Universität St.Gallen

A Topic Model for 10-K Management Disclosures

Matthias Fengler and Tri Minh Phan

August 2023 Discussion Paper no. 2023-07

School of Economics and Political Science,
Department of Economics

University of St.Gallen

Editor: Mirela Keuschnigg
University of St.Gallen
School of Economics and Political Science
Department of Economics
Müller-Friedberg-Strasse 6/8
CH-9000 St.Gallen
Email seps@unisg.ch

Publisher: School of Economics and Political Science
Department of Economics
University of St.Gallen
Müller-Friedberg-Strasse 6/8
CH-9000 St.Gallen

Electronic Publication: http://www.seps.unisg.ch

A Topic Model for 10-K Management Disclosures

Matthias Fengler and Tri Minh Phan

Author's address:

Matthias Fengler
University of St.Gallen
Dufourstrasse 50
9000 St.Gallen
matthias.fengler@unisg.ch

Tri Minh Phan
University of St.Gallen
Dufourstrasse 50
9000 St.Gallen
triminh.phan@unisg.ch

Abstract

We investigate the topics discussed in the Management's Discussion and Analysis (MD&A) section of 10-K filings from January 1994 to December 2018. In our modeling approach, we elicit the MD&A topics by clustering words around a set of anchor words that broadly define a potential topic. From the topics, we extract two hidden loading series from the MD&As - a measure of topic prevalence and a measure of topic sentiment. The results are three-fold. First, the topics we find are intelligible and distinctive but are potentially multi-modal, which may explain why classical topic models applied to 10-K filings often lack interpretability. Second, topic prevalence and sentiment tend to follow trends which, by and large, can be rationalized historically. Third, sentiment affects topics heterogeneously, i.e., in topic-specific ways. Adding to the extant document-level techniques, our study demonstrates the potential benefits of using a nuanced topic-level approach to analyze the MD&A.

Keywords

10-K files, MD&A, natural language processing, topic modeling

JEL Classification

C55, G30, M41

1 Introduction

The Management Discussion and Analysis (MD&A) section is arguably the most read part of a company's 10-K and 10-Q filing (Tavcar, 1998). It is expected to provide the management's assessment of topics like firm operations, performance goals, capital resources, and liquidity conditions in "... a balanced presentation that includes both positive and negative information about these topics " (FASAB, 2022, Chapter 15, p. 4). The significance of the MD&A section is underscored by its obligatory presence in the annual reports and quarterly filings of listed companies. Consequently, the MD&A section has become an important object of analysis, with studies investigating its informational value to investors and focusing on the MD&A's content level (how much it says) and content sentiment (how it sounds); see Feldman et al. (2010), Li (2010), Mayew et al. (2015), Caserio et al. (2019), among others.

In this paper, we add a new angle of analysis by studying the MD&A's substantive content (*what* it says). Put differently, we study the *topics* of the MD&A section. By the term topic, we refer to a collection of related words that describe certain subjects of firm disclosures, such as sales, expenses, profits, operations, liquidity, investment, financing, litigation, employment, taxes, and accounting.[1] To achieve this goal, we first elicit the sets of topic words that are associated with the suggested topics, based on the MD&A text corpus of 10-K filings from January 1994 to December 2018.[2] In a quantitative second step, we measure the topic words in terms of two topic-specific textual indicators: (*i*) a topic loading, which measures the prevalence of a specific topic; and (*ii*) a topic sentiment, which measures the optimism/pessimism of the language pertaining to the topic.[3] Our modeling concept can therefore be characterized as adopting a topic-level approach. It thus differs from the extant literature, in which textual indicators are determined on the basis of the entire document, e.g., the whole MD&A section. By utilizing our indicators, we aim to explore three questions: (*i*) What are the talking points of the topics? (*ii*) How do the MD&A topical indicators, i.e., the topic loading and the topic sentiment, vary over time? (*iii*) How are the topical indicators and firm fundamentals correlated?

As the first result of our approach, the topics we obtain are distinctly interpretable. Starting from a short list of initial words, which we call *anchor words*, we expand the anchor

[1] Although the US Securities and Exchange Commission (SEC) requires several topics to be covered in the MD&A (SEC, 2003), the requirements are general. Nevertheless, our topics include all the required topics regulated by the SEC, see Section 2.1 for more details.

[2] We choose to focus on 10-K filings rather than 10-Q filings because the stock prices are more responsive to the information in the 10-K filings (Griffin, 2003).

[3] We refer to our textual measure as "sentiment" to express the polarity of tone, i.e., positivity (optimism) and negativity (pessimism), although we acknowledge that a certain stream of literature prefers the term "tone" in order to set it apart from notions of sentiment which are not necessarily fact-based; see, e.g., Baker and Wurgler (2007).

words into comprehensive lists of coherent words,[4] which we refer to as *topic word lists*. We find that the MD&A topics are potentially multimodal, in the sense that a topic can cover several different aspects or subtopics. For instance, we find that the topic of the firm's financing activities contains words referring to borrowing activities (like "credit facility", "borrowing", "senior note") but also words referring to the firm's equity (like "common shares", "shares", etc.). While this appears very natural, as there may be multiple aspects that could embody a topic like corporate financing, the result is still unusual, because statistical topic models typically assume unimodal distributions of words over topics (Park et al., 2019). We provide the comprehensive word lists in Table 7 with the anticipation that these lists will prove valuable for future research focused on the textual composition of MD&A documents.

In the next quantitative steps, we obtain the topic loadings by projection and the topic sentiment by a dictionary-based approach. We then aggregate the topic loading and sentiment indicators in the cross section of firms to examine how the prominence of each topic and its associated sentiment varies over time.

Our research thus documents the historical evolution of the topics contained in MD&A documents over a time span of 25 years. As we discuss in detail, the content of the topics as well as their attached sentiment display very distinct patterns and characteristics over the sample period. For example, the economic state impacts the topics heterogeneously. For concreteness, the financial crisis appears to have a strong impact on the topic of financing activities but only marginally affects the topic of firm investment; in contrast, the dot-com crisis has opposite effects on these two topics. As another example, certain topics, in particular those of sales, profitability, operations, and liquidity, exhibit strong seasonality on an annual basis. Given that US firms usually release their 10-K filings at about the same time every year, this result potentially suggests the usage of generic language in the MD&A compilation.

Finally, by using regression analyses, we link topic information with firm fundamentals. We find that the topics tend to exhibit systematic variation over firm fundamentals. Firms with good performance talk more and more optimistically not only about performance-related topics but also about other topics such as liquidity. Moreover, firms in financial distress are pessimistic not only about financing activities but also about their profitability and investments. We also discover that several firm characteristics do influence topic-level sentiment, but their significance becomes obscured when assessed on the total MD&A sentiment, i.e., the document-level sentiment. For instance, the accrual-on-asset ratio is significantly correlated with the sentiment of the sales and accounting topics, yet it shows no significant relationship to the total MD&A sentiment.

[4] By "coherent", we mean that the additional words, obtained by the expansion, describe the meaning as the initial words.

This work gives insights about the historical evolution of MD&A topics and their corresponding sentiment. Our evidence suggests a substantial amount of heterogeneity across topics, both over time and in the cross-sectional dimension. In particular, the significance of MD&A topics and their sentiment appears to be shaped in distinct ways by historical events, as well as the economic state and firm fundamentals. Without a detailed examination of individual topics, these facts would go unnoticed.

Related literature and contributions

We contribute to the extant literature in two ways. First, we introduce a topic model that takes into account the word semantics[5] specific to the MD&A section. Consequently, the topics generated by our model are easily understandable and closely align with human comprehension. Second, to the best of our knowledge, we are the first to study a topic-specific sentiment measure for MD&A documents as opposed to document-level sentiment.

A growing stream of literature analyzes the textual content of corporate disclosures and connects it to firm characteristics and financial market conditions (Tetlock, 2007; Engelberg, 2008; Henry, 2008; Tetlock et al., 2008; Li, 2010; Loughran and McDonald, 2011a,b; Jiang et al., 2019; Li et al., 2021; Chen et al., 2022; Duan and Yao, 2022). Li (2010) applies the Naive Bayes algorithm to categorize forward-looking sentences in the MD&A sections in 10-K and 10-Q filings into sentimental categories (*positive* versus *negative*). The study finds that the forward-looking sentiment of the MD&A is positively correlated with future earnings. Loughran and McDonald (2011b) introduce a sentiment dictionary designed for financial texts. Applying it to 10-K filings, they find a positive relationship between firm manager sentiment and stock market returns. In a similar vein, Jiang et al. (2019) construct an aggregate manager sentiment index based on 10-K/Q filings and conference calls.

Studying another aspect of the MD&A section, Brown and Tucker (2011) find that its content has become more uniform over the years; they thus conclude that its informational value for stock markets may have eroded. Instead, Cohen et al. (2020) argue that the information in the MD&A section is still important and overlooked by investors. In contrast to our work, these studies are all performed on the document level, meaning that they aim to represent the textual indicators (e.g., sentiment, similarity) of the entire document. We instead elicit two topic-level indicators, namely a topic loading and a topic sentiment. This allows us to examine the finer aspects of a document, thereby aiding in the identification and the capture of potential variations in information across different topics.

[5] Word semantics is referred to as the relationship between words as opposed to word syntax, the arrangement of words in a sentence.

We also contribute to the burgeoning literature on the topic modeling of economic and financial text corpora. One of the earliest attempts to do this is the study by Bao and Datta (2014), who develop a variation of the *Latent Dirichlet Allocation* model (LDA), first proposed by Blei et al. (2003), to quantify risk types from the risk disclosures in 10-K filings. Jegadeesh and Wu (2017) apply the LDA to quantify the economic content in Federal Reserve communication statements (the Federal Open Market Committee minutes). Based on the LDA, Dyer et al. (2017) examine which topics are responsible for the increase in length of 10-K filings over time. Bellstam et al. (2021), by using the LDA, construct a text-based measure of innovation and find that the measure robustly forecasts firm performance. Recently, Brown et al. (2020) have used the LDA model to produce a set of meaningful topics capable of predicting financial misreporting.

Thus, most studies on economic and financial documents rely on the LDA model. However, because of the underlying bag-of-words concept, the LDA model suffers from its intensive computational costs on huge data sets and its lack of word semantics (Mikolov et al., 2013b). Therefore, topics detected by the LDA model tend to be dominated by high-frequency words (words that appear many times in a document and in many documents) if the prior parameters are not specified carefully (Wallach et al., 2009). These properties can limit its usefulness.

Following recent suggestions by Cong et al. (2019) and Li et al. (2021), instead of applying the LDA, we base our topic model on a similarity-based clustering algorithm and the Word2Vec model of Mikolov et al. (2013a). Word2Vec is a neural-network-based natural language model that learns a semantic vector representation of a word or phrase by looking at its relationship with other words of the vocabulary generated from a text corpus, that are found in its neighborhood. The result of this model is a dense matrix of word representations. Word2Vec is increasingly used in Natural Language Processing tasks because it addresses the weaknesses of count-based word-representation methods. As a major advantage, it accounts for the semantics of words in the vocabulary by learning from a corpus of documents. As a result of this learning, words with similar meanings tend to be grouped together in the word vector space. This fact makes Word2Vec suitable for detecting topics in a document because topics tend to be formed from words within a close context.

To incorporate more interpretability in the model, we adopt the concept of *anchor words* to guide the model to learn topics that align with human understanding.[6] We, therefore,

[6] The idea of including prior information is one that has various origins in semi-supervised machine learning algorithms for natural language processing. It is used, for example, in the bootstrapping literature (Thelen and Riloff, 2002), in prototype-based learning (Haghighi and Klein, 2006), and in nonnegative matrix factorization (Choo et al., 2013) but increasingly also in topic modeling (Arora et al., 2012; Lund et al., 2017; Cong et al., 2019) and lexicon definition (Li et al., 2021). There have also been efforts to incorporate prior anchor word information into variants of the LDA; see, for example, Jagarlamudi et al. (2012) and

form the topics by a clustering algorithm that retrieves the words close to these anchor words using the word vectors of the Word2Vec model. To mitigate the subjectivity of the anchor word suggestion, we utilize the word lists introduced in Appendix C of Li (2010). In contrast to the extant literature, we introduce a method to optimally choose the cluster size via the *coherence-coverage* trade-off, making the modeling data-driven. Additionally, to address compound words within the MD&A corpus (those composed of multiple individual words), we apply a method proposed by Mikolov et al. (2013b) to identify and learn phrases. As a result of our modeling, the set of topic words found is readily interpretable and less contaminated by noisy words. The topics formed by our model are also separate and coherent, in the sense that one can understand the main theme of each topic from the topic words alone.

2 The topic modeling framework

The proposed topic model consists of four stages: (i) propose lists of anchor words, which are the sets of initial words that specify human-interpretable topics; (ii) expand the anchor word lists to topic word lists; (iii) compute document-wise topic loadings; and (iv) estimate topic-wise sentiment. These four steps rely on the following three technical building blocks: (i) a phrase-learning model, which detects phrases in the corpus; (ii) a Word2Vec model, which maps words into a vector space that captures the semantic orientation of the language used in the corpus; and (iii) a sentiment projection.

2.1 Formation of anchor words via the phrase-learning model

The starting point for our model is the suggestion of lists of anchor words, each of which is intended to represent a topic discussed in the MD&A documents. The purpose is to inject initial expert knowledge about the MD&As into our learning algorithm so that it starts allocating words to the desired topics. This will reinforce the interpretability of the topics learned. To limit subjectivity, we use the word lists and categories offered in Appendix C of Li (2010). This set of word lists spans eleven categories and includes words from the following topics: *Sales/Revenue, Cost/Expense, Profit/Loss, Operations, Liquidity, Investment, Financing, Litigation, Employment, Regulation/Tax*, and *Accounting* (see Table 1). The concept of using anchor words, although not very recent, is gradually becoming used in topic modeling to seed prior knowledge into topic models (Arora et al., 2012; Jagarlamudi et al., 2012; Lund et al., 2017; Cong et al., 2019; Eshima et al., 2020; Li et al., 2021). We use the anchor word lists for two specific purposes: (i) to validate the phrase-

Eshima et al. (2020).

learning model; and (ii) as semantic anchors for the clustering algorithm that identifies the topic word lists. The first purpose will be detailed in Appendix B, while the latter will be described in Section 2.2.

A major challenge of mining economic and financial texts is that they contain a large number of phrases. A phrase is a compound of two or more words (i.e., single tokens in a text) whose meaning is not fully described by its component words; as examples, consider "market condition" or "capital expenditure". Indeed, our anchor word list in Table 1 features a high number of phrases. Additionally, in order to implement the word embedding, one needs to build up the vocabulary of the corpus, in which, again, many phrases would be omitted if they were not handled properly. Therefore, detecting phrases is an essential task in our modeling approach.

The classical way of handling phrases is to use n-grams.[7] This approach, however, drastically increases the size of the vocabulary; at the same time, the computational burden of the model increases because meaningless phrases are also taken into account (Mikolov et al., 2013b). To handle this issue, we train the phrase-learning model introduced by Mikolov et al. (2013b) to detect phrases (bigrams and trigrams) appearing in the corpus. Generally, a phrase-learning model has two major benefits over an n-gram model. First, it automatically learns phrases in the corpus, with no human intervention, thus avoiding subjectivity. Secondly, it enriches the vocabulary selectively by only adding to the vocabulary phrases with plausible meanings. This makes the word embedding absorb more refined word semantics, and at the same time, it lowers noise compared to n-gram models. We explain all the details, including the hyperparameter selection, in Appendix B.

After obtaining the complete vocabulary including single words and phrases detected by the phrase-learning model, we map each word and phrase in the vocabulary into a dense vector representation using the Word2Vec model of Mikolov et al. (2013a). Word2Vec represents a word as a vector and it captures the semantics of the word, in the sense that two words that bear a similar meaning also have a close representation in the vector space. To achieve this goal, the model maximizes the similarity of two words that appear together within a context window and minimizes the similarity of words that do not appear together. We train the Word2Vec model directly on the MD&A corpus after several text normalization steps; all details of text normalization and model hyperparameters are relegated to Appendix C. Both the Word2Vec model and the phrase-learning model are trained using the *gensim* package, an open-source Python library with C++ backend (Řehůřek and Sojka, 2010). After phrase-learning and word vectorization, words and phrases both have a (single) vector representation and from now on we simply refer to them as *words*. The word vectors are the main ingredients for the next stage, the formation

[7] An n-gram is a sequence of n adjacent words. A bigram and a trigram correspond to the cases of $n = 2$ and $n = 3$, respectively.

Topic 1: Sales/Revenue	Topic 2: Cost/Expense	Topic 3: Profit/Loss	Topic 4: Operations	Topic 5: Liquidity	Topic 6: Investment	Topic 7: Financing	Topic 8: Litigation	Topic 9: Employment	Topic 10: Regulation/Tax	Topic 11: Accounting
sale	cost	profit	operation	liquidity	investment	financing	litigation	employee relation	regulation	accounting method
revenue	expense	income	production	interest coverage	general capital expenditure	debt	lawsuit	retention	environment law	accounting estimation assumption
market condition	reserve for contingent liability	performance result	general business	cash balance	M&A	equity		hiring	income tax	auditing
market position	asset impairment	margin		working capital condition	divestiture	dividend		union relation	government relation	internal control
consumer demand competition pricing new contract	goodwill impairment				discontinued operation	repurchase				

Table 1: Lists of anchor words proposed by Li (2010). Li's category of "Miscellaneous" is not considered here. Words in these lists are made singular to align with the lemmatization in our textual preprocessing step. The anchor words produced by the phrase-learning model are given in Table 6 in Appendix B.

of the topic word lists via a guided clustering algorithm.

2.2 Formation of topic word lists via a coherence-coverage trade-off

The next step is to expand the anchor word lists to create the corresponding topic word lists. The expansion is done via a similarity-based clustering algorithm applied to the word embeddings obtained by the Word2Vec model in Section 2.1. To implant the initial expert knowledge, we use the anchor words as "semantic anchors" and search for semantically related words. This process requires us to address the difficult question of determining the "relatedness of words". For word embeddings, where vectors are proxies for the meaning of words, it is natural to employ a distance measure on the word vectors as a measure of semantic relatedness (Kiela et al., 2015; Kruszewski and Baroni, 2015).

One approach for expanding the anchor word lists into topic word lists is the *similarity threshold* approach, according to which two words are considered to be similar when their cosine similarity[8] is larger than a given threshold (Rekabsaz et al., 2017). This result, however, may depend on the corpus on which the Word2Vec model is trained. Therefore, in this paper, we determine an optimal cluster size by maximizing a *topic coherence-coverage trade-off*. Topic coherence is measured by *Normalized Pointwise Mutual Information* (Bouma, 2009; Lau et al., 2014; Dieng et al., 2020), which decreases (increases) when the cluster size is larger (smaller). Topic coverage, in turn, is measured by the proportion of topic word counts to the total number of words in the corpus. This quantity increases (decreases) given a larger (smaller) cluster size.

Details of this optimization are given in Appendix D. As a result of the optimization process, the cluster size of 10 neighboring words is chosen. To ensure robustness, we verify that the optimal cluster size and its nearby values (sub-optimal cluster sizes) yield qualitatively similar regression results (see Section 6.2). To control for words appearing in several topics,[9] we assign these words to the topic which has the closest anchor word to them.[10]

[8] The cosine similarity between two vectors x and y is defined as $\text{sim}(x, y) = \frac{\langle x,y\rangle}{\|x\|\|y\|}$, where $\langle x, y\rangle$ is the inner product of two vectors x, y, and $\|x\|$ is the Euclidean norm of a vector x. The cosine similarity measure is a natural choice because Word2Vec learns words that are adjacent to each other in terms of cosine similarity.

[9] This property is referred to as the *separability* of topic modeling (Arora et al., 2012).

[10] For example, suppose word w_T appears in three topics T_1, T_2, and T_3 with the corresponding anchor words w_1, w_2, and w_3. If the cosine similarity between w_T and w_1 is the largest among the three anchor words, we assign w_T to topic T_1. If a word is chosen by many anchor words in a topic, we base the measurement on the average cosine similarity.

2.3 Estimation of topic loadings

The next step in our proposed clustering algorithm is to gauge how much a document talks about a given topic. To this end, for each MD&A document, we compute its *topic loading* by means of projections. The topic loadings can be interpreted as the extent to which each topic is covered in a given document.

Fundamentally, the document-term matrices, which are obtained as a result of the word list formation, capture the desired information about a given topic. A document-term matrix has rows representing the documents of a corpus and columns representing words. Each entry of the matrix shows the occurrences of a word in the corresponding document. These matrices, however, are very unwieldy because they have a high number of dimensions. Instead, we seek to summarize the information optimally in a 1-dimensional space. With this intention, *Singular Value Decomposition* (SVD) is a natural candidate to address this challenge.

Indeed, SVD, and matrix factorizations in general, play an enormous role in topic modeling (Dumais, 2004; Arora et al., 2012; Choo et al., 2013, 2015). In like manner, Cong et al. (2019) use SVD to estimate topic loadings from the topic-specific document-term matrices. By applying SVD to a topic-specific document-term matrix N, one obtains the 1-dimensional subspace that maximizes the cumulative magnitudes of the projections of all rows of N on this subspace. The topic loadings are the magnitudes of these projections. We follow these ideas here and defer the technical details to Appendix F.

Because of these properties, the size of the projections, formally given by $\sigma_1 u_1$,[11] ranks the rows of the document-term matrix. Thus, they can be interpreted as the extent to which a document talks about a given topic, or its *topic loading*. High loadings imply a high prevalence of the topic, while loadings close to zero imply that the text and the topic words are close to orthogonal and thus unrelated. On the other hand, the entries in v_1 rank the magnitudes of the projections of the columns of the document-term matrix N onto the first left singular vector u_1 (up to the scale of the first singular value σ_1). In our case, each column in N carries the frequencies of each topic word in the corpus (word count over the corresponding document length). Therefore, the entries of v_1 serve as a measurement of the importance of the topic words over the entire corpus. We exploit this fact when visualizing the word clouds in Figure 1 of Section 4.1.

That said, comparing the magnitudes across different topics is more subtle. This is because topics differ in size and the topic loadings are obtained from different document-term matrices. Different loadings may thus occur owing to the size counts and size vari-

[11] σ_1, u_1, and v_1 are the first singular value, the first right and the first left singular vectors, respectively, of a topic-specific document-term matrix N.

ation of the topics. Nevertheless, because the loadings are all scaled by the first singular value, which is the maximizer of each of these row-wise projections, we still read the magnitudes of the factor loadings across topics as an indicator of the prevalence of the given topics across the corpus. Note that we apply SVD to the *normalized* document-term matrix, in which each row is divided by the total word counts of the corresponding document, instead of the raw document-term matrix, to make the interpretation robust against differing document lengths.

2.4 Estimation of topic sentiment by lexicon projection

The final stage of our proposed model is to incorporate sentiment information into the document-wise topic loadings of Section 2.3. Sentiment analysis in economics and finance is commonly based on unsupervised learning, typically a lexicon projection, and relies on a pre-defined sentiment dictionary (Tetlock, 2007; Loughran and McDonald, 2011b; Jiang et al., 2019; Chen et al., 2022). The Loughran-McDonald (LM) dictionary (Loughran and McDonald, 2011b) has emerged as the standard because it is designed for economic and financial texts, see, e.g., Feldman et al. (2010), Dougal et al. (2012), Garcia (2013), Jiang et al. (2019). Despite the dominance of lexicon projections in the literature, recent advances suggest supervised learning approaches (Chen et al., 2022).

Here, we adopt the approach of Jiang et al. (2019), with a small modification, to compute the document sentiment score.[12] For an MD&A document d of a given firm, we locate the topic words of topic j. Then, within a window of five words around the identified topic word, we search for sentimentally-charged words as defined by the LM dictionary. This scanning is restricted to within sentences (determined by periods) to prevent information spillover between adjacent sentences. The positive (negative) score of topic j in document d, $s^{+}_{j,d}$ ($s^{-}_{j,d}$), is computed as the sum of the topic-specific positive (negative) word counts divided by the total word count of document d. The sentiment score of topic j in document d is

$$s_{j,d} = \frac{s^{+}_{j,d} - s^{-}_{j,d}}{l_d} \tag{1}$$

where l_d is the total word count in document d. In this way, we aim to capture sentiment information only in the vicinity of the topic words.

Besides topic sentiment scores, we compute the overall sentiment score of the MD&A documents, independently of the topic loading information, which provides a helpful

[12] There are a variety of approaches to measuring a sentiment score. See also Antweiler and Frank (2004); Loughran and McDonald (2011b); Chen et al. (2022).

sanity check.

3 The text corpus

The 10-K filings can be downloaded directly from the webpage of the SEC or The Notre Dame Software Repository for Accounting and Finance (SRAF).[13] The latter page also provides additional resources for textual data analysis, such as stopword lists and the LM dictionary. The SRAF data of the 10-K and 10-Q filings are in text-file format, with HTML tags having been removed. We focus our analysis on 10-K filings only because the information content is acknowledged to be more significant than that of 10-Q filings (Griffin, 2003). We design our own extractor to excerpt the MD&A section out of each 10-K files, following the advice laid out by Loughran and McDonald (2016), and we manage to extract 68% of all the 10-K files in the corpus.[14] We discard documents that have fewer than 250 words in the MD&A section. After these purges, we retain $124,133$ MD&A documents spanning the period 1994:01 to 2018:12.[15]

After extracting the MD&A documents from the 10-K filings, we execute several standard steps for text normalization. In doing this, we take particular care to properly process negations because ignoring negations changes the polarity of a statement and leads the sentiment analysis astray (Mukherjee et al., 2021); for details about the text normalization and the settings of the phrase-learning and Word2Vec models, see Appendix C.

Finally, we match the MD&A data with fundamental data from the CRSP/Compustat merged database. The matched data set includes $6,065$ stocks (PERMNO numbers) and spans 26 years from 1994:01 to 2018:12. Table 2 reports the data loss incurred during the extraction and processing steps. Compare with Li (2010) and Loughran and McDonald (2011b), we find that we are similarly successful in these steps. In Li (2010), the data include all 10-K and 10-Q filings from 1994:01 to 2007:12. Adjusting for the different time span and discarding the 10Q files, the sample sizes match. In Loughran and McDonald (2011b), the survival rate of firm-year observations amounts to 30.8% in a sample comprising 121,217 10-K and 10-K405 files from 1994:01 to 2008:12, while ours is 27.4%, which is comparable.

[13] https://sraf.nd.edu/

[14] Loughran and McDonald (2011b) first match the 10-K files to CRSP data, then extract the MD&A sections from the matched 10-K files. With a sample spanning from 1994 to 2008, they obtain roughly 49.55% successfully extracted MD&A from the match 10-K filings.

[15] From now on, we use the format of *yyyy:mm* to indicate the time of month *mm* in year *yyyy*.

	1994	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**Total**
No. of 10-K filings	1910	2213	4293	6640	6861	6716	6578	6225	6670	8433	8524	8997	8821	8524	8641	9785	9095	8750	8333	7998	7955	7842	7452	7157	7003	181,416
No. of identified MD&A	900	1150	2573	4138	4337	4442	4408	4386	4563	5711	5551	5499	5199	4983	5483	6879	6560	6613	6365	6136	6097	5942	5669	5378	5171	124,133
Merge with firm fundamental data	356	539	1015	1677	1791	1868	1812	1942	2019	2549	2514	2524	2437	2368	2393	2425	2323	2309	2245	2177	2160	2202	2158	1998	1818	49,619
Successfully extracted ratio (%)	47.1	52.0	59.9	62.3	63.2	66.1	67.0	70.5	68.4	67.7	65.1	61.1	58.9	58.5	63.5	70.3	72.1	75.6	76.4	76.7	76.6	75.8	76.1	75.1	73.8	68.4
Survival ratio (%)	18.6	24.4	23.6	25.3	26.1	27.8	27.5	31.2	30.3	30.2	29.5	28.1	27.6	27.8	27.7	24.8	25.5	26.3	26.9	27.2	27.2	28.1	29.0	27.9	26.0	27.4

Table 2: Effects of data extraction and processing steps on the MD&A sample size. The *No. of identified MD&A* shows the number of MD&A documents after the extraction and the longer-than-250-word filter. *Successfully extracted ratio* is the ratio between the number of identified MD&A documents and the initial 10-K filings. *Survival ratio* is the ratio between the ultimate number of MD&A documents and initial 10-K filings.

4 The content of the MD&A section

4.1 What talking points do the topics have?

Among the most crucial qualities of a topic model is its capability to reveal the diverse aspects encompassed by a topic. We learn these aspects by studying the topic word lists obtained from our model, which we offer in Table 7 in Appendix E. This table provides the words for two configurations of cluster sizes: the optimal configuration of 10, on which we base the main analysis, and a slightly larger variant of 15. For ease of exposition, we also present the word lists of the first configuration by means of topic word clouds in Figure 1. Words represented with larger letters are relatively more important than words depicted in smaller letters, as explained in Section 2.3.

As can be inferred from Figure 1, the most important words in the topic *Sales/Revenue* are "revenue_result", "market_environment", "revenue_related", "contract_include", and "pricing_structure".[16] Besides covering firm sales and revenues, this topic includes matters of competition (words like "competition", "competitive_environment", "competitive_pressure"), consumers and customers (with words like "consumer_preference", "consumer_spending", "customer_demand"), the current state of the economy (words like "economic_condition", "economic_environment", "downturn"), and the pricing strategy (with words like "pricing_level", "pricing_structure", "pricing_product"). These aspects all relate to the ability of a firm to generate revenue.

The topic *Cost/Expense* covers three aspects of a firm's costs and expenses. The first aspect relates to operating costs and expenses, with words like "cost", "expense", and "expense_associate". It is worth noting that this topic does not cover the costs and expenses of two further specific activities, namely wages (covered by *Employment*) and costs due to taxation (covered by *Regulation/Tax*). This result emphasizes the robustness of the proposed topic model in forming intuitive and cohesive topics. The second aspect dealt with by the topic is asset impairment with words such as "impairment_asset" (impairment of assets), "impairment_charge" (impairment charge), and "impairment_goodwill" (impair-

[16] The fact that all these terms are phrases underscores the importance of the phrase-learning algorithm (see Appendix B).

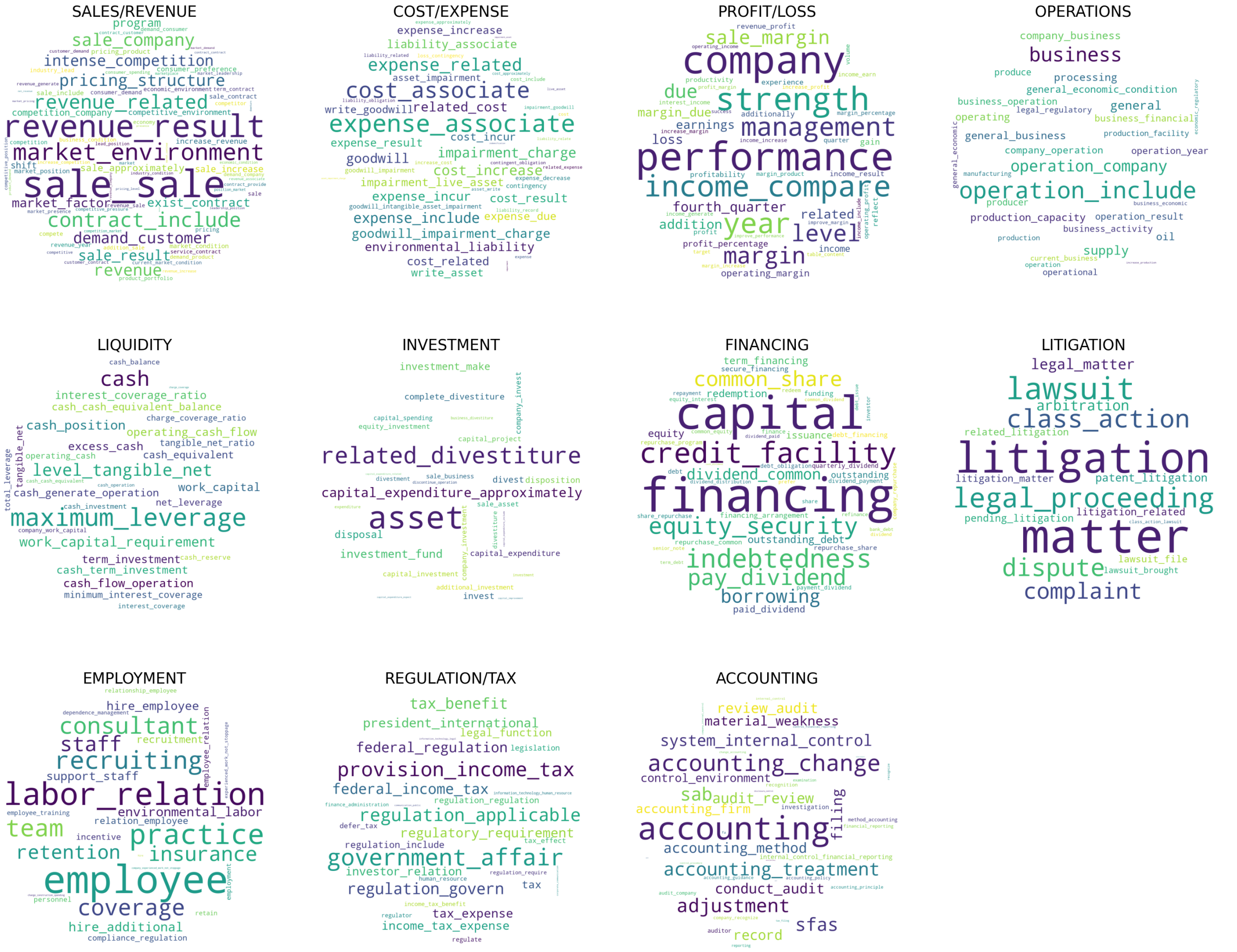

Figure 1: The topic word clouds of eleven topics, with the cluster size of 10, as shown in Table 7 (the non-underlined words). These words and phrases are generated by the phrase-learning model and the similarity-based clustering algorithm. The larger a word is in each cloud, the more important that word is in the corpus, compared to other words in the same cloud (topic). The importance of words is determined by the first right singular vector, $v_j^{(1)}$, as explained in Section 2.3.

ment of goodwill). This appears reasonable because the dollar value of an impairment is the difference between the asset's carrying value and its fair market value. Interestingly, in the topic word list for *Cost/Expense*, we observe further related words like "write_asset" (write-off of assets) or "write_goodwill" (write-off of goodwill), which also refer to impairment. We note that the anchor word list of *Cost/Expense* does not involve the word "write". This finding shows that the proposed topic model is capable of detecting associated words that are beyond the initial anchor word lists. The last aspect covered by this topic refers to the firm's liabilities, reflected in words such as "liability_obligation", "liability_record", and "liability_related".

The topic *Profit/Loss* describes a firm's performance, i.e., its income and losses. The central words are "performance" and "income_compare". Additional salient words for this topic are "income_generate", "increase_profit", "interest_income", "loss", "margin_product", and "sale_margin". The topic *Operations* includes words describing production and manufacturing. Besides words directly relating to the anchor words of this topic, such as "operation_include" and "business", words like "supply", "oil", and "production_capacity" also appear as material words of *Operations*. Further important words, such as "manufacturing", "producer", and "production_facility", are all closely related to this topic.

The topic *Liquidity* discusses various facets of a company's liquidity, such as its cash holdings, interest coverage, and working capital. The topic words for this topic have equal importance, as can be seen from the fairly homogeneous word size in the *Liquidity* word cloud. This may reflect the fact that firm managers use diverse language to talk about firm liquidity. Two of the most important words of *Liquidity* are "maximum_leverage" and "level_tangible_net" (level of tangible net worth). Note that these words do not directly relate to the anchor words of this topic. The appearance of "maximum_leverage" (ratio) makes sense in the context of firm liquidity, given that this word has a high cosine similarity score to "interest_coverage" in the anchor word list of *Liquidity*. In particular, the interest coverage ratio, which is defined as firm operating income divided by interest expenses, can be seen as a measure of firm leverage. The second most important word for the *Liquidity* topic is "level_tangible_net" which describes a firm's level of tangible assets. Tangible net worth includes the physical assets of the firm, which can be easily converted to cash and thus serve as a source of firm liquidity. Therefore, the appearance of this word is conforming with a topic about firm liquidity. Besides these words, this topic also describes a firm's immediate liquidity, with words like "cash_cash_equivalent_balance" (cash and cash equivalent balance), "cash_generate_operation" (cash-generating operations), and "work_capital_requirement" (working capital requirement).

The topic *Investment* focuses on both divestment and investment. This is seen from topic words like "divest", "disposition", and "disposal", which describe firm divestment. Moreover, we find associated words like "asset", "sale_business" (sales of businesses), and "sale_asset" (sales of assets). It is worth noting that these words, despite featuring "sale", and without bearing any obvious relation to the *Investment* anchor words, are allocated by our model to the *Investment* topic rather than *Sales/Revenue*. Nevertheless, this result is intuitive given that these words refer to the firm's activities of selling assets rather than making sales. The other aspect of this topic describes the firm's investment decisions with words such as "capital_spending", "capital_investment", "investment_fund", and "equity_investment".

The topic *Financing* characterizes the firm's financial resources, including debt and equity. Similar to *Liquidity*, two out of the three most important words of this topic, "capital" and "credit_facility", do not directly relate to the topic anchor words. Note that "capital" belongs to the topic *Financing*, while words like "capital_expenditure" and "capital_spending" belong to *Investment*. This, however, appears correct as *Investment* describes the firm's activities pertaining to its productive assets, whereas *Financing* talks about the firm's capital structure. Besides "credit_facility", we detect words in *Financing* which do not appear or directly relate to the topic anchor words, such as "borrowing", "senior_note", and "redeem". These all describe funding via debt. As regards the firm's equity, we uncover further words that are beyond the initial anchor words like "common_share", "investor", and "share", among others.

With only two anchor words in the *Litigation* topic, namely "litigation" and "lawsuit", our model successfully detects a number of important topic words like "arbitration", "complaint", "dispute", and "legal_matter", which are within the scope of litigation but outside the initial anchor words. Our model enriches the anchor word list of the *Employment* topic in a similar manner. Significant words for this topic are "compliance_regulation", "personnel", "staff", "insurance", and "incentive". These important words do not directly relate to the anchor words, yet they are spotted out by our model. There are two main aspects in *Regulation/Tax*, namely regulation (with words like "government_affair", "legislation", and "corporate_communication") and taxation (with words like "defer_tax", "provision_income_tax", "tax_benefit", and "tax_expense"). We also find that "income_tax" and "tax_expense" are not allocated to the topics *Profit/Loss* and *Cost/Expense*, respectively. This allocation, however, appears plausible as these words describe tax-related income and tax-related expenses while *Profit/Loss* and *Cost/Expense* discuss operating income and operating expenses. Finally, the topic *Accounting* mentions accounting and auditing activities within a firm as well as accounting and auditing standards with the words "sab" (Staff Accounting Bulletin) and "sfas" (Statement of Financial Accounting Standards).

There are two key takeaways from the above discussions. First, our model successfully expands the initial anchor word lists to create comprehensive topic word lists. A significant number of words detected by our modeling approach do not directly involve the anchor words in an obvious way, but still turn out to be highly important words in their respective topics. This finding emphasizes the importance of the topic word formation step in Section 2.2. At the same time, it shows that parsimonious anchor word lists like ours allow serendipitous results to be found. The second takeaway is that some words that seem to be similar at first glance are assigned to two different topics by our model. Overall, we can conclude that the words have been effectively categorized into the respective topics. This is attributed to the accurate identification of semantic contexts achieved by the Word2Vec model. We give a more profound discussion of the word distribution in the next section.

4.2 How are the topic words distributed?

To provide a visual representation of the distribution of topic words, we show the projection of the topic word vectors onto the 2-dimensional space through a scatter plot in Figure 2.[17] As a first observation, we note that the topics are well-separated. Secondly, several topics have sub-topics, as can be seen from the sub-clusters of the same color. Furthermore, as is also mentioned in Section 4.1, some topics are woven into each other. In particular, the three topics *Sales/Revenue* (royal blue), *Profit/Loss* (green), and *Operations* (red) appear very close to each other. It should be recalled, however, that we are studying 300-dimensional word vectors, which are projected into the plane. Therefore, even though some topics are visually woven into each other in the plane, this does not necessarily mean that the topics are close to each other in the original space. Nevertheless, to a certain degree, this relatedness can be rationalized on semantic grounds in a number of cases.

For example, the part of *Sales/Revenue* (royal blue) relating to firm sales and revenue, with the words "sale", "sale_result", and so on in the lower left of Figure 2, is close to the part of *Profit/Loss* (green) talking about firm profit and income (words like "operating_income", "sale_profit", "profit_margin", etc.). This adjacency is comprehensible as they both describe firm performance. It is worth noting that the part of *Profit/Loss* (green) in the center of Figure 2 is next to the left part of *Regulation/Tax* (cyan), which is dominated by words relating to taxes. This part of *Profit/Loss* (green), which is close to words about taxes of *Regulation/Tax* (cyan), features many words related to "income", but is naturally far from the part talking about firm profitability.

The topics *Cost/Expense* (orange), *Liquidity* (purple), *Regulation/Tax* (cyan), and *Accounting* (blue) are well-separated in the vector space. Specifically, *Cost/Expense* (orange) contains three distinct clusters: one (in the left of Figure 2) talks about firm costs and expenses, another (in the right of the figure) about liabilities, and the third (in the upper-left corner of the figure) about asset impairment. In *Liquidity* (purple), words about interest coverage are located separately from those describing the firm's short-term assets (cash and working capital). The *Regulation/Tax* topic (cyan), as the name suggests, mentions two distinct aspects, the first being regulation and legislative restrictions, and the other being taxation. The latter is adjacent to the income-related cluster of the *Profit/Loss* topic (green) as mentioned above. Similarly, *Accounting* (blue) displays accounting-related aspects, such as accounting methods, accounting principles, and a part that is about internal controlling and auditing.

[17] We use Stochastic Neighbor Embedding to project the high-dimensional word vectors to the 2-dimensional space for visualization. This technique tends to produce a better low-dimensional representation than classical linear dimension reduction approaches such as principle component analysis (Van der Maaten and Hinton, 2008).

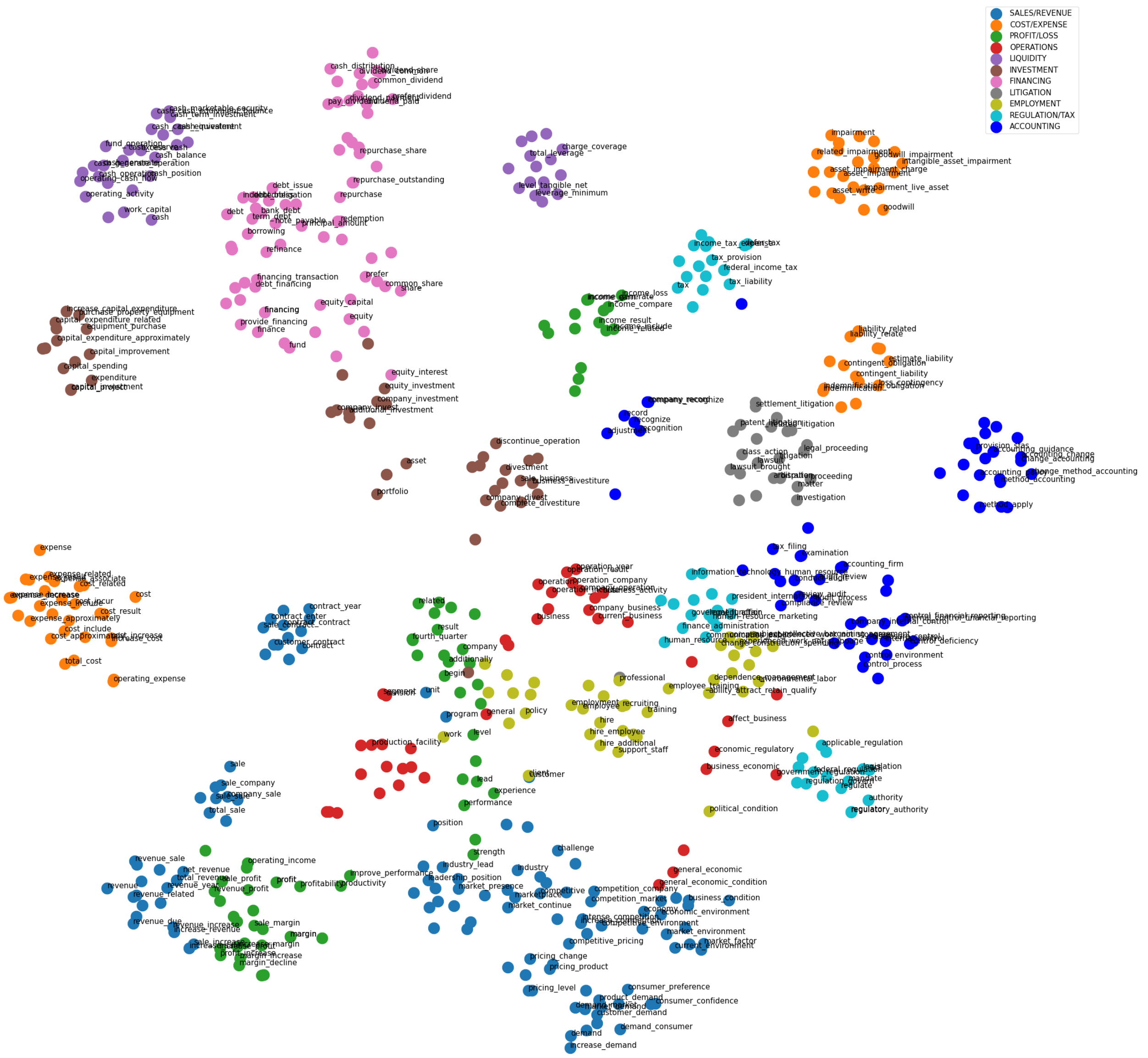

Figure 2: A visualization of the word vectors of the topic words in a 2-dimensional space. The dimensionality reduction is achieved by Stochastic Neighbor Embedding (Van der Maaten and Hinton, 2008). Because of the complexity, only 25 random words of each topic are reported in the graph.

Finally, the topics *Employment* (olive) and *Litigation* (grey) are very well separated. This appears naturally because both topics tend to use only a specific set of closely related words within the MD&A sections, offering little semantic overlap with other topics. Similarly, the part of *Investment* in the upper left of Figure 2 about capital expenditure is close to the topic *Financing* about credit and debts, which is in the upper part of the figure. These two topics are, however, well separated.

From these observations, two conclusions can be drawn. First, because of the underlying Word2Vec model, topics appear close when the word collections of those topics are related

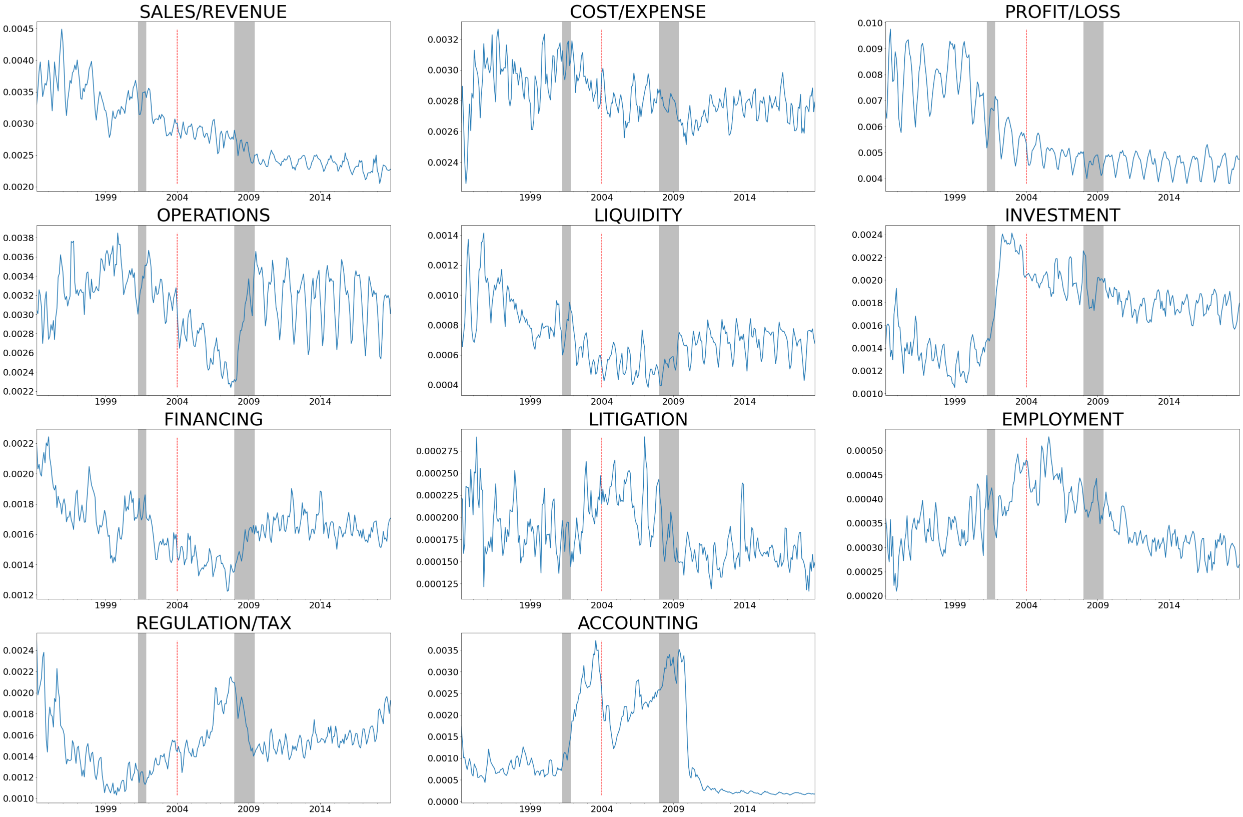

Figure 3: Monthly topic loading time series, with the cluster size of 10. These loading indices are constructed by aggregating the topic loadings of all firms that submit their 10-K filings in a given month. The time series is smoothed by a moving average over the four previous months. The vertical grey bars indicate the economic recessions defined by the NBER. The dashed vertical red lines indicate the time of 2004:01 when the 2003 SEC regulation became effective. The sample spans the period 1994:01 to 2018:12.

to each other in their semantic orientation (e.g., *Sales/Revenue* and *Profit/Loss*). Secondly, topics may contain more than one sub-topic, and words describing these sub-topics may be located distantly from each other in the vector space. Thus, topics like *Cost/Expense* are, in some sense, multimodal. For these two reasons, a cluster analysis based on classical and purely data-driven methods would be challenging and potentially misleading. For example, words like "income_earn" (income and earnings) can easily be merged with the word "income_tax" although these two words depict two different objects. Our approach, by contrast, uncovers well-defined topic words, which sets it apart from results obtained from classical LDA.

5 How does the MD&A content change over time?

5.1 Topic loading time series

In this section, we study the time series of topic loadings and topic sentiment. We start the analysis by presenting the time series of the topic loading indices. The index of a topic loading is constructed by averaging the topic loadings of all firms that submit their 10-K filings in a given month. Following Jiang et al. (2019), we smooth the time series by a moving average over the four previous months but we do not standardize them in order to preserve their interpretation as explained in Section 2.3. Figure 3 shows the indices of the eleven topics from 1994:01 to 2018:12. Besides the economic recessions, as defined by the National Bureau of Economic Research (NBER), another event of interest is the release of the SEC's guidance on the content discussed in the MD&A section, which became effective on December 29th, 2003.[18] With the 2003 SEC regulation, the SEC required firms to increase the informativeness of the MD&A section. As a result, this regulatory update has become is an intriguing subject for studying changes in MD&A contents (Li, 2010; Brown and Tucker, 2011).

All the topics exhibit substantial variation in time. Starting with the topics *Sales/Revenue, Profit/Loss*, and *Liquidity*, we observe a downward trend in the loadings, especially after the release of the 2003 SEC regulation. As discussed in Section 2.3, this trend implies that firm managers have deprioritized these topics over the sample period. The loadings of *Operations* recede shortly after the dot-com bubble until before the financial crisis, during which the *Operations* loadings rise rapidly. This observation implies that firms increasingly discuss their operations (production, manufacturing, etc.) during and after the financial crisis. It also suggests that managers, under the impression of the crisis, are induced to provide more detailed discussions of firm operations to investors and stakeholders.

Investment, by contrast, is most affected by the dot-com crisis from 2001:04 to 2001:11. In particular, there is a sharp increase in the *Investment* loadings starting at the beginning of the dot-com bubble, implying that firms, on average, are talking more about their investment activities during this time. A potential explanation of this pattern could be the specific business climate of the time. As documented by Ljungqvist and Wilhelm Jr (2003), a large proportion of the IPO proceeds of the years before were employed to finance daily operating activities rather than debt repayment, capital expenditures, or investment plans. Against the backdrop of this diversion of funds, firm managers may have felt compelled to elaborate more on their investment plans, and this, in turn, is reflected

[18] For brevity we will hereafter refer to the SEC's regulatory guidance that was issued in 2003 as the "2003 SEC regulation".

in the strong increase in the *Investment* loadings during the dot-com crisis.

There are further remarkable patterns. The loadings of *Financing* gradually decrease until shortly before the financial crisis. During the crisis, the discussion of this topic increases strongly, and it subsequently stays at about the same level. The increase in the *Financing* loadings during the financial crisis coincides with the spike in loans issued by banks to US commercial and industrial firms, a pattern that is described by Ivashina and Scharfstein (2010) as the "draw now, just in case" phenomenon among US corporates. Accordingly, acting under the concern that banks might restrict their access to their lines of credit facility, financially constrained firms withdrew funds during the financial crisis (Campello et al., 2010). These observations may explain why the MD&As display higher *Financing* loadings after the financial crisis.

Firm managers increasingly discuss the *Litigation* and *Employment* topics in their MD&As until about 2005, after which the prominence of these topics gradually diminishes. The topic loadings of *Regulation/Tax* decline until the dot-com crisis, but increase again before the financial crisis, reaching their peak in 2008. The implementation of various measures to enforce tax compliance in the mid-2000s could be a possible driver of these dynamics. The financial crisis sharply drove down the contents relating to regulation and taxes, and these have remained stable since that time.

There are two peaks in the loading series for *Accounting*. After the dot-com crisis, firms dedicated more space to the accounting and auditing themes in their MD&A until the 2003 SEC regulation became effective. This trend starts in 2001, possibly in the aftermath of the Enron and WorldCom accounting scandals and the ensuing regulatory and litigious environment (Brown and Tucker, 2011). Note also that, in this period, the topic loadings of *Litigation* increase visibly. The increase in the *Accounting* loadings in this period could also be a result of the Sarbanes-Oxley Act, which became effective in 2002. This act requires firm managers to observe certain practices in financial record keeping and reporting.[19] In the period between the 2003 SEC regulation and the financial crisis, we observe a slight slump in the *Accounting* loading. Since the financial crisis, the loadings of this topic have dropped sharply and are now almost negligible.

Two further remarks on the topic loading time series are in order. First, the 2003 SEC regulation appears to have had limited impact on the topic contents, except perhaps on the *Accounting* topic. This is indicated by the significant shift in the topic loading series when the guidance is implemented. Instead, the business cycle, and particularly recessions, have a larger impact on the loadings, to the extent that major disruptions in many topics

[19] It should be noted that, to guarantee the implementation of the Sarbanes-Oxley Act after the accounting frauds, the SEC requires "...written statements, under the oath, from CEOs and CFOs regarding the accuracy of their companies' financial statements" (Donaldson, 2003). Consequently, firm managers might be urged to explain their accounting practices in more detail.

appear during or around these two crises. Second, topics like *Sales/Revenue*, *Profit/Loss*, *Operations*, and *Liquidity* exhibit strong annual seasonality patterns from 2009 onwards. Given that firms usually release their 10-K filings at the same time every year, this cyclicity suggests that the contents of these topics are similar between the years. This could be because firm managers tend to use boilerplate language to describe firm conditions relating to these topics. Several studies, which focus on the language used in the 10-K files in general and the MD&A section in particular, support the hypothesis of boilerplate language (SEC, 2003; Feldman et al., 2010; Brown and Tucker, 2011).[20] This may explain why the contents are so persistent on an annual basis.

5.2 Topic sentiment time series

We now examine the firm-wise topic sentiment scores (see Section 2.4 for their measurement). As with the topic loading time series, we take the average of the topic sentiment scores of all firms that release their MD&A documents in a given month and present the 4-month moving averages. Following Baker and Wurgler (2006, 2007) and Jiang et al. (2019), we standardize the series to obtain zero means and unit variances. Each series, then, captures the aggregate manager sentiment regarding the corresponding topic. The plots are shown in Figure 4.

The sentiment scores of *Sales/Revenue* and *Operations* decline during the two recessions and exhibit an upward trend between these two recessions. These observations suggest that, during the dot-com crisis, firms express pessimistic views about their sales and operations. The subsequent recovery in the sentiment of these two topics coincides with the bullish S&P 500 index between the two recessions. The topic *Cost/Expense* exhibits a dramatic decline during the dot-com crisis, suggesting that firms become very pessimistic about their operational costs during this period. Indeed, there is evidence suggesting that during this time firms try to manage their operational costs, by, for example, offshoring parts of their business activities (Milberg and Winkler, 2010). This evidence suggests that firms faced cost-related managerial challenges during the dot-com crisis, potentially leading to the pessimistic sentiment about this topic. After the sharp drop, the sentiment of *Cost/Expense* gradually recovers until the financial crisis and then stays at the same level until 2018:12. Thus, the consequences of the financial crisis for firm costs appear to differ strongly from those of the dot-com crisis despite the severe recession.

Similarly, during the dot-com crisis, the management sentiment about firm capital expenditure, investment, and divestment activities, which are covered by *Investment*, strongly drops. The *Investment* sentiment recovers and appears to be only slightly affected by

[20] These studies, however, make use of samples from 1994 to 2006. To the best of our knowledge, there is no research that studies boilerplate language in a sample from 2006 onwards.

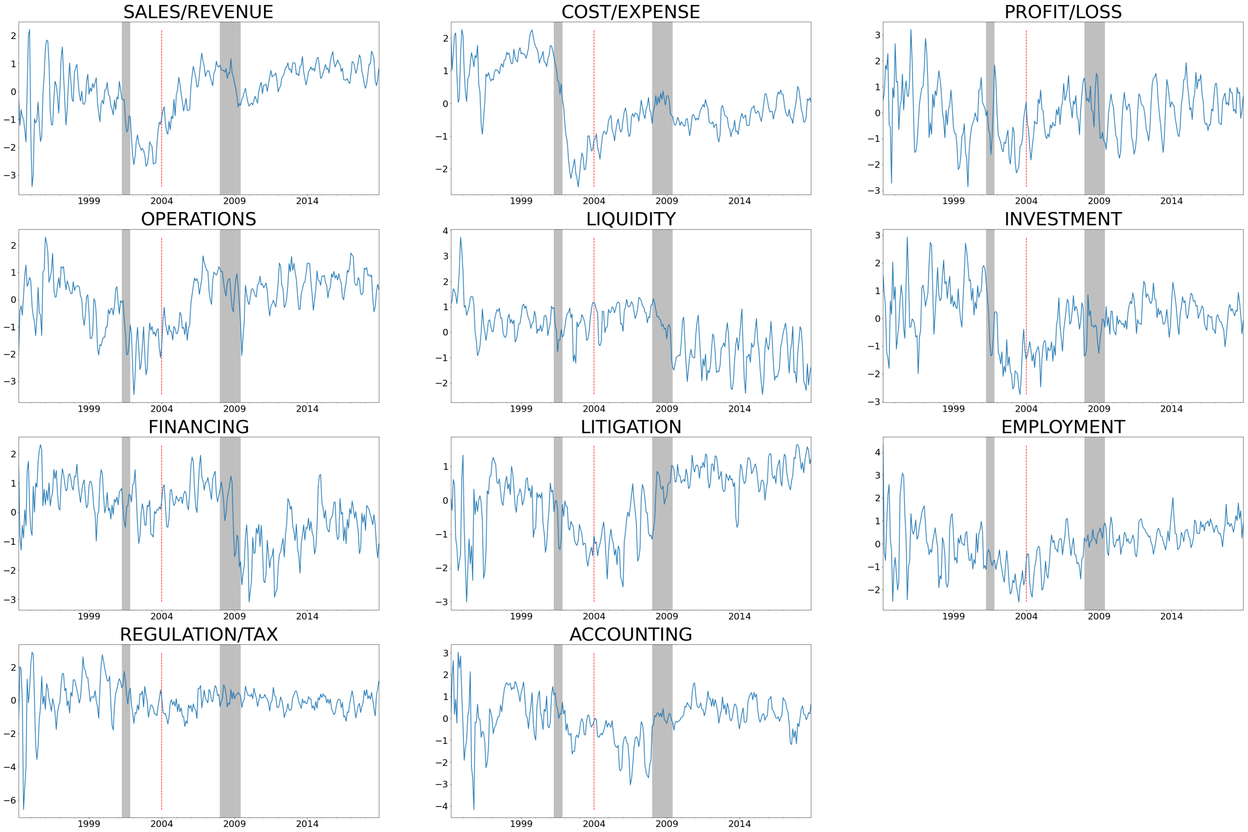

Figure 4: Monthly topic sentiment time series, with the cluster size of 10. These sentiment indices are constructed by aggregating the topic sentiment of all firms that file their 10-K filings in a given month. The time series is smoothed by a moving average over the four previous months, and standardized to have zero means and unit variances. The vertical grey bars indicate the economic recessions provided by the National Bureau of Economic Research (NBER). The dashed vertical red lines indicate the time of 2004:01 when the 2003 SEC regulation became effective. The sample spans the period 1994:01 to 2018:12.

the financial crisis. This is in contrast to *Liquidity* and *Financing*, for which firms express a more negative sentiment during the financial crisis. We conjecture that this is because, during the financial crisis, the access of firms to lines of credit facility was limited (Campello et al., 2011); recall that "credit_facility" is an important topic word in the *Financing* topic, see Table 7. Furthermore, following Campello et al. (2011), lines of credit facility (included in *Financing*) and cash holdings (described by the topic *Liquidity*) are generally negatively related during the financial crisis, implying that firms in financial distress suffer from access to liquidity. With the joint declines in both *Liquidity* and *Financing* sentiment presented in Figure 4, our topic sentiment measure likely reflects this fact.

Investment, *Litigation*, and *Accounting* exhibit a relatively low-sentiment period between the two recessions. Recalling Figure 3, we note that this period also witnesses relatively high *Investment*, *Litigation*, and *Accounting* loadings; this implies that managers not only write more but are also more pessimistic about the investment, litigation, and accounting aspects of their firms between the two recessions. Once again, in part, this phenomenon may be driven by the accounting scandals of 2001 and 2002.

Further observations can be made. First, it is worth noting that the two recessions affect topic sentiment differently. *Sales/Revenue* and *Operations* are the two topics that are hurt by both crises. *Cost/Expense* and *Investment* are only negatively affected by the dot-com crisis, while *Liquidity* and *Financing* are only influenced by the financial crisis. Second, *Profit/Loss, Operations* and *Liquidity* also exhibit a strong annual seasonality from 2004 onward that is similar to the loadings time series (see Figure 3). This suggests that sentimentally-charged contents regarding these topics may also consist of a great deal of generic language.

6 MD&As and firm fundamentals

6.1 Variables of firm fundamentals

To investigate the determinants of topic loadings and topic sentiment, we make use of financial ratios retrieved from the CRSP/Compustat merged database. To be specific, seven fundamental variables for firms are included. Following Li (2010), we discuss the variables according to the type of information they describe.

AT_TURN - Asset Turnover, i.e., the ratio between sales and averaged total assets based on the most recent two quarters. This ratio is used as an indicator that shows how efficiently a firm uses its assets to generate revenue. If a firm operates efficiently, it generates relatively more revenue as a fraction of the firm's assets. By contrast, a low asset turnover ratio implies that the firm is incapable of generating many sales over its assets. The topic *Sales/Revenue* conveys information about firm sales in this study. Therefore, a positive relationship between asset turnover and the *Sales/Revenue* sentiment is expected. Regarding *Sales/Revenue* loadings, it appears plausible that a firm that generates a high asset turnover talks more expansively about sales topics. Thus, we also expect a positive relationship between asset turnover and *Sales/Revenue* loadings.

ROA - Return on Assets, i.e., operational income before depreciation over averaged total assets based on the most recent two quarters. This ratio represents firm performance, and is different from *AT_TURN* in that ROA additionally shows how effectively a firm manages its costs and expenses. Li (2010) empirically shows that this ratio has a positive relationship to the sentiment of forward-looking statements in the MD&A section. However, a negative relationship between the forward-looking sentiment in the MD&A section and this ratio is also possible because earnings are mean-reverting. Furthermore, Li (2008) and Bloomfield (2008) argue that less profitable firms often have longer and more complicated sentences than more profitable firms. This phenomenon is explained by the obfuscation behavior of firm managers in compiling the MD&A. Because our topic loading measure is subject to document-length normalization, long obfuscating statements may contain relatively less useful topic information. Accordingly, we expect a positive relationship between *ROA* and the *Profit/Loss* loadings.

ACC - Accrual Ratio, i.e., accruals over total assets based on the most recent two periods. Firm accruals are documented to have a negative impact on the firm's future performance (Sloan, 1996). As pointed out by Li (2010), a positive relationship is also possible if managers try to obfuscate in the MD&A content about accruals.

CAPITAL_RATIO - Capital Ratio, i.e., total long-term debt over the sum of total long-term debt, common/ordinary equity, and preferred stocks. This ratio is a solvency measure of a firm. If this ratio is high, it suggests that the firm may face a solvency risk. In our set of topics, *Financing* covers debt information. Therefore, the managers are expected to report concerns about the firm's solvency in this topic. We thus expect a negative sentiment when this ratio is high. Besides that, financially distressed firms may have to give up investment projects with positive net-present-value because they have limited opportunities to obtain external financing (Purnanandam, 2008). If this is reflected in the MD&A documents, a positive relationship between this variable and the *Financing* and *Investment* loadings is expected.

ME (*FIRM_SIZE*) - Market Capitalization, i.e., the logarithm of market value equity. This is an indicator of firm size. Li (2010) suggests a negative relationship between firm size and forward-looking sentiment when firms are cautious about political and legal costs, as suggested by the political cost theory (Watts and Zimmerman, 1986). However, under the political power theory (Siegfried, 1972), large firms wield more political influence than small firms and may be able to negotiate their tax burden or drive legislation in their favor (Belz et al., 2019). As a result, large firms possibly have a more positive MD&A sentiment and, specifically, higher *Regulation/Tax* loadings.

B/M - Book-to-Market ratio, i.e., the book-value over market-value equity, and *FIRM_AGE* - Firm age, measured by the number of years since the firm's first appearance on the CRSP database. These two explanatory variables serve as proxies for growth options. In particular, growth (low book-to-market ratio) and young firms tend to face more environmental uncertainties (Smith Jr and Watts, 1992; Anthony and Ramesh, 1992). As a result, these firms tend to be more cautious and less optimistic in their MD&A, which may lead to a negative relationship between the *B/M* ratio, age, and the MD&A sentiment. With regard to topic loadings, Muslu et al. (2015) suggest that growth and young firms, which encounter more uncertainties, need to report more information to reassure investors. However, as our topics span many facets, a variety of relationships between these variables (B/M and $FIRM\_AGE$) and the topic loadings are conceivable. Furthermore, as firms in the early stage of their life cycle may expect future increases in profitability (Warusawitharana, 2018), younger firms may discuss their profits and losses more in the MD&A, and thus a positive relationship between this variable and the *Profit/Loss* topic loading can occur.

The descriptive statistics for the text-related variables, the topic sentiment scores and topic loading scores, and the above fundamental variables are presented in Appendix G. Because the topic loadings are positive, we normalize them to $[0, 1]$.[21] All other explanatory variables are standardized to have zero mean and unit variance for ease of interpretation.

6.2 Regression analysis

The level (measured by the topic loadings) and the sentiment (measured by the topic sentiment) are among the most useful attributes to be unveiled by researchers in text analytics (Li, 2010). Therefore, we now study the determinants of the disclosure level (topic loadings) and sentiment (topic sentiment). To investigate which factors drive the MD&A disclosures, we design a set of regression models in which topic loadings and topic sentiment are regressed on firm fundamentals controlled by cross-sectional and time dummies. More specifically, the model is

$$\begin{aligned} Y_{j,i,t} =& \alpha + \beta_1 AT\_TURN_{i,t} + \beta_2 ROA_{i,t} + \beta_3 ACC_{i,t} + \beta_4 CAPITAL\_RATIO_{i,t} \\ &+ \beta_5 FIRM\_SIZE_{i,t} + \beta_6 BM_{i,t} + \beta_7 FIRM\_AGE_{i,t} \\ &+ Sector\_dummies + Quarter\_dummies + Year\_dummies + u_i + \epsilon_{i,t} \end{aligned} \tag{2}$$

where $Y_{j,i,t} \in \{F_{j,i,t}, s_{j,i,t}\}$ is either the loading or sentiment of topic j in firm i's MD&A published in year t. We estimate the models using Random Effect Ordinary Least Squares, in which u_i is the firm-specific random effect. Because Das and Shroff (2002) argue that the behavior of accounting information may differ across reporting quarters, we also include quarterly dummies.

6.2.1 Determinants of topic loadings

Table 3 presents the regression results for Equation (2) with $Y_{j,i,t} = F_{j,i,t}$, the topic loadings. In general, firms with a good (bad) performance tend to have richer (poorer) content in the MD&A about performance-related topics, as exhibited by the positive coefficients of *AT_TURN* and *ROA* in *Sales/Revenue*, *Cost/Expense*, *Profit/Loss*, and *Operations*. More

[21] For normalization, we use $\frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$, $i = 1, 2, \ldots, 11$.

	Sales/Revenue	**Cost/Expense**	**Profit/Loss**	**Operations**	**Liquidity**	**Investment**	**Financing**	**Litigation**	**Employment**	**Regulation/Tax**	**Accounting**
AT_TURN	0.065***	0.0028***	0.0062***	0.0019***	0.0042***	-0.0052***	0.0025***	0.0002	0.0018*	0.0005	-0.0001
	(3.1734)	(3.3772)	(3.4180)	(2.5474)	(3.1331)	(-3.2352)	(3.4064)	(0.5012)	(1.8143)	(0.7706)	(-0.2570)
ROA	0.0054***	-0.0012	0.0078***	0.0009	0.0014**	0.0015**	-0.0003	-0.0001	-0.0015**	0.0070***	0.0015**
	(3.9562)	(-1.4782)	(6.6472)	(1.4664)	(2.1998)	(2.0163)	(-0.6312)	(-0.5262)	(-2.3181)	(9.8226)	(2.2657)
ACC	-0.0001	-0.0001	-0.0004	-0.0002	-0.0005*	0.0010	-0.0003	-0.0001	-0.0000	0.0001	0.0005
	(-0.3154)	(-0.9895)	(-0.5830)	(-0.8604)	(-1.6998)	(1.3403)	(-0.9619)	(-0.9870)	(-0.1690)	(0.2952)	(1.0706)
CAPITAL_RATIO	-0.0075***	-0.0021***	-0.0025**	0.0001	0.0001	0.0015***	0.0044***	0.0001	-0.0029***	-0.0017***	-0.0016**
	(-6.8037)	(-3.1129)	(-2.5093)	(0.2570)	(0.3038)	(2.6633)	(10.2870)	(0.2531)	(-6.0621)	(-2.6995)	(-2.2270)
FIRM_SIZE	-0.0088***	-0.0044***	-0.0233***	-0.0103***	-0.0078***	-0.0026	-0.0040***	0.0001	-0.0002	0.0047**	-0.0063**
	(-3.6205)	(-2.9161)	(-5.5570)	(-7.4812)	(-4.3674)	(-1.4377)	(-4.7506)	(0.1389)	(-0.1422)	(1.9879)	(-2.2516)
B/M	-0.0003	-0.0002	-0.0002	0.0002**	-0.0003**	0.0001	-0.0002***	-0.0001***	0.0002***	0.0001**	-0.0000
	(-1.4441)	(-1.2712)	(-0.6370)	(2.4239)	(-2.1328)	(0.8261)	(-3.9138)	(-4.7556)	(5.6223)	(1.9958)	(-0.3300)
FIRM_AGE	-0.0026	-0.0084***	0.0084***	0.0047**	0.0015	0.0030***	0.0021	0.0055***	-0.0007	0.0006	0.0009
	(-1.3548)	(-7.9714)	(3.0155)	(2.2884)	(1.0352)	(3.0455)	(1.5129)	(7.1120)	(-1.7349)	(0.3348)	(0.5387)
No. Obs	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619
R^2	0.1219	0.1051	0.1643	0.0834	0.0413	0.0530	0.0846	0.0083	0.0353	0.0408	0.0209

Table 3: This table reports the Random Effect Ordinary Least Squares (RE OLS) results of regression model (2) when $Y_{j,i,t} = F_{j,i,t}$ and the cluster size is 10 (most similar words),

$$\begin{aligned} F_{j,i,t} =& \alpha + \beta_1 AT\_TURN_{i,t} + \beta_2 ROA_{i,t} + \beta_3 ACC_{i,t} + \beta_4 CAPITAL\_RATIO_{i,t} + \beta_5 FIRM\_SIZE_{i,t} \\ &+ \beta_6 BM_{i,t} + \beta_7 FIRM\_AGE_{i,t} + Sector\_dummies + Quarter\_dummies + Year\_dummies + u_i + \epsilon_{i,t} \end{aligned}$$

where $F_{j,i,t}$ is the topic loading of the topic *Sales/Revenue, Cost/Expense, Profit/Loss, Operations, Liquidity, Investment, Financing, Litigation, Employment, Regulation/Tax* and *Accounting*, j, of firm i at time t. The set of independent variables is described in Section 6.1. All independent variables are standardized for the sake of interpretation. The regression coefficients, two-way clustered (by year and firm) t-statistics (in parentheses), and R^2 are reported. The coefficients of the intercepts, sector dummies, and year dummies are not reported to save space. The data sample spans the period 1994:01 to 2018:12. $*$, $**$, and $***$ denote significance at 10%, 5% and 1% respectively.

specifically, firms with a high asset turnover ratio (high *AT_TURN*) tend to provide more details about most topics (sales, expenses, profits, operations, liquidity, and financing), except for their investment plans. Firms with a high asset-turnover ratio successfully transform their assets into current sales. Therefore, firm managers may talk less about their potentially profitable investment plans, thus possibly leading to low *Investment* loadings. As regards the relationship between a firm's profitability and its MD&A contents, our findings suggest that more (less) profitable firms, as reflected in higher (lower) *ROA*, talk more (less) about their current sales and profits, but not about their costs. This result aligns with the results of Li (2010) and meets our expectation that more (less) profitable firms may want to provide more (fewer) details about their success. Not surprisingly, firms with a high financial distress risk (high *CAPITAL_RATIO*) tend to focus more of their MD&A content on the investment (capital-related activities) and financing (debt- and borrowing-related activities) topics than the others.

Firm size (measured by the logarithm of market capitalization) has a negative impact on the loadings of many topics in the MD&A such as the topics of sales, costs, profitability, operations, liquidity, financing, and accounting. This implies that larger firms discuss these topics less than smaller firms. In contrast, big firms focus their MD&A content more on taxation than smaller firms. The higher loadings could be in accordance with the political power theory that suggests that big firms try to express their power to drive taxation in their favor (Siegfried, 1972; Belz et al., 2019). Alternatively, it could be that big firms have a more complex tax structure, which requires more careful documentation.

B/M and *FIRM_AGE*, as expected, have mixed effects on the MD&A topic loadings. Growth firms, which are young or have a low *B/M* ratio, talk less about their operations. This could be explained by the fact that in the early growth phase of its lifetime, a firm may have a smaller product portfolio, which makes the descriptions in its MD&A less complicated. Another possible explanation could be that growth firms possess fewer physical assets and smaller production capacities than value firms, and thus have fewer incentives to talk about their operations. Furthermore, firms with a higher book-to-market ratio have a propensity to write less about their liquidity, debts (financing), and legal activities, but more about their taxes and employment. We also find that young firms provide more information about their costs/expenses but less information about their profitability, operations, investment, and litigation.

As a robustness check, we also present the regression results with the cluster size of 15 of the most similar words in Table 9 in Appendix H. The regression results, in this case, are similar to those of the case presented, both in values and in signs, thereby showing robustness to the choice of the cluster size.

6.2.2 Determinants of the topic sentiment

	TMS	Sales/Revenue	Cost/Expense	Profit/Loss	Operations	Liquidity	Investment	Financing	Litigation	Employment	Regulation/Tax	Accounting
AT_TURN	0.0863***	0.0396***	0.0454***	0.0939***	0.0160*	0.0003	0.0419***	0.0171*	-0.0006	-0.0080	0.0058	0.0178**
	(3.4352)	(4.4575)	(3.8228)	(3.6314)	(1.7723)	(0.0420)	(4.3640)	(1.6979)	(-0.06404)	(-0.6220)	(0.6736)	(2.4499)
ROA	0.1024***	0.0327***	0.0296**	0.0480***	0.0642***	0.0375***	0.0297***	0.0377***	0.0187**	0.0266***	0.0233***	0.0081
	(6.2531)	(2.6774)	(2.0606)	(4.1067)	(4.2385)	(3.8355)	(4.0040)	(4.9077)	(2.0713)	(3.1398)	(3.6541)	(1.1036)
ACC	-0.0161	-0.0381***	-0.0363	-0.0074	-0.0103	-0.0065	-0.0120	-0.0092*	0.0035	-0.0055	-0.0060	-0.0215***
	(-0.8999)	(-2.6646)	(-1.2944)	(-0.9270)	(-1.3373)	(-0.9406)	(-1.4147)	(-1.7324)	(0.4682)	(-1.0684)	(-1.0814)	(-8.4983)
CAPITAL_RATIO	-0.0445***	-0.0022	-0.0495***	-0.0420***	0.0096	-0.0180*	-0.0470***	-0.0447***	0.0189**	-0.0044	-0.0133	-0.0148**
	(-3.7892)	(-0.2082)	(-5.8397)	(-5.1551)	(0.8585)	(-1.8604)	(-4.6930)	(-5.7722)	(2.5141)	(-0.6640)	(-1.4645)	(-2.0388)
FIRM_SIZE	0.1511***	0.0818***	0.0250	0.1441***	0.0817***	0.0410***	-0.0007	0.0569***	0.0175	0.0229**	0.0071	0.0102
	(6.0081)	(6.3364)	(0.9736)	(10.921)	(4.6424)	(5.0426)	(-0.0532)	(5.2882)	(1.1913)	(2.3642)	(0.8370)	(0.8124)
B/M	-0.0041	-0.0104***	-0.0067**	0.0012	0.0025***	0.0008	0.0006	0.0011	0.0020***	0.0063***	0.0039***	-0.0020***
	(-1.2977)	(-6.1915)	(-2.0035)	(0.7548)	(3.5555)	(1.1629)	(0.5176)	(1.1611)	(3.0010)	(13.524)	(9.5997)	(-4.2500)
FIRM_AGE	-0.0433**	-0.0045**	-0.569***	0.0108	-0.0203*	0.0225**	-0.0494***	0.0201	-0.0434***	-0.0033	-0.0128	-0.0320***
	(-2.1323)	(-2.1739)	(-3.0943)	(0.4943)	(-1.7312)	(2.2570)	(-3.6125)	(1.5776)	(-2.9482)	(-0.3034)	(-1.4090)	(-3.2706)
No. Obs	49169	49169	49169	49169	49169	49169	49169	49169	49169	49169	49169	49169
R^2	0.0290	0.0075	0.0069	0.0195	0.0086	0.0030	0.0038	0.0047	0.0010	0.0010	0.0007	0.0018

Table 4: This table reports the Random Effect Ordinary Least Squares (RE OLS) results of regression model (2) when $Y_{j,i,t} = s_{j,i,t}$ and the cluster size is 10 (most similar words),

$$\begin{aligned} s_{j,i,t} =& \alpha + \beta_1 AT\_TURN_{i,t} + \beta_2 ROA_{i,t} + \beta_3 ACC_{i,t} + \beta_4 CAPITAL\_RATIO_{i,t} + \beta_5 FIRM\_SIZE_{i,t} \\ &+ \beta_6 BM_{i,t} + \beta_7 FIRM\_AGE_{i,t} + Sector\_dummies + Quarter\_dummies + Year\_dummies + u_i + \epsilon_{i,t} \end{aligned}$$

where $s_{j,i,t}$ is the total MD&A sentiment (*TMS*) and the sentiment of the topics *Sales/Revenue, Cost/Expense, Profit/Loss, Operations, Liquidity, Investment, Financing, Litigation, Employment, Regulation/Tax* and *Accounting*, j, of firm i at time t. The set of independent variables is described in Section 6.1. All independent variables are standardized for the sake of interpretation. The regression coefficients, two-way clustered (by year and firm) t-statistics (in parentheses), and R^2 are reported. The coefficients of the intercepts and year dummies are not reported to reserve more space. The data sample spans the period 1994:01 to 2018:12. $*$, $**$, and $***$ denote significance at 10%, 5% and 1% respectively.

We now examine the determinants of the total MD&A sentiment and topic-specific sentiment. Table 4 reports the results of Equation (2) when $Y_{j,i,t} = s_{j,i,t}$, that is, the regression of the MD&A sentiment scores on the corresponding firm fundamentals. The entries under the heading *TMS* within the table refer to the regression outcome of the total MD&A sentiment, which is derived from the entire document, disregarding specific topics. By using *TMS*, we aim to show that, with our approach, more detailed insights can be uncovered, in the sense that some variables may have a statistically significant impact not on the sentiment of the entire MD&A but on specific topics instead. To the best of our knowledge, our work is the first paper to examine the impact of firm fundamentals and MD&A sentiment at the topic level.

Total MD&A sentiment regression

Starting with the *TMS* regression, we see that the estimate for *AT_TURN* is significant at the 1% level; this shows that firms efficiently generating their sales from their assets tend to express positive sentiment in the MD&A section. Firm performance (*ROA*) is also positively related to management sentiment (the estimated coefficient for *ROA* is significant at the 1% level). These results confirm the hypotheses documented by Li (2010) about the positive relationship between sentiment and *AT_TURN* and *ROA*. Remarkably, *ACC* has no significant impact on *TMS*, yet negatively affects the sentiment of *Sales/Revenue* and *Accounting*; we will return to this observation in the next subsection.

A high capital ratio implies a high risk of financial distress. This is reflected in the MD&A by the significantly negative coefficient for *CAPITAL_RATIO*. Furthermore, we discover that *FIRM_SIZE* and *FIRM_AGE* have different effects on *TMS*. Bigger (smaller) firms tend to be more (less) optimistic in their MD&A. As regards *FIRM_AGE*, we find that younger (older) firms are more (less) positive in their MD&A. This may be because young firms are incentivized to attract and reassure their investors and therefore tend to talk more positively in their MD&A (Muslu et al., 2015). We do not find a significant correlation between *B/M* and *TMS*. However, like *ACC*, *B/M* is found to be significant in the topic-specific regressions, which emphasizes the advantage of our method in studying the MD&A at the topic level. We discuss the effects of *B/M* on topic-level sentiment in detail in the next paragraphs.

Topic-specific sentiment regressions

Looking into the details of the regression results for the individual topics, we can observe that *Sales/Revenue, Cost/Expense, Profit/Loss*, and *Financing* are the top four topics that are most closely related to the firm fundamental variables, based on the R^2 and the number of significant coefficients. Firms with a good asset turnover ratio show their positive sentiment regarding not only sales/revenue topics in the MD&A, but also expenses, profitability, operations, liquidity, and investment topics. We find a strongly significant relationship between the earnings ratio (*ROA*) and the *Profit/Loss* topic sentiment. Not surprisingly, firms with a high earnings ratio are also optimistic about their profitability. In conjunction with the topic loadings regression results for *Sales/Revenue* and *Profit/Loss* on *ROA* (see Table 3), we find that firms with good (bad) performance talk more (less), both in level and sentiment, about performance-related topics. This finding shows that managers correctly reflect the firm's performance conditions in the MD&A even when they are struggling to make revenue and profits.

ACC affects the sentiment of *Sales/Revenue* and *Accounting* negatively. Thus firms with a higher accrual ratio tend to be pessimistic about their sales and accounting topics. As reported earlier, *ACC* has no impact on document-level sentiment *TMS*. This result underlines the importance of investigating the sentiment disclosure at the topic level. Additionally, the negative correlation between *ACC* and the accounting topic sentiment suggests that firm managers are likely to understand the negative relationship between accruals and earnings (Sloan, 1996) and to report accruals truthfully in the MD&A documents. This result adds to Li (2010) who finds that forward-looking statements deliver the pessimism of firm managers when a firm's accrual-on-asset ratio is high. We discover here that the sentiment of the *Sales/Revenue* and *Accounting* topics carries this information as well.

We also find a negative and highly significant relationship between the capital ratio and *Financing* and *Investment*. Recalling that *Financing* covers the firm's credit and debt situation, we see that firm managers truthfully describe the financial distress conditions in the MD&A. Combining this finding with the topic loadings regression results of *Financing* on *CAPITAL_RATIO* in Section 6.2.1 unveils deeper insights about firm behavior under financial pressure: Firms in financial distress not only discuss their financing and investing conditions more (recall the significantly positive coefficients in the *Investment* and *Financing* columns of Table 3), but also express their pessimism about these matters. Put differently, firms with a high risk of financial distress tend to provide more information but with a pessimistic sentiment. This suggests that firms tend to correctly describe their performance and financial conditions to the public.

Firm size and age are also important determinants of topic sentiment in the MD&A. While *FIRM_SIZE* has positive relationships with the sentiment of *Sales/Revenue*, *Profit/Loss*, *Operations*, *Liquidity*, *Financing*, and *Employment*, *FIRM_AGE* is negatively related to *Sales/Revenue*, *Cost/Expense*, *Investment*, *Litigation*, and *Accounting*. Interestingly, these two dimensions of firm characteristics impact the MD&A topics differently. Big firms tend to be optimistic about their operation-related topics such as revenue-making activities, profitability, and production. Besides that, big firms are also found to be positive about their liquidity and financing activities. We also observe that big firms tend to be more optimistic about their human resources. This could be explained by their more formalized recruitment processes (Barber et al., 1999) and their need for a larger labor force to execute complex business operations. With negative coefficients for many topics, young firms, as reflected by low *FIRM_AGE*, are more optimistic in their MD&As. This is consistent with the extant literature, which hypothesizes that young firms are likely to encounter economic uncertainties and hence have incentives to attract and reassure their investors (Anthony and Ramesh, 1992; Smith Jr and Watts, 1992; Muslu et al., 2015).

The book-to-market ratio *B/M* has mixed effects on the topic sentiment. We find that value firms, which have a high book-to-market ratio, tend to be pessimistic about their revenues and expenses. As discussed by Novy-Marx (2011), value firms possibly have difficulties in generating revenues and managing expenses, a fact that might be reflected in their MD&A. Merz and Yashiv (2007) document a relationship between labor and the market value of firms. In particular, they find that firms, in order to maximize their market value, need to decide on the optimal number of workers to recruit and the optimal investment in physical assets. With a significantly positive coefficient between *B/M* and the *Employment* sentiment, we find that value firms tend to be positive about their labor force. Besides that, we also figure that value firms are positive about the regulation/tax perspective.

To sum up, the regression results for *B/M* and *ACC* underscore the advantages of a topic-level approach over its document-level counterpart, which is current practice. Both variables are found to be insignificant in the *TMS* regression. Our refinement using topic-level sentiment allows us to expose further economic relations that one would be unlikely to learn from the current document-level approaches.

7 Conclusion

This paper reveals the topics contained in the MD&A section of 10-K filings from 1994:01 to 2018:12. We proceed in two steps. First, we retrieve the topic words in a data-driven manner. In the second step, we construct two topic-specific textual indicators: (i) topic loadings and (ii) topic sentiment.

In contrast to much of the literature, we rely on a Word2Vec concept as our textual model. As a result, our topics are immediately telling and intelligible. Moreover, our model successfully categorizes words that seem similar at face value but actually belong to different topics. The topics formed by the model are found to be multimodal, meaning that one topic could feature multiple aspects. This finding is plausible given that there are commonly several facets that reflect a topic in a corporate finance context.

The time-series of the topic loadings and topic sentiment estimated by our model uncover substantial time variation in the MD&A content. Topics differ from each other in terms of both prevalence and sentiment, and are affected heterogeneously by different economic episodes. Our regression analyses show strong correlations between topics and certain firm characteristics. Together with the variation of the topic loading series, this result emphasizes the advantages of a topic-level approach such as ours over methods that aggregate quantitative measures of text at a total document level.

We expect our topic word lists to be useful for further textual analysis of MD&A documents. A potential limitation, however, is that the word lists are conceptualized as fixed. This may be perceived as a critical assumption because the composition of words for a given MD&A topic may itself be subject to change over time. Such alterations might be driven by linguistic evolution, regulation, and language standardization, or by technological progress and cultural change. Future research may want to investigate this important aspect.

References

Anthony, J. H. and Ramesh, K. (1992). Association between accounting performance measures and stock prices: A test of the life cycle hypothesis, *Journal of Accounting and Economics* **15**(2-3): 203–227.

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

Arora, S., Ge, R. and Moitra, A. (2012). Learning topic models–going beyond SVD, *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, IEEE, pp. 1–10.

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns, *The Journal of Finance* **61**(4): 1645–1680.

Baker, M. and Wurgler, J. (2007). Investor sentiment in the stock market, *Journal of Economic Perspectives* **21**(2): 129–152.

Bao, Y. and Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures, *Management Science* **60**(6): 1371–1391.

Barber, A. E., Wesson, M. J., Roberson, Q. M. and Taylor, M. S. (1999). A tale of two job markets: Organizational size and its effects on hiring practices and job search behavior, *Personnel Psychology* **52**(4): 841–868.

Bellstam, G., Bhagat, S. and Cookson, J. A. (2021). A text-based analysis of corporate innovation, *Management Science* **67**(7): 4004–4031.

Belz, T., von Hagen, D. and Steffens, C. (2019). Taxes and firm size: Political cost or political power?, *Journal of Accounting Literature* **42**: 1–28.

Bird, S., Klein, E. and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*, " O'Reilly Media, Inc.".

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation, *The Journal of Machine Learning Research* **3**(Jan): 993–1022.

Bloomfield, R. (2008). Discussion of "Annual report readability, current earnings, and earnings persistence", *Journal of Accounting and Economics* **45**(2-3): 248–252.

Blum, A., Hopcroft, J. and Kannan, R. (2020). *Foundations of Data Science*, Cambridge University Press.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction, *Proceedings of GSCL* **30**: 31–40.

Brown, N. C., Crowley, R. M. and Elliott, W. B. (2020). What are you saying? Using topic to detect financial misreporting, *Journal of Accounting Research* **58**(1): 237–291.

Brown, S. V. and Tucker, J. W. (2011). Large-sample evidence on firms' year-over-year MD&A modifications, *Journal of Accounting Research* **49**(2): 309–346.

Campello, M., Giambona, E., Graham, J. R. and Harvey, C. R. (2011). Liquidity management and corporate investment during a financial crisis, *The Review of Financial Studies* **24**(6): 1944–1979.

Campello, M., Graham, J. R. and Harvey, C. R. (2010). The real effects of financial constraints: Evidence from a financial crisis, *Journal of Financial Economics* **97**(3): 470–487.

Caserio, C., Panaro, D. and Trucco, S. (2019). Management discussion and analysis: a tone analysis on us financial listed companies, *Management Decision* **58**(3): 510–525.

Chen, C. Y.-H., Fengler, M. R., Härdle, W. K. and Liu, Y. (2022). Media-expressed tone, option characteristics, and stock return predictability, *Journal of Economic Dynamics and Control* **134**. Forthcoming.

Choo, J., Lee, C., Reddy, C. K. and Park, H. (2013). Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization, *IEEE transactions on visualization and computer graphics* **19**(12): 1992–2001.

Choo, J., Lee, C., Reddy, C. K. and Park, H. (2015). Weakly supervised nonnegative matrix factorization for user-driven clustering, *Data mining and knowledge discovery* **29**: 1598–1621.

Cohen, L., Malloy, C. and Nguyen, Q. (2020). Lazy prices, *The Journal of Finance* **75**(3): 1371–1415.

Cong, L. W., Liang, T. and Zhang, X. (2019). Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information, *Interpretable, and Data-driven Approach to Analyzing Unstructured Information (September 1, 2019)* .

Das, S. and Shroff, P. K. (2002). Fourth quarter reversals in earnings changes and earnings management, *Available at SSRN 308441* .

Dieng, A. B., Ruiz, F. J. R. and Blei, D. M. (2020). Topic modeling in embedding spaces, *Transactions of the Association for Computational Linguistics* **8**: 439–453.
URL: *https://aclanthology.org/2020.tacl-1.29*

Donaldson, W. H. (2003). Testimony concerning implementation of the Sarbanes-Oxley Act of 2002, *Before the Senate Committee on Banking, Housing and Urban Affairs* .

Dougal, C., Engelberg, J., Garcia, D. and Parsons, C. A. (2012). Journalists and the stock market, *The Review of Financial Studies* **25**(3): 639–679.

Duan, J.-C. and Yao, X. (2022). Media sentiments for enhanced credit risk assessment.

Dumais, S. T. (2004). Latent Semantic Analysis, *Annual Review of Information Science and Technology* **38**(1): 188–230.

Dyer, T., Lang, M. and Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation, *Journal of Accounting and Economics* **64**(2-3): 221–245.

Engelberg, J. (2008). Costly information processing: Evidence from earnings announcements, *AFA 2009 San Francisco Meetings Paper*.

Eshima, S., Imai, K. and Sasaki, T. (2020). Keyword assisted topic models, *arXiv preprint arXiv:2004.05964* .

FASAB (2022). *Handbook of Federal Accounting Standards and Other Pronouncements, as Amended*, 21 edn, Washington, DC.

Feldman, R., Govindaraj, S., Livnat, J. and Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals, *Review of Accounting Studies* **15**(4): 915–953.

Garcia, D. (2013). Sentiment during recessions, *The Journal of Finance* **68**(3): 1267–1300.

Griffin, P. A. (2003). Got information? Investor response to Form 10-K and Form 10-Q EDGAR filings, *Review of Accounting Studies* **8**(4): 433–460.

Haghighi, A. and Klein, D. (2006). Prototype-driven learning for sequence models, *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pp. 320–327.

Henry, E. (2008). Are investors influenced by how earnings press releases are written?, *The Journal of Business Communication (1973)* **45**(4): 363–407.

Hoffman, M., Bach, F. and Blei, D. (2010). Online learning for Latent Dirichlet Allocation, *Advances in Neural Information Processing Systems* **23**.

Ivashina, V. and Scharfstein, D. (2010). Bank lending during the financial crisis of 2008, *Journal of Financial Economics* **97**(3): 319–338.

Jagarlamudi, J., Daumé III, H. and Udupa, R. (2012). Incorporating lexical priors into topic models, *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 204–213.

Jegadeesh, N. and Wu, D. (2017). Deciphering fedspeak: The information content of fomc meetings, *Monetary Economics: Central Banks–Policies & Impacts eJournal* .

Jiang, F., Lee, J., Martin, X. and Zhou, G. (2019). Manager sentiment and stock returns, *Journal of Financial Economics* **132**(1): 126–149.

Kiela, D., Hill, F., Clark, S. et al. (2015). Specializing word embeddings for similarity or relatedness., *EMNLP*, Citeseer, pp. 2044–2048.

Kruszewski, G. and Baroni, M. (2015). So similar and yet incompatible: Toward the automated identification of semantically compatible words, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 964–969.

Lau, J. H. and Baldwin, T. (2016). The sensitivity of topic coherence evaluation to topic cardinality, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–487.

Lau, J. H., Newman, D. and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 530–539.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence, *Journal of Accounting and Economics* **45**(2-3): 221–247.

Li, F. (2010). The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach, *Journal of Accounting Research* **48**(5): 1049–1102.

Li, K., Mai, F., Shen, R. and Yan, X. (2021). Measuring corporate culture using machine learning, *The Review of Financial Studies* **34**(7): 3265–3315.

Ljungqvist, A. and Wilhelm Jr, W. J. (2003). IPO pricing in the dot-com bubble, *The Journal of Finance* **58**(2): 723–752.

Lochter, J. V., Silva, R. M. and Almeida, T. A. (2022). Multi-level out-of-vocabulary words handling approach, *Knowledge-Based Systems* **251**: 108911.

Loughran, T. and McDonald, B. (2011a). Barron's red flags: Do they actually work?, *Journal of Behavioral Finance* **12**(2): 90–97.

Loughran, T. and McDonald, B. (2011b). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *The Journal of Finance* **66**(1): 35–65.

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey, *The Journal of Accounting Research* **54**(4): 1187–1230.

Lund, J., Cook, C., Seppi, K. and Boyd-Graber, J. (2017). Tandem anchoring: A multiword anchor approach for interactive topic modeling, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 896–905.

Mayew, W. J., Sethuraman, M. and Venkatachalam, M. (2015). MD&A disclosure and the firm's ability to continue as a going concern, *The Accounting Review* **90**(4): 1621–1651.

Merz, M. and Yashiv, E. (2007). Labor and the market value of the firm, *American Economic Review* **97**(4): 1419–1431.

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

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

Milberg, W. and Winkler, D. E. (2010). Trade crisis and recovery: Restructuring of global value chains, *World Bank Policy Research Working Paper* (5294).

Mukherjee, P., Badr, Y., Doppalapudi, S., Srinivasan, S. M., Sangwan, R. S. and Sharma, R. (2021). Effect of negation in sentences on sentiment analysis and polarity detection, *Procedia Computer Science* **185**: 370–379.

Muslu, V., Radhakrishnan, S., Subramanyam, K. and Lim, D. (2015). Forward-looking MD&A disclosures and the information environment, *Management Science* **61**(5): 931–948.

Newman, D., Lau, J. H., Grieser, K. and Baldwin, T. (2010). Automatic evaluation of topic coherence, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 100–108.

Novy-Marx, R. (2011). Operating leverage, *Review of Finance* **15**(1): 103–134.

Park, H., Park, T. and Lee, Y.-S. (2019). Partially collapsed Gibbs sampling for Latent Dirichlet Allocation, *Expert Systems with Applications* **131**: 208–218.

Pillai, S. U., Suel, T. and Cha, S. (2005). The Perron-Frobenius theorem: Some of its applications, *IEEE Signal Processing Magazine* **22**(2): 62–75.

Pröllochs, N., Feuerriegel, S. and Neumann, D. (2015). Enhancing sentiment analysis of financial news by detecting negation scopes, *2015 48th Hawaii International Conference on System Sciences*, IEEE, pp. 959–968.

Purnanandam, A. (2008). Financial distress and corporate risk management: Theory and evidence, *Journal of Financial Economics* **87**(3): 706–739.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora, *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50.

Rekabsaz, N., Lupu, M. and Hanbury, A. (2017). Exploration of a threshold for similarity based on uncertainty in word embedding, *European Conference on Information Retrieval*, Springer, pp. 396–409.

Röder, M., Both, A. and Hinneburg, A. (2015). Exploring the space of topic coherence measures, *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 399–408.

SEC (2003). Interpretation: Commission guidance regarding management's discussion and analysis of financial condition and results of operations, *Securities Act Release* (33-8350): 34–48960.

Siegfried, J. J. (1972). *The relationship between economic structure and the effect of political influence: Empirical evidence from the federal corporation income tax program*, The University of Wisconsin-Madison.

Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings?, *Accounting Review* pp. 289–315.

Smith Jr, C. W. and Watts, R. L. (1992). The investment opportunity set and corporate financing, dividend, and compensation policies, *Journal of Financial Economics* **32**(3): 263–292.

Tavcar, L. R. (1998). Make the MD&A more readable, *The CPA Journal* **68**(1): 10.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market, *The Journal of Finance* **62**(3): 1139–1168.

Tetlock, P. C., Saar-Tsechansky, M. and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals, *The Journal of Finance* **63**(3): 1437–1467.

Thelen, M. and Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts, *Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002)*, pp. 214–221.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE., *Journal of Machine Learning Research* **9**(11).

Wallach, H. M., Mimno, D. M. and McCallum, A. (2009). Rethinking LDA: Why priors matter, *Advances in Neural Information Processing Systems*, pp. 1973–1981.

Warusawitharana, M. (2018). Profitability and the lifecycle of firms, *The BE Journal of Macroeconomics* **18**(2).

Watts, R. L. and Zimmerman, J. L. (1986). *Positive accounting theory*, Prentice-Hall Inc.

A Notation

Notation	Description
D	The set of all MD&A documents in the corpus
S	The set of all segments in MD&A documents in the corpus
W	The vocabulary built from the MD&A corpus of documents
C	The total word count in the MD&A corpus of documents
T_j	The set of words representing topic j
T	$T = \{T_1, ..., T_{11}\}$, the collection of topics
N_j	The document-term matrix of topic j, with dimension $\|D\| \times \|T_j\|$
$\tilde{N}_j$	The normalized document-term matrix of topic j, with dimension $\|D\| \times \|T_j\|$
F_j	A vector of loadings of topic j of the MD&A documents
$\|\mathcal{S}\|$	Cardinality of the set $\mathcal{S}$

Table 5: Notation used in the paper.

B Construction of anchor word lists

This paper uses the word lists introduced by Li (2010) as the anchor word lists. These word lists, however, are not ready for training the Word2Vec model as they contain undetected phrases which have the potential to cause an out-of-vocabulary error (Lochter et al., 2022).[22] To overcome this issue, we first learn the potential phrases in the suggested word lists using the phrase-learning model of Mikolov et al. (2013b).

In the model, the ability of the phrase-learning model to detect plausible phrases is governed by a threshold s. The higher (lower) this threshold is, the fewer (more) potential phrases are detected by the phrase-learning model. To calibrate this threshold, we proceed as follows. A set of thresholds is proposed as $s \in \{0.001, 0.01, 0.05, 1, 2, 5\}$. For each value of the threshold, we train the phrase-learning model using the MD&A corpus. After that, the suggested anchor words borrowed from Li (2010) are tokenized.[23] We then evaluate the phrase-learning model by reconstructing the tokenized anchor word lists. More

[22] As this section describes how the phrase-learning model is used, in this particular section, we distinguish single *words* from *phrases*, which are compositions of two or more single words.

[23] Tokenization is the action that separates a text into a list of single words (tokens). For example, the word list for the topic of firm costs is tokenized as {cost, expense, reserve, for, contingent, liability, asset, impairment, goodwill, impairment}.

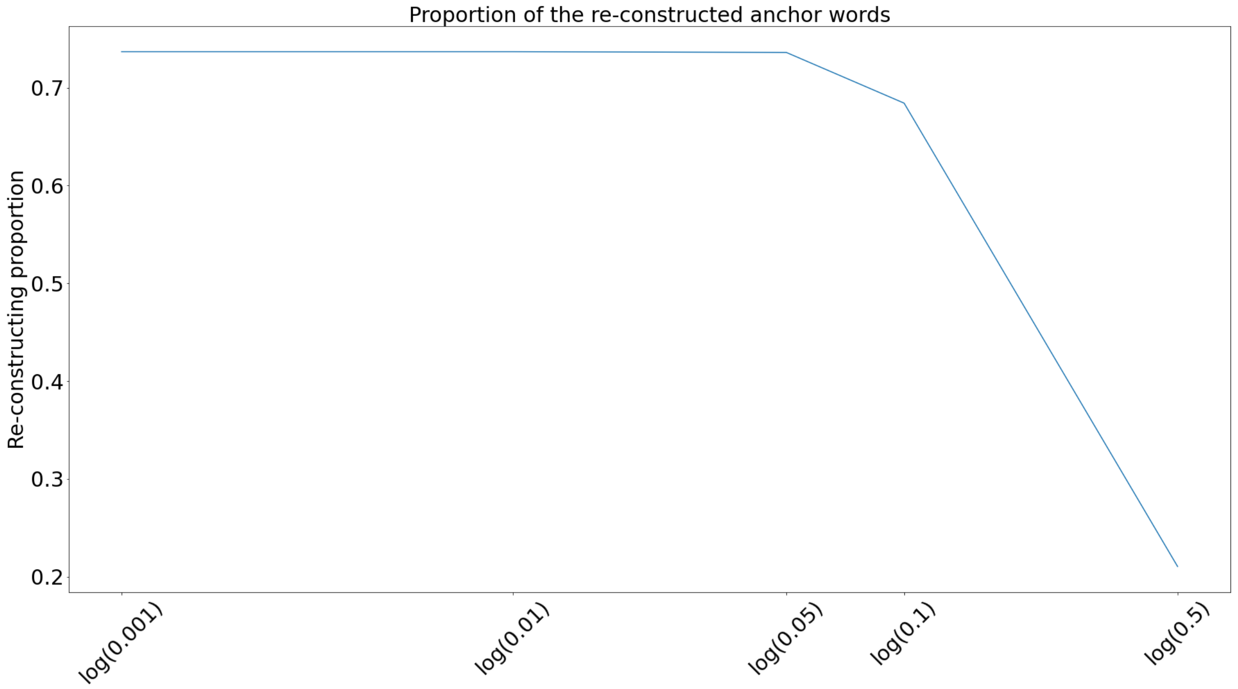

Figure 5: The proportion of the reconstructed anchor words with different values of the threshold s, with $s \in \{0.001, 0.01, 0.05, 1, 2, 5\}$ (in the log-scale on the x-axis). The corresponding proportions of reconstructed words are $\{0.7368, 0.7368, 0.7361, 0.6842, 0.2105\}$, which are given on the y-axis.

specifically, the threshold that reproduces most words and phrases from the tokenized anchor words is considered to be the optimal threshold.

Following Mikolov et al. (2013b), we run the phrase-learning model on our MD&A corpus twice, with the same threshold for both rounds, to detect trigrams and 4-grams in the corpus. Figure 5 reports the proportion of reconstructed words from the initial anchor word lists. The proportion of reconstructed words is the ratio between the number of words and phrases detected by the phrase-learning model and the number of initial anchor words. The proportions of reconstructed words when $s = 0.001$ and $s = 0.01$ are identical. However, we choose $s = 0.01$ over $s = 0.001$ as the optimal value in an attempt to reduce noise.

It is worth noting that there are several phrases in the initial anchor word lists given in Table 1 which the model fails to detect. For these phrases, we decide on a case-by-case basis. For example, the model is unable to detect the phrase “market position”, and the two words “market” and “position” alone do not fully deliver the meaning. Consequently, we decide to exclude both of these individual words from the anchor word list. On the other hand, in “reserve for contingent liability”, the model merely detects the phrase “contingent liability”. In this case, the phrase “contingent liability” still delivers the context of the topic about firm’s costs and expenses. Therefore, we keep that phrase in the anchor word list for topic 2. Phrases treated in a similar way are “new contract” (keep "contract"), "working capital condition" (keep "working capital"), "general capital expenditure" (keep "capital expenditure"), "employee relation" (keep "employee"), "union relation" (keep "union"), and "accounting method" (keep "accounting").

The results of this step are (i) the reconstructed anchor word lists given in Table 6; and (ii) the corpus containing potential words and phrases learned by the phrase-learning model.

C Text processing

Before passing the documents into the phrase-learning and Word2Vec models, we implement two main textual normalization steps: (i) replacing contractions (i.e., converting "don't" into "do not", etc.); (ii) removing noise (i.e., single letters, numbers, special characters, punctuation marks,[24] multiple whitespaces, and breaklines).

We further discard stopwords, which are words that appear very often in the text but have negligible meaning. We use the stopword lists provided by Loughran-McDonald (the LM stopwords) for this process. These lists differ from the stopword list provided by the *nltk* Python library (Bird et al., 2009) in that the LM stopword lists are specifically designed for financial applications and are more detailed. It should be noted that in the LM stopword lists of names, "Sale" and "Cash" are considered as stopwords. However, they are meaningful in the business and financial context, so we keep them in our vocabulary. We further discard the words "Inc.", "Co.", "Ltd.", "Mr.", "Mrs.", and "Ms." from the vocabulary. Finally, the texts are lemmatized to remove the inflectional endings of words. We do not stem the words, to preserve the meanings of words within a word family.

After normalization, the MD&A documents are used to train the phrase-learning model and the Word2Vec model. The documents are split into sentences before being input into the two models in order to prevent the information from one sentence spilling over to nearby sentences when training the models.[25] For training the phrase-learning and Word2Vec models, we discard words that appear in fewer than 15 documents. Conse-

[24] While punctuations like comma (,), colon (:), semi-colon (;), etc. are removed, periods (.) are kept because they serve as sentence delimiters.

[25] Consider the following paragraph, which includes two sentences, in the 10-K filing of SUMMIT SECURITIES INC (CIK number is 0000868016) in 1994, "Management believes that cash flow from operating activities and financing activities will be sufficient for the Company to conduct its business and meet its anticipated obligations as they mature during fiscal 1994. The Company has not defaulted on any of its obligations since its founding in 1990". If the entire paragraph, instead of separated sentences, is fed into the phrase-learning (or Word2Vec) model instead of separated sentences, the word "fiscal" at the end of the first sentence will be considered to be close to the word "company" at the beginning of the second sentence, resulting in into improper handling. Because of this, we split the documents into sentences before training the phrase-learning and Word2Vec models.

Topic 1: Sales/Revenue	Topic 2: Cost	Topic 3: Profit	Topic 4: Operations	Topic 5: Liquidity	Topic 6: Investment	Topic 7: Financing	Topic 8: Litigation	Topic 9: Employment	Topic 10: Regulation	Topic 11: Accounting
sale	cost	profit	operation	liquidity	investment	financing	litigation	employee relation	regulation	accounting method
revenue	expense	income	production	interest coverage	capital expenditure	debt	lawsuit	retention	law	accounting
market condition	contingent liability	margin		cash	disvestiture	equity		hiring	income tax	auditing
consumer demand	asset impairment			working capital	discontinued operation	dividend		union	government relation	internal control
competition	goodwill impairment					repurchase				
pricing contract										

Table 6: List of anchor words produced by the phrase-learning model with the optimal threshold of 0.01. The process for reproducing these lists of anchor words is described in detail in Section 2.1 and Appendix B. The original anchor words of Li (2010) are presented in Table 1.

quently, the vocabulary built from our corpus has $1,884,140$ words. Following Mikolov et al. (2013b), the negative sampling parameter is 15 and the context window is 5.

As stressed by Mukherjee et al. (2021), we handle negations carefully before proceeding to the sentiment analysis. To this end, words in a document that are contained in the LM sentiment dictionary are searched. We are only concerned with these words because they are considered for sentiment estimation. We further determine whether, within a certain window, there are negation terms appearing around these sentimentally charged words. If sentimentally charged words appear together with a negation term within the considered window, the "not_" prefix is added to the word. For example, consider the following sentence in the 10-K filing of SUMMIT SECURITIES INC (CIK number is 0000868016) in 1994, "The Company has not **defaulted** on any of its obligations since its founding in 1990". The word "defaulted" appears in the Loughran-McDonald dictionary as a negative word. Because it is proceeded by the negation "not", we record it as a positive statement and the new term **not_defaulted** is added to the positive word list of the LM dictionary. The negation terms we consider are "not", "no", "none", "neither", "nor", and "never". Following Pröllochs et al. (2015), the length of the window around a sentimentally charged word is five on either side.

D Topic coherence-coverage trade-off

The data-driven optimization of the cluster size involves two concepts, topic coherence[26] and topic coverage. The former relates to the extent to which words within a topic are close to each other so that humans can easily identify the topic by its word list. Because classical machine-learning based topic models do not provide guarantees of topic interpretability, many studies rely on topic coherence measures as tools for model selection (Newman et al., 2010; Lau et al., 2014; Röder et al., 2015). According to Newman et al. (2010), *Pointwise Mutual Information* (PMI) is the best-performing coherence measure in the sense that this measure has the closest Spearman correlation to human-judged measurements. In our work, we use the Normalized-PMI (NPMI) with mean aggregation instead of PMI to obtain a quantity of a scale similar to the second criterion, the topic coverage. The PMI and NPMI scores of two words w_m and w_n are computed as,

$$\begin{aligned}\text{PMI}(w_m, w_n) &= \log\frac{p(w_m, w_n)}{p(w_m)p(w_n)}\\ \text{NPMI}(w_m, w_n) &= \frac{\text{PMI}(w_m, w_n)}{-\log p(w_m, w_n)}\,.\end{aligned}$$

[26] In philosophy, one of the theoretical definitions of *coherence* is that "A set of statements or facts is said to be coherent if they support each other" (Röder et al., 2015, p.1).

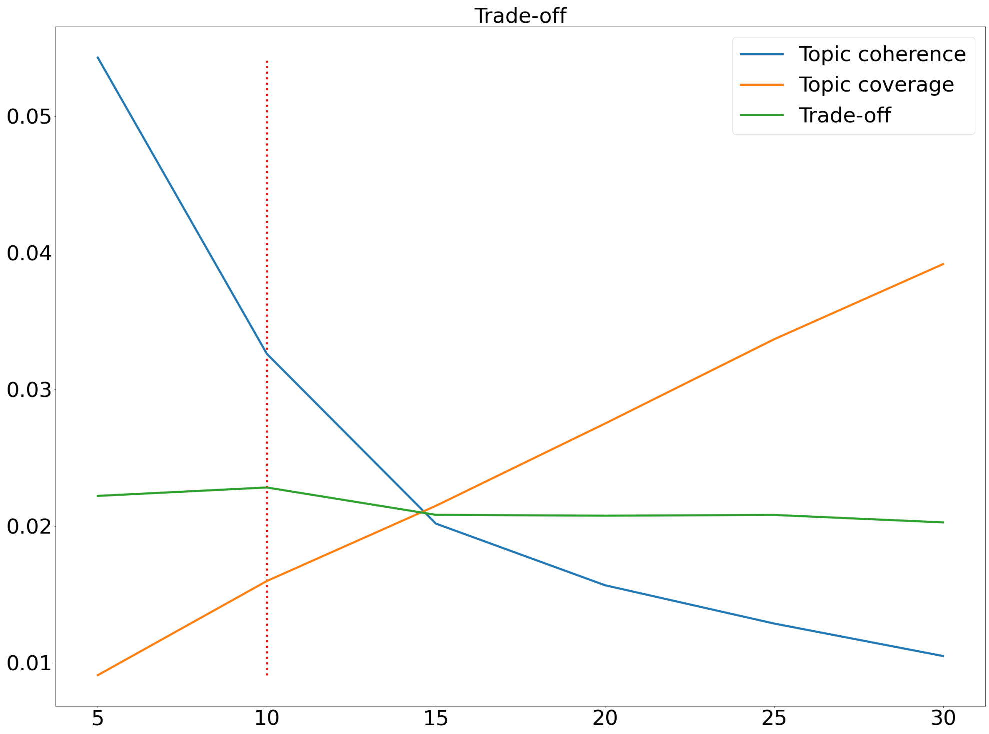

Figure 6: Topic coverage, topic coherence, and the trade-off quantity, G_k, for different cluster sizes $k \in \{5, 10, 15, 20, 25, 30\}$. The blue and orange lines correspond to the arithmetic averages of topic coverage and topic coherence overall topics. G_k is given by the green line. The maximum value of G_k is 0.0228, achieved at $k = 10$, which is highlighted by the vertical dotted red line. The values of G_k are $\{0.0222, 0.0228, 0.0208, 0.0208, 0.0208, 0.0203\}$ corresponding to $k \in \{5, 10, 15, 20, 25, 30\}$.

While a larger cluster size allows more words to be included, a longer topic word list will have a lower topic coherence because more variant words are included (Lau and Baldwin, 2016). Therefore, a larger cluster size results in lower topic coherence. We measure topic coherence as,

$$Coh_j^{(k)} = \frac{2}{|T_j^{(k)}|(|T_j^{(k)}| - 1)} \sum_{w_m, w_n \in T_j^{(k)};\, i > j} \mathrm{NPMI}(w_m, w_n) \; ,$$

where $Coh_j^{(k)}$ is the topic coherence of topic j; $T_j^{(k)}$ indicates the set of words in topic j and $|T_j^{(k)}|$ is the number of topic words in that topic. The superscript k reminds us that the quantity depends on the cluster size k.

Topic coverage measures the probability of a topic given a corpus. Inspired by the LDA model (Hoffman et al., 2010), we compute topic coverage as the ratio between the total topic word count and total word count, i.e.,

$$Cov_j^{(k)} = \frac{C_j^{(k)}}{C} \; ,$$

where $Cov_j^{(k)}$ is the topic coverage of topic j; $C_j^{(k)}$ is total word count of topic j, computed on all documents of the corpus; and C is the total word count of the entire MD&A corpus. With this measure, longer (shorter) topic word lists will have a wider coverage, because more words are taken into account.

Given these two competing measures, we aim to balance the topic coherence/topic coverage trade-off. As our primary objective is to create highly coherent topics that encompass a broad range of information, we maximize the following quantity,

$$G^{(k)} = \sqrt{\sum_{j=1}^{|T|} Coh_j^{(k)} \times \sum_{j=1}^{|T|} Cov_j^{(k)}}\,.$$

For each cluster size k, we obtain the optimal topic word lists by searching for the words similar to each anchor word in each topic. The multiplicative structure is motivated by Dieng et al. (2020). Topic coherence, topic coverage, and the trade-off quantity $G^{(k)}$ are computed based on these topic word lists.

We use a grid of $k \in \{5, 10, 15, 20, 25, 30\}$ for the purpose of optimizing this hyperparameter. Figure 6 shows the trade-off between topic coherence and topic coverage for varying values of the cluster size. The optimal value is 10. Thus, for each anchor word, we choose the top 10 closest words in the vocabulary to that anchor word. The closeness is measured by cosine similarity. The full topic word lists are reported in Table 7, in Appendix E.

E Topic word lists

	No.\topic words	Topic words
Sales/Revenue	82	addition_sale, business_condition, challenge, change_market_condition, company_sale, compete, competition, competition_company, competition_market, competitive, competitive_environment, competitive_market, competitive_position, competitive_pressure, competitive_pricing, competitor, condition, consumer_confidence, consumer_preference, consumer_spending, contract_contract, contract_customer, contract_enter, contract_include, contract_provide, contract_year, current_environment, current_market_condition, customer, customer_contract, customer_demand, demand, demand_company, demand_consumer, demand_customer, demand_market, demand_product, downturn, economic_condition, economic_environment, economy, exist_contract, growth_opportunity, increase_competition, increase_demand, increase_revenue, increase_sale, industry, industry_condition, industry_demand, industry_lead, intense_competition, lead_position, leadership_position, macroeconomic_condition, market, market_continue, market_demand, market_environment, market_factor, market_leadership, market_opportunity, market_presence, market_pricing, market_share, marketplace, net_revenue, position, position_lead, position_market, presence, pressure, pricing_change, pricing_level, pricing_product, pricing_structure, product_demand, product_portfolio, product_pricing, program, revenue, revenue_associate, revenue_attributable, revenue_company, revenue_due, revenue_earn, revenue_generate, revenue_increase, revenue_related, revenue_result, revenue_sale, revenue_year, sale_approximately, sale_company, sale_contract, sale_include, sale_increase, sale_result, sale_sale, service_contract, service_revenue, shift, term_contract, total_revenue, total_sale, trend, unit, year_contract, sale, revenue, market_condition, market_position, consumer_demand, competition, pricing, contract
Cost/Expense	53	additional_cost, asset_impairment, asset_impairment_charge, asset_impairment_charge_related, asset_write, cash_impairment_charge, contingency, contingent_obligation, cost, cost_approximately, cost_associate, cost_cost, cost_due, cost_include, cost_increase, cost_incur, cost_related, cost_result, environmental_liability, estimate_liability, expense, expense_approximately, expense_associate, expense_decrease, expense_due, expense_include, expense_increase, expense_incur, expense_relate, expense_related, expense_result, expense_year, goodwill, goodwill_impairment, goodwill_impairment_charge, goodwill_intangible_asset_impairment, impairment, impairment_asset, impairment_charge, impairment_goodwill, impairment_goodwill_intangible_asset, impairment_live_asset, impairment_loss, increase_cost, indemnification, indemnification_obligation, intangible_asset_impairment, liability_associate, liability_obligation, liability_record, liability_relate, liability_related, live_asset, live_asset_impairment, loss_contingency, operating_cost, operating_expense, potential_liability, record_liability, related_cost, related_expense, related_impairment, restructuring_charge, total_cost, write_asset, write_goodwill, cost, expense, contingent_liability, asset_impairment, goodwill_impairment
Profit/Loss	61	addition, additionally, attributable, begin, company, due, earnings, experience, finally, fourth_quarter, gain, gain_sale, improve_margin, improve_performance, income_compare, income_earn, income_generate, income_include, income_increase, income_loss, income_related, income_result, increase_margin, increase_profit, interest_income, lead, level, loss, management, margin, margin_decline, margin_due, margin_increase, margin_percentage, margin_product, margin_result, margin_sale, net_income, net_sale, operating_income, operating_margin, operating_profit, previously, productivity, profit, profit_increase, profit_margin, profit_percentage, profitability, quarter, reflect, related, result_company, revenue_margin, revenue_profit, sale_margin, sale_profit, specifically, strength, success, target, turn, volume, year, profit, income, performance, result, margin
Operations	31	affect_business, business, business_activity, business_economic, business_financial, business_include, business_operation, company_business, company_operation, current_business, division, economic_regulatory, general, general_economic, general_economic_condition, government_regulation, increase_production, legal_regulatory, manufacturing, market_economic, oil, operating, operation_company, operation_include, operation_result, operation_year, operational, output, processing, produce, producer, production_capacity, production_company, production_facility, production_increase, production_volume, segment, supply, operation, production, general_business
Liquidity	33	cash, cash_balance, cash_cash_equivalent, cash_cash_equivalent_balance, cash_equivalent, cash_flow_operation, cash_fund, cash_generate, cash_generate_operation, cash_investment, cash_marketable_security, cash_operation, cash_position, cash_reserve, cash_term_investment, charge_coverage, charge_coverage_ratio, charge_coverage_ratio_minimum, company_work_capital, coverage_ratio, ebitda_interest, ebitda_maximum, excess_cash, fund_operation, interest_coverage_ratio, invest_cash, level_tangible_net, leverage_minimum, maximum_leverage, minimum_interest_coverage, net_leverage, operating_activity, operating_cash, operating_cash_flow, tangible_net, tangible_net_ratio, term_investment, total_leverage, work_capital, work_capital_requirement, interest_coverage, cash_balance, work_capital
Investment	30	additional_investment, asset, business_divestiture, capital_expenditure_approximately, capital_expenditure_expect, capital_expenditure_include, capital_expenditure_related, capital_improvement, capital_investment, capital_project, capital_spending, company_divest, company_invest, company_investment, complete_divestiture, discontinue_operation, disposal, disposition, divest, divestiture_asset, divestiture_business, divestment, equipment_purchase, equity_investment, estate, exit, expect_capital_expenditure, expenditure, increase_capital_expenditure, invest, investment_fund, investment_make, portfolio, purchase_property_equipment, related_divestiture, related_investment, result_divestiture, sale_asset, sale_business, security, table_content, investment, capital_expenditure, divestiture
Financing	50	annual_dividend, bank_debt, borrowing, capital, cash_distribution, cash_dividend, common, common_dividend, common_equity, common_share, company_repurchase, credit_facility, debenture, debt, debt_financing, debt_issue, debt_obligation, dividend_common, dividend_distribution, dividend_paid, dividend_payment, dividend_share, equity_capital, equity_interest, equity_security, finance, financing, financing_arrangement, financing_company, financing_transaction, fund, funding, indebtedness, investor, issuance, line_credit, note_payable, option, outstanding, outstanding_debt, paid_dividend, pay_dividend, payment_dividend, prefer, prefer_dividend, principal_amount, provide_financing, purchase_share, quarterly_dividend, redeem, redemption, refinance, repayment, repurchase_common, repurchase_outstanding, repurchase_program, repurchase_share, repurchase_share_common, secure_financing, senior_debt, senior_note, share, share_repurchase, term_debt, term_financing, warrant, financing, debt, equity, dividend, repurchase
Litigation	19	arbitration, class_action, class_action_lawsuit, complaint, dispute, investigation, lawsuit, lawsuit_brought, lawsuit_file, legal, legal_matter, legal_proceeding, litigation, litigation_matter, litigation_related, litigation_settlement, matter, patent_infringement, patent_litigation, pending_litigation, proceeding, related_litigation, settlement_litigation, suit, litigation, lawsuit
Employment	30	ability_attract_retain_qualify, change_construction_spending, client, company_experienced_work_not_stoppage, compliance_regulation, consultant, coverage, dependence_management, employee, employee_training, employment, environmental_labor, experienced_work_not_stoppage, hire, hire_additional, hire_employee, incentive, insurance, labor_relation, labor_union, ongoing, personnel, policy, political_condition, practice, professional, recruiting, recruitment, relation_employee, relationship_employee, retain, salary, staff, subject_collective_bargaining_agreement, support_staff, team, training, work, work_stoppage_strike, employee_relation, retention, hire
Regulation/Tax	34	act, applicable_regulation, authority, communication_public, corporate_communication, defer_income_tax, defer_tax, federal_income_tax, federal_regulation, finance_administration, government_affair, human_resource, human_resource_information_technology, human_resource_marketing, income_tax_benefit, income_tax_expense, income_tax_provision, information_technology_human_resource, information_technology_legal, investor_relation, legal_affair, legal_function, legislation, mandate, net_tax, president_international, provision_income_tax, regulate, regulation_applicable, regulation_govern, regulation_include, regulation_regulation, regulation_require, regulator, regulatory_authority, regulatory_requirement, tax, tax_benefit, tax_effect, tax_expense, tax_income, tax_information_technology, tax_liability, tax_provision, tax_treasury, regulation, income_tax, government_relation
Accounting	42	accounting_change, accounting_firm, accounting_guidance, accounting_policy, accounting_principle, accounting_standard, accounting_treatment, adjustment, audit_company, audit_process, audit_review, audit_tax, auditor, change_accounting, change_method_accounting, company_internal_control, company_recognize, company_record, complete_audit, compliance_review, conduct_audit, control_deficiency, control_environment, control_financial_reporting, control_procedure, control_process, disclosure_control, disclosure_control_procedure, eitf, examination, fa, filing, financial_reporting, internal_control_financial_reporting, internal_control_procedure, internal_control_system, material_weakness, method_accounting, method_apply, method_recognize, method_record, policy_procedure, provision_sfas, recognition, recognize, record, reporting, review_audit, sab, sfas, sop, system_internal_control, table_content_accounting, tax_filing, tax_return, valuation, accounting_method, accounting, audit, internal_control

Table 7: Lists of topic words derived from the anchor word lists and the clustering algorithm using cosine similarity and choices of the optimal cluster size. Underlined words are topic words only appearing when the cluster size is 15. The other words appear in both cases when the cluster size is 10 or 15. The number of topic words, when the cluster size is 10, is reported. Using the cluster size of 15 adds 177 words in total to the 10-cluster-size topic word lists.

F Estimation of topic loadings

Define for each topic j a $|D| \times |T_j|$ document-term matrix N_j where $j = 1, ..., |T|$; here $|T| = 11$. This matrix contains the word counts of all topic-j words for all MD&A documents. Thus, each row of N_j is a vector in the $|T_j|$-dimensional space. By SVD, we obtain

$$N_j = U_j \Sigma_j V_j^\top \ ,$$

where U_j and V_j have orthonormal columns called respectively the left and right singular vectors, and Σ_j is a diagonal matrix with positive real entries, the singular values, which are ordered in a descending manner. According to Theorem 3.1 in Blum et al. (2020), the first k columns of V_j constitute the best-fit k-dimensional subspace of the $|D|$ data points in a $|T_j|$-dimensional space.[27] Thus, the first right singular vector, $v_j^{(1)}$, is the optimal 1-dimensional subspace that captures the most information of N_j. Therefore, $v_j^{(1)}$ reflects the importance of topic words in topic j.

As a result of the 2003 SEC regulation, the length of MD&A documents increases over time (Brown and Tucker, 2011). To account for the heterogeneity in document length, we divide each row of N_j by the total word count in the document. This yields a normalized matrix $\tilde{N}_j$, to which we apply the SVD. The optimal summary of the information contained in $\tilde{N}_j$ is the length of the projection of the initial document-term matrix on, $v_j^{(1)}$, i.e.,

$$F_j = (\tilde{N}_j v_j^{(1)})^{|\cdot|} \ ,$$

where F_j is a $|D|$-dimensional vector whose entries are the loadings of topic j in all MD&A documents; $X^{|\cdot|}$ is the element-wise absolute-value operator of the matrix X. Because the matrix $\tilde{N}_j$ is the document-term matrix, its entries are non-negative. By the Perron–Frobenius theorem (Pillai et al., 2005), all elements of $v_t^{(1)}$ have the same sign and the absolute norm is added to convert them into positive numbers. In this way, we compute the topic loadings for all eleven topics.

[27] The best-fit k-dimensional subspace of a set of data points is the k-dimensional subspace such that the sum of squares of the perpendicular distances from the data points to the subspace is minimized.

G Descriptive statistics

Variables	Mean	Std	Min	25%	Median	75%	Max
Sentiment							
Overall	-0.00420	0.00495	-0.05044	-0.00700	-0.00381	-0.00112	0.04018
Sales/Revenue	-0.00030	0.00096	-0.01980	-0.00070	0.00000	0.00000	0.01539
Cost/Expense	-0.00075	0.00106	-0.01626	-0.00126	-0.00056	0.00000	0.01111
Profit/Loss	-0.00116	0.00210	-0.02326	-0.00211	-0.00088	0.00000	0.04000
Operations	-0.00033	0.00089	-0.01980	-0.0006	0.00000	0.00000	0.00971
Liquidity	-0.00002	0.00022	-0.00592	0.00000	0.00000	0.00000	0.00761
Investment	-0.00022	0.00070	-0.01887	-0.00039	0.00000	0.00000	0.01333
Financing	-0.00012	0.00055	-0.01887	-0.00029	0.00000	0.00000	0.01471
Litigation	-0.00027	0.00067	-0.02518	-0.00033	0.00000	0.00000	0.00395
Employment	-0.00004	0.00034	-0.01070	0.00000	0.00000	0.00000	0.01205
Regulation/Tax	-0.00003	0.00052	-0.00658	0.00000	0.00000	0.00000	0.01613
Accounting	-0.00014	0.00049	-0.01197	-0.00027	0.00000	0.00000	0.01333
Loadings							
Overall	0.00297	0.00164	0.00000	0.00182	0.00272	0.00381	0.01941
Sales/Revenue	0.00288	0.00149	0.00000	0.00186	0.00264	0.00362	0.01988
Cost/Expense	0.00491	0.00269	0.00000	0.00300	0.00420	0.00620	0.03329
Profit/Loss	0.00256	0.00166	0.00000	0.00142	0.00223	0.00331	0.02702
Operations	0.00050	0.00071	0.00000	0.00000	0.00027	0.00070	0.01415
Liquidity	0.00185	0.00120	0.00000	0.00100	0.00168	0.00250	0.02364
Investment	0.00137	0.00088	0.00000	0.00077	0.00121	0.00176	0.01904
Financing	0.00019	0.00034	0.00000	0.00000	0.00002	0.00026	0.00808
Litigation	0.00038	0.00045	0.00000	0.00007	0.00027	0.00054	0.00995
Employment	0.00183	0.00136	0.00000	0.00087	0.00162	0.00252	0.02334
Regulation/Tax	0.00118	0.00198	0.00000	0.00010	0.00021	0.00150	0.04648
Accounting	0.00328	0.00162	0.00000	0.00213	0.00305	0.00414	0.01840
Firm fundamentals							
AT_TURN	0.97607	0.94943	0.00000	0.36900	0.80200	1.31400	43.7420
ROA	0.06230	0.20504	-4.33800	0.02100	0.09300	0.15800	3.68700
ACC	0.06513	0.23339	-3.01500	0.00700	0.04600	0.09800	28.9860
CAPITAL_RATIO	0.26933	0.27243	-4.27700	0.00500	0.21500	0.45450	4.87100
FIRM_SIZE	5.95531	1.97936	0.45882	4.50988	5.90920	7.25765	13.6284
B/M	0.90413	32.7176	0.00000	0.29500	0.52900	0.85700	5152.55
FIRM_AGE	21.6485	13.7870	0.44110	11.0603	19.2110	28.5178	58.9562
Q1	0.71410	0.45185	0.00000	0.00000	1.00000	1.00000	1.00000
Q2	0.09539	0.29375	0.00000	0.00000	0.00000	0.00000	1.00000
Q3	0.08847	0.28399	0.00000	0.00000	0.00000	0.00000	1.00000

Table 8: Descriptive statistics of the text-related variables (topic sentiment and topic loadings) and the firm fundamentals. The text-related variables are measured with the cluster size of 10. There are 49619 observations.

H Additional regression results

This section presents the regression results of Equation (2). Here we create the topic word lists from the 15 most similar words as measured by cosine similarity.

	Sales/Revenue	**Cost/Expense**	**Profit/Loss**	**Operations**	**Liquidity**	**Investment**	**Financing**	**Litigation**	**Employment**	**Regulation/Tax**	**Accounting**
AT_TURN	0.0067***	0.0032***	0.0075***	0.0021***	0.0043***	-0.0067***	0.0019***	0.0003	-0.0002	0.0005	-0.0001
	(3.2444)	(3.6160)	(3.4738)	(2.7439)	(3.1122)	(-3.4856)	(3.5559)	(0.7170)	(-0.4364)	(0.7422)	(-0.2124)
ROA	0.0056***	-0.0011	0.0088***	0.0009	0.0012*	0.0002	-0.0016**	-0.0002	-0.0004	0.0070***	0.0015**
	(4.1316)	(-1.4065)	(6.7519)	(1.4426)	(1.9143)	(0.2501)	(-2.4996)	(0.7972)	(-0.9573)	(9.7746)	(2.2953)
ACC	-0.00003	-0.0002	-0.0005	-0.0002	-0.0005*	0.0008	-0.0004	-0.00005	-0.00006	0.00005	0.0005
	(-0.0939)	(-1.1895)	(-0.6484)	(-0.8226)	(-1.6624)	(1.0510)	(-1.1892)	(-0.3680)	(-0.5453)	(0.2496)	(1.0920)
CAPITAL_RATIO	-0.0064***	-0.0020***	-0.0027***	0.0003	0.0000	0.0014**	0.0026***	-0.00007	-0.0026***	-0.0017***	-0.0016**
	(-7.1456)	(-3.0845)	(-2.5861)	(0.6375)	(0.0017)	(2.0672)	(7.1814)	(-0.2180)	(-7.3380)	(-2.6468)	(-2.2359)
FIRM_SIZE	-0.0084***	-0.0048***	-0.0258***	-0.0100***	-0.0078***	0.0072***	-0.0039***	-0.0001	0.0002	0.0048**	-0.0064**
	(-3.6401)	(-2.9885)	(-5.6258)	(7.3061)	(-4.2827)	(3.0736)	(-4.3253)	(-0.1824)	(0.2055)	(2.0022)	(-2.2806)
B/M	-0.0004**	-0.0001	-0.0004	0.0002**	-0.0003**	0.0004*	-0.0003***	-0.0001***	0.0004***	0.0001*	0.00000
	(-2.0492)	(-0.7469)	(-0.9524)	(2.4446)	(-2.2526)	(2.0527)	(-3.6734)	(-4.5813)	(7.5378)	(1.9180)	(0.0238)
FIRM_AGE	-0.0023	-0.0087***	0.0088***	0.0055***	0.00120	0.00140	-0.00004	0.0053***	-0.0007	0.0007	0.0010
	(-1.2337)	(-7.9235)	(2.6091)	(2.6176)	(0.8303)	(0.5894)	(-0.0274)	(6.9282)	(-0.4754)	(0.3538)	(0.5724)
No. Obs	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619
R2	0.1405	0.1090	0.1666	0.0831	0.0436	0.0424	0.1021	0.0095	0.0437	0.0412	0.0215

Table 9: This table reports the Random Effect Ordinary Least Squares (RE OLS) results of regression model (2) when $Y_{j,i,t} = F_{j,i,t}$ and the cluster size is 15 (most similar words),

$$\begin{aligned} F_{j,i,t} =& \alpha + \beta_1 AT\_TURN_{i,t} + \beta_2 ROA_{i,t} + \beta_3 ACC_{i,t} + \beta_4 CAPITAL\_RATIO_{i,t} + \beta_5 FIRM\_SIZE_{i,t} \\ &+ \beta_6 BM_{i,t} + \beta_7 FIRM\_AGE_{i,t} + Quarter\_dummies + Sector\_dummies + Year\_dummies + u_i + \epsilon_{i,t} \end{aligned}$$

where $F_{j,i,t}$ is the topic loading of the topics *Sales/Revenue, Cost/Expense, Profit/Loss, Operations, Liquidity, Investment, Financing, Litigation, Employment, Regulation/Tax* and *Accounting*, j, of firm i at time t. The set of independent variables is described in Section 6.1. All independent variables are standardized for the sake of interpretation. The regression coefficients, two-way clustered (by year and firm) t-statistics (in parentheses), and R^2 are reported. The coefficients of the intercepts, sector dummies, and year dummies are not reported to save space. The data sample spans the period 1994:01 to 2018:12. $*$, $**$, and $***$ denote significance at 10%, 5% and 1% respectively.

	Overall	**Sales/Revenue**	**Cost/Expense**	**Profit/Loss**	**Operations**	**Liquidity**	**Investment**	**Financing**	**Litigation**	**Employment**	**Regulation/Tax**	**Accounting**
AT_TURN	0.0863***	0.0462***	0.0400***	0.0910***	0.0179**	0.053	0.0681***	0.0185*	0.0011	0.0008	0.0050	0.0257***
	(3.4352)	(4.8506)	(3.7019)	(3.3676)	(1.9912)	(0.7378)	(4.7624)	(1.7256)	(0.1208)	(0.0586)	(0.5706)	(2.9119)
ROA	0.1024***	0.0366***	0.0346**	0.0467***	0.0642	0.0456***	0.0288***	0.0462***	0.0219**	0.0258***	0.0282***	0.0034
	(6.2531)	(3.0161)	(2.4778)	(3.9491)	(4.2715)	(4.9041)	(3.2062)	(5.1464)	(2.4353)	(3.0564)	(4.3425)	(0.5834)
ACC	-0.0161	-0.0371**	-0.0344	-0.0075	-0.0103	-0.0039	-0.0139	-0.0086*	0.0015	-0.0053	-0.0027	-0.0021***
	(-0.8999)	(-2.4475)	(-1.2475)	(-0.9164)	(-1.5646)	(-0.7615)	(-1.6382)	(-1.8955)	(0.1903)	(-0.8756)	(-0.4241)	(-8.8473)
CAPITAL_RATIO	-0.0445***	-0.0013	-0.0489***	-0.0404***	0.0045	-0.0122	-0.0535***	-0.0432***	0.0181**	0.0042	-0.0143*	-0.0125**
	(-3.7892)	(-0.1167)	(-5.5260)	(-5.1220)	(0.4064)	(-1.4032)	(-5.9570)	(-5.3057)	(2.2877)	(0.6031)	(-1.6861)	(-2.0231)
FIRM_SIZE	0.1511***	0.0799***	0.0270	0.1515***	0.0819***	0.0496***	0.0179	0.0656***	0.0113	0.0331***	0.0005	0.0400***
	(6.0081)	(5.2886)	(1.0903)	(11.864)	(4.9700)	(5.5505)	(1.3802)	(5.4633)	(0.7869)	(2.9726)	(0.0564)	(3.2480)
B/M	-0.0041	-0.0088***	-0.0072**	0.0022	0.0002	0.0010*	-0.0016	0.0013	0.0006	0.0028***	0.0040***	-0.00004
	(-1.2977)	(-4.4350)	(-2.1901)	(1.3292)	(0.2029)	(1.6691)	(-0.9961)	(1.1989)	(0.6887)	(4.7748)	(7.1568)	(-0.1267)
FIRM_AGE	-0.0433**	-0.0462**	-0.0575***	0.0091	-0.0225*	0.0216*	-0.0521***	0.0303**	-0.0484***	-0.0034	-0.0205**	-0.0344***
	(-2.1323)	(-2.2753)	(3.3483)	(0.4426)	(-2.2480)	(1.9293)	(-3.8587)	(2.4812)	(-3.4897)	(-0.3327)	(-2.4163)	(-2.6187)
No. Obs	49169	49169	49169	49169	49169	49169	49169	49169	49169	49169	49169	49169
R2	0.0290	0.0078	0.0068	0.0198	0.0090	0.0040	0.0060	0.0063	0.0013	0.0013	0.0009	0.0021

Table 10: This table reports the Random Effect Ordinary Least Squares (RE OLS) results of regression model (2) when $Y_{j,i,t} = s_{j,i,t}$ and the cluster size is 15 (most similar words),

$$\begin{aligned} s_{j,i,t} =& \alpha + \beta_1 AT\_TURN_{i,t} + \beta_2 ROA_{i,t} + \beta_3 ACC_{i,t} + \beta_4 CAPITAL\_RATIO_{i,t} + \beta_5 FIRM\_SIZE_{i,t} \\ &+ \beta_6 BM_{i,t} + \beta_7 FIRM\_AGE_{i,t} + Quarter\_dummies + Sector\_dummies + Year\_dummies + u_i + \epsilon_{i,t} \end{aligned}$$

where $s_{j,i,t}$ is the total MD&A sentiment (*TMS*) and the sentiment of the topics *Sales/Revenue, Cost/Expense, Profit/Loss, Operations, Liquidity, Investment, Financing, Litigation, Employment, Regulation/Tax* and *Accounting*, j, of firm i at time t. The set of independent variables is described in Section 6.1. All independent variables are standardized for the sake of interpretation. The regression coefficients, two-way clustered (by year and firm) t-statistics (in parentheses), and R^2 are reported. The coefficients of the intercepts and year dummies are not reported to save space. The data sample spans the period 1994:01 to 2018:12. $*$, $**$, and $***$ denote significance at 10%, 5% and 1% respectively.