentiment that belong to the given context, divided by the total number of clauses extracted from the filing. A clause has negative (positive) sentiment if it contains more words that are in the LM negative (positive) sentiment dictionary than the LM positive (negative) sentiment dictionary. A clause has neutral sentiment if it has neither positive nor negative sentiment; this may be because no LM dictionary words were contained in the clause, or because there were an equal number of negative and positive words in the clause.
Controls	
log(Market value)	Natural log of the share price at date 0 (filing date) times the number of shares outstanding at date 0, per CRSP.
log(BTM)	Natural log of the book value of equity (from Compustat) divided by the market value as defined above.
log(Share turnover)	Natural log of the average volume over trading days [-252, -6] divided by the shares outstanding at date 0 (filing date).
Pre-event FF alpha	The alpha from an FF 3-factor model applied to trading days [-252, -6].
I(Nasdaq)	An indicator if the firm is listed on the Nasdaq stock exchange, per CRSP.

Tables

Table 1: Sample Construction

	Filings	
	Documents	Documents dropped
Unique 10-K filings	188,030	
Unique 10-K405 filings	20,139	
Total filings	208,169	
10-K with MD&A	93,551	-94,479
10-K405 with MD&A	14,045	-6,094
Total files with MD&As	107,596	

Sample restriction	MD&As	MD&As dropped	Extractions (clauses)	Extractions dropped
MD&A has extractions from OpenIE	105,921	1,675	48,576,229	
Filing matched to the Loughran McDonald data library	103,137	2,784	47,317,492	1,258,737
First filing per year	102,079	1,058	47,023,707	293,785
At least 180 days after last filing	101,877	202	46,942,952	80,755
CIK In CRSP Compustat Merged	56,460	45,417	31,219,059	15,723,893
Data available in Compustat	49,812	6,648	28,110,347	3,108,712
Market cap available in CRSP	49,411	401	27,896,026	214,321
Price on t-1 >= $3	41,693	7,718	23,988,897	3,907,129
Return & volume has >= 60 obs from trading days [-252,-6]	40,489	1,204	23,344,479	644,418
NYSE, AMEX, or NASDAQ listed	40,476	13	23,336,694	7,785
Book to market available and positive	39,466	1,010	22,734,045	602,649
At least 2000 words in the 10-K	39,357	109	22,730,774	3,271
At least 250 words in the MD&A	35,362	3,995	22,669,186	61,588

This table reports the sample selection and the number of documents and clauses dropped by each sample restriction.

Table 2: Context Frequencies

Panel A: Most and least frequent contexts by clause count (excluding Ungrouped text)

	Most frequent contexts	Number of clauses	Number of documents		Least frequent contexts	Number of clauses	Number of documents
1	Increases in accounts	337,697	32,100	128	Depreciation and amortization	92,892	19,481
2	Loans issued	279,540	18,325	129	Company expectations	92,490	20,344
3	Mixed business activities	266,742	18,069	130	Risk factor disclosures 2	87,818	23,780
4	Revenue recognition	260,647	28,908	131	Prices	76,571	20,986
5	Sales of goods or assets	253,431	29,193	132	Deferred tax	73,184	16,587
6	Interest rates	248,814	26,361	133	Economic and business conditions	69,642	20,402
7	Funds and financing activities	247,055	17,051	134	New accounting standards	65,962	17,518
8	General business description	241,394	25,927	135	Partners in partnership	61,671	13,074
9	Increases in income or revenue	225,729	27,029	136	Cautionary statements	55,234	23,860
10	Tax	221,030	27,630	137	Partnerships	35,387	5,272

Panel B: Most and least frequent contexts by document count (excluding Ungrouped text)

	Most frequent contexts	Number of clauses	Number of documents		Least frequent contexts	Number of clauses	Number of documents
1	Increases in accounts	337,697	32,100	128	Funds and financing activities	247,055	17,051
2	Increases in performance	220,800	31,038	129	Energy	160,500	17,027
3	Operating performance	172,906	30,441	130	Leases	123,008	16,892
4	Income statement items	167,842	30,417	131	Deferred tax	73,184	16,587
5	Decreases and offsets in performance	203,605	30,125	132	Discussion of accounting procedures	182,118	16,257
6	Cash flows	192,820	29,837	133	Negative accounting outcomes	103,508	15,124
7	Financing and investment	180,431	29,267	134	Contracting with other entities	99,712	14,695
8	Sales of goods or assets	253,431	29,193	135	Accounting standards	108,039	13,485
9	Expenses and provisions	176,610	29,180	136	Partners in partnership	61,671	13,074
10	Decreases in expenses or performance	218,754	29,105	137	Partnerships	35,387	5,272

Table 2 (Continued): Context Frequencies

Panel C: Most and least frequent contexts by percent of negative clauses within context

	Most frequent contexts	Number of clauses	Percent of clauses		Least frequent contexts	Number of clauses	Percent of clauses
1	**Losses**	187013	94.77%	128	**Income statement items**	3239	1.93%
2	**Decreases in value or performance**	108445	93.49%	129	Dates with events	2462	1.86%
3	**Negative accounting outcomes**	78293	75.64%	130	Reporting periods	1696	1.68%
4	**Accounting losses**	79245	68.05%	131	Headers	980	1.45%
5	**Risk factor disclosures**	82938	63.03%	132	"Company" + mixed accounts	2082	1.33%
6	**Decreases and offsets in performance**	65922	32.38%	133	Increases with time reference	2152	1.17%
7	Unrelated statements 6	37606	30.06%	134	Cash headings	1172	1.07%
8	**Economic impact on company**	39218	27.01%	135	Dates	268	0.58%
9	**Decreases in different measures**	36761	23.71%	136	"Increase" + unrelated statements	515	0.36%
10	**Economic and business conditions**	15730	22.59%	137	Percents in year	407	0.24%

Panel D: Most and least frequent contexts by percent of positive clauses within context

	Most frequent contexts	Number of clauses	Percent of clauses		Least frequent contexts	Number of clauses	Percent of clauses
1	**Tax**	66248	29.97%	128	"Notes" + unrelated statements	1217	1.27%
2	Unrelated statements 6	20927	16.73%	129	"Decrease" + unrelated statements	1063	1.25%
3	**Accounting standards**	17981	16.64%	130	Headers	671	0.99%
4	**Partners in partnership**	9368	15.19%	131	**Changes in interest and forex rates**	1095	0.95%
5	"Company" + unrelated statements 2	11194	13.50%	132	**Costs or expenses**	1001	0.80%
6	"We" or "our" + change statements	17447	13.24%	133	Cash headings	667	0.61%
7	**Economic and business conditions**	8828	12.68%	134	Dates	179	0.39%
8	**Changes in operating measures**	16781	12.24%	135	**Losses**	586	0.30%
9	**Growth**	19165	11.81%	136	**Decreases in value or performance**	272	0.23%
10	**General business description**	27256	11.29%	137	Percents in year	191	0.11%

Panels A and B present the most and least common contexts in our sample of 22,669,186 clauses across 35,362 documents. Panel A presents the frequency of contexts by the count of clauses in the data, while Panel B presents the frequency of contexts by the number of documents with at least 1 clause from the context. Both Panels A and B only present business-relevant contexts. Panels C and D present the most and least common contexts to contain positive and negative sentiment, as a percentage of clauses within the context, based on the LM dictionary. Bolded contexts in Panels C and D are business-relevant.

Table 3: Univariate Statistics

Variable	Obs	Mean	SD	5%	Median	95%
Sentiment measures						
Negative, Full 10-K, LM parser	35,362	1.55%	0.45%	0.81%	1.54%	2.29%
Negative, Full 10-K, Our parser	35,362	1.30%	0.48%	0.58%	1.27%	2.13%
Negative, MD&A, Our parser	35,362	1.22%	0.59%	0.42%	1.14%	2.32%
Positive, Full 10-K, LM parser	35,362	0.68%	0.18%	0.44%	0.65%	1.01%
Positive, Full 10-K, Our parser	35,362	0.64%	0.19%	0.38%	0.61%	0.97%
Positive, MD&A, Our parser	35,362	0.65%	0.29%	0.26%	0.61%	1.16%
Extraction measures						
Clauses per MD&A	35,362	641.1	457.9	75.0	548.0	1,511.0
Negative clauses per MD&A	35,362	36.6	34.8	2.0	27.0	105.0
Positive clauses per MD&A	35,362	20.1	16.9	1.0	16.0	52.0
Dependent variables						
Event period excess return	35,362	-0.36%	7.65%	-11.47%	-0.27%	10.26%
Event period abnormal volume	35,361	0.493	3.848	-0.771	-0.059	3.062
Post-event return volatility	35,362	0.160	0.131	0.000	0.143	0.331
Material weakness count, t+1	23,034	0.153	0.782	0	0	1
Control variables						
log(Market value)	35,362	12.72	1.72	10.14	12.60	15.74
log(BTM)	35,362	-7.63	0.926	-9.21	-7.527	-6.35
log(Share turnover)	35,362	1.37	1.09	-0.553	1.45	2.98
Pre-event FF alpha	35,362	0.08%	2.50%	-2.91%	0.04%	3.17%
I(Nasdaq)	35,362	59.50%	4.91%	0	1	1

The sentiment measures represent the proportion of words from the given LM dictionary (Negative or Positive). Extraction measures are presented as the raw number of clauses per 10-K MD&A. The sample consists of 35,362 firm-years.

Table 4: Context Underlying MD&A Tone

	Negative MD&A Tone (1)		Positive MD&A Tone (2)	
Part A: High sentiment contexts				
Context: Accounting				
Cautionary statements	0.082	***	0.018	**
Context: Business operations				
Economic impact on company	0.029	***	0.017	***
Employee matters	0.037	***	0.015	***
Market condition and competition	0.013	**	0.039	***
Operating performance	0.033	***	0.014	***
Context: Changes				
Reduction in accounts	0.023	***	0.024	***
Decreases and offsets in performance	0.071	***	0.016	***
Part B: Contexts skewed toward negative				
Context: Accounting				
Accounting losses	0.180	***	0.004	
Expenses and provisions	0.023	***	-0.002	
Losses	0.136	***	-0.005	**
Negative accounting outcomes	0.123	***	-0.003	
Noncurrent assets	0.016	***	-0.011	***
Context: Business operations				
Customers	0.017	***	-0.011	***
Economic and business conditions	0.029	***	0.011	*
Loans	0.026	***	-0.005	**
Loans issued	0.008	***	0.000	
Management expectations	0.068	***	-0.004	
Market risk	0.010	***	-0.018	***
Risk factor disclosures	0.252	***	-0.008	**
US Regulatory	0.011	***	.	
US-centric statements	0.014	***	-0.005	**
Context: Changes				
Decreases in value or performance	0.140	***	-0.004	
Decreases in different measures	0.029	***	0.005	
Part C: Contexts skewed toward positive				
Context: Accounting				
Accounting standards	0.005		0.006	**
Cash flows	-0.001		0.013	***
Income statement items	-0.005		0.024	***
Interest income or expense	.		0.013	***
New accounting standard	-0.010	*	0.014	***
Tax	-0.009	***	0.025	***
Context: Business operations				
Continuation or going concern	-0.008		0.066	***
Contracting with other entities	-0.031	***	0.015	***
Expected outcomes	0.005		0.014	***
Financing and investment	-0.039	***	0.012	***
General business description	-0.007	***	0.015	***
Growth	-0.016	***	0.035	***
Investments	-0.010	**	0.013	***
Investments and horizons	-0.031	***	0.018	***

Table 4 (Continued): Context Underlying MD&A Tone

	Negative MD&A Tone (1)	Positive MD&A Tone (2)
Leases	-0.007 **	0.004 **
Partners in partnership	-0.004	0.011 ***
Products	0.003	0.005 **
Context: Changes		
Change in sales	-0.024 ***	0.018 ***
Changes in operating measures	-0.006	0.010 ***
Increases in performance	-0.018 ***	0.020 ***
Part D: Low sentiment contexts		
Context: Accounting		
Accounting processes	-0.012 **	-0.010 **
Depreciation and amortization	-0.036 ***	-0.011 ***
Large expenses	-0.034 ***	-0.021 ***
Revenue recognition	-0.022 ***	-0.009 ***
Context: Business operations		
Credit facilities	-0.019 ***	-0.015 ***
Energy	-0.012 ***	-0.003 **
Options and ESOs	-0.051 ***	-0.019 ***
Sales of goods or assets	-0.010 ***	-0.007 ***
Subsidiaries	-0.029 ***	-0.014 ***
Context: Changes		
Decreases in expenses or performance	-0.013 ***	-0.024 ***
Increases in accounts	-0.026 ***	-0.010 ***
Part E: All other contexts		
Context: Accounting		
Account details	-0.018 ***	.
Accounting assumptions	-0.043 ***	0.002
Accounting policies	.	0.000
Costs or expenses	.	0.003
Deferred tax	0.007	-0.001
Discussion of accounting procedures	.	-0.007 **
Fair value measurement	-0.020 ***	-0.004
Securities and securities filings	-0.010 ***	-0.001
Context: Business operations		
Company expectations	-0.007	0.000
Contracting	.	-0.002
Debt transactions	0.002	-0.005 *
Expenses	-0.002	-0.008 *
Financing	-0.030 ***	0.002
Funds and financing activities	-0.021 ***	0.001
Interest income or expense operations	-0.001	-0.004
Interest rates	-0.021 ***	0.004 *
Mixed business activities	-0.015 ***	0.003
Obligations and covenants	-0.006	-0.005
Partnerships	-0.003	-0.001
Prices	-0.001	-0.011 *
Pricing	-0.008	0.005
Risk factor disclosures 2	0.009	-0.004
Share transactions	-0.005	-0.006 ***

Table 4 (Continued): Context Underlying MD&A Tone

	Negative MD&A Tone (1)		Positive MD&A Tone (2)	
Context: Changes				
Changes in expenses	0.008		-0.004	
Changes in interest and forex rates	0.002		-0.024	***
Changes in revenue and expenses	-0.016	***	0.000	
Increase in expenses	-0.010		-0.003	
Increases in income or revenue	-0.005		0.001	
Part F: Controls				
log(Market value)	.		0.075	***
log(BTM)	0.218	***	-0.019	
log(Share turnover)	-0.017		-0.115	***
Pre-event FF alpha	-0.342		-1.276	**
I(Nasdaq)	-0.180	***	-0.221	***
Part G: Ungrouped Text				
# of high sentiment clusters	5		5	
# of contexts skewed toward negative	10		10	
# of contexts skewed toward positive	7		7	
# of low sentiment contexts	9		9	
# of other contexts	24		24	
Controls	Included		Included	
FF48 Industry FE	Included		Included	
Adjusted R^2	0.526		0.204	
# of negative and significant contexts	54		40	
# of positive and significant contexts	38		39	
Baseline with randomized context assignment				
Adjusted R^2	0.211		0.096	
# of negative and significant contexts	3		3	
# of positive and significant contexts	6		1	

Columns (2) and (3) report LASSO regressions including all 137 contexts, with coefficient values multiplied by 1,000 for readability. All contexts are not restricted to any sentiment. Columns (1) and (4) present the expected signs for columns (2) and (3), respectively, based on hand coding a sample of 20 clauses from each context. All regressions are based on 35,362 observations. P-values are indicated as follows: * indicates p<0.10, ** indicates p<0.05, and *** indicates p<0.01. A period indicates that the variable was dropped in the regression by the LASSO procedure.

Table 5: Predicting Filing Period Excess Return

Sentiment examined:	Negative sentiment		Positive Sentiment		Neutral
Variable	(1)	(2)	(3)	(4)	(5)
Negative, MD&A, Our parser	-0.241 ***				
Positive, MD&A, Our parser			-0.130		
Accounting policies					
Accounting processes		0.226 **		.	.
Discussion of accounting procedures		0.566 ***		0.190	0.197 ***
Fair value measurement		0.591 ***		.	0.032
Tax		.		0.144 **	0.043
Accounting standards					
Accounting standards		0.326 ***		0.146 *	0.006
General and balance sheet discussion					
Negative accounting outcomes		-0.009 **		.	0.180 *
Income statement discussion					
Accounting losses		-0.098 ***		0.257	-0.014
Debt, Equity, and investment					
Debt transactions		0.072 **		0.053	.
Expectations / future					
Company expectations		.		-0.338 *	-0.102 ***
Continuation or going concern		-0.532 ***		.	.
Expected outcomes		.		-0.289 **	.
Risk factor disclosures		-0.215 ***		.	-0.196 **
Operations					
Products		-0.233		-0.542 ***	-0.125 ***
US-centric statements		.		-0.253 **	.
Structure					
Contracting with other entities		.		.	0.086 ***
Changes					
Changes in expenses		.		.	-0.069 **
Decreases in different measures		0.139 **		.	.
Increase in expenses		.		.	0.093 **
Decreases and offsets in performance		.		.	0.179 ***
Ungrouped					
Modal weak statements		-0.429 ***		-0.011	-0.147 **
Percents in year		.		-1.937 **	0.032
Dates with unrelated statements		-0.034 *		.	-0.046 **
Time references + "our"		.		0.827 ***	.
Unrelated statements 4		.		.	0.028 **
"Our" + unrelated statements		.		-0.293 **	-0.035
"Statements" + unrelated statements		-0.110 **		.	.
"We" + unrelated statements 2		.		-0.456 ***	-0.144 ***
"We/our" + operations statements		-0.225		.	-0.165 ***
Controls					
log(Market value)	0.002 ***	0.001 ***	0.002 ***	0.002 ***	0.002 ***
log(BTM)	0.001 **	.	0.001 *	0.000	.
log(Share turnover)	-0.005 ***	-0.004 ***	-0.005 ***	-0.004 ***	-0.004 ***
Pre-event FF alpha	0.012	.	0.012	.	0.001
I(Nasdaq)	0.001	0.000	0.001	.	.
FF48 Industry FE	Included	Included	Included	Included	Included
Adjusted R^2	0.009	0.014	0.009	0.012	0.014
# of negative and significant contexts		6		6	8
# of positive and significant contexts		6		2	5
Double LASSO adjustment					
Adjusted R^2		0.013		0.010	0.013
# of negative and significant contexts		6		5	6
# of positive and significant contexts		7		3	7

Columns (1) and (3) report linear regressions, while columns (2), (4), and (5) report LASSO regressions including all 137 contexts. All contexts are restricted to only the sentiment specified in the column. Only contexts that are significant at p<0.05 for at least one regression from columns (2), (4), and (5) are included. All regressions are based on 35,362 observations. P-values are indicated as follows: * indicates p<0.10, ** indicates p<0.05, and *** indicates p<0.01. A period indicates that the variable was dropped in the regression by the LASSO procedure. The bottom section presents the adjusted R-squared and number of significant contexts by sign (at p<0.05) when using a Double LASSO procedure as in Belloni, Chernozhukov and Hansen [2014] as described in Section 3.1.

Table 6: Predicting Filing Period Abnormal Volume

Sentiment examined:	Negative sentiment		Positive Sentiment		Neutral
Variable	(1)	(2)	(3)	(4)	(5)
Negative, MD&A, Our parser	-1.253				
Positive, MD&A, Our parser			-28.69 ***		
Accounting policies					
Discussion of accounting procedures		43.33 ***		17.56	6.234 *
Revenue recognition		.		.	3.493 ***
Accounting standards					
Accounting standards		.		5.489	-4.338 ***
Debt, Equity, and investment					
Obligations and covenants		11.71		.	5.413 **
Expectations / future					
Expected outcomes		14.60 **		-1.092	3.814 **
Risk factor disclosures 2		.		63.51 ***	.
Changes					
Decreases in different measures		-2.311		.	-4.208 **
Decreases in expenses or performance		10.68 **		.	1.085
Ungrouped					
Increases with time reference		.		-8.245	0.762 **
Timing		7.446 **		31.43 *	3.283
Date references with accounting content		7.163 **		.	0.562
Dates with events		.		1.755	5.790 **
"Change" + unrelated statements		.		60.26 ***	.
"Changes" or "differences"		36.85 ***		.	6.081 *
"Company" + unrelated statements 2		-0.951		-17.34 **	.
"Decrease" + unrelated statements		.		87.05 ***	.
"Estimates" + unrelated statements		35.23 ***		.	3.424
"Increase" or "decrease" + unrelated statements		.		.	-2.033 **
Controls					
log(Market value)	-0.039 **	-0.044 ***	-0.033 **	-0.040 ***	-0.050 ***
log(BTM)	0.069 ***	0.027 *	0.069 ***	0.034 **	0.030 *
log(Share turnover)	-0.021	-0.004	-0.025	-0.009 *	-0.030 **
Pre-event FF alpha	1.710 **	0.800 **	1.677 **	0.823 **	0.959 **
I(Nasdaq)	0.001	.	-0.006	.	.
FF48 Industry FE	Included	Included	Included	Included	Included
Adjusted R^2	0.002	0.005	0.002	0.005	0.007
# of negative and significant contexts		0		1	3
# of positive and significant contexts		7		3	5
Double LASSO adjustment					
Adjusted R^2		0.003		0.002	0.004
# of negative and significant contexts		1		1	3
# of positive and significant contexts		7		2	7

Columns (1) and (3) report linear regressions, while columns (2), (4), and (5) report LASSO regressions including all 137 contexts. All contexts are restricted to only the sentiment specified in the column. Only contexts that are significant at p<0.05 for at least one regression from columns (2), (4), and (5) are included. All regressions are based on 35,362 observations. P-values are indicated as follows: * indicates p<0.10, ** indicates p<0.05, and *** indicates p<0.01. A period indicates that the variable was dropped in the regression by the LASSO procedure. The bottom section presents the adjusted R-squared and number of significant contexts by sign (at p<0.05) when using a Double LASSO procedure as in Belloni, Chernozhukov and Hansen [2014] as described in Section 3.1.

Table 7: Predicting Post-event Return Volatility

Sentiment examined:	Negative sentiment		Positive Sentiment		Neutral
Variable	(1)	(2)	(3)	(4)	(5)
Negative, MD&A, Our parser	1.585 ***				
Positive, MD&A, Our parser			0.404		
Accounting policies					
Accounting policies		0.307 **		-0.155	-0.058
Accounting processes		-0.616 **		-0.390	-0.047
Discussion of accounting procedures		-0.777 ***		-0.561 *	-0.336 ***
Revenue recognition		0.232		-0.547 **	.
Tax		.		-0.344 ***	-0.115 **
Accounting standards		.		0.511 **	0.053
General and balance sheet discussion					
Deferred tax		1.863 ***		0.289	0.419 **
Negative accounting outcomes		0.921 ***		0.837	1.743 ***
Income statement discussion					
Accounting losses		0.611 ***		-0.025	0.011
Depreciation and amortization		.		-0.621	-0.199 **
Expenses and provisions		0.181		.	-0.159 **
Income statement items		0.011		0.430 *	-0.121 **
Interest income or expense		.		.	-0.154 **
Large expenses		.		.	-0.176 **
Debt, Equity, and investment					
Financing		.		.	-0.202 ***
Funds and financing activities		0.706 *		-0.622 **	-0.160 **
Loans		-0.586 **		-1.050	-0.182 **
Expectations / future					
Expected outcomes		.		1.231 ***	0.056
Macro					
Economic and business conditions		0.541 *		0.343	-0.492 ***
Interest rates		-0.661 ***		-0.425	-0.066
US Regulatory		.		.	-0.094 ***
Operations					
Customers		0.150		-0.729 **	-0.153 ***
Energy		0.198		.	-0.095 ***
General business description		0.389		.	0.160 **
Mixed business activities		.		.	-0.259 ***
Options and ESOs		-0.219		.	-0.105 **
Pricing		.		-0.911 ***	-0.053
Outcomes					
Growth		-0.816 **		-0.719 ***	-0.306 ***
Interest income or expense operations		-0.887 **		-0.233	-0.124 **
Operating performance		0.061		.	-0.086 **
Structure					
Partnerships		.		.	-0.130 ***
Subsidiaries		1.180 ***		.	0.012
Changes					
Change in sales		-0.044		.	-0.131 ***
Changes in interest and forex rates		.		.	-0.319 ***
Changes in revenue and expenses		-0.829 ***		-0.233	0.007
Decreases in value or performance		0.263 ***		.	2.363 ***
Decreases in different measures		0.345 **		.	0.067
Decreases in expenses or performance		1.059 ***		-0.416	0.368 ***
Increase in expenses		.		.	-0.348 ***
Increases in accounts		0.526 *		-0.148	-0.114 **
Increases in income or revenue		.		-0.189	-0.114 ***
Increases in performance		0.043		.	-0.166 **
Decreases and offsets in performance		0.724 ***		1.314 ***	0.266

Table 7 (Continued): Predicting Post-event Return Volatility

Ungrouped					
Cash headings		.		.	-0.446 ***
Company information with name		.		.	-0.113 ***
Dollar amounts		.		1.277 ***	0.020
Headers		0.700		-0.320	-0.120 ***
Mentions of management		0.285 *		-0.521 **	-0.168 ***
Percents in year		-1.649		.	-0.075 **
Timing		.		.	-0.380 ***
"We" + verb + date		.		0.114	0.371 ***
Dates		1.018		-3.712	-0.230 **
Dates with unrelated statements		.		0.362	-0.124 **
Reporting periods		.		.	-0.153 **
Time references + ""company""		.		.	-0.058 ***
Unrelated events in time		0.089		0.466 **	0.029
Location names		.		-0.332	-0.114 ***
Mixed accounting terms		.		.	-0.214 ***
Unrelated statements 4		0.399 **		0.127	-0.055 *
Unrelated statements 5		0.242		0.580 **	.
Unrelated statements about banks		0.493 **		0.562 ***	0.029
"Change" + unrelated statements		-0.168		-0.143	-0.158 ***
"Changes" or "differences"		-0.298		-1.081 **	0.088
"Company" + unrelated statements 1		0.491 **		.	-0.057 ***
"Company" + unrelated statements 2		.		0.779 ***	.
"Company" + unrelated statements 3		.		0.479 **	.
"Company" + unrelated statements 4		-0.098 *		.	-0.076 ***
"Decrease" + unrelated statements		-0.408 **		4.210 ***	0.286
"Full" or "Total" + unrelated statements		.		.	-0.275 ***
"Increase" + unrelated statements		.		-0.463	-0.215 **
"Increase" or "decrease" + unrelated statements		0.045		-0.194	-0.110 **
"Net" + unrelated statements		0.159		.	-0.277 ***
References to "consolidated statements"		0.228		1.739 ***	0.016
"We" + unrelated statements 1		0.982 **		1.681 ***	.
"We" + unrelated statements 2		.		0.554 *	-0.139 **
"We/our" + operations statements		.		0.961 **	-0.090
Controls					
log(Market value)	-0.017 ***	-0.016 ***	-0.017 ***	-0.016 ***	-0.016 ***
log(BTM)	-0.010 ***	-0.009 ***	-0.009 ***	-0.008 ***	-0.010 ***
log(Share turnover)	0.015 ***	0.014 ***	0.016 ***	0.015 ***	0.015 ***
Pre-event FF alpha	0.065 **	0.045 **	0.066 **	0.045 **	0.062 ***
I(Nasdaq)	0.008 ***	0.008 ***	0.008 ***	0.007 ***	0.009 ***
FF48 Industry FE	Included	Included	Included	Included	Included
Adjusted R^2	0.089	0.093	0.084	0.088	0.098
# of negative and significant contexts		8		8	46
# of positive and significant contexts		13		13	6
Double LASSO adjustment					
Adjusted R^2		0.092		0.088	0.098
# of negative and significant contexts		7		7	49
# of positive and significant contexts		15		12	5

Columns (1) and (3) report linear regressions, while columns (2), (4), and (5) report LASSO regressions including all 137 contexts. All contexts are restricted to only the sentiment specified in the column. Only contexts that are significant at $p<0.05$ for at least one regression from columns (2), (4), and (5) are included. All regressions are based on 35,362 observations. P-values are indicated as follows: * indicates $p<0.10$, ** indicates $p<0.05$, and *** indicates $p<0.01$. A period indicates that the variable was dropped in the regression by the LASSO procedure. The bottom section presents the adjusted R-squared and number of significant contexts by sign (at $p<0.05$) when using a Double LASSO procedure as in Belloni, Chernozhukov and Hansen [2014] as described in Section 3.1.

Table 8: Predicting Future Material Weaknesses

Sentiment examined:	Negative sentiment		Positive Sentiment		Neutral
Variable	(1)	(2)	(3)	(4)	(5)
Negative, MD&A, Our parser	-1.065				
Positive, MD&A, Our parser			-5.759 ***		
Accounting policies					
Accounting processes		-4.66 *		11.86 ***	-0.911
Cautionary statements		-23.60 **		.	-1.760 **
Fair value measurement		2.134		-7.956 **	0.000
Revenue recognition		6.296 **		-3.333	0.491
General and balance sheet discussion					
Account details		.		.	2.073 ***
Deferred tax		-9.305 ***		-9.982 ***	0.203
Securities and securities filings		-3.571 **		18.26 ***	-0.980 *
Income statement discussion					
Depreciation and amortization		-2.031		-14.24 **	-1.715 ***
Expenses and provisions		.		-7.492 ***	.
Income statement items		2.255		8.956 ***	.
Interest income or expense		-5.765 **		-1.836	-0.578
Large expenses		1.963		-9.294 **	0.092
Losses		1.327 ***		-0.469	2.948
Debt, Equity, and investment					
Debt transactions		-0.678		.	3.050 ***
Obligations and covenants		1.020		.	-1.482 **
Expectations / future					
Company expectations		3.086		.	1.546 **
Expected outcomes		-2.752		-0.708	-2.479 ***
Risk factor disclosures		.		-11.11 **	1.395
Risk factor disclosures 2		-1.810		.	2.593 **
Macro					
Market risk		-3.947 **		.	.
Operations					
Energy		-8.354 ***		-6.723 ***	-2.418 ***
Expenses		.		-1.785	2.016 ***
Investments		-0.283		8.416 ***	-0.676
Prices		-2.287		2.782	-2.297 **
Products		7.940 ***		1.158	1.307 ***
Sales of goods or assets		.		-2.143	-1.485 ***
Operating performance		.		8.029 ***	1.224 **
Changes					
Changes in expenses		-1.709		12.61 **	-1.090 *
Changes in revenue and expenses		-4.993 **		8.130 ***	-0.542
Decreases in value or performance		-1.981 **		.	-6.488 *
Decreases in expenses or performance		14.50 ***		.	0.043
Increase in expenses		28.41 ***		8.071	-0.396
Increases in income or revenue		5.335 *		-4.421 **	.
Increases in performance		-4.956 ***		.	0.132
Reduction in accounts		.		-16.10 ***	0.564

Table 8 (Continued): Predicting Future Material Weaknesses

Ungrouped					
Cash headings		21.06 **		-0.483	-0.832 *
Dollar amounts		.		18.52 ***	2.262 ***
Increases with time reference		-10.40 *		-0.158	-1.316 **
Modal weak statements		2.472 **		10.95 ***	.
Percents in year		-5.976		30.21 *	0.802 ***
Timing		.		11.20 **	-0.661
Dates with events		-5.662		.	1.451 **
Dates with unrelated statements		.		.	0.952 **
Reporting periods		-15.41 **		.	0.019
Location names		10.38 ***		3.572	.
Unrelated statements 1		6.438 ***		.	1.604 ***
Unrelated statements 3		.		0.203	0.739 **
Unrelated statements 4		-1.156		6.407 ***	-0.076
Unrelated statements 6		.		-0.440	2.060 **
Unrelated statements about banks		-5.158 ***		-1.686	-0.102
"Changes" or "differences"		11.48 **		.	0.331
"Company" + unrelated statements 2		9.258 ***		.	2.247 **
"Company" + unrelated statements 4		-3.655 ***		.	0.030
"Decrease" + unrelated statements		-6.324 **		23.40 **	0.395
"Full" or "Total" + unrelated statements		.		.	2.520 ***
"Increase" or "decrease" + unrelated statements		8.568 **		-0.310	3.962 ***
"Interest" + unrelated statements		.		.	0.789 **
"We/our" + operations statements		.		.	3.690 ***
Controls					
log(Market value)	-0.051 ***	-0.045 ***	-0.050 ***	-0.044 ***	-0.044 ***
log(BTM)	-0.032 ***	-0.025 ***	-0.033 ***	-0.026 ***	-0.027 ***
log(Share turnover)	0.030 ***	0.021 ***	0.029 ***	0.023 ***	0.021 ***
Pre-event FF alpha	-0.084	.	-0.094	.	.
I(Nasdaq)	0.017	0.009	0.015	0.009	0.017 *
FF48 Industry FE	Included	Included	Included	Included	Included
Adjusted R^2	0.019	0.028	0.019	0.027	0.033
# of negative and significant contexts		13		9	8
# of positive and significant contexts		12		12	19
Double LASSO adjustment					
Adjusted R^2		0.026		0.025	0.031
# of negative and significant contexts		13		9	8
# of positive and significant contexts		12		12	20

Columns (1) and (3) report linear regressions, while columns (2), (4), and (5) report LASSO regressions including all 137 contexts. All contexts are restricted to only the sentiment specified in the column. Only contexts that are significant at p<0.05 for at least one regression from columns (2), (4), and (5) are included. All regressions are based on 35,362 observations. P-values are indicated as follows: * indicates p<0.10, ** indicates p<0.05, and *** indicates p<0.01. A period indicates that the variable was dropped in the regression by the LASSO procedure. The bottom section presents the adjusted R-squared and number of significant contexts by sign (at p<0.05) when using a Double LASSO procedure as in Belloni, Chernozhukov and Hansen [2014] as described in Section 3.1.

Table 9: Other Sentiment Measures

Panel A: Predicting Filing Period Excess Return

Sentiment Approach used: Variable	Henry 2008 (1)	Harvard GI (2)	FinBERT (3)	LM (4)
Negative sentiment, document-level				
Coefficient	0.145	0.280 ***	-0.043 ***	-0.241 ***
Adjusted R^2	0.009	0.009	0.010	0.009
Negative sentiment, context-level				
# of negative and significant contexts	9	6	7	6
# of positive and significant contexts	4	5	7	6
Adjusted R^2	0.012	0.012	0.014	0.014
Positive sentiment, document-level				
Coefficient	0.109 *	0.112 **	0.017 **	-0.130
Adjusted R^2	0.009	0.009	0.009	0.009
Positive sentiment, context-level				
# of negative and significant contexts	7	4	4	6
# of positive and significant contexts	10	6	5	2
Adjusted R^2	0.014	0.013	0.011	0.012
Neutral sentiment, context-level				
# of negative and significant contexts	10	11	7	8
# of positive and significant contexts	2	2	7	5
Adjusted R^2	0.015	0.015	0.014	0.014

Panel B: Predicting Filing Period Abnormal Volume

Sentiment Approach used: Variable	Henry 2008 (1)	Harvard GI (2)	FinBERT (3)	LM (4)
Negative sentiment, document-level				
Coefficient	-7.086	-7.582	-1.662 ***	-1.253
Adjusted R^2	0.002	0.002	0.002	0.002
Negative sentiment, context-level				
# of negative and significant contexts	1	1	2	0
# of positive and significant contexts	4	8	8	7
Adjusted R^2	0.004	0.006	0.005	0.005
Positive sentiment, document-level				
Coefficient	-12.46 ***	-3.796	-1.888 ***	-28.69 ***
Adjusted R^2	0.002	0.002	0.003	0.002
Positive sentiment, context-level				
# of negative and significant contexts	0	0	0	1
# of positive and significant contexts	5	4	3	3
Adjusted R^2	0.006	0.005	0.004	0.005
Neutral sentiment, context-level				
# of negative and significant contexts	2	2	2	3
# of positive and significant contexts	5	10	6	5
Adjusted R^2	0.007	0.007	0.007	0.007

Table 9 (Continued) : Other Sentiment Measures

Panel C: Predicting Post-event Return Volatility

Sentiment Approach used: Variable	Henry 2008 (1)	Harvard GI (2)	FinBERT (3)	LM (4)
Negative sentiment, document-level				
Coefficient	1.940 ***	0.200	0.158 ***	1.585 ***
Adjusted R^2	0.087	0.085	0.087	0.089
Negative sentiment, context-level				
# of negative and significant contexts	7	7	2	8
# of positive and significant contexts	14	12	13	13
Adjusted R^2	0.091	0.091	0.092	0.093
Positive sentiment, document-level				
Coefficient	-0.900 ***	0.129	-0.072 ***	0.404
Adjusted R^2	0.087	0.084	0.085	0.084
Positive sentiment, context-level				
# of negative and significant contexts	8	20	7	8
# of positive and significant contexts	8	14	10	13
Adjusted R^2	0.090	0.094	0.090	0.088
Neutral sentiment, context-level				
# of negative and significant contexts	6	14	26	46
# of positive and significant contexts	19	17	11	6
Adjusted R^2	0.095	0.097	0.097	0.098

Panel D: Predicting Future Material Weaknesses

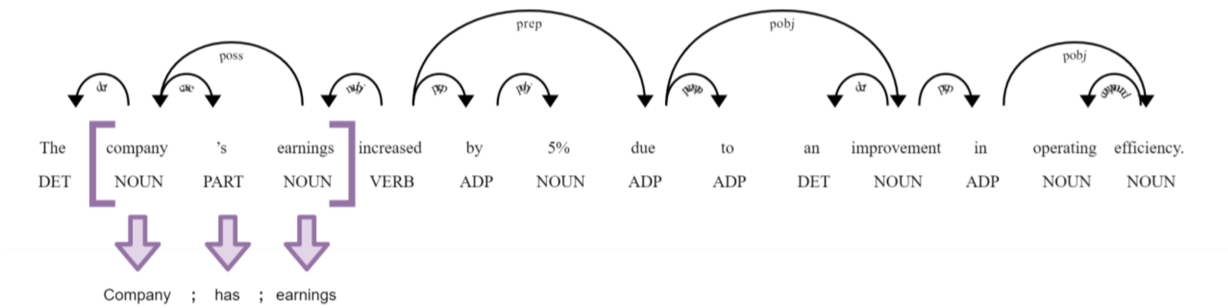
Sentiment Approach used: Variable	Henry 2008 (1)	Harvard GI (2)	FinBERT (3)	LM (4)
Negative sentiment, document-level				
Coefficient	-6.000 ***	-1.158	-0.023 **	-1.065
Adjusted R^2	0.019	0.019	0.019	0.019
Negative sentiment, context-level				
# of negative and significant contexts	9	12	10	13
# of positive and significant contexts	12	11	15	12
Adjusted R^2	0.027	0.030	0.027	0.028
Positive sentiment, document-level				
Coefficient	-0.716	-0.695	-0.311 ***	-5.759 ***
Adjusted R^2	0.019	0.019	0.019	0.019
Positive sentiment, context-level				
# of negative and significant contexts	9	18	5	9
# of positive and significant contexts	11	13	4	12
Adjusted R^2	0.027	0.031	0.024	0.027
Neutral sentiment, context-level				
# of negative and significant contexts	7	13	9	8
# of positive and significant contexts	17	16	22	19
Adjusted R^2	0.032	0.033	0.034	0.033

Panels A, B, C, and replicate the tests from Tables 5, 6, 7, and 8, respectively. In each panel, columns (1), (2), and (3) reports results using the Henry [2008] dictionary, Harvard General Inquirer dictionary, and FinBERT, respectively. Column (4) shows the original results presented in Tables 5, 6, 7, and 8 using the LM dictionary. Document level tests report the coefficient on the sentiment and adjusted R^2 as in columns (1) and (3) of Tables 5 to 8. Contet-level tests report the number of negative and significant coefficients at $p<0.05$, the number of positive and significant coefficients at $p<0.05$, and the adjusted R^2 as in columns (2), (4), and (5) of Tables 5 to 8. All regressions are based on 35,362 observations. For coefficients, P-values are indicated as follows: * indicates $p<0.10$, ** indicates $p<0.05$, and *** indicates $p<0.01$.

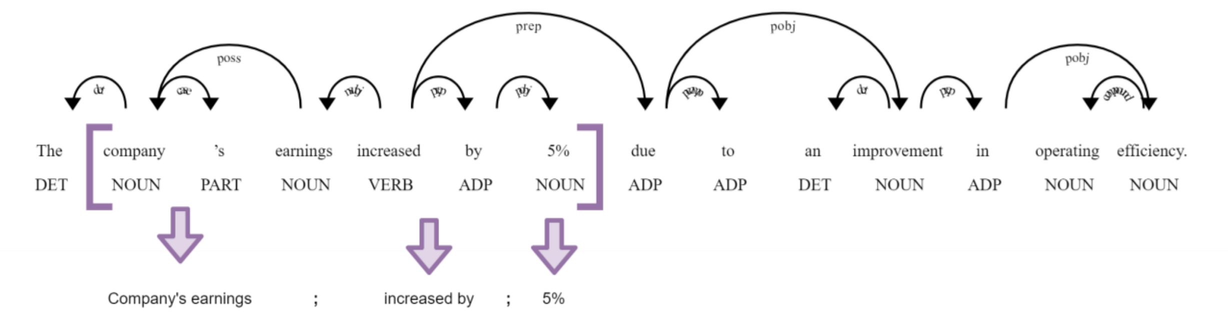
Figures

Figure 1: Extractions of a single sentence

Panel A: First extraction



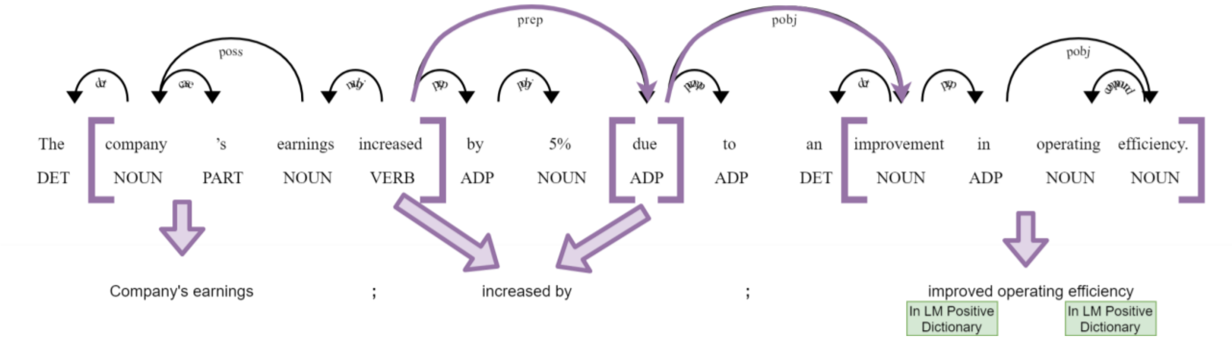
Panel B: Second extraction



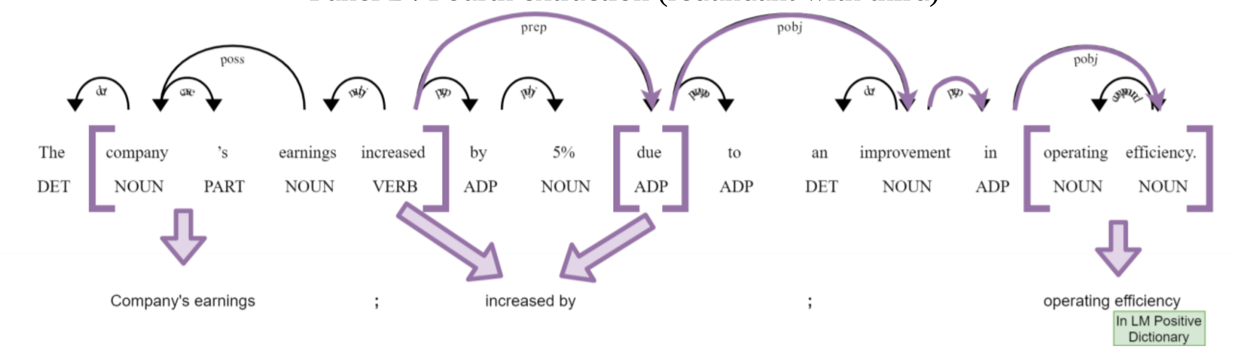
(Continued…)

Figure 1 (Continued)

Panel C: Third extraction

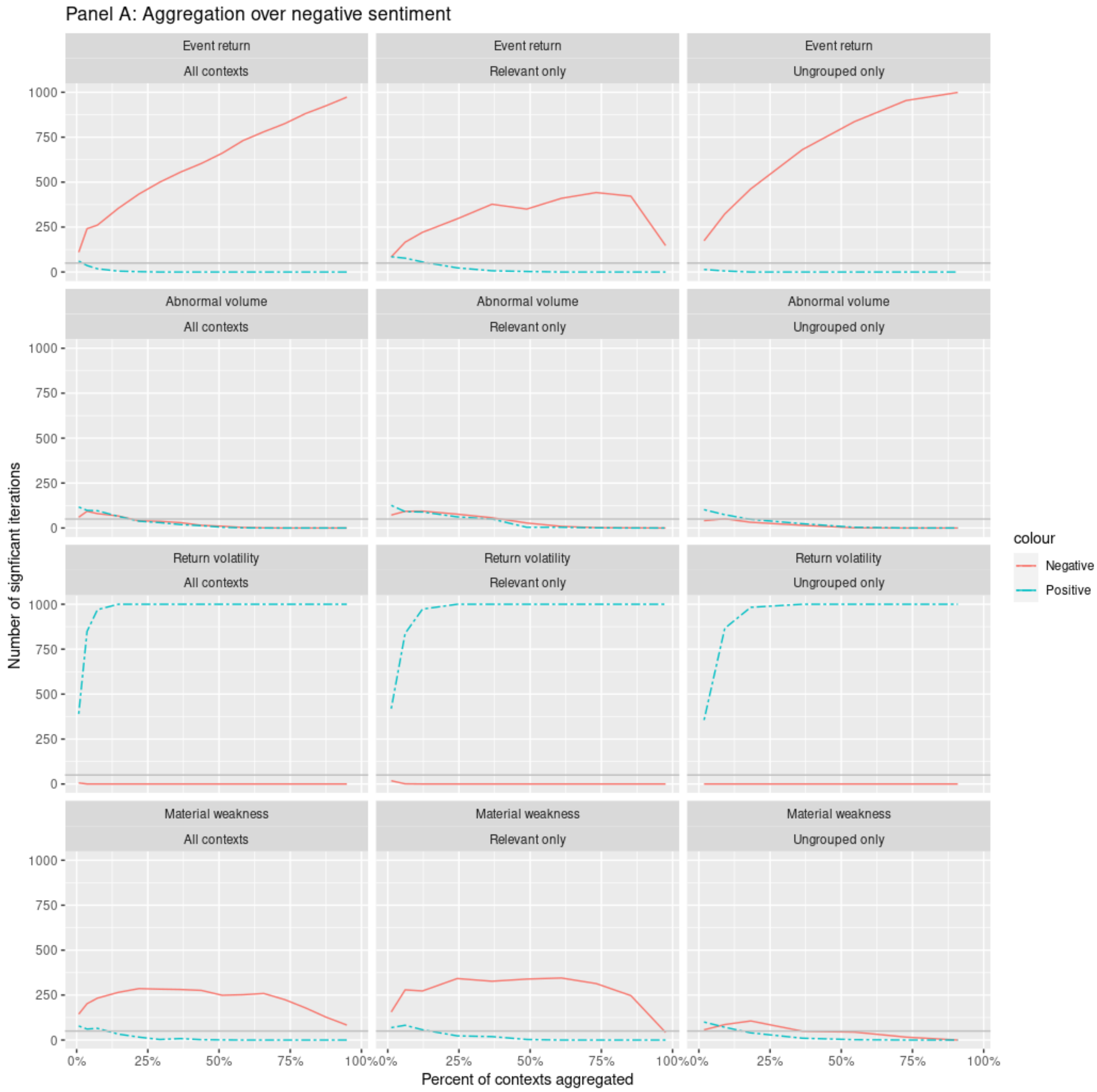


Panel D: Fourth extraction (redundant with third)



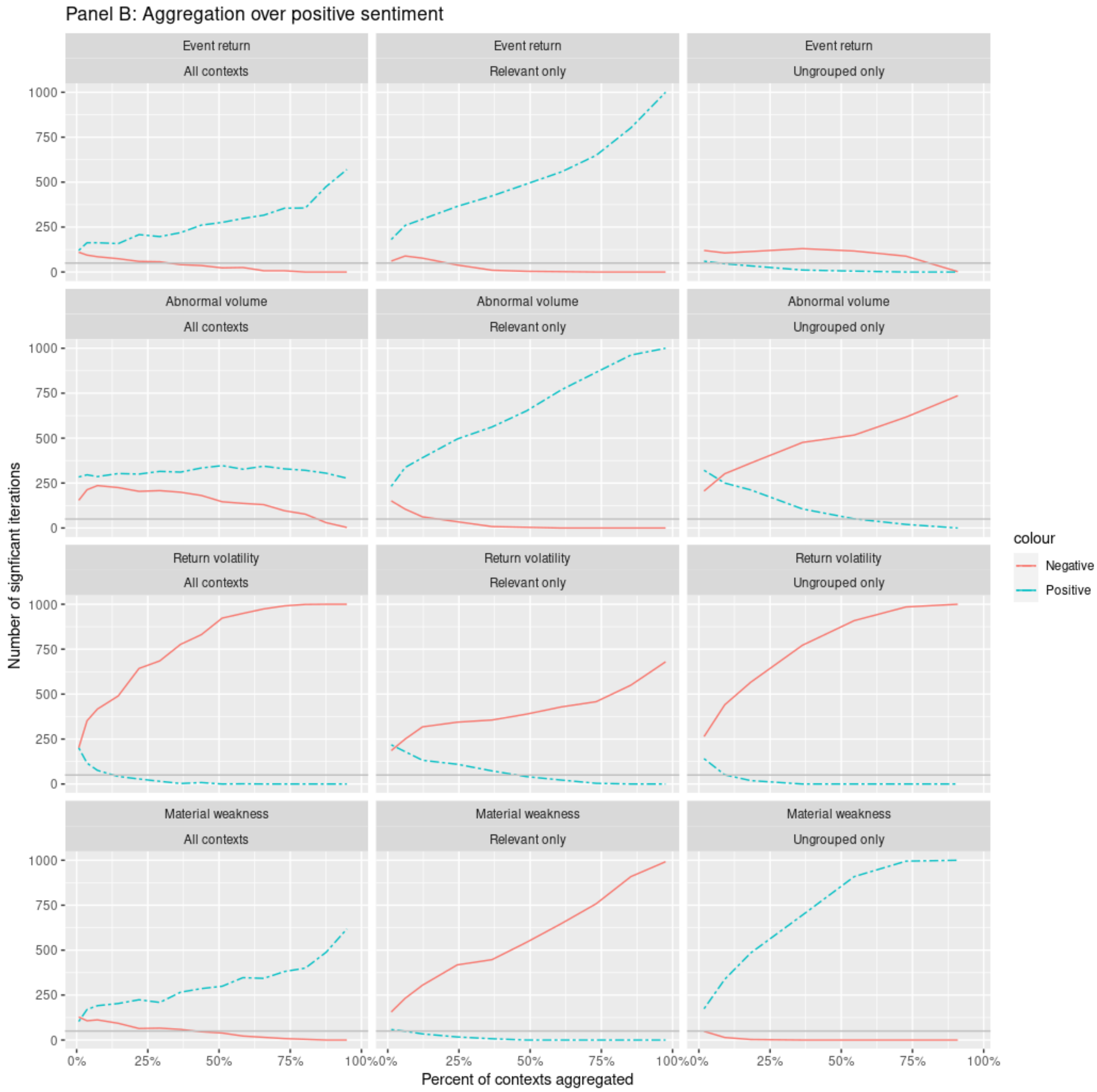
This figure shows all four triples constructed by Open IE for the sentence: "The company's earnings increased by 5% due to an improvement in operating efficiency." The black arrows along the top of the sentence show the parse tree that underlies Open IE's computation, while the purple arrows show the path taken by Open IE in constructing each triple. At the bottom of each panel is the triplet constructed by Open IE of the form (subject; relation; object). Words in the LM dictionaries, if any, are highlighted below the word in the triple.

Figure 2: Simulations of Aggregation



(Continued…)

Figure 2 (Continued)



These graphs show the results of simulations varying the level of aggregation of a sentiment measure. We randomly aggregate *n* coefficients across contexts at various levels of *n*, at 2, 5, 10, and each multiple of ten less than the number of contexts. We conduct this random process 1,000 times for each *n*. Panel A (Panel B) shows the number of times the aggregated coefficient is statistically significant across 1,000 iterations when aggregating across negative (positive) sentiment measures. The first, second, third, and fourth rows of each panel replicate, respectively, the results of Tables 5, 6, 7, and 8. The first column of each panel presents results allowing aggregation over all 137 contexts we identified. The second column only allows for aggregation over the 82 contexts that relate to *accounting*, *business operations*, or *changes*. The third column only allows for aggregation over the 55 *ungrouped* contexts. Coefficients were considered as significant if their p-value is less than 0.05.