Understanding Sentiment Through Context

Richard M. Crowley
School of Accountancy
Singapore Management University

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

April 25, 2021

Preliminary and Incomplete

Abstract
We examine the extent to which results based on financial sentiment of U.S. annual reports are conditional on the underlying context from which financial sentiment is derived, as well as the extent to which financial sentiment is related to the underlying context of the annual report. To achieve this, we construct a measure of *context* that is based on the grammar, syntax, and content of sentences in each report. We then apply sentiment measures to the phrases within each context to examine how sentiment is related to each context, and under which contexts financial sentiment works as expected or not for a variety of prediction problems. We show that sentiment encompasses a wide variety of contexts, and that positive and negative sentiment respond to different contexts. In addition, we show that there is significant noise in predicting various outcomes (stock return, volume, volatility, and material weaknesses). Specifically, only select contexts drive the primary results of each analysis, and these select contexts vary by the outcome being predicted. Furthermore, under some contexts we find results opposite to expected predictions, indicating a nontrivial amount of systematic noise or error in sentiment classification.

1. Introduction

This paper examines the use of the bag-of-words method, especially the use of the word lists sentiment[1] dictionaries of Loughran and McDonald (2011), in accounting and finance research. 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 et al. 2017; El-Haj et al. 2019). 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 meanings from the document. Given the popularity and simplicity of this method, we believe researchers will find it useful to know how well it works in general and in specific contexts.

The word lists used in the literature vary from a few attribute-specific keywords to a dictionary with over 100 attributes. The former includes Li (2006) and Loughran et al. (2009). 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. The latter, such as Tetlock (2007) and Kothari et al. (2009), uses the Harvard IV General Inquirer word lists that include over 100 attributes. The rise in the popularity of textual analysis in accounting and finance research has led to the development of finance-specific word lists by Henry (2008) and Loughran and McDonald (2011). The Loughran-McDonald (henceforth LM) word lists have since become the most used word lists for analyzing financial documents.[2] Besides general-purpose word lists, various studies have created custom word lists to capture context-specific attributes: Managerial deception or extreme negative and positive words (Larcker and

[1] We use the terms “sentiment” and “tone” interchangeably throughout this paper.

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

Zakolyukina 2012), competition (Li et al. 2013), financial constraints (Bodnaruk et al. 2015), corporate culture (Audi et al. 2016), firm complexity (Loughran and McDonald 2020), and extreme language (Bochkay et al. 2020).

The key assumption of the bag-of-words method is that each word is independent. Hence, it ignores text order, sentence structure, and grammar when “calculating” the meaning of sentences. Obviously, this assumption does not reflect how language works. Two alternative methods have been used to overcome this shortcoming: Naïve Bayes and topic modelling.

The Naïve Bayes method 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 label 1,000 stock message board postings and then use them to train a Naive Bayes algorithm to classify posting tone. Similarly, Li (2010b) and Huang et al. (2014), among others, use pre-labelled training data to “teach” Naive Bayes models to interpret the content of 10-K filings and analyst research reports, respectively. Azimi and Agrawal (2021) use 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.[3]

Topic modelling is an unsupervised machine learning technique, which looks for patterns in how words covary within and across documents in a bag of words manner. Dyer, Lang, and Stice-Lawrence (2017) use Latent Dirichlet allocation (LDA) to identify the major topics that led to an increase in the length of 10-K reports over time. Huang, Lehavy, Zang, and Zheng (2018) quantify the information intermediary role of analysts by applying LDA to extract the common

[3] Siano and Wysocki (2020) apply the BERT language model, which is developed by Google and pre-trained on unlabeled data, to capture context rather than words. Yang, Uy and Huang (2020) train a BERT language model specifically for financial contexts, termed FinBERT.

topics being discussed in both earnings conference calls and analyst reports. Brown, Crowley, and Elliott (2020) use LDA to obtain a set of semantically meaningful topics for predicting intentional misreporting.

Despite the availability of alternative methods that take into account word dependency, the bag-of-words approach is still the most popular for textual analysis (El-Haj et al. 2019, Figure 1). Given its widespread application in accounting and finance research, we believe it is important to investigate whether the bag-of-words approach works as intended in general and in specific contexts. Moreover, we want to develop a new approach for analyzing contextual meaning.

Our approach includes four steps. First, we process each sentence in the document using open information extraction (hereafter, Open IE). Open IE is a natural language processing method that summarises a sentence into relation triples in the form of (subject; relation verb; object). Second, we subset Open IE triples to remove redundant extractions. This is done at the sentence level and aims to keep a set of extractions that are each as short as possible, yet without dropping triples that include any accounting/finance terms or words from the LM positive and negative word lists. We use the accounting/finance terms in Campbell Harvey's hypertextual finance glossary and NYSSCPA's Accounting Terminology Guide.[4] Third, we concatenate the Open IE triples to form "extractions" and apply the Universal Sentence Encoder algorithm across all extractions to get a 512-dimensional representation of the phrases' meanings (Cer et al. 2018). Lastly, we cluster all phrases across the 512-dimensional vector space using Mini-Batch K-Means (a variant of k-means that uses less memory by batching vectors). This process is optimized using the Gap statistic (Tibshirani et al. 2001), which provides an objective criterion

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

for an optimal number of clusters. Finally, we group the Open IE extractions using the 131 clusters provided by the Mini-Batch K-Means algorithm.

We conduct our analysis using the MD&A section of 35,362 10-K filings for the period from 1994 through 2018. The above process generates 131 clusters, with each cluster containing an average of 173,047 Open IE extractions (ranging from 40,803 to 403,925 extractions). We manually label each cluster according to the common themes of the extractions in the cluster and rate the presence of positive and negative sentiment in each cluster based on our reading of a random sample of extractions in the cluster. Moreover, we validate the clustering process using an intrusion task. In particular, we take three words from one cluster and one from another (the intruder), randomize the order, and test if the intruder can be picked out.

We first examine whether the Loughran and McDonald (2011) sentiment measures are related to the clusters in the direction we predicted according to the sentiments of the extractions in the clusters. We regress the LM negative and positive measures of the MD&A section on the 131 clusters, using the least absolute shrinkage and selection operator (lasso) regression method with 10-fold cross-validation. We find that 96 and 88 clusters exhibit significant explanatory power for the LM negative and positive sentiments, respectively. There are more estimated coefficients with the expected signs in the LM negative sentiment regression than in the positive sentiment regression, consistent with the LM sentiment measures being better at capturing negative tone than positive tone. Finally, we divide the clusters into four groups based on their association with the LM sentiment measures: high sentiment, skewed toward negative sentiment, skewed toward positive sentiment, and low sentiment. We find that out of the 31 (14) clusters where we expect to find negative sentiment, only 14 (4) are positively related to LM negative (positive) sentiment.

Next, we examine the ability of the LM tone 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 131 sentiment-by-cluster variables. Three sentiment-based variables are created for each cluster to measure the percentage of LM negative, positive, and neutral sentiment extractions in the cluster, respectively.

The filing-period excess return regressions show that nine clusters have a negative effect and seven have a positive impact on excess return in the negative sentiment regression. Similarly, seven and six clusters exhibit a statistically positive and negative coefficient estimates in the positive sentiment regression. In the neutral sentiment regression, 18 clusters are significantly associated with filing-period excess returns. Taken together, these findings suggest that sentiments do not capture variation in context across the 131 clusters. Moreover, we find that sentiment for three clusters is content driven (i.e., the signs of the estimated coefficients are the same for both positive and negative sentiments), 13 are sentiment driven (i.e., the sign is negative for negative sentiment and positive for positive sentiment), and 12 are sentiment driven but the signs are counter to expectations (i.e., the sign is positive for negative sentiment and negative for positive sentiment).

The abnormal volume and stock volatility regressions tell a similar story. While we find that most sentiment-driven clusters have a positive effect on filing-period abnormal trading volume, the effect is concentrated in only seven clusters. For post-filing stock volatility, we find a large number of sentiment-driven clusters pointing in both positive and negative directions. In other words, while sentiment generally increases stock volatility, higher sentiment content in some clusters reduces it. This finding is different from that 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 clusters predict more material weaknesses, while seven clusters predict less. Similarly, more clusters positively predict material weaknesses (ten versus six) for positive sentiment. This latter finding is opposite to our expectation and inconsistent with the findings documented in Loughran and McDonald (2011).

If our approach randomly assigns extractions (of phrases/sentences) into the 131 clusters, we would expect the estimated coefficients on these clusters to have the same sign and similar magnitude as the LM sentiment measures. However, our approach does not assign the extractions to the clusters randomly. Instead, it preserves syntax and grammar, and the clusters are formed according to the similarity of the "context" among the extractions. Hence, our findings, that most of the estimated coefficients on the sentiment-based clusters are different in sign and magnitude than those on the LM sentiment measures, are consistent with context mattering in understanding sentiment.

This study makes two contributions to the textual analysis literature in accounting and finance. First, we evaluate the impact of a critical assumption of the bag-of-words method that words are independent, by examining whether the Loughran-McDonald (2011) sentiment measures work as intended regardless of context. We document evidence consistently showing that context matters in understanding the meaning of sentiment. Second, the approach we developed for this analysis demonstrates a potential approach for parsing contextual meaning. It is similar in spirit to the topic modelling alternative, in that it can be used to agnostically assign text to groups based on some definition of meaning. However, our context approach is finer grained, able to accurately classify short snippets of text (parts of sentences), whereas topic modeling excels at labeling large sections of text (paragraphs to full documents).

Section 2 describes the methodology behind our approach. Section 3 presents the sample and research design. Section 4 reports the empirical findings and Section 5 concludes.

2. Methodology

In order to develop our measure of context, we collect all annual 10-K reports from 1994 to 2018. We process each annual report using the python parser developed in Brown, Crowley and Elliott (2020), including using the same Management Discussion and Analysis (MD&A) regex-based extraction method. As shown in Table 1, we have processed 208,169 annual reports (188,030 10-K filings and 20,139 10-K405 filings), netting a total of 107,596 MD&A sections. We construct our context approach on this full set of MD&A filings, though we restrict its application and our analysis to the 35,362 MD&A filings that match all of the requirements listed in Table 1. We rely on MD&A sections of annual reports due to the computational complexity of our approach.[5]

After extracting all MD&A filings, we parse them using Stanford NLP's open information extraction algorithm, Open IE (Gabor, Premkumar and Manning 2015). OpenIE is a method used to extract "relation triples," i.e., snippets of text from sentences of the form (subject; relation verb; object). OpenIE 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

[5] Running OpenIE across all MD&A filings takes ~6.5 days on a 6 core processor (using 11 threads). Processing all extractions using Universal Sentence Encoder is efficiently done on a GPU (GTX 1060) in around 2 hours. The Mini-Batch K-Means procedure takes around 1 week to run. The remaining operations described in this section take only minutes to run. All three parts of this scale somewhat linearly (or worse) with the number of sentences processed; as such, what takes around 2 weeks to run on MD&As would take around 4 months to run on full text 10-K filings on the same workstation.

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). The second step is to extract self-contained 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 clauses that are able to stand on their own as sentences. The final step is then to segment the clauses into the subject, relation verb, and object portions of the triples. This is done entirely using the dependency tree using a set of six linguistic patterns.

As an example, consider the following phrase: "The company's earnings increased by 5% due to an improvement in operating efficiency." This sentence has a few key takeaways: 1) it is discussing earnings, 2) earnings increased by 5%, and 3) the 5% increase is due to operating efficiency. The OpenIE extractions for this sentence 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 perfectly match the three key takeaways from the sentence. As such, we can see that OpenIE 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 OpenIE method: it frequently generates excess extractions with slight differences in wording. We manually handle this issue in the third step of our methodology.

Applying OpenIE 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. In order 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 processed 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 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. While both accounting and finance glossaries predominantly contain terms that are 1 word long, both contain phrases as terms as well. For phrases (2 or more words), we first determine which would already be flagged based on the individual word terms within each dictionary and discard them. 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 content; 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 our 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 OpenIE 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.

At this point we keep all remaining extractions throughout, but we still need to reduce the dimensionality of these extractions in order to be able to make sense of them more broadly. To accomplish this, we use Universal Sentence Encoder (USE) along with Mini-Batch K-Means, both algorithms developed at Google. We use the Deep Averaging Network (DAN) variant of USE[6] by Cer et al. (2018). This model takes a snippet of text, and, based on both word order and the words themselves, maps the snippet to a 512-dimensional vector space, where each dimension of each vector is bounded between -1 and +1. While the dimensions themselves are 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 extractions, USE is a natural choice. This method has been used in the accounting literature by Crowley, Huang, and

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

Lu (2020), in order to show the similarity in meaning between tweets from executives and their CEOs.

After mapping all 48,576,229 extractions to USE's 512-dimensional vector space, we then apply a clustering method to gather together extractions that are similar in meaning. Since USE relies on Euclidean distance to measure similarity, we use a variant of K-Means, as it clusters based on Euclidean distance. The 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 200GB), 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 and run it at a variety of cluster counts. We then optimize the number of clusters using a simulated bootstrapping technique based on Tibshirani et al. (2001), in order to construct their Gap statistics measure. The criterion for the Gap statistic is intuitive – an optimal number of clusters, n, is the lowest n such that the error at n clusters is within a certain bound from the error at n+1 clusters, adjusted for the variation in error at n+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 131 was the optimal number of clusters under the Gap statistic.

Lastly, using the output of the Mini-Batch K-Means algorithm at 131 clusters, we assign each extraction to a cluster based on the closeness of the cluster centers. These cluster assignments constitute the final measure that is used throughout our tests.

2.1 Labelling Clusters

To interpret the clusters, we start by hand-labeling each cluster. To do this, we randomly pull 10 extractions from each cluster, interpreting them to determine a label. For any cluster that was ambiguous, we pull an additional 30 random extractions to examine. The output of this process is presented in brief in Appendix A, showing two of the ten extractions used for labeling. At the same time as the labeling exercise, we also hand-code a broader classification (presented in Appendix A as well), along with whether we notice any positive or negative sentiment contained in the extractions.[7]

Based on our hand-coding, we find there to be eight primary types of clusters. A particularly relevant set of clusters focused on *Accounting* makes up 36 of the 131 clusters, covering topics such as accounting assumptions, assets, cash flows, various tax matters, profitability or losses, and revenue. Another relevant set of clusters, *Business Operations*, includes 37 clusters covering everything from company descriptions to financial services, manufacturing, risk, leasing, and R&D. We also document 8 clusters related to *Contracting* and 6 clusters related to *Regulation*. In total, these business-focused clusters comprise 87 of our 131 clusters. The remaining clusters tend to focus more on grammar or language constructs. We find that 9 clusters are fairly generic and related to *Changes*, while 20 clusters relate to other generic grammar constructs such as dates, dollar amounts, usage of we/our (first person plural forms), instructions or references, and modal strong phrases. These types of clusters are largely attributable to short extractions from OpenIE that capture ancillary details and are likely unavoidable. We also identify 8 clusters related to mentions of specific years – we intend to remove these from the next version of the model by masking all years and dates before providing the data to OpenIE. Lastly, we find 7 clusters which we term *Ungrouped*. These clusters pick up extractions that do not otherwise fit anywhere. To some extent, at least some of these clusters are unavoidable as natural language itself is not naturally clustered, and thus there is likely to always be some extractions that do not

[7] We note that for this to be a reliable measure, more extractions should be examined per cluster and more and independent coders should be used. We intend to do this in a future draft of the study.

match the rest. On the other hand, certain refinements to our next iteration of the model may remove some of these ungrouped clusters.

Table 2, Panel A (Panel B) presents the most and least frequent clusters based on the number of extractions in the cluster (number of documents with at least 1 cluster in the extraction). While the ungrouped text is the most common individual extraction type, all ungrouped clusters combined comprise only 8% of the sample. The next most common clusters discuss basic company information, interest rates, sales, and other metrics, future uncertainty, and geographic locations. The least frequent clusters focus on fine-grained issues like tax, effective rules, and accounting policies, as well as boilerplate text. Based on the number of documents represented, ungrouped text is again widespread, as expected. More interestingly, we find that almost all MD&A sections discuss increasing metrics, future uncertainty, and expenses, and inclusion of modal strong statements is also quite common. The least widespread topics include oil and gas, partnerships, tax positions, and loan impairment. Also infrequent are clusters about certain years, though this is likely mechanical (since the year 2006 is mostly only mentioned in filings from 2006, whereas non-year topics are spread more evenly across years).

Panels C through F introduce the primary twist we apply to our cluster measure: subsetting the clusters based on the LM sentiment dictionaries. Panel C shows which clusters contain the highest and lowest number of negative extractions, where a negative extraction is defined as an extraction with more terms in the LM negative dictionary than it has in the LM positive dictionary. We see that the Losses cluster has by far the most common negative context, with more than double the number of negative extractions as compared to the second highest cluster. Other commonly discussed clusters under the LM negative dictionary are about uncertainty, impairment, decreases and declines, and insurance. The least common contexts include boilerplate, tax positions, partnerships, FASB statements, and attributions for increases. Panel D looks at a related construct: the clusters in which negative sentiment makes up the highest and lowest proportion. There is a lot of similarity across Panels C and D, though discussions of net figures, accounting policies, legal or compliance matters, or valuation are frequently flagged as negative. On the flip side, the clusters least commonly flagged as negative includes increase in expenses, which seems inconsistent with the goal of the negative dictionary.

Panels E and F present similar statistics for positive extractions, where a positive extraction is defined as an extraction with more terms in the LM positive dictionary than it has in the LM negative dictionary. We see that, rather unexpectedly the top two clusters are tax rates and effective rules and policies. Both of these do not appear to be positive topics, but they are easily explained by the inclusion of one word in the LM positive dictionary: “effective.” Within these clusters, the word “effective” is picking up phrases like “effective tax rate,” “effective date” and “Rule is effective,” to the tune of over 100,000 such extractions. This may in part explain the lack of effectiveness of positive LM sentiment, as this is likely adding a lot of noise to the measurement of positive sentiment.

Looking at the other clusters, however, we see that the LM positive dictionary is picking up some useful clusters as well, such as “increases and improvements” and profitability. In terms of the least common clusters, accounting standard issuance, losses, cash flows, boilerplate, and cost/expense mentions all rarely are flagged as positive, which is as expected.

2.2 Validation

To validate our clusters, we conduct an intrusion task following Brown, Crowley and Elliott (2020). The task is presented as a series of multiple-choice questions, asking “Which phrase does not belong?” along with presenting four alternatives. Three of these alternatives are from the same cluster, while a fourth, the intruder, is from a different cluster. If the clusters are intelligible, then the test taker should do better than chance at identifying ithe intruder. We intend to run a full-scale test of this experimentally at a later date. One researcher and one research assistant have taken the instrument thus far, averaging 72% correct out of 50 questions each. Compared to the experimental results from Brown, Crowley and Elliott (2020), this is quite a high score, though we caution that the sample size is too small to extrapolate from.

Our second validation approach is a regression approach. Using a regression structure defined in Section 3.1 below, we regress the 10-K MD&A LM sentiment scores on the sentiment restricted components of our clusters, aggregated at the filing level. We find that our extraction-based negative sentiment measures capture 82.2% of the MD&A-based LM negative sentiment score, as indicated by adjusted R^2, while our extraction-based positive sentiment measures capture 67.4% of the MD&A-based LM positive sentiment score. Taken together, this shows that our extractions capture the majority of the content of the filing, when defining content as the context in which LM sentiment words are contained.

3. Sample and Research Design

Table 1 describes the sample selection process. The sample covers the period from 1994 through 2018. 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. The sample drops to 101,877 after removing filings that cannot be parsed by OpenIE, 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 ourselves. We obtain material weakness counts from Audit Analytics.

Table 3 reports summary statistics of various sentiment measures, our extractions, and the dependent variables and controls used in our regressions. As the Loughran McDonald data library only provides full-text sentiment scores, we present our parser's full-text scores alongside the full-text scores from Loughran and McDonald's data to allay any concerns about them being too different. While the mean and median for our negative dictionary is slightly lower, the mean and median for positive sentiment are quite similar. Untabulated correlations show that our negative sentiment measure 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 53.8% for negative and positive sentiments, respectively, and likewise have univariate statistics that deviate a bit more. This is as expected, since the MD&A talks about a potentially different set of issues and contexts as compared to the full-text filings.

For our extractions, we find an average of 641.1 extractions per MD&A. As such, there is quite a lot of context in the average MD&A. When filtering extractions based on the LM sentiment dictionaries, we find that there are 36.6 negative and 20.1 positive extractions per MD&A, on average. Since extractions remove a lot of non-content bearing words, we see that the ratio of sentiment-containing extractions to all extractions is much higher than the word-based sentiment measures. Lastly, summary statistics for the dependent and control variables are presented. These variables are defined in Appendix C.

3.1 Empirical approach

Throughout our tests, we use a consistent framework in constructing our regressions. For our first tests investigating the relationship between LM sentiment and context, we use regressions of the following form:

$$Sentiment_{f,t} = \alpha + \sum_{i=1}^{131} \beta_i \, Cluster_{i,f,t} + \gamma \cdot Controls_{f,t} + \delta \cdot Industry\ FE + \varepsilon. \qquad (1)$$

To control for potential issues stemming from multicollinearity, we estimate equation (1) using the least absolute shrinkage and selection operator (i.e., lasso regression) (Tibshirani 1996). 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:

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

we are instead minimizing the following:

$$\min_{\beta,\gamma,\delta\in\mathbb{R}} \frac{1}{N} \ |\varepsilon|_2^2 + \lambda \left[\sum |\beta|_1 + \sum |\gamma|_1 + \sum |\delta|_1\right]. \qquad (3)$$

The additional term in equation (3) as compared to equation (2) 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 what is determined via the 10-fold cross validation. To derive p-values, we reimplement the resulting lasso model in a linear model.

The *Cluster* measures are defined as the number of extractions in a given cluster in a given MD&A divided by the total number of extractions in that same MD&A. 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 a [-252,-6] trading day window, where day 0 is the 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 clusters in predicting four outcome variables from Loughran and McDonald (2011). We first use the following regression structure to replicate results from Loughran and McDonald (2011):

$$DV_{f,t} = \alpha + \beta MD\&A\ Sentiment_{f,t} + \gamma \cdot Controls_{f,t} + \delta \cdot Industry\ FE + \varepsilon. \qquad (4)$$

The dependent variables in this specification are one of the following: filing period excess return, filing period abnormal volume, post-event return volatility, or future material weaknesses. The control variables and fixed effects are the same as with equation (1).

We then use the regression below to examine the LM sentiment measures across different contexts. The regression specification parallels equation (4), except it adds in our cluster measures conditional on the sentiment of the regression:

$$DV_{f,t} = \alpha + \beta_0 MD\&A\ Sentiment_{f,t} + \sum_{i=1}^{131} \beta_i Cluster_{i,f,t} | Sentiment + \qquad (5)$$

$$\gamma \cdot Controls_{f,t} + \delta \cdot Industry\ FE + \varepsilon$$

The dependent variables, controls, and fixed effects are all the same as equation (4). The sentiment-based cluster variables are measured as the number of extractions within a cluster labeled as the given sentiment, divided by the number of extractions in the document. We also run equation (5) for neutral sentiment; for neutral sentiment we drop the *MD&A Sentiment* measure, and we define our cluster measure to be neutral for any extraction that is not labeled as positive or negative.

4. Empirical Findings

4.1. Loughran-McDonald Sentiments and Content of the Clusters

In the first set of the analysis, we validate the LM tone measures based on the contents of the clusters by regressing LM sentiments of the MD&A section on the 131 clusters. Table 4 Columns 1 and 4 show the expected sign of clusters in the regression according to our reading of the Open IE extractions included in the cluster.

Column 2 summaries the estimation results of the LM negative sentiment regression with control variables and fixed effects included following equation (1). The estimated coefficients on 96 of the 131 clusters are significantly different from zero at 5% or less (58 negative and 38 positive). Moreover, 13 of the significantly positive coefficients are consistent with the prediction reported under Column 1 (e.g., clusters for “loan impairment” and “net figures”). However, 8 of the estimated coefficients are statistically negative, which are inconsistent with our expectation (e.g., clusters for “increases in expenses” and “credit facilities and agreements”). Column 3 reports the estimated coefficients on the clusters from the LM positive sentiment regression. Of the 131 clusters, 49 and 39 have statistical negative and positive coefficients, respectively, at the 5% significance level. While five clusters have the predicted positive coefficients (e.g., clusters for “decreases and increases in financials” and “increases and improvements”), five exhibit a negative coefficient which is opposite to our expectation shown in Column 4 (e.g., clusters for “increases in metrics” and “expense change details”). Taken together, the results reported in Columns 2 and 3 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).

Table 4 is partitioned into five groups (A-E), according to the results under Columns 2 and 3. We label the first group “high sentiment,” which include 20 clusters that are positively related to both LM negative and positive sentiments. They may contain a higher level of sentiment, but potentially non-directionally on average. For example, the “Net figures,” “Markets (product, regional, financial)”, and “Future uncertainty” clusters are in this group. The second group makes up 19 clusters that load positively for negative sentiment and negatively or insignificantly for positive sentiment, indicating that these clusters skew towards negative LM sentiment content. “Legal/Compliance”, “Estimates”, and “Losses” are examples of clusters in this group. In contrast, the third group has 22 clusters that load negatively or insignificantly for negative sentiment and positively for positive sentiment, suggesting that the clusters skew towards positive LM sentiment content. Members of this group include “Interest income”, “Investments”, and “Increases and improvements.” The fourth group, called “low sentiment,” comprises of 32 clusters. These clusters indicate less sentiment across the board. Some appear to be more matter-of-fact issues, including the “Subsidiaries”, “Contracts”, and “Fair value” clusters. The remaining group includes all clusters that are either insignificant for one sentiment and significantly negative for the other, or insignificant for both sentiments.

4.2. Loughran-McDonald Sentiment Measure of Content Clusters

To shed light on the ability of the LM sentiment measures to capture contextual meaning, we run four regressions from Loughran and McDonald (2011), with the LM sentiment measures of our 131 clusters added to the regressions. In particular, we create three sentiment-based variables for each cluster. The first variable is equal to the percentage of Open IE extractions in the MD&A with positive LM tone in each cluster. The second and third variables equal to the percentages of extractions that have LM negative and LM neutral tones in a cluster, respectively. We run regressions on each set of these 131 sentiment-based variables (i.e., positive, negative, or neutral), the LM sentiment measure of the MD&A section, the control variables used in section 4.1, as well as Fama-French (1997) 48 industry fixed effects (as used in Loughran and McDonald 2011). We consider four dependent variables: filing period excess returns, filing period abnormal volume, post-filing return volatility, and future material weaknesses.

4.2.1. Filing Period Excess Returns

We first examine the stock market reaction to the content of the 131 clusters, 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.

Table 5 reports the estimated coefficients from two linear regressions following equation (4) and three lasso regressions following equation (5) in which filing period excess return is regressed on the 131 variables capturing the percentage of LM negative tone (Column 2), LM positive tone (Column 4), and LM neutral tone (Column 5) in the corresponding clusters.

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 16 of the 131 contexts at the 5% level (nine negative and seven positive). Given that the variables are measuring the negative tone of the content in the clusters, the seven positive estimates are inconsistent with the intended purpose of the negative LM sentiment measure. Similarly, Column (3) finds that positive sentiment is not association with excess return (the estimated coefficient is significant at the 10% level). In comparison, Column (4) reports that only seven of the estimated coefficients on the 131 variables capturing the positive sentiment of the clusters are statistically positive (versus six negative). In other words, seven clusters with a higher percentage of optimistic extractions are associated with higher excess returns in the filing date event window, but six are associated with lower excess returns. Moreover, the results in Column (5) (partially tabulated) show that 18 of the estimated coefficients are statistically significant (six negative and 12 positive), suggesting that the market reacts to the content of these 18 clusters, even though the LM measures consider the tone of these clusters to be neutral.

An alternative way to interpret the findings is to examine the signs of the estimated coefficients on each cluster across the three regressions reported in Columns (2), (4), and (5). If the signs are the same for positive and negative (and neutral) sentiments, the significant market reaction is driven by the content, rather than the sentiment, of the cluster. In other words, the sentiment of the cluster's content does not matter for the cluster's effect on market reaction. If the sign is negative for negative sentiment and positive for positive sentiment (as well as either positive or negative for neutral sentiment), the result is sentiment driven and consistent with the LM measures working as intended. Finally, if the sign is positive for negative sentiment and negative for positive sentiment (either positive or negative for neutral sentiment), the result is sentiment driven but opposite to what the LM measure is supposed to capture.

Referring back to the results in Table 5, we find three content driven clusters (“accounting assumptions,” “Selling,” and “Years (2000--2002)”), 12 sentiment driven clusters, and nine clusters that are sentiment driven but counter to our expectations. Since only 12 of the 131 clusters are sentiment driven, it indicates that the LM sentiment measures do not perform well under most contexts.

4.2.2. Filing Period Abnormal Volume

The second dependent 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.

Table 6 summarizes the estimation results. Columns (1) and (3) report that negative sentiment has no impact on filing-period abnormal volume, but positive sentiment has a negative effect. In comparison, five clusters exhibit significant coefficient estimates for negative sentiment (Column 2), four for positive sentiment (Column 4), and six for neutral sentiment (Column 5, partially tabulated) at the 5% level. The signs of the significant estimated coefficients are predominantly positive (8 versus 1 negative). Of these clusters, only one, “Years (2008--2017),” is content driven rather than sentiment driven. The rest of the significant clusters all appear to be sentiment driven. In other words, a higher percentage of negative and positive extractions in specific clusters is associated with a larger abnormal trading volume around the filings of the 10-K reports. That being said, the effect is concentrated in a small number of contexts.

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.

Table 7 reports the estimation of two linear regressions for replication and three lasso regressions. Columns (1) and (3) both show the expected sign on the sentiment measures—as expected, more sentiment leads to more volatility. Column (2) indicates that the estimated coefficients on 19 of the 131 clusters are significant at the 5% level (six negative and 13 positive) for negative sentiment. Similarly, Column (4) shows that 19 clusters are statistically significant for positive sentiment (ten have negative coefficients versus nine positive). In addition, Column (5) finds that 50 of the estimated coefficients are statistically significant (34 negative and 16 positive). This suggests that the post-filing return volatility is reacting to the content of those 50 clusters, even though the LM sentiment measures considered the tones of these cluster neutral.

Loughran and McDonald (2011) document that a higher percentage of positive or negative words is associated with a larger stock volatility in the post 10-K filing period [+6, +252], which we replicated in Columns (1) and (3). In contrast, our results are mixed. After removing the seven content-driven clusters, for negative sentiment, our findings are consistent: we find 11 positive clusters that are sentiment driven, versus only five negative clusters. For positive sentiment, however, we find six clusters of each sign that are sentiment-driven. This suggests that while sentiments induce trading and volatility, the content in some clusters reduces them.

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.

Table 8 reports the estimation of two linear regressions for replication alongside three lasso regressions. Column (1) finds that negative sentiment has no power predicting material weaknesses. In contrast, column (2) indicates that 22 of the estimated coefficients on the clusters are significant at the 5% level (ten negative and 12 positive) for negative sentiment. On the other hand, Column (3) shows a moderating effect of positive sentiment on material weaknesses. Column (4) indicates that 19 of the coefficients are statistically significant (seven negative and 12 positive). After removing the three content-driven clusters ("Company description, operations," "Oil and gas," and "Years (2008--2017)"), we still find more positive than negative coefficients (ten and six, respectively) for positive sentiment. This is counter to both our expectations as well as the sign on positive sentiment in Column (3).

4.2.5. Variation of Content across Outcome Variables

If the effect of sentiment is driven by a consistent reason, then we expect that the same set of clusters 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 "Geographic location"—this is not necessarily a topic that most investors would have an interest in, but it may indicate some aspect of complexity within the firm. In fact, for negative sentiment, geographic location is only relevant for material weaknesses, where complexity is likely a factor. Second, "Accounting assumptions" is significant in explaining excess return, abnormal volume, and material weaknesses for negative sentiment. As such, this cluster is reasonably consistent – this is ideally how sentiment would work on all clusters if its interpretation was not dependent on the economic context being examined.

For negative sentiment, there is only one context that is significant across all regressions, “Years (2008--2017),” which is also has the expected sign in every regression. Overall, however, 54 different clusters are statistically significant at a 5% or lower level across the regressions, yet 42 of these are significant for only one dependent variable. Another ten are significant in two regressions, while “Accounting assumptions” is significant for three dependent variables. Consequently, there appears to be little commonality in the reasoning as to why sentiment is working in these contexts. It may well be acting as a completely separate construct under each dependent variable.

For positive sentiment, the results are no better. Out of the 46 different significant clusters, 35 of them are only significant for one dependent variable, while 11 are significant for two dependent variables. Not a single cluster is significant across three or all four dependent variables.

5. Conclusion

We examine the bag-of-words method in analyzing the sentiment of the MD&A section of 10-K filings under different contexts. We use open information extraction and Universal Sentence Encoder to obtain short extractions from MD&As and use them to create 131 clusters that capture different contexts in the MD&As. First, we find that the Loughran and McDonald (2011) sentiment measures do not capture the expected sentiment in most of these clusters. Second, the positive, negative, and neutral sentiments of the clusters 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 the LM positive and negative sentiment measures. Finally, we show that different sets of clusters exhibit predictive power for different outcome variables. Taken together, this study indicates that the bag-of-words method is inadequate for capturing contextual meaning, which can be attributed to the fact that the method assumes words are independent and, hence, context does not matter.

References

Antweiler, W. and M. Frank. (2004). Is all that talk just noise? Information content of internet stock message boards. *Journal of Finance*, 59(3), 1259–1294.

Audi, Robert, Tim Loughran, and Bill McDonald. "Trust, but verify: MD&A language and the role of trust in corporate culture." *Journal of Business Ethics* 139, no. 3 (2016): 551-561.

Azimi, Mehran and Anup Agrawal. 2021. Is positive sentiment in corporate annual reports informative? Evidence from deep learning. *Review of Asset Pricing Studies*, forthcoming.

Bochkay, Khrystyna, Jeffrey Hales, and Sudheer Chava. 2020. Hyperbole or Reality? Investor Response to Extreme Language in Earnings Conference Calls. *The Accounting Review* 95(2), 31-60.

Brown, Nerissa C., Richard M. Crowley, and W. Brooke Elliott. "What are you saying? Using topic to detect financial misreporting." *Journal of Accounting Research* 58, no. 1 (2020): 237-291.

Bodnaruk, A., Loughran, T., and McDonald, B. (2015). Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, 50(4), 623–646.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant et al. "Universal sentence encoder." arXiv preprint arXiv:1803.11175 (2018).

Crowley, Richard M., Wenli Huang, and Hai Lu. "Executive Tweets." Rotman School of Management Working Paper (2020).

Dyer, Travis, Mark Lang, and Lorien Stice-Lawrence. "The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation." *Journal of Accounting and Economics* 64, no. 2-3 (2017): 221-245.

El-Haj, Mahmoud, Paul Rayson, Martin Walker, Steven Young, and Vasiliki Simaki. 2019. “In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse.” Journal of Business Finance and Accounting 46: 265-306.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns of stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Fama, Eugene F., and Kenneth R. French. "Industry costs of equity." *Journal of Financial Economics* 43, no. 2 (1997): 153-193.

Angeli, Gabor, Melvin Jose Johnson Premkumar, and Christopher D. 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 (Volume 1: Long Papers), pp. 344–354. 2015.

Gentzkow, M., B. T. Kelly, and M. Taddy. 2017. Text as data. NBER Working Paper 23276.

Huang, A., A. Zang, and R. Zheng. (2014). Evidence on the information content of text in analyst reports. *The Accounting Review* 89(6), 2151–2180.

Huang, Allen H., Reuven Lehavy, Amy Y. Zang, and Rong Zheng. "Analyst information discovery and interpretation roles: A topic modeling approach." Management Science 64, no. 6 (2018): 2833-2855.

Henry, E. "Are Investors Influenced by How Earnings Press Releases Are Written?" *Journal of Business Communication* 45(4) (2008), pp. 363–407.

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." *Available at SSRN: https://ssrn.com/abstract=890586* (March 2006).

Li, F. "Annual Report Readability, Current Earnings, and Earnings Persistence." *Journal of Accounting and Economics* 45 (2008): 221–47.

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 Naıve Bayesian Machine Learning Approach." *Journal of Accounting Research* 48 (2010b):1049–102.

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

Loughran, Tim, and Bill McDonald. "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks" *Journal of Finance* 66 (2011): 35-65.

Loughran, Tim, and Bill McDonald. "The use of word lists in textual analysis." *Journal of Behavioral Finance* 16, no. 1 (2015): 1-11.

Loughran, Tim, and Bill McDonald. "Textual analysis in accounting and finance: A survey." *Journal of Accounting Research* 54, no. 4 (2016): 1187-1230.

Loughran, Tim, and Bill McDonald. "Textual Analysis in Finance." *Annual Review of Financial Economics* 12 (2020): 357-375.

Loughran, Tim, and Bill McDonald. "Measuring firm complexity." *Available at SSRN 3645372* (2020).

Loughran, Tim, Bill 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.

Sculley, David. "Web-scale k-means clustering." In *Proceedings of the 19th international conference on World wide web*, pp. 1177–1178. 2010.

Siano, Federico and Peter D. Wysocki. “Transfer Learning and Textual Analysis of Accounting Disclosures: Applying Big Data Methods to Small(er) Data Sets.” *Accounting Horizons*, Forthcoming (March 20, 2020). Available at SSRN: https://ssrn.com/abstract=3560355

Tetlock, Paul C. “Giving content to investor sentiment: The role of media in the stock market,” *Journal of Finance* 62 (2007), 1139–1168.

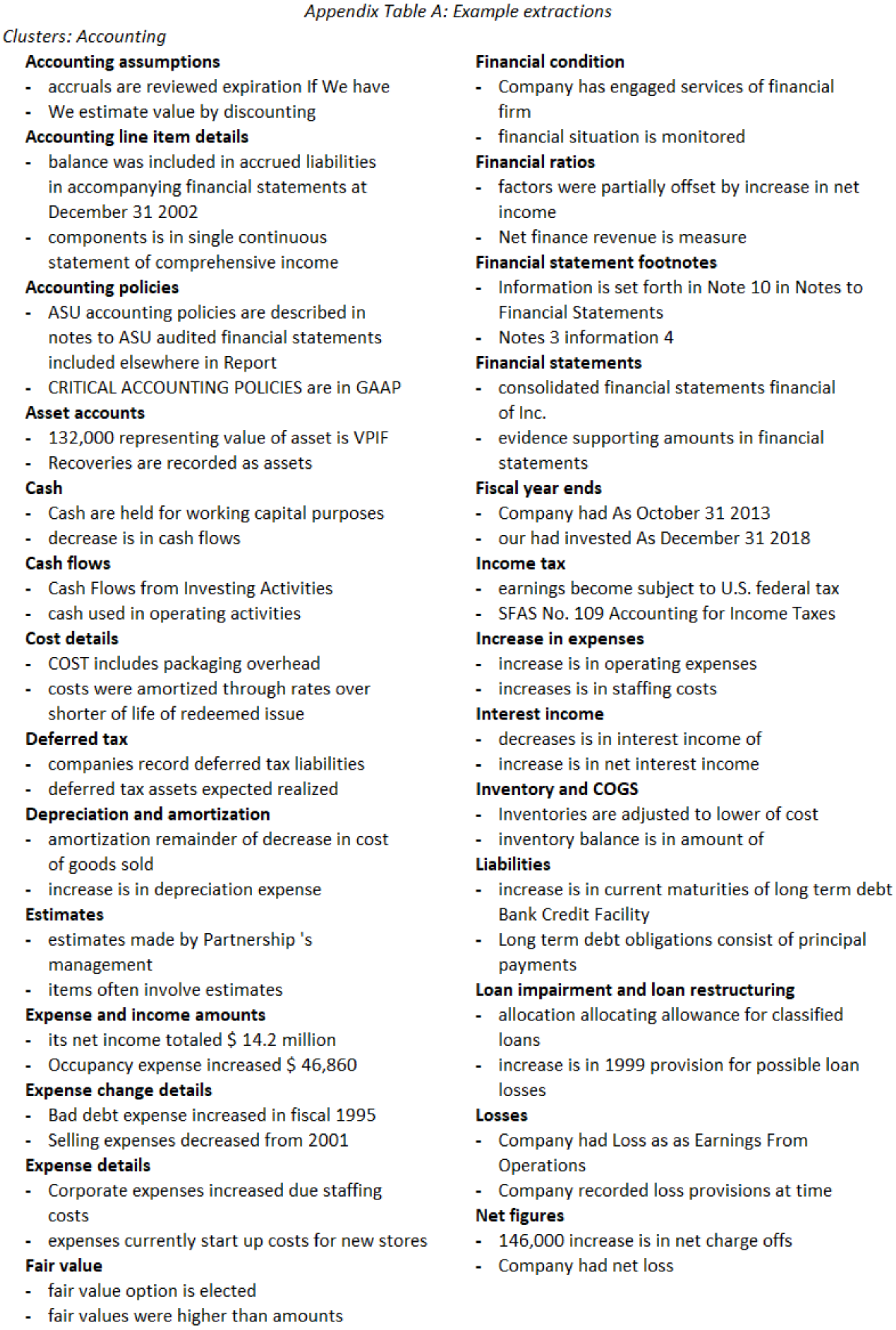
Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58, no. 1 (1996): 267-288.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society*, Series B, 63, part 2, 411-423.

Walker, Martin, Steven Young, Andrew Moore, Vasiliki Athanasakou, Paul Rayson, and Thomas Schleicher. 2020. Who’s the Fairest of Them All? A Comparison of Methods for Classifying Tone and Causal Reasoning in Earnings-related Management Discourse. Working Paper, Lancaster University Management School.

Yang, Yi, Mark Christopher Siy UY, and Allen Huang. "Finbert: A pretrained language model for financial communications." *arXiv preprint arXiv:2006.08097* (2020).

Appendix A: Cluster examples



Appendix Table A: Example extractions

Clusters: Accounting

Accounting assumptions
- accruals are reviewed expiration If We have
- We estimate value by discounting

Accounting line item details
- balance was included in accrued liabilities in accompanying financial statements at December 31 2002
- components is in single continuous statement of comprehensive income

Accounting policies
- ASU accounting policies are described in notes to ASU audited financial statements included elsewhere in Report
- CRITICAL ACCOUNTING POLICIES are in GAAP

Asset accounts
- 132,000 representing value of asset is VPIF
- Recoveries are recorded as assets

Cash
- Cash are held for working capital purposes
- decrease is in cash flows

Cash flows
- Cash Flows from Investing Activities
- cash used in operating activities

Cost details
- COST includes packaging overhead
- costs were amortized through rates over shorter of life of redeemed issue

Deferred tax
- companies record deferred tax liabilities
- deferred tax assets expected realized

Depreciation and amortization
- amortization remainder of decrease in cost of goods sold
- increase is in depreciation expense

Estimates
- estimates made by Partnership 's management
- items often involve estimates

Expense and income amounts
- its net income totaled $ 14.2 million
- Occupancy expense increased $ 46,860

Expense change details
- Bad debt expense increased in fiscal 1995
- Selling expenses decreased from 2001

Expense details
- Corporate expenses increased due staffing costs
- expenses currently start up costs for new stores

Fair value
- fair value option is elected
- fair values were higher than amounts

Financial condition
- Company has engaged services of financial firm
- financial situation is monitored

Financial ratios
- factors were partially offset by increase in net income
- Net finance revenue is measure

Financial statement footnotes
- Information is set forth in Note 10 in Notes to Financial Statements
- Notes 3 information 4

Financial statements
- consolidated financial statements financial of Inc.
- evidence supporting amounts in financial statements

Fiscal year ends
- Company had As October 31 2013
- our had invested As December 31 2018

Income tax
- earnings become subject to U.S. federal tax
- SFAS No. 109 Accounting for Income Taxes

Increase in expenses
- increase is in operating expenses
- increases is in staffing costs

Interest income
- decreases is in interest income of
- increase is in net interest income

Inventory and COGS
- Inventories are adjusted to lower of cost
- inventory balance is in amount of

Liabilities
- increase is in current maturities of long term debt Bank Credit Facility
- Long term debt obligations consist of principal payments

Loan impairment and loan restructuring
- allocation allocating allowance for classified loans
- increase is in 1999 provision for possible loan losses

Losses
- Company had Loss as as Earnings From Operations
- Company recorded loss provisions at time

Net figures
- 146,000 increase is in net charge offs
- Company had net loss

Appendix Table A (Continued): Example extractions

Profitability

- Gross margin is calculated by cost of sales from revenue
- Gross profit demand for high volume specialty wheels

Results and outcomes

- annual results differ sharply from our results
- results differ from actuarial assumptions

Revenue details

- Company 's retail operations segment revenue
- Gas marketing revenue purchased for resale

Revenue mentions

- Revenue increased compared 2010
- revenue stream decreased 6.5 %

Revenue recognition

- comparison can result in increase in final recognition
- Revenue earned is recognized

Sales

- increase is due gross profit from sales volumes
- sales have raised in gross proceeds

Tax positions and related uncertainty

- FASB Interpretation No. 48 Accounting for in Income Taxes
- Uncertainty is in Income Taxes

Tax rates

- effective tax rate reflect reversal of reserves
- effective tax rate was 32.1 %

Valuation

- Accruals are recorded based expected levels of performance
- average claim values are assumptions for reserve

Clusters: Business operations

Capital sources

- Company has financed improvements from capital by securing funds under mortgages
- equity financing creating means

Common and preferred stock

- $ 2.8 million repurchase 755,400 shares of Stock
- company start trading on New York Stock Exchange

Company description, accounting/quantitative

- 29 were Of 61 owned clinics constructed by Our
- Adjusted EBITDA is metric used by our management

Company description, financial

- Company about regulatory capital ratios
- Company may purchase up at prices

Company description, operations

- company at molybdenum facilities
- Company formulates policies governing

Company details

- Company had borrowings outstanding
- Company has Board of Directors

Connections

- connection is with affiliations
- connection is with Ayres ALM200 aircraft program

Costs and expense mentions

- compensation cost is recognized
- costs are capitalized

Credit facilities and agreements

- Company utilized Sabine Mining Company facilities
- Credit Agreement provided Company

Customers

- customer financing was offered
- customers funding for customers capital needs

Detailed company info with name

- Delta Companies Group shifted Delta Group focus
- Internet Business Consulting Inc. come to acceptable terms

Detailed company info without name

- Company acquired from former franchisees restaurants
- Company continue pay to Acquisition and of Shopping Centers Company stockholders

Financial services details

- Atlantic Central Bankers Bank were net seller of funds at December 31 2012
- bank line has total capacity

Foreign currency and exchange rates

- changes is in foreign currency exchange rates
- currency are translated

Future uncertainty

- certain circumstances could cause borrowers
- failure can occur at stage of trials

Geographic locations

- California outstandings are distributed as follows
- economy is in U.S.

Appendix Table A (Continued): Example extractions

Insurance

- Emerging Issues Task Force Issue No. 94 3 for Certain Employee Termination Benefits
- Insurance Solutions Group Protection segment group life to employers

Interest

- Interest accruals are continued
- Interest is payable

Interest rates

- decrease was result of decline in interest rates investments held
- Fluctuations is in interest rates

Investment portfolios

- Company had real estate construction portfolio
- Company held to maturity portfolios

Investments

- Fund 's investment is in Brazos Sportswear , Inc.
- its investment is in Bank

Leasing

- Company utilized capital leases for purchase
- Company will own 100 % of Leases

Liquidity

- our convertible subordinated debt is sensitive to in price of our common stock
- we may pursue equity financing may not able

Loan details

- interest reserve be established for loan term
- loans are secured in amounts

Management decisions

- management estimates For assets held in business
- management is attempting maintain

Markets (product, region, financial)

- assumptions reflect consideration of participants
- Canon looking at opportunities to expand into markets

Oil and gas

- 275,000 barrel refinery purchase terminal
- decline is in worldwide oil drilling activity

Operating agreements

- Company entered into agreement
- FTS assumes Meldisco agreement

Operating cost details

- 3M benefited from material sourcing cost projects
- EMI charged Trust with adjustment for costs to copyright renewals

Operations and transactions

- discussed recapitalization transaction is with its shareholders
- operations is in Alberta

Payments to others

- Notes require annual principal payments
- payment is is accrued over life of Hercules Loan

Pricing

- average rate decreased 9 basis points
- future rate be significant variable in rates for services provided

Products and manufacturing

- Companys ability manufacture products
- Plastic Products Business pass on resin cost for products

Research and development

- NanoViricides technology is now maturing with facility
- Patent applications are maintained in secrecy in U.S.

Risk

- Market risk is potential loss may occur as result changes in market of particular instrument
- Risks will continue hold

Securities

- $ 175,000 holders of Trust I Securities
- QSPE uses proceeds from issuance of securities

Selling

- our products sold are subject to VAT of 17 % of sales price
- we selling acquired during current year

Clusters: Contracting

Acquisitions

- Company results from date of Acquired Systems acquisition
- Mineral property acquisition costs are including licenses

Contracts

- BusinessPower Supply ResourcesPower Contracts is in Item 1
- negotiated contract savings is with certain vendors

Credit agreements and covenants

- bargaining agreements are scheduled
- compliance is with covenants

Legal/Compliance

- Forgent received Subpoena Duces Tecum
- legal fees were related to usual transaction for Company

Appendix Table A (Continued): Example extractions

Obligations

- covenants require maintenance of various ratios
- elements are whether separable from other of contractual relationship

Partnerships

- gold remain primary metals market exposure for Partnership
- operating partnership through right to receive

Shareholder related

- income represents AFUDC Equity
- PELS make distributions in dividends

Subsidiaries

- Bank of 100 % beneficial interests in subsidiaries
- Company formed subsidiary

Clusters: Regulation

Accounting standard issuance and adoption

- FASB issued SFAS No. 144
- FSP 157 1 amends SFAS 157

Accounting standard issuance timing

- FASB issued In January 2003
- FASB issued In July 2001

Accounting standards

- disclosures GAAP required under United States
- GAAP have met through capital generated GAAP initial public offering

Effective rules and policies

- ASU is effective for beginning
- FAS is effective for years with adoption

FASB Statements

- Financial Accounting Standards Board Earnings Per Share
- Financial Accounting Standards Board provisions In December 2007

Regulation

- Federal regulations is in United States
- U.S. Food and Drug Administration expect continue

Clusters: Changes

Changes 1

- change is in assumptions
- change is in volume

Changes 2

- Changes are subject
- Changes be long term

Decreased values

- decrease is in advances to suppliers
- decrease is in average borrowings

Decreases and declines

- decline is attributable
- decreases is in average yields on primarily asset categories

Decreases and increases in financials

- Advances increased to 71.2 percent of total assets
- Decreases were slightly offset by buying costs

Increase attributions

- increase is consistent
- increase was attributable

Increases and improvements

- improvement is attributable to increase in revenues
- increase was primarily attributable to increase in shipments

Increases in metrics

- G&A increased in aggregate
- increase is in accrued liabilities of 279,000

Reductions and decreases

- Discounts were principal drivers during fiscal 2007
- principal element was elimination of 25 employees

Clusters: Grammatical patterns

Boilerplate: financial condition

- Management for Discussion and Analysis of Condition
- Management has Discussion and Analysis of Condition

Company actions on a given date

- Company invested In Fiscal 2000
- Company reported During 1999

Dollar amounts (millions)

- $ 19.6 million consist of costs associated
- SOX held $ 1.7 million corridors classified

Dollar amounts, equity and debt

- $ 2.0 million outstanding principal accrued on Company 's Consolidated Statements of credit line
- Company assumed $ 6.5 million in floating rate debt securities

Appendix Table A (Continued): Example extractions

Dollar amounts, small

- Golden Gate advanced $ 753,381
- Minority interest decreased $ 180,000

Dollar changes (millions)

- $ 1.8 million reduction is in personnel related costs
- $ 27.2 million increase is in brokerage

Events on a given day/month/year

- Jobs Creation Act was signed in October 2004
- The summarizes LTXs obligations at July 31 2004

Explanations

- advance notice is given
- data are capitalized principally

First person plural: charges

- our paid one time fees
- We record cost Throughout year

First person plural: dates

- we are also subject Under July 2007 Facility
- We determined During 2002

First person plural: operations

- our agreement is with PPG Industries
- our Federal division is excluded

Fiscal quarters

- FIN No. is effective for Tyco in first quarter
- managing director was appointed in quarter

Future requirements

- We are required estimate
- We may have In addition

Instructions to readers

- first statement should present In two statement approach
- guidance is in ASC Topic

Modal strong statements

- ASU requires that
- Farmers Branch will need depleted

Numeric amount descriptions

- amount is estimable
- percentage exceeds for year

Percentage in year

- 36 % is in 2003
- 37.2 % is in 2001

Periods and fiscal years

- as same period is in prior year
- derivative instrument is then re-valued at date

Reference to tables

- table presents our contractual obligations
- table presents results based location

Specific dates

- September 30 , 2015 rate of 3.0 %
- total obligation remaining as June 29 2003

Clusters: Timeframes

Company actions (1992-1998)

- Company continued through 1996
- Company entered On March 15 1996

Years (1992--1998)

- 1995 was first full year
- 8 % costs 1998

Years (1995-1998)

- Fund sold In March 1997
- increase experienced in 1998

Years (2000--2002)

- increases is in 2001
- Power supply expenses increased in 2000

Years (2002--2005)

- expense is in 2005
- increased sales is in 2005

Years (2006)

- cash needs raised during year 2006
- Laserscope was acquired in July 2006

Years (2007-2013)

- 31.3 % is in 2009
- decrease is in 2011

Years (2008--2017)

- increase year ended December 31 2010
- year interest expense ended October 31 2012

Clusters: Ungrouped

First person plural ungrouped text 1

- our customers or our industry have difficulties in future
- our goal build our commercial infrastructure

First person plural ungrouped text 2

- we anticipated repaying
- we consider trade accounts receivable

First person plural ungrouped text 3

- we expect target
- We have As such

Ungrouped text 1

- Board appointed Stephen Keen
- Goodwill is with indefinite useful lives

Ungrouped text 2

- each represents component of We enterprise
- trends is in industry

Ungrouped text 3

- confidence is low
- step is performed

Ungrouped text 4

- 11,000 were outstanding
- Computer models were used

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 et al. (2001). The Gap statistic at *k* clusters for *B* simulated samples, 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, is calculated as follows:

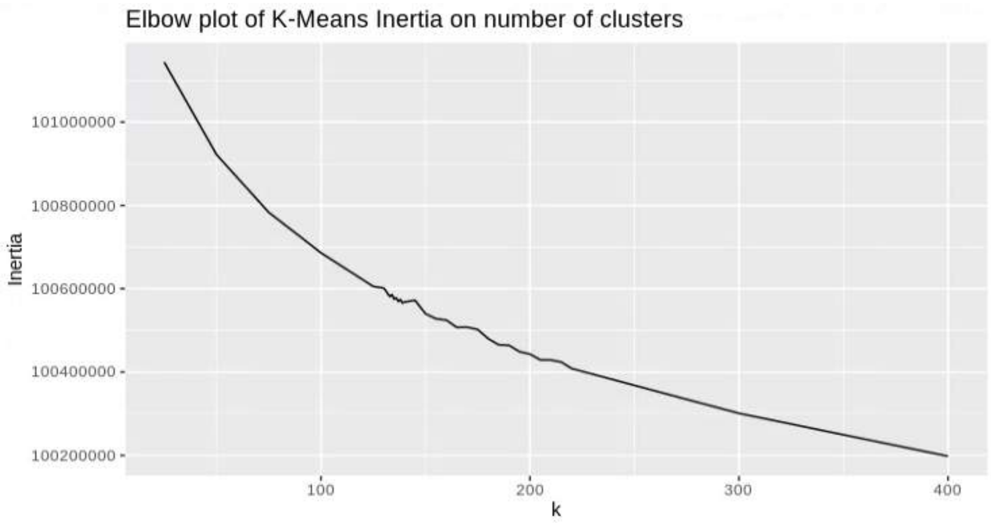
$$Gap(k) = \left(\frac{1}{B}\right)\sum_{r=1}^{B} log\left(W^*_{k,r}\right) - log(W_k), and$$

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

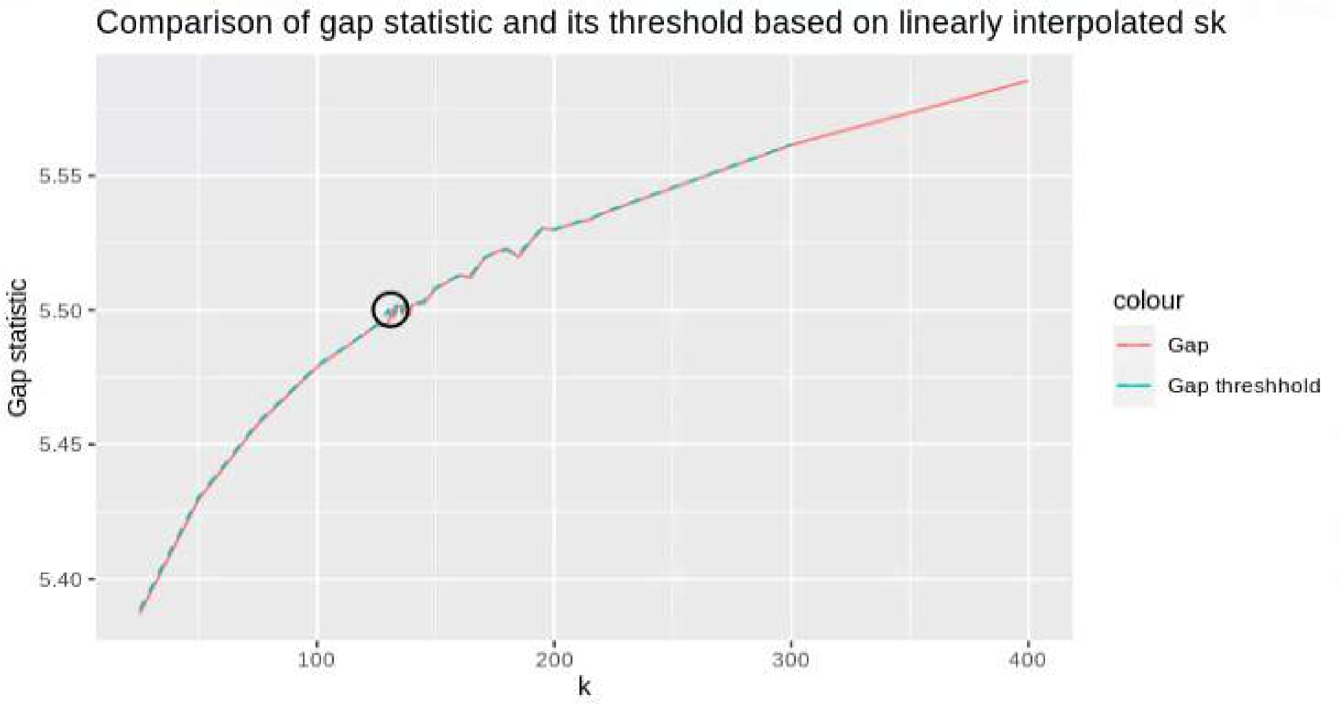
To choose the optimal *k* based on the Gap statistic, we follow Tibshirani et al. (2001) and select the lowest *k* such that $Gap(k) \geq Gap(k+1) - s_{k+1}$.

As each iteration of Mini-Batch K-Means is computationally expensive to run, we start our optimization with a grid search at k values of 100, 200, 300, and 400. Based on an elbow plot of the resulting inertia values, we determined that the optimal number of clusters was unlikely to be much above 200.[8] The full elbow plot for all values of *k* that we tried in this grid search and subsequent searches is presented below.

[8] Elbow plots are commonly used alongside the "elbow method" for determining optimal numbers of K-Means clusters. The elbow method is the commonly held idea that the optimal number of clusters occurs around a kink or "elbow" in the plot of inertia values on cluster counts.



We then run another grid search at a spacing of 25 from 25 up to 225, which further pinpointed the expected optimal region to be between 125 and 225. We subsequently run a grid search with a spacing of 5 covering 125 through 225. At this point we then ran a synthetic sample of 10 Mini-Batch K-Means samples with 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). We then used the derived s_k from this sample to estimate if there was a plausibly optimal number of clusters via the Gap statistic; we find that $s_{200} = 0.000070$. This procedure pinpointed 135 as potentially optimal. Lastly, we ran a final grid search with a spacing of 1 from 130 to 140 to see which *k* is optimal. Based on this simulation, a *k* of 131 appears to be optimal, with $Gap(131) = 5.500126$ and $Gap(132) = 5.497125$. We also bootstrap a sample to determine that $s_{132} = 0.000069$. To ensure that our limited bootstrapping methodology was not biasing against find a lower optimal number of clusters, we re-estimate the difference between $Gap(k)$ (Gap statistic) and $Gap(k+1) - s_{k+1}$ (Gap threshold) using a linear interpolation of s_k values using s_{132} and s_{200}. We plot the two components of this expression below and can confirm that 131 clusters (circled in the graph) are optimal, as it is the first point at which the Gap statistic is higher than the Gap threshold.



Appendix C: 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.
Independent variables	
Cluster	The number of extractions in a filing that belong to the given cluster, divided by the total number of extractions in the filing.
Cluster|Sentiment	The number of extractions in a filing that belong to the given cluster and which have the specified sentiment, divided by the total number of extractions in the filing. An extraction 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. An extraction has neutral sentiment

	if it has neither positive nor negative sentiment; this may be because no LM dictionary words were contained in the extraction, or because there were an equal number of negative and positive words in the extraction.
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	

		MD&As		Extractions
Sample restriction	MD&As	dropped	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

Table 2: Context Frequencies

Panel A: Most and least frequent clusters by extraction count

		Number of	Number of			Number of	Number of
	Most frequent clusters	extractions	documents		Least frequent clusters	extractions	documents
1	Ungrouped text 1	403,925	32,561	122	Deferred tax	77,518	16,937
2	Company details	355,895	27,136	123	Accounting policies	77,412	19,105
3	Detailed company info without name	350,898	28,819	124	Years (2006)	71,084	7,517
4	Interest rates	340,763	27,715	125	FASB Statements	64,746	17,430
5	Explanations	334,716	31,932	126	Partnerships	62,996	9,834
6	Sales	329,224	27,332	127	Effective rules and policies	58,695	16,105
7	Increases in metrics	321,923	32,012	128	Company actions (1992-1998)	55,488	12,432
8	Future uncertainty	318,033	31,484	129	Boilerplate: financial condition	50,388	27,401
9	Geographic locations	309,011	30,454	130	Accounting standard issuance timing	41,626	13,675
10	Increases and improvements	305,329	32,024	131	Tax positions and related uncertainty	40,803	11,719

Panel B: Most and least frequent clusters by document count

		Number of	Number of			Number of	Number of
	Most frequent clusters	extractions	documents		Least frequent clusters	extractions	documents
1	Ungrouped text 1	403,925	32,561	122	Years (1992--1998)	100,427	15,621
2	Increases and improvements	305,329	32,024	123	Future requirements	91,030	15,454
3	Increases in metrics	321,923	32,012	124	Loan impairment and loan restructuring	130,533	13,944
4	Explanations	334,716	31,932	125	Accounting standard issuance timing	41,626	13,675
5	Ungrouped text 3	275,829	31,661	126	Company actions (1992-1998)	55,488	12,432
6	Future uncertainty	318,033	31,484	127	Tax positions and related uncertainty	40,803	11,719
7	Ungrouped text 4	217,815	31,363	128	Years (2002--2005)	140,952	11,366
8	Results and outcomes	184,278	31,148	129	Partnerships	62,996	9,834
9	Modal strong statements	253,436	31,071	130	Oil and gas	157,806	8,732
10	Expense details	187,732	30,587	131	Years (2006)	71,084	7,517

Table 2 (Continued): Context Frequencies

Panel C: Most and least frequent clusters by extraction count, negative extractions only

		Number of	Percent of			Number of	Percent of
	Most frequent clusters	extractions	extractions		Least frequent clusters	extractions	extractions
1	Losses	193,450	94.7%	122	Accounting standard issuance timing	3,076	6.1%
2	Future uncertainty	95,884	30.1%	123	Percentage in year	2,891	7.1%
3	Loan impairment and loan restructuring	78,366	60.0%	124	Effective rules and policies	2,419	3.8%
4	Decreases and declines	75,254	54.6%	125	Tax rates	1,992	3.1%
5	Insurance	50,742	17.1%	126	Cash flows	1,262	0.8%
6	Decreases and increases in financials	48,913	28.2%	127	Increase attributions	987	1.0%
7	Detailed company info without name	46,536	13.3%	128	FASB Statements	899	1.0%
8	Ungrouped text 1	44,588	11.0%	129	Partnerships	505	0.9%
9	Net figures	40,662	31.9%	130	Tax positions and related uncertainty	318	0.3%
10	Asset accounts	36,899	13.7%	131	Boilerplate: financial condition	225	0.5%

Panel D: Most and least frequent clusters by percent of extractions within cluster, negative extractions only

		Number of	Percent of			Number of	Percent of
	Most frequent clusters	extractions	extractions		Least frequent clusters	extractions	extractions
1	Losses	193,450	94.7%	122	Revenue recognition	3,319	2.9%
2	Loan impairment and loan restructuring	78,366	60.0%	123	Increases in metrics	8,563	2.7%
3	Decreases and declines	75,254	54.6%	124	Company description, operations	4,785	2.2%
4	Net figures	40,662	31.9%	125	Increase in expenses	3,319	2.0%
5	Accounting policies	24,363	31.5%	126	Tax rates	899	1.0%
6	Future uncertainty	95,884	30.1%	127	Cash flows	987	1.0%
7	Decreases and increases in financials	48,913	28.2%	128	Effective rules and policies	505	0.9%
8	Legal/Compliance	36,074	27.7%	129	Increase attributions	1,262	0.8%
9	Valuation	32,983	21.0%	130	Accounting standard issuance timing	225	0.5%
10	Operating cost details	34,750	20.8%	131	Percentage in year	318	0.3%

Table 2 (Continued): Context Frequencies

Panel E: Most and least frequent clusters by extraction count, positive extractions only

		Number of	Percent of			Number of	Percent of
	Most frequent clusters	extractions	extractions		Least frequent clusters	extractions	extractions
1	Tax rates	64,001	72.6%	122	Costs and expense mentions	1,350	0.8%
2	Effective rules and policies	53,947	91.9%	123	Connections	1,296	1.6%
3	Increases and improvements	51,075	16.7%	124	Reference to tables	1,208	1.2%
4	Detailed company info without name	38,300	10.9%	125	Financial statement footnotes	1,187	0.9%
5	Ungrouped text 1	34,045	8.4%	126	Accounting policies	933	1.2%
6	Research and development	31,233	11.7%	127	Cash flows	683	0.7%
7	Future uncertainty	29,583	9.3%	128	Losses	445	0.2%
8	Insurance	28,670	9.7%	129	Boilerplate: financial condition	395	0.8%
9	Income tax	23,036	11.0%	130	Percentage in year	80	0.1%
10	Company description, accounting/quantitative	22,234	9.8%	131	Accounting standard issuance timing	32	0.1%

Panel F: Most and least frequent clusters by percent of extractions within cluster, positive extractions only

		Number of	Percent of			Number of	Percent of
	Most frequent clusters	extractions	extractions		Least frequent clusters	extractions	extractions
1	Effective rules and policies	53,947	91.9%	122	Instructions to readers	2,059	1.3%
2	Tax rates	64,001	72.6%	123	Reference to tables	1,208	1.2%
3	Increases and improvements	51,075	16.7%	124	Accounting policies	933	1.2%
4	Research and development	31,233	11.7%	125	Financial statement footnotes	1,187	0.9%
5	Income tax	23,036	11.0%	126	Costs and expense mentions	1,350	0.8%
6	Detailed company info without name	38,300	10.9%	127	Boilerplate: financial condition	395	0.8%
7	Operating cost details	18,025	10.8%	128	Cash flows	683	0.7%
8	Subsidiaries	21,572	10.2%	129	Losses	445	0.2%
9	Profitability	15,212	9.9%	130	Accounting standard issuance timing	32	0.1%
10	Company description, accounting/quantitative	22,234	9.8%	131	Percentage in year	80	0.1%

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	1.33%	0.49%	0.59%	1.30%	2.17%
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						
Extractions per MD&A	35,362	641.1	457.9	75.0	548.0	1,511.0
Negative extractions per MD&A	35,362	36.6	34.8	2.0	27.0	105.0
Positive extractions 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

Table 4: Context Underlying MD&A Tone

	Handcoded	Negative	Positive	Handcoded
	prediction	MD&A Tone	MD&A Tone	prediction
	(1)	(2)	(3)	(4)
Part A: High sentiment clusters				
Clusters: Accounting				
Loan impairment and loan restructuring	+	0.107 ***	0.006 **	
Net figures	+	0.029 ***	0.014 ***	
Results and outcomes	+	0.100 ***	0.027 ***	
Clusters: Business operations				
Company details		0.012 ***	0.008 ***	
Detailed company info without name		0.022 ***	0.026 ***	
Detailed company info with name		0.003 *	0.003 ***	
Future uncertainty	+	0.147 ***	0.022 ***	
Management decisions		0.017 ***	0.013 ***	
Markets (product, region, financial)	+	0.022 ***	0.028 ***	
Operating cost details		0.039 ***	0.032 ***	
Clusters: Business operations				
Accounting standard issuance and adoption		0.013 **	0.007 *	
Clusters: Changes				
Decreases and increases in financials	+	0.045 ***	0.017 ***	+
Reductions and decreases		0.072 ***	0.022 ***	+
Clusters: Grammatical patterns				
Dollar amounts, equity and debt		0.017 ***	0.005 *	
First person plural: operations		0.014 **	0.017 ***	
Future requirements		0.100 ***	0.027 ***	
Clusters: Timeframes				
Years (2007-2013)		0.005 **	0.005 **	
Clusters: Ungrouped				
First person plural ungrouped text 1		0.023 ***	0.004 **	
First person plural ungrouped text 3		0.021 ***	0.017 ***	
Ungrouped text 4		0.034 ***	0.008 **	
Part B: Clusters skewed toward negative				
Clusters: Accounting				
Asset accounts	+	0.032 ***	-0.003	
Estimates		0.019 ***	-0.024 ***	
Inventory and COGS		0.012 **	-0.028 ***	
Liabilities	+	0.026 ***	0.001	
Losses	+	0.156 ***	-0.003 **	
Valuation		0.034 ***	-0.006	
Clusters: Business operations				
Company description, operations		0.011 **	-0.011 ***	
Customers		0.008 ***	-0.007 ***	
Foreign currency and exchange rates		0.012 ***	-0.003 **	
Operations and transactions		0.013 ***	0.001	
Products and manufacturing		0.016 ***	0.002	
Selling		0.015 ***	-0.002	
Clusters: Contracting				
Legal/Compliance	+	0.165 ***	-0.019 ***	
Clusters: Regulation				
Regulation		0.020 ***	-0.003	

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

	Handcoded	Negative	Positive	Handcoded
	prediction	MD&A Tone	MD&A Tone	prediction
	(1)	(2)	(3)	(4)
Clusters: Changes				
Decreases and declines	+	0.143 ***	-0.002	
Clusters: Grammatical patterns				
Fiscal quarters	+	0.016 ***	-0.008 **	+
Modal strong statements		0.032 ***	-0.005 *	
Clusters: Timeframes				
Years (1995-1998)	+	0.019 ***	0.003	+
Years (2000--2002)		0.025 ***	-0.003 *	
Part C: Clusters skewed toward positive				
Clusters: Accounting				
Cash		-0.004	0.016 ***	
Income tax		-0.010 **	0.008 ***	
Interest income	+	-0.012 **	0.018 ***	+
Tax rates		-0.013 **	0.064 ***	
Clusters: Business operations				
Capital sources		-0.035 ***	0.013 ***	
Company description, accounting/quantitative		.	0.016 ***	
Financial services details		0.000	0.003 **	
Investment portfolios		-0.039 ***	0.006 **	
Investments		-0.012 ***	0.008 ***	
Leasing	+	-0.002	0.004 ***	+
Payments to others		-0.001	0.010 ***	
Research and development		-0.013 ***	0.015 ***	
Clusters: Contracting				
Shareholder related		-0.014 **	0.016 ***	
Credit agreements and covenants		-0.010	0.021 ***	
Clusters: Regulation				
Effective rules and policies		-0.039 ***	0.063 ***	
FASB Statements		-0.017 ***	0.009 **	
Clusters: Changes				
Increases and improvements		-0.046 ***	0.065 ***	+
Clusters: Grammatical patterns				
Periods and fiscal years		-0.050 ***	0.009 ***	
Clusters: Timeframes				
Years (1992--1998)		0.002	0.029 ***	
Years (2002--2005)		-0.002	0.004 **	
Clusters: Ungrouped				
Ungrouped text 1		-0.010 ***	0.011 ***	
Ungrouped text 2		0.004	0.009 ***	
Part D: Low sentiment clusters				
Clusters: Accounting				
Cash flows		-0.050 ***	-0.020 ***	
Expense details		-0.029 ***	-0.009 **	
Fair value		-0.017 ***	-0.008 ***	
Financial ratios		-0.034 ***	-0.005 **	
Financial statement footnotes		-0.013 ***	-0.007 **	
Financial statements		-0.024 ***	-0.018 ***	
Fiscal year ends		-0.025 ***	-0.021 ***	
Increase in expenses	+	-0.029 ***	-0.011 **	

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

	Handcoded prediction	Negative MD&A Tone	Positive MD&A Tone	Handcoded prediction
	(1)	(2)	(3)	(4)
Revenue details		-0.016 ***	-0.007 ***	
Revenue recognition		-0.032 ***	-0.008 ***	
Clusters: Business operations				
Common and preferred stock		-0.017 ***	-0.011 ***	
Company description, financial		-0.053 ***	-0.015 ***	
Credit facilities and agreements	+	-0.019 ***	-0.012 ***	
Geographic locations	+	-0.021 ***	-0.009 ***	
Liquidity		-0.037 ***	-0.012 ***	
Oil and gas		-0.007 ***	-0.004 ***	
Clusters: Contracting				
Acquisitions		-0.049 ***	-0.012 ***	
Contracts		-0.023 ***	-0.008 ***	
Subsidiaries		-0.032 ***	-0.016 ***	
Clusters: Changes				
Changes 1	+	-0.036 ***	-0.008 *	
Decreased values		-0.011 **	-0.025 ***	
Increase attributions		-0.040 ***	-0.049 ***	
Increases in metrics	+	-0.035 ***	-0.018 ***	+
Clusters: Grammatical patterns				
Company actions on a given date		-0.035 ***	-0.008 **	
Dollar changes (millions)	+	-0.016 ***	-0.005 **	+
Events on a given day/month/year		-0.019 ***	-0.018 ***	
Explanations		-0.017 ***	-0.015 ***	
Instructions to readers		-0.033 ***	-0.013 ***	
Specific dates		-0.012 **	-0.024 ***	
Clusters: Timeframes				
Company actions (1992-1998)		-0.059 ***	-0.013 **	
Years (2008--2017)		-0.016 ***	-0.012 ***	
Clusters: Ungrouped				
First person plural ungrouped text 2		-0.017 ***	-0.012 ***	
Part E: All other clusters				
Clusters: Accounting				
Accounting assumptions		-0.002	-0.015 ***	
Accounting line item details		-0.001	0.004	
Accounting policies		-0.005	-0.001	
Cost details		.	0.004	
Deferred tax		0.004	0.001	
Depreciation and amortization	+	-0.023 ***	-0.005	
Expense and income amounts		-0.006	-0.008 ***	
Expense change details	+	.	-0.008 **	+
Financial condition		-0.014 ***	-0.004	
Profitability	+	-0.001	0.001	+
Revenue mentions		-0.002	0.003	
Sales		-0.005 ***	0.000	
Tax positions and related uncertainty	+	0.010	0.008	
Clusters: Business operations				
Connections		-0.010	-0.047 ***	
Costs and expense mentions		-0.007	-0.005	
Insurance		-0.018 ***	.	
Interest		0.000	-0.011 ***	

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

	Handcoded prediction	Negative MD&A Tone	Positive MD&A Tone	Handcoded prediction
	(1)	(2)	(3)	(4)
Interest rates	+	-0.018 ***	0.002	+
Loan details		-0.009 ***	-0.002	
Operating agreements		-0.011 **	-0.004	
Pricing	+	-0.037 ***	0.002	+
Risk	+	.	-0.016 ***	
Securities	+	-0.003	-0.002	
Clusters: Contracting				
Obligations		-0.010 *	0.003	
Partnerships	+	-0.003	0.001	+
Clusters: Regulation				
Accounting standard issuance timing		-0.033 **	0.002	
Accounting standards		-0.008	0.002	
Clusters: Changes				
Changes 2		.	-0.007 **	
Clusters: Grammatical patterns				
Boilerplate: financial condition		.	-0.044 ***	
Dollar amounts (millions)		-0.021 ***	0.001	
Dollar amounts, small		-0.011 **	-0.003	
First person plural: charges		.	-0.010 **	
First person plural: dates		-0.008	-0.007	
Numeric amount descriptions		0.000	-0.004	
Percentage in year		-0.015 ***	0.005	
Reference to tables		-0.043 ***	-0.005	
Clusters: Timeframes				
Years (2006)		-0.002	0.001	
Clusters: Ungrouped				
Ungrouped text 3	+	0.009 *	-0.002	
Part F: Controls				
log(Market value)		-0.055 ***	0.065 ***	
log(BTM)		0.221 ***	-0.023	
log(Share turnover)		0.004	-0.118 ***	
Pre-event FF alpha		.	-1.256 **	
I(Nasdaq)		-0.181 ***	-0.212 ***	
FF48 Industry FE		Included	Included	
Adjusted R^2		0.5031	0.24	

Columns (2) and (3) report lasso regressions including all 131 clusters, with coefficient values multiplied by 1,000 for readability. All clusters 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 10 extractions from each cluster. 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 Event Period Excess Return

Clusters conditional on:	Negative sentiment		Positive Sentiment		Neutral
Variable	(1)	(2)	(3)	(4)	(5)
Negative, MD&A, Our parser	-0.241 ***	.			
Positive, MD&A, Our parser			-0.266 *	-0.128 **	
Clusters: Accounting					
Accounting assumptions		0.340 **		0.568 *	0.178 ***
Cash flows		0.657 **		.	.
Expense details		.		0.987 ***	-0.088 **
Fiscal year ends		-0.405		-0.652 **	-0.065
Interest income		-0.164 **		.	.
Net figures		-0.193 **		.	.
Profitability		0.165 **		.	0.014
Revenue details		-0.163 **		0.126	.
Tax rates		.		0.179 ***	.
Valuation		0.382 ***		.	.
Clusters: Business operations					
Capital sources		-0.218 **		-0.131	.
Costs and expense mentions		0.252		1.635 ***	-0.003
Interest rates		.		0.390 ***	0.017 *
Selling		-0.943 ***		-0.509 *	-0.122 **
Clusters: Contracting					
Contracts		.		-0.679 ***	0.084 **
Subsidiaries		.		-0.733 **	-0.027
Clusters: Regulation					
Accounting standard issuance and adoption		0.538 ***		-1.125 ***	0.032
Effective rules and policies		.		0.136 **	.
Clusters: Grammar patterns					
Company actions on a given date		-0.306 **		.	-0.055 *
Dollar amounts, equity and debt		0.390 ***		.	0.009
Future requirements		-0.697 ***		-1.398 ***	-0.129
Periods and fiscal years		0.008		0.337 **	0.140 ***
Reference to tables		.		-1.335 **	.
Clusters: Timeframes					
Years (1995-1998)		-0.248 **		-0.064	-0.139 ***
Years (2000--2002)		0.142 ***		0.599 ***	0.069 ***
Years (2008--2017)		-0.234 **		.	-0.016
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	0.015	0.009	0.013	0.016

Columns (1) and (3) report linear regressions, while columns (2), (4), and (5) report lasso regressions including all 131 clusters. All clusters are restricted to only the sentiment specified in the column. Only clusters that are significant at p<0.05 for at least one regression from columns (2) and (4) 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.

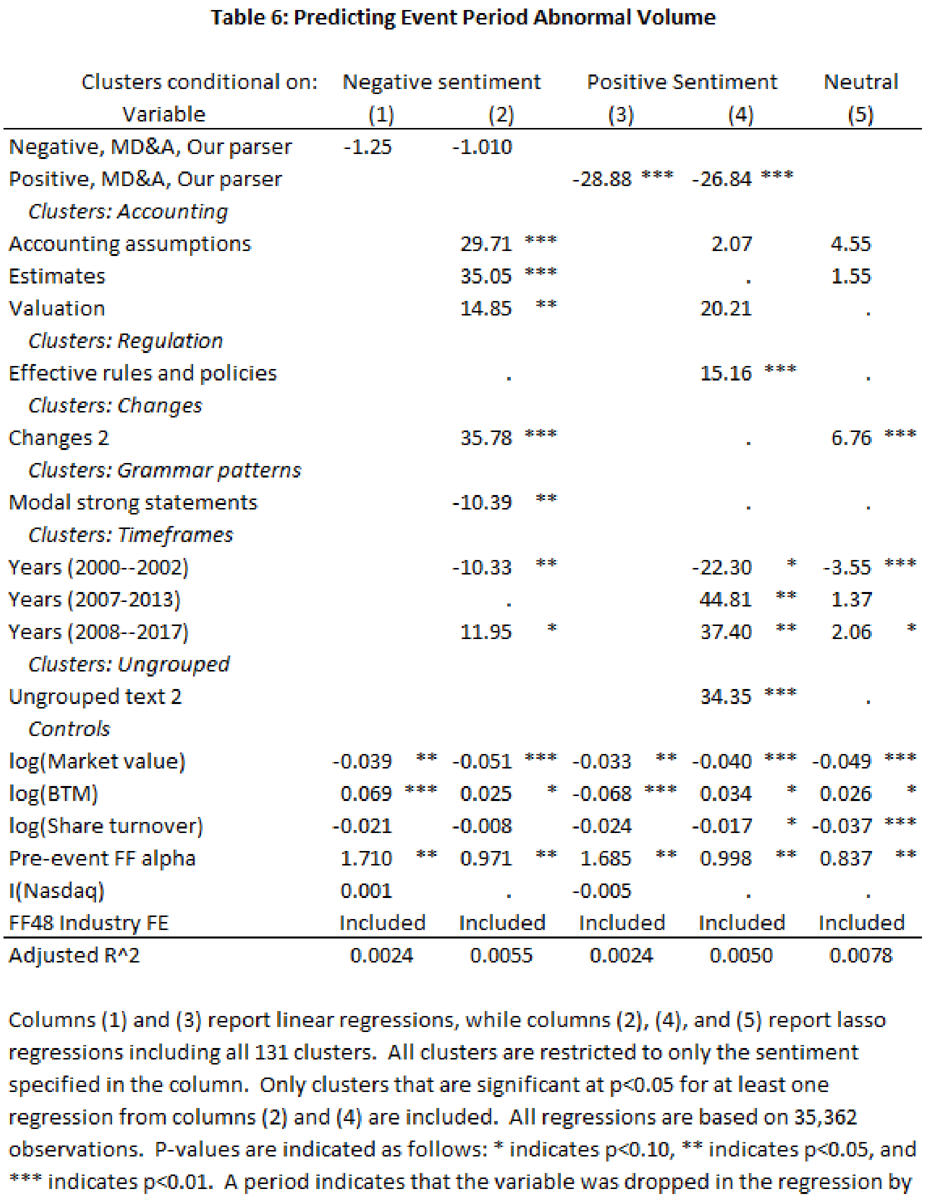


Table 6: Predicting Event Period Abnormal Volume

Clusters conditional on:	Negative sentiment				Positive Sentiment				Neutral	
Variable	(1)		(2)		(3)		(4)		(5)	
Negative, MD&A, Our parser	-1.25		-1.010							
Positive, MD&A, Our parser					-28.88	***	-26.84	***		
Clusters: Accounting										
Accounting assumptions			29.71	***			2.07		4.55	
Estimates			35.05	***			.		1.55	
Valuation			14.85	**			20.21		.	
Clusters: Regulation										
Effective rules and policies			.				15.16	***	.	
Clusters: Changes										
Changes 2			35.78	***			.		6.76	***
Clusters: Grammar patterns										
Modal strong statements			-10.39	**			.		.	
Clusters: Timeframes										
Years (2000--2002)			-10.33	**			-22.30	*	-3.55	***
Years (2007-2013)			.				44.81	**	1.37	
Years (2008--2017)			11.95	*			37.40	**	2.06	*
Clusters: Ungrouped										
Ungrouped text 2							34.35	***	.	
Controls										
log(Market value)	-0.039	**	-0.051	***	-0.033	**	-0.040	***	-0.049	***
log(BTM)	0.069	***	0.025	*	-0.068	***	0.034	*	0.026	*
log(Share turnover)	-0.021		-0.008		-0.024		-0.017	*	-0.037	***
Pre-event FF alpha	1.710	**	0.971	**	1.685	**	0.998	**	0.837	**
I(Nasdaq)	0.001		.		-0.005		.		.	
FF48 Industry FE	Included		Included		Included		Included		Included	
Adjusted R^2	0.0024		0.0055		0.0024		0.0050		0.0078	

Columns (1) and (3) report linear regressions, while columns (2), (4), and (5) report lasso regressions including all 131 clusters. All clusters are restricted to only the sentiment specified in the column. Only clusters that are significant at p<0.05 for at least one regression from columns (2) and (4) 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

Table 7: Predicting Post-event Return Volatility

Clusters conditional on:	Negative sentiment		Positive Sentiment		Neutral
Variable	(1)	(2)	(3)	(4)	(5)
Negative, MD&A, Our parser	1.585 ***	1.026 ***			
Positive, MD&A, Our parser			0.501 **	0.436 **	
Clusters: Accounting					
Accounting assumptions		.		-1.355 **	-0.124
Accounting policies		0.557 ***		0.376	0.008
Deferred tax		1.853 ***		.	0.440 ***
Depreciation and amortization		0.249		-1.688 **	-0.187 **
Financial condition		-0.615 ***		.	.
Financial ratios		-0.735 **		-0.007	-0.232 ***
Inventory and COGS		0.213		1.138 **	.
Loan impairment and loan restructuring		-0.290 **		-0.273	.
Net figures		0.499 ***		-0.019	-0.242 **
Tax rates		0.436		-0.396 ***	-0.068
Clusters: Business operations					
Company details		0.527 ***		.	.
Costs and expense mentions		.		2.527 ***	-0.012
Credit facilities and agreements		0.319 **		.	.
Detailed company info with name		-0.160 **		-0.004	-0.089 ***
Financial services details		0.220 **		0.497 **	0.053 *
Markets (product, region, financial)		0.360 **		.	.
Operating agreements		.		0.781 **	.
Pricing		-0.172 *		-0.629 **	-0.136 **
Research and development		0.538 ***		.	0.056
Clusters: Contracting					
Acquisitions		-0.446 *		-0.949 ***	-0.247 ***
Clusters: Changes					
Changes 2		-0.516 *		-0.915 **	.
Increases and improvements		.		-0.331 ***	-0.260 ***
Clusters: Grammar patterns					
First person plural: charges		0.789 ***		.	-0.202 **
Future requirements		.		1.303 **	0.129
Modal strong statements		0.222		0.778 **	0.140
Periods and fiscal years		.		-0.484 **	-0.400 ***
Reference to tables		1.146 **		0.688	-0.032
Specific dates		0.825 **		0.293	0.281 ***
Company actions (1992-1998)		-0.293 *		-0.665 ***	-0.157 ***
Clusters: Timeframes					
Years (1992--1998)		-0.727 ***		-0.857 **	-0.213 ***
Years (2002--2005)		0.861 *		1.555 ***	0.178 ***
Years (2007-2013)		9.264 ***		7.965 ***	0.484 ***
Years (2008--2017)		0.747 **		.	-0.188 ***
Clusters: Ungrouped					
First person plural ungrouped text 2		.		1.443 **	-0.147 **
Ungrouped text 1		-0.467 ***		-0.066	-0.039 *

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

Controls					
log(Market value)	-0.017 ***	-0.017 ***	-0.017 ***	-0.017 ***	-0.016 ***
log(BTM)	-0.010 ***	-0.010 ***	-0.009 ***	-0.009 ***	-0.009 ***
log(Share turnover)	0.015 ***	0.014 ***	0.016 ***	0.015 ***	0.015 ***
Pre-event FF alpha	0.065 **	0.045 ***	0.066 **	0.046 ***	0.057 ***
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	0.097	0.085	0.090	0.101

Columns (1) and (3) report linear regressions, while columns (2), (4), and (5) report lasso regressions including all 131 clusters. All clusters are restricted to only the sentiment specified in the column. Only clusters that are significant at p<0.05 for at least one regression from columns (2) and (4) 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.

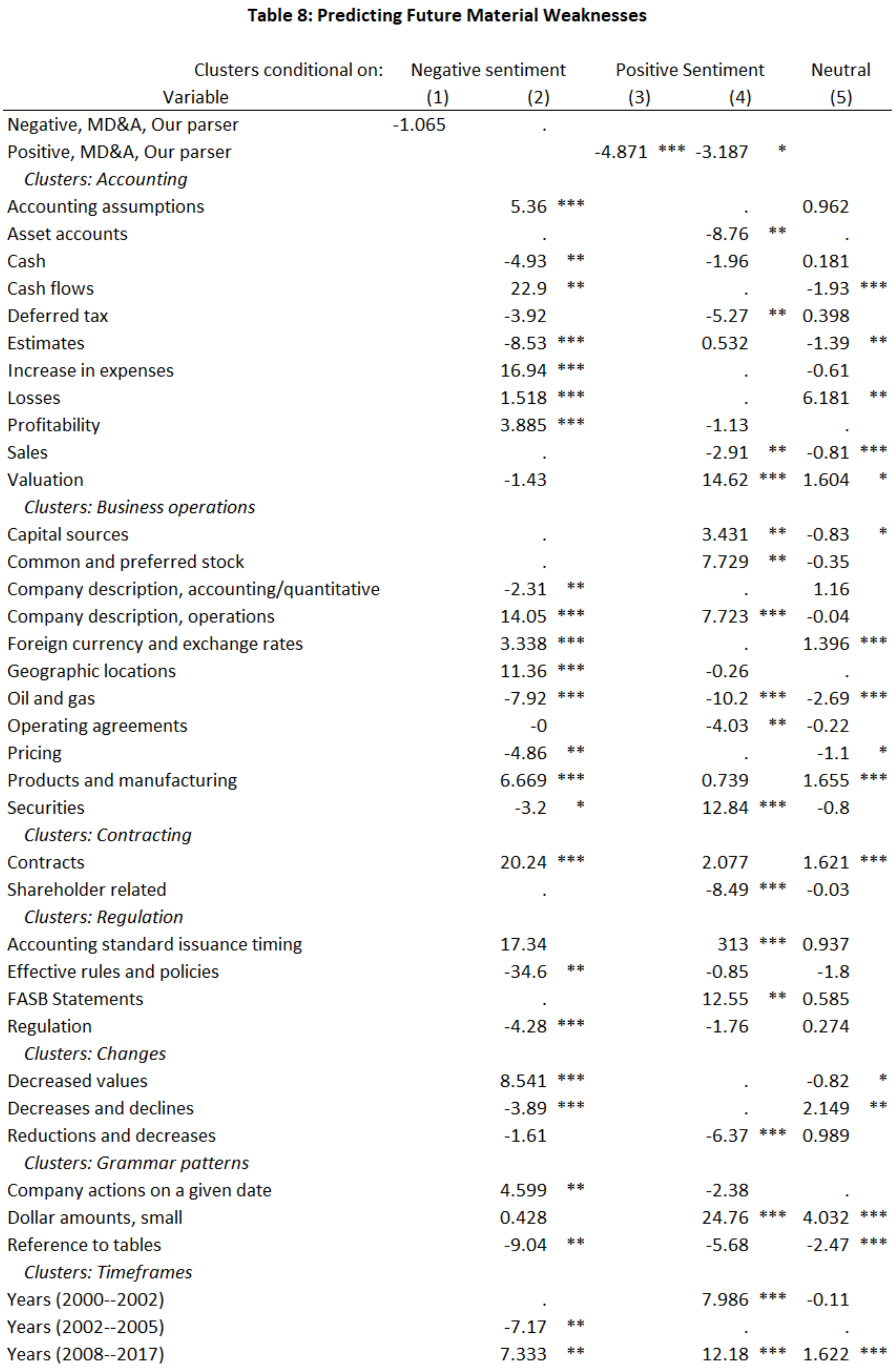


Table 8: Predicting Future Material Weaknesses

Clusters conditional on:	Negative sentiment		Positive Sentiment		Neutral
Variable	(1)	(2)	(3)	(4)	(5)
Negative, MD&A, Our parser	-1.065	.			
Positive, MD&A, Our parser			-4.871 ***	-3.187 *	
Clusters: Accounting					
Accounting assumptions		5.36 ***		.	0.962
Asset accounts		.		-8.76 **	.
Cash		-4.93 **		-1.96	0.181
Cash flows		22.9 **		.	-1.93 ***
Deferred tax		-3.92		-5.27 **	0.398
Estimates		-8.53 ***		0.532	-1.39 **
Increase in expenses		16.94 ***		.	-0.61
Losses		1.518 ***		.	6.181 **
Profitability		3.885 ***		-1.13	.
Sales		.		-2.91 **	-0.81 ***
Valuation		-1.43		14.62 ***	1.604 *
Clusters: Business operations					
Capital sources		.		3.431 **	-0.83 *
Common and preferred stock		.		7.729 **	-0.35
Company description, accounting/quantitative		-2.31 **		.	1.16
Company description, operations		14.05 ***		7.723 ***	-0.04
Foreign currency and exchange rates		3.338 ***		.	1.396 ***
Geographic locations		11.36 ***		-0.26	.
Oil and gas		-7.92 ***		-10.2 ***	-2.69 ***
Operating agreements		-0		-4.03 **	-0.22
Pricing		-4.86 **		.	-1.1 *
Products and manufacturing		6.669 ***		0.739	1.655 ***
Securities		-3.2 *		12.84 ***	-0.8
Clusters: Contracting					
Contracts		20.24 ***		2.077	1.621 ***
Shareholder related		.		-8.49 ***	-0.03
Clusters: Regulation					
Accounting standard issuance timing		17.34		313 ***	0.937
Effective rules and policies		-34.6 **		-0.85	-1.8
FASB Statements		.		12.55 **	0.585
Regulation		-4.28 ***		-1.76	0.274
Clusters: Changes					
Decreased values		8.541 ***		.	-0.82 *
Decreases and declines		-3.89 ***		.	2.149 **
Reductions and decreases		-1.61		-6.37 ***	0.989
Clusters: Grammar patterns					
Company actions on a given date		4.599 **		-2.38	.
Dollar amounts, small		0.428		24.76 ***	4.032 ***
Reference to tables		-9.04 **		-5.68	-2.47 ***
Clusters: Timeframes					
Years (2000--2002)		.		7.986 ***	-0.11
Years (2002--2005)		-7.17 **		.	.
Years (2008--2017)		7.333 **		12.18 ***	1.622 ***

Table 8 (Continued): Predicting Future Material Weaknesses

Clusters: Ungrouped					
Ungrouped text 1		-0.86		5.395 ***	.
Ungrouped text 2		.		6.617 ***	0.279
Controls					
log(Market value)	-0.051 ***	-0.043 ***	-0.050 ***	-0.042 ***	-0.042 ***
log(BTM)	-0.032 ***	-0.024 ***	-0.033 ***	-0.024 ***	-0.029 ***
log(Share turnover)	0.030 ***	0.020 ***	0.030 ***	0.019 ***	0.023 ***
Pre-event FF alpha	-0.084	.	-0.091	.	.
I(Nasdaq)	0.017	0.010	0.016	0.009	0.013
FF48 Industry FE	Included	Included	Included	Included	Included
Adjusted R^2	0.019	0.030	0.019	0.027	0.037

Columns (1) and (3) report linear regressions, while columns (2), (4), and (5) report lasso regressions including all 131 clusters. All clusters are restricted to only the sentiment specified in the column. Only clusters that are significant at p<0.05 for at least one regression from columns (2) and (4) are included. All regressions are based on 23,034 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.