A Data-driven and Interpretable Topic Model for Management Disclosures

Matthias Fengler* Minh Tri Phan†

Version: June 14, 2022

Abstract

Topic models have been increasingly applied in economics and finance. Due to computational limitations and issues in interpretability, the most popular topic model, the Latent Dirichlet Allocation (LDA), exhibits mixed results in economic and financial research. This paper introduces a topic modeling framework that extracts the hidden topics discussed in the Management Discussion and Analysis (MD&A) section in 10-K filings in a data-driven and interpretable way. The results of our paper are three-fold. Firstly, the hidden topics discovered by our model are conveniently interpretable and distinguishable in the financial context. Secondly, the time series of hidden topic loadings grant insights into time variation of topic prevalence. Finally, regression analyses show variations of the relationships between firm characteristics and the MD&A compilation behavior of firm managers. We find that firm managers truthfully report the firm performance and accrual conditions in their MD&A.

Keywords: MD&A, phrase-learning model, topic modeling, Word2vec

*School of Economics and Political Science, Department of Economics, University of St. Gallen, Bodanstrasse 6, 9000 St. Gallen, Switzerland. Email: matthias.fengler@unisg.ch.

†School of Economics and Political Science, Department of Economics, University of St. Gallen, Bodanstrasse 6, 9000 St. Gallen, Switzerland. Email: triminh.phan@unisg.ch.

1 Introduction

In recent years, with the development of computer hardware as well as *Machine Learning* (ML) models,[1] various new sources of data have become accessible and exploitable. With these developments, textual data, have increasinly demonstrated their potential in economics and finance (Tetlock, 2007; Engelberg, 2008; Henry, 2008; Tetlock et al., 2008; Loughran and McDonald, 2011a,b; Chen et al., 2022).

In this paper, we propose a machine-learning-based methodology for topic modeling and examinee the structure of topics contained in the *Management Discussion and Analysis* (MD&A) section of 10K filings thereby gaining more knowledge about the trend of topic variations over time, and the determinants of MD&A disclosure contents. As a key section in the annual textual communication between firms and investors (Tavcar, 1998) where firm managers express their perspectives of the firm, such as the liquidity situation, outlook of earnings, or the views about future operations and economic environment. In general, the content included in the MD&A is standardized by various regulations of the *Securities and Exchange Commission* (SEC) mandating the inclusion of a certain number of topics. However, due to the flexibility granted to the firm management board in compiling this document, they have the leeway to discuss the included topics with their intentions and this language. Consequently, the value of topics discussed in the MD&A with respect to investors is still being questioned. Our paper introduces a novel method that is based on both machine learning models and prior information about to uncover the content of the MD&A.

While many papers address the behaviour of market participants to the information extracted from the management disclosures (Henry, 2008; Loughran and McDonald, 2011b; Price et al., 2012; Jiang et al., 2019), few papers actually exploit the contents of the management disclosures. In this paper, we introduce a method to retrieve the contents discussed in the MD&A sections and use them to address three main questions: (i) What does the MD&A talk about? (ii) How do the MD&A topics change over time? (iii) Does the MD&A reflect the firm's fundamental conditions? To this end, we develop a data-driven framework to discover the latent content of the MD&A sections of US firms' 10-K reports from 1994 to 2018. As an attempt to make the MD&A section more transparent, SEC (2003) provide guidance to compile and interpret the MD&A section. Answering the above-mentioned questions will shed light not only on the hidden content of the MD&A but also reveal the effectiveness of SEC's efforts in providing investors with a standardized communication channel from firms.

[1] We use the definition of the term "*machine learning model*" of (Mitchell et al., 1997). A machine learning model is a computer program that "*learns* from experience E with respect to some class of tasks T and performance P, if its performance at tasks in T, as measured by P, improves with experience E".

The contribution of our work to the current literature is three-fold. Firstly, we propose a framework to determine the topic structures of the MD&A that, on the one hand, minimizes human interventions, on the other hand, adds more human interpretability. Several previous studies attempt to determine topics (or in some papers, referred to as *aspects*) by human-encoding (Brown and Tucker, 2011; Li, 2010) which faces the subjectivity in topic determination as well as diminishes scalability of their proposed models. Meanwhile, some others use the LDA model to automatically learn topics (topic words, topic probabilities) from a document that suffers the aforementioned obstacles (lack of word semantics, high dimensionality, or domination of high-frequency words). Our model is introduced to overcome these challenges, leaving most of the work for computers but still preserves much of human interpretability. Secondly, based on the outcomes of the proposed model, the time series of topic contents and topic sentiment capture some remarkable events in history (e.g. Enron's scandal in 2001, the dot-com and financial recessions). Finally, by the regression analysis, we find that firm characteristic are reflected in the MD&A document. In particular, we find that well-performed firms tend to express an optimistic tone in their MD&A, while firms with a high accrual-on-asset ratio tend to be pessimistic about their sale and accountings topics.

Various topic modeling methods have been presented and widely applied, the Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and Latent Semantic Analysis (LSA) (Dumais, 2004). Despite their success in empirical applications, few attempts have been made to apply them in the financial and economic world. One of the earliest attempts to do that is described in a paper by Jegadeesh and Wu (2017) who apply the LDA model to quantify the economic content in Federal Reserve communication statements, the Federal Open Market Committee minutes, and examine the reaction of financial markets toward these pieces of information. Similarly, by employing the LDA model, Bellstam et al. (2019) detect that a text-based measure of innovations is able to robustly forecast the firm performances up to four years into the future. Recently, Brown et al. (2020) propose a new text-based measure of the content consisting of financial statement disclosures via the LDA model, and empirically suggest that topics discussed in the annual statement releases are good signals to detect intentional financial misreporting. In general, the LDA model is mainly applied in financial and economic research to model topics in a document and to construct text-based measures for further analytical purposes. However, due to the bag-of-word background, the LDA model suffers, (i) intensive computational cost on very huge data sets, and (ii) the lack of word semantics[2] (Mikolov, Sutskever, Chen, Corrado and Dean, 2013). Furthermore, topics detected from the LDA model tend to be dominated by high-frequency words (they are words that appear many times in a document and in many documents) if the prior parameters are not specified carefully (Wallach et al., 2009). Consequently, applying the LDA model in the financial textual data is still

[2]Word semantics is referred as the relationship between words.

limited and inefficient (Cong et al., 2019).

To overcome the challenges of the current topic models and enhance their applications in economics and finance, we develop a topic model which is generated by employing a similarity-based clustering algorithm that is based on the novel Word2vec model, developed by (Mikolov, Sutskever, Chen, Corrado and Dean, 2013). To incorporate more interpretability to the model, we introduce the concept of *anchor words* that guides the model to learn and output the topics that are close to the human sense. Word2vec is a neural-network-based model that learns the semantic representation of words by looking at their relationship to other words in the vocabulary generated from a collection of documents. The ultimate result of this model is a dense matrix of word representations that has word vectors as its rows. This model is widely used nowadays in Natural Language Processing (NLP) tasks because it solves the weaknesses of the count-based word representing 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 make Word2vec suitable for detecting topics in a document because topics tend to be formed from words with similar meanings.[3] To further provide the model guidance to formulate topics with human interpretability, a set of anchor words that carry prior expert knowledge about MD&A documents is introduced. Associated with a clustering algorithm, this process will search for words that deliver close meanings with the anchor words. Besides the choice of anchor words, this process is fully automatic, thus reducing the researchers' bias. To alleviate the subjectivity of the anchor word suggestion, we utilize exactly the same word lists introduced in Appendix C of Li (2010). To handle compound words (i.e. words are constituted by more than one single word), we apply a method introduced by Mikolov, Sutskever, Chen, Corrado and Dean (2013) to learn phrases in the MD&A corpus. The set of topic words, as the result of this framework, is found to be highly interpretable in the management financial context, less contaminated by noisy words. Topics formed by our topic model are also separate and coherent in the sense that one can easily realize the main theme of each topic by the corresponding topic words.

After obtaining the topic word lists, we gauge the prevalence of topics in the MD&A by computing the topic loadings. Using these topic loadings, time series of the topics discussed in the MD&A is constructed and investigated to uncover the time variations of the MD&A content over time in a similar manner to Jiang et al. (2019). We find that the 2003 SEC regulation has limited effects on the MD&A contents. This result contributes to the current literature about the content of the MD&A, which generally agrees that firm managers tend to use boilerplate and generic language to compile the MD&A. We com-

[3]This statement is neither theoretically nor empirically justified. However, with the wide agreement with this statement, many topic models are designed based on clustering algorithms.

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 $\lvert D\rvert \times \lvert T_j\rvert$
$\tilde{N}_j$	The normalized document-term matrix of topic j, with dimension $\lvert D\rvert \times \lvert T_j\rvert$
F_j	A vector of loadings of topic j of the MD&A documents
$\lvert\mathcal{S}\rvert$	Cardinality of the set $\mathcal{S}$

Table 1: Notation used in the paper.

pute the topic-specific sentiment indexes, in which the topic-specific sentiment scores are computed by the difference between the predefined positive and negative words which are adjacent to the determined topic words.

We also examine the relationship between the MD&A contents and firm financial and fundamental indicators. We find that well-performed firms tend to express an optimistic tone in their MD&A, while firms with a high accrual-on-asset ratio tend to be pessimistic about their sale and accountings topics.

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 tone. These four steps rely on the 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 tone projection. These technical building blocks are described in detail in Appendix A. Important notation is collected in Table 1.

2.1 Formation of anchor words via the phrase-learning model

Starting point of our model is the suggestion of lists of anchor words, each of which is supposed to represent a topic discussed in the MD&A documents. The purpose is to inject initial expert knowledge about the MD&As into our ML algorithms so that they will start allocating words into the desired topics. This suggestion will reinforce the interpretability of topics learned by our model. To limit subjectivity, we use the word lists and categories offered in the Appendix C of Li (2010). This set of word lists spans over 11 categories and includes words from the following topics: *Sale/Revenue, Cost/Expense, Profit/Loss, Operation, Liquidity, Investment, Financing, Litigation, Employment, Regulation/Tax*, and *Accounting* (see Table 2). The concept of using anchor words, although not very recent, is gradually used in topic modeling: Arora et al. (2012) introduce this idea to guarantee the separability of their topic model. With a similar purpose, Cong et al. (2019) use anchor words to accommodate pre-existing knowledge in their topic model. In our work, the anchor word lists are used with two purposes: (i) to validate the phrase-learning model; and (ii) to use them as semantic anchors for the clustering algorithm that identifies the topic word lists. The former purpose will be detailed in the rest of this section, while the latter will be described in Section 2.2.

A major challenge of mining economic and financial texts is that they contain many 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 (Mikolov, Sutskever, Chen, Corrado and Dean, 2013); as examples consider "market condition" or "capital expenditure". Indeed, our anchor word list in Table 2 features a lot 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 using n-grams.[4] This approach, however, drastically increases the size of the vocabulary (Mikolov, Sutskever, Chen, Corrado and Dean, 2013); at the same time, the computational burden of the model increases because meaningless phrases will also be taken into account. To handle this issue, we train the phrase-learning model introduced in Appendix A.1 to detect phrases (bigrams or trigrams) appearing in the corpus. Generally, a phrase-learning model has two major benefits over n-gram models. Firstly, it automatically learns phrases in the corpus with no human intervention, thus avoiding subjectivity. Secondly, it enriches the vocabulary selectively by only adding phrases with plausible meanings to the vocabulary. This advantage makes the word embedding[5] absorb more refined word semantics and at the same time lowers

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

[5] The word embedding is a vector representation that captures the semantics of a word in the sense that

Topic 1: Sale/Revenue	Topic 2: Cost/Expense	Topic 3: Profit/Loss	Topic 4: Operation	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	disvestiture	dividend		union relation	government relation	internal control
consumer demand competition pricing new contract	goodwill impairment				discontinued operation	repurchase				

Table 2: Lists of anchor words proposed by Li (2010). Li's category of "Miscellaneous" is not considered. Words in these lists are singularized to align with lemmatization in our textual preprocessing step.

noise compared to n-gram models. An important hyperparameter in training the phrase-learning model is the threshold at which a pair of words is considered a phrase because that, in turn, determines the size of the vocabulary in the next stages. We explain our selection process in Appendix B.

After obtaining the complete vocabulary including single words and plausible phrases as detected by the phrase-learning model, we map each word and phrase in the vocabulary into a dense vector representation using the Word2vec model (Mikolov, Chen, Corrado and Dean, 2013), which is explained in Appendix A.2. Because after phrase-learning and word vectorization, words and phrases both have a vector representation, we from now on simply refer to single words and phrases as *words*. The word vectors are the main ingredients for the next stage, the formation of the topic word lists via a guided clustering algorithm.

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

The next step in our topic model is to expand the anchor word lists to 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 and phrases as "semantic anchors" and search for similar terms. This process requires us to adress the difficult question of determining the "relatedness of words". In word embeddings, where vectors are proxies for the meaning of words, it is natural to employ distances and proximity measures on the word vectors as measures of semantic relatedness (Kiela et al., 2015; Kruszewski and Baroni, 2015).

An approach to expand 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[6] is larger than the threshold. Based on experiments in document retrieval, Rekabsaz et al. (2017) suggest a similarity threshold of about 0.71 for a word embedding in a 300-dimensional space. This result, however, may depend much on the corpus on which the Word2vec model is trained. Therefore, in this paper, we take an alternate route by determining an optimal *cluster size*, i.e., the number of words which are similar to each anchor word. We do so by maximizing a *topic coherence-coverage trade-off*. Topic coherence is measured by the *Normalized Pointwise Mutual Information* measure (Bouma, 2009; Lau

two words that bear a similar meaning also have a close representation in the vector space.

[6]The cosine similarity between two vectors x and y is computed 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 vector x. The cosine similarity measure is a natural choice in this case because Word2vec itself learns words that are adjacent to each other in terms of cosine similarity.

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 detail in Appendix C. To ensure robustness, we verify that the optimal cluster size and its nearby values (sub-optimal cluster sizes) yield qualitatively similar results.

2.3 Estimation of topic loadings

The next step of 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 *topic loadings*, which can be interpreted as the extent to which each topic is covered in a given document.

Fundamentally, the document-term matrices,[7] which are obtained as a result of the word list formation, capture the desired information about a given topic. These matrices, however, are very unwieldy due to their high dimensions. Instead, we seek to summarize the information optimally in a 1-dimensional space. With this intention, *Singular Value Decomposition* (SVD) appears as a natural candidate to address this challenge.

Indeed, the SVD, and matrix factorizations in general, play an important role in topic modeling. Dumais (2004) develops the *Latent Semantic Analysis* (LSA) by applying the SVD to the document-term matrix to find low-rank representative vectors of documents and words. Topics are formed by applying a clustering algorithm to these representative vectors. The vectorization step, however, is done by the Word2vec model in our case, which is known to be more efficient than LSA (Mikolov, Chen, Corrado and Dean, 2013). Moreover, LSA assumes a single topic per document (Arora et al., 2012), which is not applicable with MD&A documents. Arora et al. (2012) suggest using the non-negative matrix factorization instead of SVD for topic modeling and prove that their algorithm overcomes the single-topic-per-document assumption of LSA. However, their algorithm, in turn, assumes a single anchor word per topic, which is restrictive. To overcome these challenges, Cong et al. (2019) use the SVD to estimate topic loadings from the topic-specific document-term matrices. In this way, by applying the SVD to each document-term matrix (A), one obtains the 1-dimensional subspace that maximizes the cumulative magnitudes of the projections of all rows of A on this subspace. The topic loadings are the magnitudes of these projections. We follow these ideas here and defer the technical details to Appendix E.

[7] A document-term matrix has rows representing documents in a corpus and columns representing words. Each entry of the matrix shows the occurences of a word in the corresponding document.

Because of these properties, the size of the projections, which are formally given by $\sigma_1 u_1$,[8] ranks the rows of the document-term matrix. Thus they can be readily interpreted as the extent to which a document talks about a given topic or *topic loadings*. 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 be orthogonal and thus unrelated. On the other hand, the entries in v_1 rank the magnitudes of the projections of matrix A's columns 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 A 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 convenient measurement of the topic word importance over the entire corpus. We exploit this fact when visualizing the word clouds in Section 4.1, see Figure 1.

That said, comparing the magnitudes across different topics, however, is a 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 size counts and size variation 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 factor loadings across topics as an indicator of the prevalence of given topics across the corpus. It is noticed that we apply the SVD to the normalized document-term matrix, in which each row is divided by the total word count of the corresponding document, instead of the vanilla document-term matrix to obtain the more robust interpretation against the document length.

2.4 Estimation of topic sentiment by Loughran-McDonald dictionary

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, relying 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 of its focus on the economic and financial texts, see, e.g. Feldman et al. (2010); Dougal et al. (2012); Garcia (2013); Jiang et al. (2019). Despite of the dominance of lexicon projections in the literature, recent advances suggest supervised learning approaches (Chen et al., 2022).

We adopt the approach of Jiang et al. (2019) to compute the document sentiment score with a small modification.[9] For a document d, the topic words of the topic j are located.

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

[9] There is a variety of approaches to measuring a sentiment score. See also Antweiler and Frank (2004);

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 document d of topic j, $s^{+}_{d,j}$ ($s^{-}_{d,j}$), 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 document d of topic j is

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

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

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

3 Data processing

3.1 Text data processing

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).[10] The latter page also provides additional resources for textual data analysis, such as stopword lists and the Loughran-McDonald (LM) dictionary. The data of SRAF are 10-K and 10-Q filings in text-file format from which HTML tags have been removed. We focus our analysis on 10-K filings only because the information content is widely acknowledged to be more significant compared to 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 in Loughran and McDonald (2016), and manage to extract up to 68% of all 10-K files in the corpus.[11] We discard documents that have fewer than 250 words in the MD&A. After these purges, we retain 124, 133 MD&A documents spanning from 1994:01 until 2018:12.

After extracting the MD&A documents from the 10-K filings, several standard steps of text normalization are executed. Thereby, we take particular care to properly process

Loughran and McDonald (2011b); Chen et al. (2022).

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

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

	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. 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 3: Effects of data extraction and processing steps on the MD&A sample size. The *No. of identified MD&A* shows the number of the 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 initial MD&A documents. *Survival ratio* is the ratio between the number of eventual MD&A documents and initial MD&A documents.

negations because ignoring negations changes the polarity of a statement and leads the sentiment analysis astray (Mukherjee et al., 2021); for any details about text normalization and the settings of the phrase-learning and Word2vec models, see Appendix F.

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 over 26 years from 1994:01 to 2018:12. Table 3 reports the data loss incurred during the extraction and processing steps. Comparing 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, 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%, i.e. comparable.

4 The content of the MD&A section

4.1 What talking points do topics have?

It is of the most imminent interest to unveal which information the topics cover (Table 7, Appendix D). The larger word size implies more importance for the word as explained in Section 2.3. For the ease of exposition, topic word clouds are presented in Figure 1. The topic *Sale/Revenue*'s most important words are "revenue_result, market_environment, sale_sale, revenue_related, contract_include, pricing_structure".[12] Besides covering firm sales and revenues, this topic includes matters of competition (with 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 (with words like "economic_condition, economic_environment, downturn"), and the pricing strategy (with words like "pricing_level, pricing_structure, pric-

[12]The fact that all these terms are phrases underscores the importance of the phrase-learning algorithm (see Appendix A.1)

ing_product). These aspects all relate to the abilities of a firm to generate revenues.

The topic *Cost/Expense* covers three aspects of firm costs and expenses. The first aspect relates to operating costs and expenses with words like "cost, expense, operating_cost". It is worth noticing that this topic does not cover costs and expenses from two further specific activities, namely wages (covered by *Employment*) or costs from taxation (covered by *Regulation/Tax*.). This result emphasizes the robustness of the proposed topic model in forming intuitive and cohesive topics. The second aspect the topic deals with is asset impairment ("impairment_asset, impairment_charge", "impairment_goodwill"). This makes sense because the dollar value of an impairment is the difference between the asset's carrying costs and its market value. Interestingly, in the topic word list of *Cost/Expense*, we observe further related words like "write_asset (write-off of assets)" or "write_goodwill (write-off of the goodwill)", which also refer to the 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, liability_related".

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

The topic *Liquidity* discusses various facets of a company's liquidity, e.g., cash holdings, interest coverage, and working capital. Topic words of this topic are evenly important as can be seen from the fairly homogeneous word size in the *Liquidity* word cloud, reflecting that firm managers use diverse language to talk about firm liquidity. Two most important words of *Liquidity*, which 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*. Particularly, 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 of the *Liquidity* topic is "level_tangible_net" which describes a firm's level of tangible assets. Tangible net worth includes physical assets of a firm, which can be easily converted to cash, thus serve as a source of firm liquidity. Therefore, the appearance of this word in the topic about

Figure 1: The topic word clouds of eleven topics, with the cluster size of 10. 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.

firm liquidity is reasonable. Besides these words, *Liquidity* also describes a firm's immediate liquidity, like "cash_cash_equivalent_balance" (cash and cash equivalent balance), "cash_generate_operation", "work_capital_requirement" (working capital requirement).

The topic *Investment* focuses on both divestment and investment. This is seen from topic words like "divest, disposition, disposal" describing firm divestment. Moreover, we also find associated words like "asset, sale_asset, sale_business". It is worth noticing that these words, despite featuring "sale", and without bearing any obvious relation to the *Investment* anchor words, are allocated to the *Investment* topic rather than *Sale/Revenue* by our model. Nevertheless, this result is intuitive given that these words refer to the firm's ac-

tivities to sell assets rather than operating sales. The other aspect of this topic describes firm's investment decisions with words such as "capital_spending, capital_investiment, investment_fund, equity_investment".

The topic *Financing* characterizes the firm's financial resources, including debt and equity. Similar to *Liquidity*, two out of 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, capital_spending" belong to *Investment*. This, however, makes sense as *Investment* describes firm activities of their productive assets, whereas *Financing* talks more about a firm's capital structure. Besides "credict_facility", we detect words in *Financing* which do not appear or directly relate to the topic anchor words, such as "borrow, senior_note, line_credit". They all describe funding via debt. As regard of firm's equity, we uncover further words that are beyond the initial anchor words like "common_share, investor, share", among others.

With only two anchor words in the *Litigation* topic, namely "litigation, lawsuit", our model successfully detects a number of important topic words like "arbitration, complaint, dispute, legal_matter", which are in the scope of litigation but beyond the initial anchor words. Similarly, we enrich the topic *Employment* by significant words such as "compliance_regulation, personnel, staff, insurance, incentive, environmental_labor". There are two main aspects in *Regulation/Tax*, i.e., regulation (with words like "government_affair, legislation, corporate_communication", etc.) and taxation (with words like "defer_tax, provision_income_tax, tax_benefit, tax_expense"). Finally, topic *Accounting* mentions accounting and auditing activities within a firm, as well as accounting and auditing standards such as words "sab" (Staff Accounting Bulletin), "sfas" (Statement of Financial Accounting Standards).

There are two key takeaways from the above discussions. Firstly, our model successfully expands the initial anchor word lists to comprehensive topic word lists. A significant amount of words detected by our modeling approach do not directly involve in the anchor words in an obvious way but still serve as the most important words in their respective topics. This finding once again emphasizes the importance of the topic word formation step in Section 2.2; at the same time it shows that parsimonious anchor word lists as ours allow one to find serendipitous results. The second takeaway is some words, which are seemingly similar at the first glance, are assigned to two different topics by our model as, given the topic contexts, they should be intuitively different. We figure that although they are reasonably classified, these words are relatively close in the word vector space. We give a profound discussion of the word distribution in the next section.

4.2 How are topic words distributed?

To give an impression of how topic words are distributed, we present in Figure 2 a scatter plot of a projection of the topic word vectors into a 2-dimensional space.[13] At first, it is noticed that topics are separate well. Secondly, several topics have obvious sub-topics. Furthermore, as mentioned also in Section 4.1, some topics are woven into each other. Particularly, the three topics *Sale/Revenue* (royal blue), *Profit/Loss* (green) and *Operation* (red) appear very close to each other. It should be recalled, however, that we study a 300−dimensional word vectors. Therefore, for visualization, a 300-dimensional space is projected to a 2-dimensional space. Therefore, even though some topics are visually woven to each other in the 2-dimensional space, this does not neccessarily mean the topics are close to each other in the original space. Nevertheless, this relatedness can often be rationalized on semantic grounds.

For example, the part of *Sale/Revenue* (royal blue) relating to firm sale and revenue, with words "sale", "sale_result", etc., 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.). Hence this adjacency is reasonable as they both describe firm performance. It is note-worthing to see that the part of *Profit/Loss* (green), which is in the center of Figure 2, is next to the left part of *Regulation/Tax* (cyan), which is dominated by words relating to taxes. Thus, this part of topic *Profit/Loss* (green), which is far from the part talking about firm profitability, is close to words about taxes of *Regulation/Tax* (cyan).

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 one (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 part describing firm short-term assets (cash and working capital). The *Regulation/Tax* topic (cyan), as the name suggests, mentions two distinct aspects, one about regulation and legislative restrictions, and one about taxation. The latter is adjacent to the income-related cluster of the *Profit/Loss* topic (green) as was mentioned above. Similarly, *Accounting* (blue) displays accounting-related aspects, e.g., accounting methods, accounting principles, and a part that is about internal controlling and auditing.

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

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

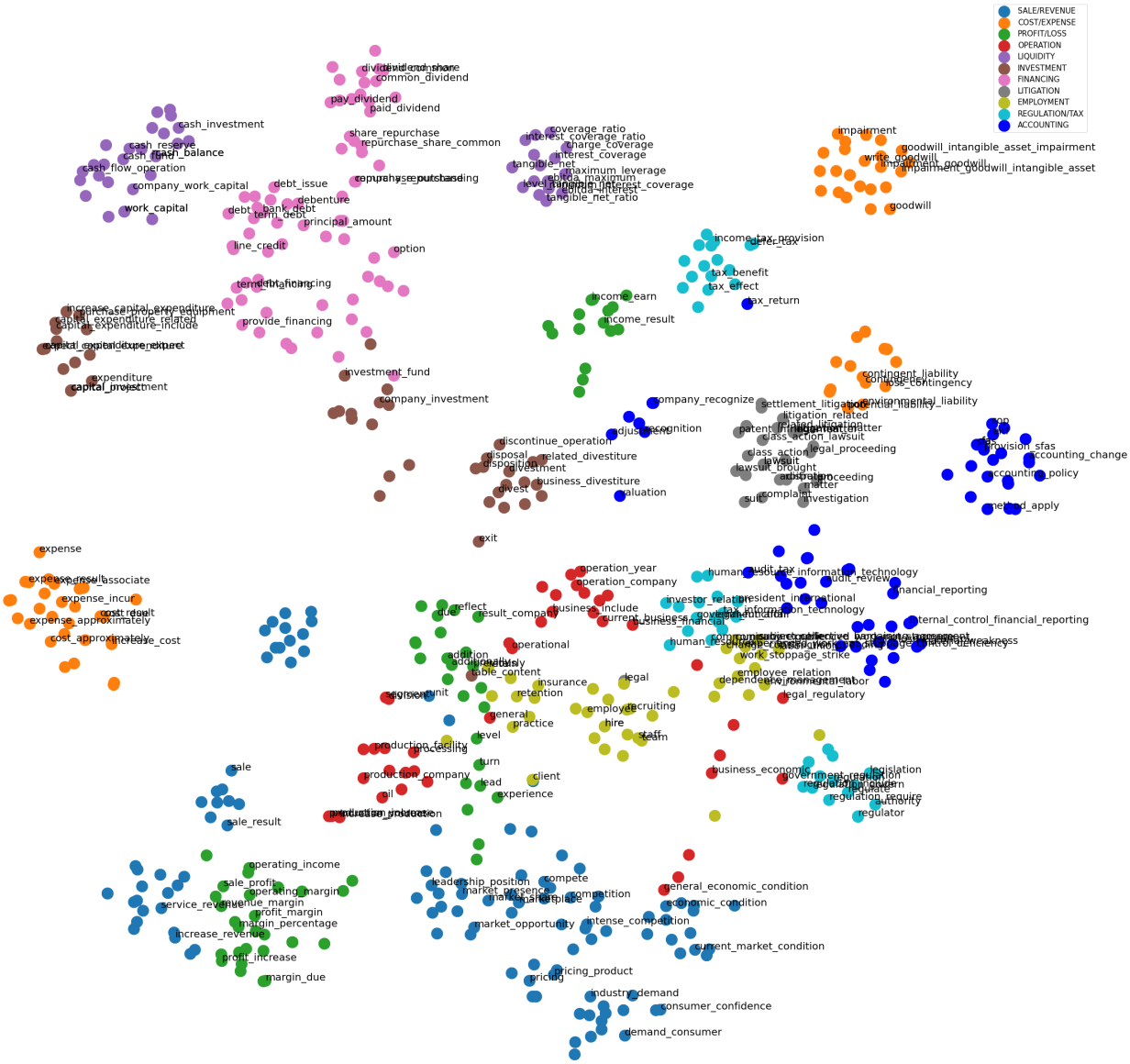

Figure 2: A visualization of the word vectors of the topic words in a 2-dimensional space. The dimensionality reduction is implemented by Stochastic Neighbor Embedding (Van der Maaten and Hinton, 2008). Because of the complexity, only 25 random words of each topic are reported.

words within the MD&A sections, offerring little semantic overlap with other topics. With the same interpretation, 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 withdrawn. Firstly, topics appear close when the word collections of topics are related in their semantic orientation to each other as discussed (e.g., *Sale/Revenue* and *Profit/Loss*). Secondly, topics may contain more than one sub-topic, and words describing these sub-topics may locate distantly in the vector

space. Thus, topics like *Cost/Expense* are, in some sense, multimodal. Because of these two reasons, a cluster analysis based on classical and purly data-driven methods is challenging and potentially misleading. For example, words like "income_earn (incomes and earnings)" can be easily merged with the word "income_tax" although these two words depict two different objects. Our approach, instead, uncovers very well-defined topic word lists.

5 How does MD&A 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 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.

All topics exhibit substantial variations in time. Starting with the topics *Sale/Revenue, Profit/Loss, Liquidity*. We observe a downward trend in the loadings, especially, after the release the the SEC's guidance on the content to be discussed in the MD&A section, which became effective on 29th, Dec 2003 (we will refer it as the "2003 SEC regulation" for short during this paper). This implies apparently that firm managers have less focused on these topics over the sample period.[14] In the opposite, other topics, e.g., *Operation* and *Investment*, have become more important in the MD&A topics over time. We can observe a marked increase in the loading time series right at the time of two recessions. In particular, after the 2007-2008 financial crisis, firms tend to discuss more their operation (production, manufacturing, etc.) in their MD&As. In contrast, the topic *Investment* receives more space in the MD&A section after the dot-com crisis (from 2001:04 to 2001:11).

There are further remarkable patterns. The loadings of *Financing* gradually decreased till shortly before the financial crisis. During the crisis, the discussion of this topic in the MD&A section increases remarkably, then stays at about the same level. This suggests that the financial crisis has had a long-term impact on the financing topic in the MD&A section. Firm managers increasingly discusse the topics *Litigation* and *Employment* in their MD&As until about 2005, after which they gradually reduce the prevalence of this topic.

[14] Recall the interpretation of topic loadings in Section 2.3, higher (lower) topic loadings imply that the topic is more (less) discussed in the MD&As.

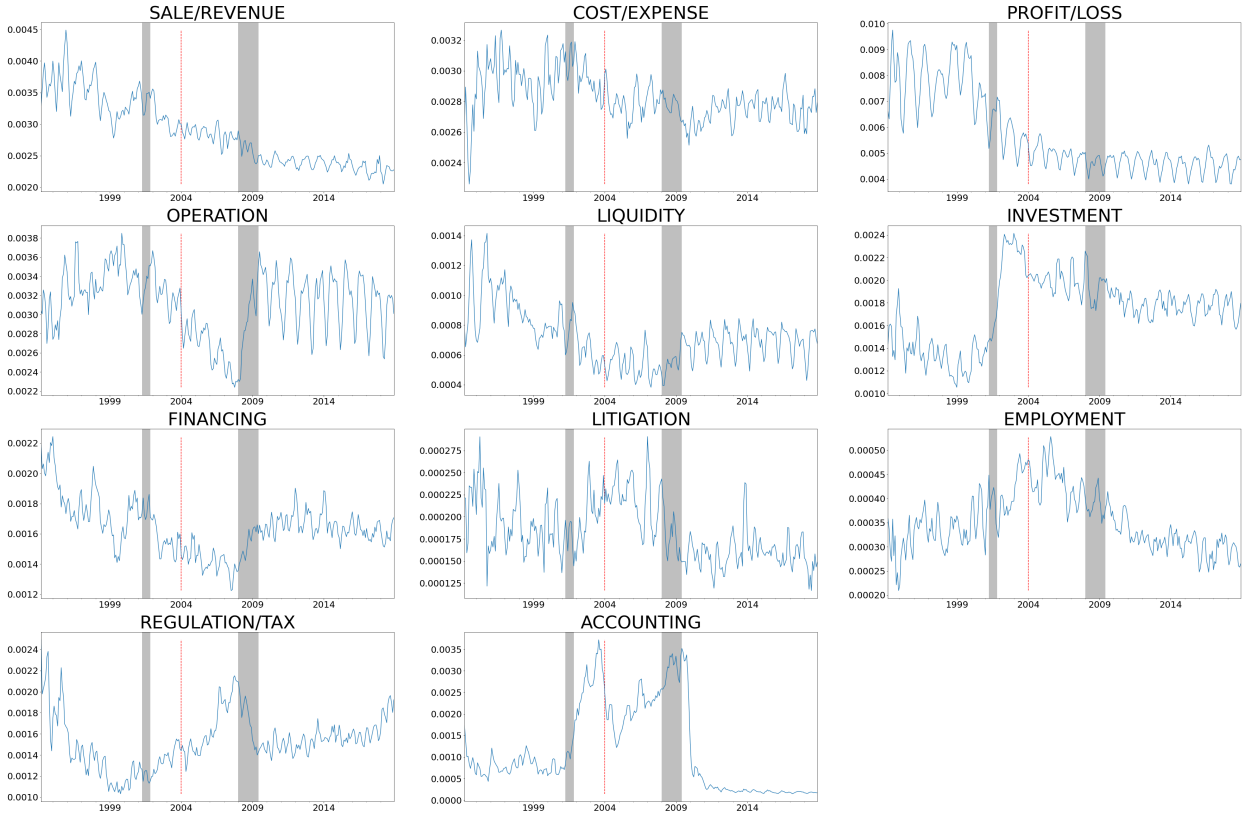

Figure 3: Monthly topic loading time series, with the cluster size of 10. These loading indices are constructed by aggregating the topic loadings of all firms that submit their 10-K filings in a given month. The time series is smoothed by a moving average over the four previous months. The vertical grey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

The topic loadings of *Regulation/Tax* decline till the dot-com crisis, but increase again before the financial crisis, reaching their peak in 2008. Apparently, the financial crisis drives down the contents relating to regulation and firm taxes sharply. After the financial crisis, the topic loadings remain at the same levels but slightly increase again in 2017-2018. There are two peaks in the loading series of the *Accounting* topic. After the dot-com crisis, firms dedicate more space to the accounting/auditing themes in their MD&A until the 2003 SEC regulation becomes effective. This trend starts in 2001, possibly in the aftermath of Enron's and WorldCom's accounting scandals and the ensuing regulatory and litigious environment (Brown and Tucker, 2011). Note that in this period, the topic loadings of *Litigation* increase visibly. In the period between the 2003 SEC regulation and the financial crisis, we observe a slight slump in *Accounting* loading. Since the financial crisis, the loadings of this topic drop sharply.

Two further remarks on the topic loading time series are in order. Firstly, the 2003 SEC regulation appears to have had limited impact on the topic contents, except perhaps the *Accounting* topic. This follows from the structural breaks in the topic loading series at the

time the guidance is enacted. Instead, the business cycle, recessions in particular, have a larger impact on the loadings to the extent that major breaks in many topics appear during or around two crises. Secondly, topics like *Sale/Revenue, Profit/Loss, Operation, Liquidity* exhibit a strong annual seasonality patterns since 2009. Given that firms release their 10-K filings at the same time every year, this cyclicity pattern suggests that the contents of these topics are similar over the years. This could be because firm managers tend to use the boilerplate language to describe firm conditions relating to these topics.[15] Therefore the contents tend to be quite persistent over years.

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 also 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 score of *Sale/Revenue* and *Operation* decline during two recessions, and they exhibit an upward trend between these two recessions. The topic *Cost/Expense* exhibits a dramatic decline during the dot-com crisis, suggesting that firms are pessimistic about their operational costs during the dot-com crisis. This trend still prolongs shortly after the crisis (until the first half of 2002), showing the consequences of the dot-com crisis persistently affects manager sentiment on the firms' operational expenses. After the sharp drop, the sentiment of *Cost/Expense* gradually recovers till the financial crisis and stay in the same level until 2018:12. Similarly, during the dot-com crisis, the management tone about firm capital expenditure and divestment activities, which are covered by *Investment*, decreases. In the financial crisis, firms express the negative tone in the topics *Liquidity* and *Financing*.

Investment, Litigation, and *Accouting* exhibit a relatively low-sentiment period between two recessions. Additionally, as shown in Figure 3, this period also witnesses a surge in the loadings of these topics as the consequence of Enron's and WorldCom's accounting scandals. This observation implies that managers tend to write more but pessimistically

[15]There are several studies and regulations with focus on the language used in the 10-K files in general and the MD&A section in particular support the hypothesis that these releases contain much boilerplate language (SEC, 2003; Brown and Tucker, 2011). However, these studies were implemented with the research sample from 1994 to 2006. To the best of our knowledge, there have been no research that directly tackles this problem in the sample from 2006 to 2018.

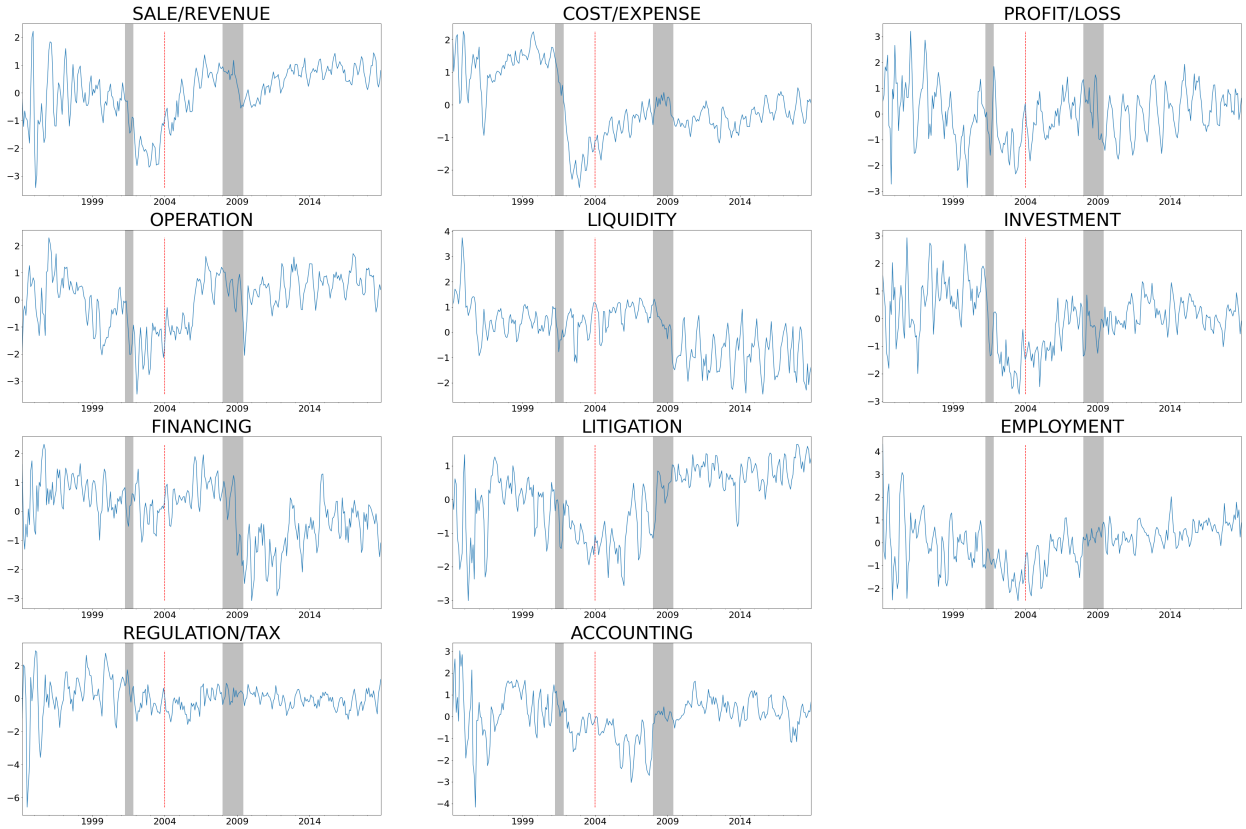

Figure 4: Monthly topic sentiment time series, with the cluster size of 10. These sentiment indices are constructed by aggregating the topic sentiment of all firms that file their 10-K filings in a given month. The time series is smoothed by a moving average over the four previous months, and standardized 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 vertical dash red lines indicate the time of 2004:01 when the 2003 SEC Guidance became effective. The sample spans from 1994:01 to 2018:12.

about the investiment, litigation, and accounting aspects of their firms during this period.

Some observations are withdrawn from the sentiment time series of the topics. Firstly, recessions affects topic sentiment differently. *Sale/Revenue* and *Operation* are the two topics being hurt by both crises. *Cost/Expense* and *Investment* are only negatively affected by the dot-com crsis, meanwhile *Liquidity* and *Financing* are only influenced by the financial crisis. Secondly, *Sale/Revenue, Profit/Loss, Operation* and *Liquidity* also exhibit a strong annual seasonality similarly to their loadings (see Firgure 3). This shows that sentimentally-charged contents regarding to these topics also consist of generic language.

6 MD&As and firm fundamentals

6.1 Variables of firm fundamentals

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

AT_TURN - Asset Turnover - The ratio between sales and averaged total assets based on the most recent two quarters. This ratio is used as an indicator showing 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 asset. In opposite, a low asset turnover ratio implies the firm is incapable of generating much sale over its assets. The topic *Sale/Revenue* conveys information about firm sales in this study. Therefore, a positive relation this ratio and the *Sale/Revenue* topic sentiment is expected.

ROA - Return on Asset - Operational income before depreciation over averaged total assets based on the most recent two quarters. This ratio represents firm performance, which is different from the *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 relation to the tone of forward-looking statements in the MD&A section. However, a negative relation between the forward-looking tone in the MD&A section and this ratio is also possible because earnings are mean-reverting. In the context of our research, *Profit/Loss* captures information about the firm current profitability. Therefore, a positive relation between ROA and the tone of this topic is expected.

ACC - Accruals - Accruals over total asset based on the most recent two periods. Firm accruals are documented to have a negative impact on firm future performances (Sloan, 1996). As pointed out by Li (2010) also, a positive relation is possible if managers try to obfuscate the MD&A content about accruals.

CAPITAL_RATIO - Capital Ratio - 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 firms. If this ratio is high, it suggests that the firm may face a solvency risk. In our set of topics, *Financing* is about debt. Therefore, the managers are expected to report the concern about the firm's solvency in this topic with a negative tone when this ratio is high, hence a negative relationship between this ratio and the tone of topic *Financing* is expected.

ME (*FIRM_SIZE*) - Market Capitalization/Equity - Logarithm of market value equity. This

is an indicator of firm size. Li (2010) suggests a negative relation between the firm size and the forward-looking tone because of the caution about political and legal costs, which aligns with the political cost theory (Watts and Zimmerman, 1986). However, according to the political power theory (Siegfried, 1972), large firms have larger political power than small firms in the sense that they can use their power to negotiate their tax burden and drive legislation to their favor (Belz et al., 2019). As a result, large firms possibly have a more positive tone in their MD&A, leading to a positive relationship between this indicator and the overall sentiment.

B/M - Book-to-Market ratio - Book-value equity over market-value equity. This ratio compares the firm value perceived by the market with the book value presented in the balance sheet. As a measure of growth opportunities, firms with high B/M ratio face less environmental uncertainties (Smith Jr and Watts, 1992), expressing less favorite forward-looking opinions in their MD&A document (Muslu et al., 2015). We expect a positive relation between this ratio and MD&A tone.

FIRM_AGE - Firm Age - The number of years since this firm's first appearance on the CRSP database. It serves as a proxy for growth options. Young (growth) firms face more economic and financial uncertainties (Anthony and Ramesh, 1992). This fact could be reflected in the MD&A section in the sense that the language of young firms tends to be more cautious. A negative relation between this factor and MD&A sentiment is expected.

Quarter dummies - Das and Shroff (2002) argue that the behavior of accounting information is different across reporting quarters. Therefore, we include quarterly dummies to investigate the potential relations between reporting quarters and the MD&A content.

Descriptive statistics of the text-related variables, topic sentiment scores and topic loading scores, and the above fundamental variables are presented in Appendix G. Due to the positive nature of the topic loadings, we normalize them to the $[0, 1]$.[16] The other variables are standardized to have zero mean and unit variance. This helps interpret the regression results.

6.2 Regression results

According to Li (2010), there are three attributes of a disclosure that attract researchers: (i) the level (how much you say, measured by the topic loadings), (ii) the tone (what do you mean, measured by the topic sentiment), and (iii) the transparency (how you say it, which is topic readability). Therefore, the relation between the disclosure level (topic loadings) and the disclosure tone (topic sentiment) with firm fundamental variables is

[16] Normalization of variable x_i to the $[0, 1]$ range is done by the formula $\frac{x_i - \min(x_i)}{\max 4 x_i - \min(x_i)}$.

worth studying. To investigate which factors drive the MD&A disclosures, we design a set of regression models in which the topic loadings and topic sentiment are regressed on some firm fundamental values controlled by year dummies. Particularly, the general regression models are,

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

$$\begin{aligned} s_{j,i,t} =& \alpha + \beta_1 AT\_TURN_{i,t} + \beta_2 ROA_{i,t} + \beta_3 ACC_{i,t} + \beta_4 CAPITAL\_RATIO_{i,t} \\ &+ \beta_5 FIRM\_SIZE_{i,t} + \beta_6 BM_{i,t} + \beta_7 FIRM\_AGE_{i,t} \\ &+ Quarter\_dummies + Sector\_dummies + Year\_dummies + \epsilon_{i,t} \end{aligned} \tag{3}$$

where $F_{j,i,t}$ and $s_{j,i,t}$ are the loading and sentiment of topic j, in firm (MD&A) i, and in year t. The dependent variables are described in detail in Section 6.1. Table 4 presents the regression results of Equation (2).

In general, firms with good performance have richer contents in the MD&A about all performance-related topics, as exhibited by the positive coefficients of *AT_TURN* and *ROA* in many regressions. In particular, firms with high sales on assets, on the one hand, tend to say more about performance-related topics, such as *Sale/Revenue, Cost/Expense, Profit/Loss, Operation*. Liquidity and financing conditions are also described more in the MD&A of these companies. On the other hand, these firms are less likely to talk about their capital-related activities (significantly negative coefficients in the *Investment* columns). With regarding the relationship between firm's profitability and the MD&A contents, our finding suggest that firms with high (low) profit, (*ROA*), talk significantly more (less) about their current sales and profitability. Firms in high financal distress risk (high *CAPITAL_RATIO*) tend to focus more MD&A contents on the investment (capital-related activities) and financing (debt- and borrowing-related activities) topics than the others.

The firm size (measure by the logarithm of market capitalization) has a negative impact on the content of the MD&A. It is significant in topics about sales, costs, profitability, operations, liquidity, financing, and accounting. Empirically, this result implies that larger firms discuss these topics less than smaller firms. However, big firms focus their MD&A content more on taxation topics than smaller firms. This finding supports the political power theory that big firms try to express their power to drive the taxation in their favor (Siegfried, 1972; Belz et al., 2019). *FIRM_AGE* has mixed effects on the MD&A topic loadings. Older (younger) firms tend to talk more (less) about their profitability, operations,

	Sale/Revenue		Cost/Expense		Profit/Loss		Operation		Liquidity		Investment		Financing		Litigation		Employment		Regulation/Tax		Accounting	
	10	15	10	15	10	15	10	15	10	15	10	15	10	15	10	15	10	15	10	15	10	15
AT_TURN	0.0065***	0.0067***	0.0028***	0.0032***	0.0062***	0.0072***	0.0019**	0.0021***	0.0042***	0.0043***	-0.0052***	-0.0067***	0.0025***	0.0019***	0.0002	0.0003	0.0018*	-0.0002	0.0005	0.0005	-0.0001	-0.0001
	(3.1734)	(3.2444)	(3.3772)	(3.6160)	(3.4180)	(3.4738)	(2.5474)	(2.7439)	(3.1331)	(3.1122)	(-3.2352)	(-3.4856)	(3.4064)	(3.5559)	(0.5012)	(0.7170)	(1.8143)	(-0.4364)	(0.7706)	(0.7422)	(-0.2570)	(-0.2124)
ROA	0.0054***	0.0056***	-0.0012	-0.0011	0.0078***	0.0088***	0.0009	0.0009	0.0014**	0.0012*	0.0015**	0.0002	-0.0003	-0.0016**	-0.0001	-0.0002	-0.0015**	-0.0004	0.0070***	0.0070***	0.0015**	0.0015**
	(3.9562)	(4.1316)	(-1.4782)	(-1.4065)	(6.6472)	(6.7519)	(1.4664)	(1.4426)	(2.1998)	(1.9143)	(2.0163)	(0.2501)	(-0.6312)	(-2.4996)	(-0.5262)	(-0.7972)	(-2.3181)	(-0.9573)	(9.8226)	(9.7746)	(2.2657)	(2.2953)
ACC	-0.0001	-3.144e-05	-0.0001	-0.0002	-0.0004	-0.0005	-0.0002	-0.0002	-0.0005*	-0.0005*	0.0010	0.0008	-0.0003	-0.0004	-0.0001	-4.871e-05	-3.516e-05	-6.193e-05	5.982e-05	5.089e-05	0.0005	0.0005
	(-0.3154)	(-0.0939)	(-0.9895)	(-1.1895)	(-0.5830)	(-0.6484)	(-0.8604)	(-0.8226)	(-1.6998)	(-1.6624)	(1.3403)	(1.0510)	(-0.9619)	(-1.1892)	(-0.9870)	(-0.3680)	(-0.1690)	(-0.5453)	(0.2952)	(0.2496)	(1.0706)	(1.0920)
CAPITAL_RATIO	-0.0057***	-0.0064***	-0.0021***	-0.0020***	-0.0025**	-0.0027***	0.0001	0.0003	0.0001	6.283e-07	0.0015***	0.0014**	0.0044***	0.0026***	7.462e-05	-6.583e-05	-0.0029***	-0.0026***	-0.0017***	-0.0017***	-0.0016**	-0.0016**
	(-6.8037)	(-7.1456)	(-3.1129)	(-3.0845)	(-2.5093)	(-2.5861)	(0.2570)	(0.6375)	(0.3038)	(0.0017)	(2.6633)	(2.0672)	(10.287)	(7.1814)	(0.2531)	(-0.2180)	(-6.0621)	(-7.3380)	(-2.6995)	(-2.6468)	(-2.2270)	(-2.2359)
FIRM_SIZE	-0.0088***	-0.0084***	-0.0044***	-0.0048***	-0.0233***	-0.0258***	-0.0103***	-0.0100***	-0.0078***	-0.0078***	-0.0026	0.0072***	-0.0040***	-0.0039***	8.297e-05	-0.0001	-0.0002	0.0002	0.0047**	0.0048**	-0.0063**	-0.0064**
	(-3.6205)	(-3.6401)	(-2.9161)	(-2.9885)	(-5.5570)	(-5.6258)	(-7.4812)	(-7.3061)	(-4.3674)	(-4.2827)	(-1.4377)	(3.0736)	(-4.7506)	(-4.3253)	(0.1389)	(-0.1824)	(-0.1422)	(0.2055)	(1.9879)	(2.0022)	(-2.2516)	(-2.2806)
B/M	-0.0003	-0.0004**	-0.0002	-0.0001	-0.0002	-0.0004	0.0002**	0.0002**	-0.0003**	-0.0003**	9.616e-05	0.0004**	-0.0002***	-0.0003***	-0.0001***	-0.0001***	0.0002***	0.0004***	0.0001**	0.0001*	-1.92e-05	1.372e-06
	(-1.4441)	(-2.0492)	(-1.2712)	(-0.7469)	(-0.6370)	(-0.9524)	(2.4239)	(2.4446)	(-2.1328)	(-2.2526)	(0.8261)	(2.0527)	(-3.9138)	(-3.6734)	(-4.7556)	(-4.5813)	(5.6223)	(7.5378)	(1.9958)	(1.9180)	(-0.3300)	(0.0238)
FIRM_AGE	-0.0026	-0.0023	-0.0084***	-0.0087***	0.0084***	0.0088***	0.0047**	0.0055***	0.0015	0.0012	0.0030***	0.0014	0.0021	-3.672e-05	0.0055***	0.0053***	-0.0020*	-0.0007	0.0006	0.0007	0.0009	0.0010
	(-1.3548)	(-1.2337)	(-7.9714)	(-7.9235)	(3.0155)	(2.6091)	(2.2884)	(2.6176)	(1.0352)	(0.8303)	(3.0455)	(0.5894)	(1.5129)	(-0.0274)	(7.1120)	(6.9282)	(-1.7349)	(-0.4754)	(0.3348)	(0.3538)	(0.5387)	(0.5724)
Q1	-0.0056	-0.0055	0.0025	0.0027	-0.0269***	-0.0287***	0.0010	0.0010	-0.0073***	-0.0072***	0.0019	0.0077*	-0.0029	-0.0013	-0.0012	-0.0013	0.0006	0.0011	0.0004	0.0004	-0.0019	-0.0019
	(-1.2100)	(-1.2298)	(1.0110)	(1.0401)	(-4.2713)	(-4.0592)	(0.4033)	(0.3804)	(-2.7728)	(-2.6587)	(0.5193)	(1.8194)	(-1.2968)	(-0.6706)	(-0.6662)	(-0.8706)	(0.3097)	(0.6724)	(0.1393)	(0.1400)	(-0.4245)	(-0.4315)
Q2	-0.0004	-0.0008	-0.0006	-0.0004	-0.0193***	-0.0192***	0.0068**	0.0069**	-0.0032	-0.0031	-7.062e-05	-0.0025	0.0012	0.0031	-0.0011	-0.0009	0.0018	0.0018	-0.0047*	-0.0047*	0.0004	0.0004
	(-0.0751)	(-0.1566)	(-0.1986)	(-0.1210)	(-3.5401)	(-3.2066)	(2.5218)	(2.5038)	(-1.3303)	(-1.3069)	(-0.0132)	(-0.4402)	(0.4819)	(1.2114)	(-0.5434)	(-0.4899)	(0.6830)	(1.2426)	(-1.6594)	(-1.6810)	(0.1141)	(0.1098)
Q3	-0.0025	-0.0035	-0.0005	-0.0001	0.0045	0.0053	0.0059**	0.0058**	-0.0017	-0.0014	-0.0029	-0.0034	0.0014	0.0023	9.862e-06	0.0001	0.0002	0.0014	0.0005	0.0005	0.0006	0.0006
	(-0.5822)	(-0.7881)	(-0.1680)	(-0.0392)	(1.1288)	(1.2197)	(2.2936)	(2.2266)	(-0.9145)	(-0.7862)	(-1.3911)	(-1.3206)	(0.7873)	(1.4261)	(0.0048)	(0.0593)	(0.0735)	(1.1360)	(0.2187)	(0.2166)	(0.1875)	(0.1809)
No. Obs	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619
R^2	0.1219	0.1405	0.1051	0.1090	0.1643	0.1666	0.0834	0.0831	0.0413	0.0436	0.0530	0.0424	0.0846	0.1021	0.0083	0.0095	0.0353	0.0437	0.0408	0.0412	0.0209	0.0215

Table 4: This table reports the Random Effect Ordinary Least Squares (RE OLS) results of regression model (2) in two cases, where the cluster size are 10 and 15 most similar words,

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

where $F_{i,t}$ is the topic loading of the topics *Sale/Revenue, Cost/Expense, Profit/Loss, Operation, Liquidity, Investment, Financing, Litigation, Employment, Regulation/Tax* and *Accounting* of firm i at time t. The set of independent variables are 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 reserve more space. The data sample spans from 1994:01 to 2018:12. $*$, $**$, and $***$ denote significance at 10%, 5% and 1% respectively.

liquidity conditions, and their legal actions, but less (more) about their expenses. Muslu et al. (2015) argue that, due to environmental uncertainties, young firms tend to express more forward-looking statements in the MD&A to assure their investors. In addition, we find that young firms provide more information about their costs/expenses. Firms with higher Book-to-Market ratio have a prospensity to write less about their liquidity, debts (financing), legal activities, but more about taxes and employment. Finally, the release time of the MD&A has minor impacts on its contents. Exceptionally, firms that release their 10-K filings in the first half of a year tend to discuss less about their profits than the other firms.

Next, we examine the determinants of the overall MD&A sentiment and topic sentiment. Tabel 5 reports the results of Equation (3), in which, we regress the MD&A sentiment scores on the corresponding firm fundamental ratios. The *Overall* column of the table shows the regression result of the overall sentiment of the MD&A documents on firm financial ratios. The coefficient on *AT_TURN* is significant at 1%, which shows that firms efficiently make more sales from their assets tend to express an optimistic tone in the MD&A sections. Firm performances (*ROA*) is positively related to the management tone (the estimated coefficient on *ROA* is significant at 1%). These results confirm the hypotheses documented in section 6.1 about the positive relationship between management sentiment and firm performances. *ACC* has no significant impact on the overall tone, yet negatively affects the tone of *Sale/Revenue* and *Accounting*. "Accruals define accounting" as stated in Oh and Penman (2020), this finding reasonably suggests that firms with the high accrual ratio tend to be pessimistic about their sale and accounting topics. Additionally, while firm managers should be aware of the firm accrual conditions, the negative correlation between this variable (*ACC*) and the MD&A overall tone hints that firm managers are well-informed and likely to understand the negative relationship between accrual components in earnings (Sloan, 1996) and truthfully report it in the MD&A documents. This result contributes to the finding of Li (2010) that not only forward-looking statements deliver the pessimism of firm managers when the firm's accrual-on-asset ratio is high, but statements expressing the current conditions also do. A high capital ratio implies a high risk of financial distress. Consistently, this situation is also reflected in the MD&A, shown by the significantly negative coefficient at the variable *CAPITAL_RATIO*. In general, according to our regression results, the firm's performance and financial conditions are correctly reflected in the MD&A via the overall tone measure. Regarding firm characteristics, big but young firms tend to express their optimism in the MD&A, and vice versa. High book-to-market firms tend to use a more negative tone in their MD&A. Finally, we find that firms that release their MD&A in the first quarter of a year like using more positive language in compiling their MD&A compared to the other firms.

	Overall	Sale/Revenue		Cost/Expense		Profit/Loss		Operation		Liquidity		Investment		Financing		Litigation		Employment		Regulation/Tax		Accounting	
		10	15	10	15	10	15	10	15	10	15	10	15	10	15	10	15	10	15	10	15	10	15
AT_TURN	0.0863***	0.0396***	0.0462***	0.0454***	0.0400***	0.0939***	0.0910***	0.0160*	0.0179**	0.0003	0.0053	0.0419***	0.0681***	0.0171*	0.0185*	-0.0006	0.0011	-0.0080	0.0008	0.0058	0.0050	0.0178**	0.0257***
	(3.4352)	(4.4575)	(4.8506)	(3.8228)	(3.7019)	(3.6314)	(3.3676)	(1.7723)	(1.9912)	(0.0420)	(0.7378)	(4.3640)	(4.7624)	(1.6979)	(1.7256)	(-0.0604)	(0.1208)	(-0.6220)	(0.0586)	(0.6736)	(0.5706)	(2.4499)	(2.9119)
ROA	0.1024***	0.0327***	0.0366***	0.0296**	0.0346**	0.0480***	0.0467***	0.0642***	0.0642***	0.0375***	0.0456***	0.0297***	0.0288***	0.0377***	0.0462***	0.0187**	0.0219**	0.0266***	0.0258***	0.0233***	0.0282***	0.0081	0.0034
	(6.2531)	(2.6774)	(3.0161)	(2.0606)	(2.4778)	(4.1067)	(3.9491)	(4.2385)	(4.2715)	(3.8355)	(4.9041)	(4.0040)	(3.2062)	(4.9077)	(5.1464)	(2.0713)	(2.4353)	(3.1398)	(3.0564)	(3.6541)	(4.3425)	(1.1036)	(0.5834)
ACC	-0.0161	-0.0381***	-0.0371**	-0.0363	-0.0344	-0.0074	-0.0075	-0.0103	-0.0130	-0.0065	-0.0039	-0.0120	-0.0139	-0.0092*	-0.0086*	0.0035	0.0015	-0.0055	-0.0053	-0.0060	-0.0027	-0.0215***	-0.0221***
	(-0.8999)	(-2.6646)	(-2.4475)	(-1.2944)	(-1.2475)	(-0.9270)	(-0.9164)	(-1.3373)	(-1.5646)	(-0.9406)	(-0.7615)	(-1.4147)	(-1.6382)	(-1.7324)	(-1.8955)	(0.4682)	(0.1903)	(-1.0684)	(-0.8756)	(-1.0814)	(-0.4241)	(-8.4983)	(-8.8473)
CAPITAL_RATIO	-0.0445***	-0.0022	-0.0013	-0.0495***	-0.0489***	-0.0420***	-0.0404***	0.0096	0.0045	-0.0180*	-0.0122	-0.0470***	-0.0535***	-0.0447***	-0.0432***	0.0189**	0.0181**	-0.0044	0.0042	-0.0133	-0.0143*	-0.0148**	-0.0125**
	(-3.7892)	(-0.2082)	(-0.1167)	(-5.8397)	(-5.5260)	(-5.1551)	(-5.1220)	(0.8585)	(0.4064)	(-1.8604)	(-1.4032)	(-4.6930)	(-5.9570)	(-5.7722)	(-5.3057)	(2.5141)	(2.2877)	(-0.6640)	(0.6031)	(-1.4645)	(-1.6861)	(-2.0388)	(-2.0231)
FIRM_SIZE	0.1511***	0.0818***	0.0799***	0.0250	0.0270	0.1441***	0.1515***	0.0817***	0.0819***	0.0410***	0.0496***	-0.0007	0.0179	0.0569***	0.0656***	0.0175	0.0113	0.0229**	0.0331***	0.0071	0.0005	0.0102	0.0400***
	(6.0081)	(6.3364)	(5.2886)	(0.9736)	(1.0903)	(10.921)	(11.864)	(4.6424)	(4.9700)	(5.0426)	(5.5505)	(-0.0532)	(1.3802)	(5.2882)	(5.4633)	(1.1913)	(0.7869)	(2.3642)	(2.9726)	(0.8370)	(0.0564)	(0.8124)	(3.2480)
B/M	-0.0041	-0.0104***	-0.0088***	-0.0067**	-0.0072**	0.0012	0.0022	0.0025***	0.0002	0.0008	0.0010*	0.0006	-0.0016	0.0011	0.0013	0.0020***	0.0006	0.0063***	0.0028***	0.0039***	0.0040***	-0.0020***	-3.882e-05
	(-1.2977)	(-6.1915)	(-4.4350)	(-2.0035)	(-2.1901)	(0.7548)	(1.3292)	(3.5555)	(0.2029)	(1.1629)	(1.6691)	(0.5176)	(-0.9961)	(1.1611)	(1.1989)	(3.0010)	(0.6887)	(13.524)	(4.7748)	(9.5997)	(7.1568)	(-4.2500)	(-0.1267)
FIRM_AGE	-0.0433**	-0.0450**	-0.0462**	-0.0569***	-0.0575***	0.0108	0.0091	-0.0203*	-0.0225**	0.0225**	0.0216*	-0.0494***	-0.0521***	0.0201	0.0303**	-0.0434***	-0.0484***	-0.0033	-0.0034	-0.0128	-0.0205**	-0.0320***	-0.0344***
	(-2.1323)	(-2.1739)	(-2.2753)	(-3.0943)	(-3.3483)	(0.4943)	(0.4426)	(-1.7312)	(-2.2480)	(2.2570)	(1.9293)	(-3.6125)	(-3.8587)	(1.5776)	(2.4812)	(-2.9482)	(-3.4897)	(-0.3034)	(-0.3327)	(-1.4090)	(-2.4163)	(-3.2706)	(-2.6187)
Q1	0.0057	0.0538	0.0429	0.0539	0.0544	-0.0675**	-0.0595*	0.0406	0.0584	0.0384	0.0128	0.0626*	0.0198	0.0357	0.0413	0.0083	0.0387	0.0060	-0.0005	-0.0132	-0.0068	0.0856**	0.0614**
	(0.0846)	(1.4369)	(1.0347)	(1.1550)	(1.2057)	(-2.1579)	(-1.8972)	(0.8503)	(1.2622)	(0.8417)	(0.2603)	(1.8631)	(0.5386)	(1.0330)	(1.1073)	(0.2682)	(1.1107)	(0.1886)	(-0.0135)	(-0.6173)	(-0.2533)	(2.2131)	(2.0067)
Q2	-0.0645	0.0031	0.0006	0.0550	0.0534	-0.0503	-0.0518	-0.0604	-0.0530	-0.0110	-0.0098	0.0398	0.0280	0.0072	-0.0036	-0.0119	0.0102	-0.0257	-0.0296	-0.0372	-0.0355	0.0301	0.0096
	(-0.9233)	(0.1044)	(0.0170)	(0.7473)	(0.7109)	(-1.4520)	(-1.4862)	(-1.5749)	(-1.5045)	(-0.2089)	(-0.1698)	(0.8038)	(0.5748)	(0.2041)	(-0.0889)	(-0.2766)	(0.2045)	(-0.7553)	(-0.7550)	(-1.3194)	(-1.1319)	(0.5575)	(0.2206)
Q3	-0.0294	0.0830**	0.0621*	0.0761**	0.0770**	-0.0368	-0.0249	-0.0297	-0.0238	-0.0771	-0.1453**	0.0037	0.0119	0.0223	0.0155	-0.0289	-0.0404	-0.0260	-0.0106	-0.0282	-0.0342	-0.0178	0.0061
No. Obs	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619	49619
R^2	0.0290	0.0075	0.0078	0.0069	0.0068	0.0195	0.0198	0.0086	0.0090	0.0030	0.0040	0.0038	0.0060	0.0047	0.0063	0.0010	0.0013	0.0010	0.0013	0.0007	0.0009	0.0018	0.0021

Table 5: This table reports the Random Effect Ordinary Least Squares (RE OLS) results of regression model (3) in two cases, where the cluster size are 10 and 15 most similar words,

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

where $s_{i,t}$ is the overall sentiment (the sentiment scores of an entire MD&A document) topic sentiment of the topics *Sale/Revenue, Cost/Expense, Profit/Loss, Operation, Liquidity, Investment, Financing, Litigation, Employment, Regulation/Tax* and *Accounting* of firm i at time t. The set of independent variables are 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 from 1994:01 to 2018:12. $*$, $**$, and $***$ denote significance at 10%, 5% and 1% respectively.

We also present the regression results of the topic sentiment scores on firm fundamental ratios in Table 5. The column names corresponding to the topics, while the 10- and 15-columns present the regression results in two cases of the cluster size, 10 and 15 words. Firstly, the regression results in the 10- and 15- columns of each topic are similar in both values and signs, implying that our topic model is robust to the choice of the cluster size. By digging into the details of the regression results, we can observe that *Sale/Revenue, Cost/Expense, Profit/Loss*, and *Financing* are the top four topics that relate to the firm fundamental variables the most, based on the R^2 and the number of significant coefficients. Firms with a good amount of sales over their total assets show their positive tone regarding not only sale/revenue topics in the MD&A, but also expenses, profitability, operation, liquidity, and investment topics. We find a strongly significant (at 1% confidence level) relation between the earning ratio (*ROA*) and the *Profit/Loss* topic sentiment. Empirically, firms with the high earning ratio are also optimistic about their profitability. We find a significant relation between the capital ratio and firm's capital and debt structures. Accordingly, this factor has negative relations to the tone of *Investment* (significant at 1%), and *Financing*, which covers the firm's credit and debt situations (significant at 1%). This finding hints that the firm managers truthfully describe the financial distress conditions in the MD&A.

Firm size and age are also important determinants of topic sentiment in the MD&A. However, they have quite opposite effects on the topics. While firm size has positive relations to the topic sentiments, the counterparts of the firm age are mostly negative. Big firms tend to be optimistic about their revenues, expenses, profitability, operations, meanwhile, long-existing firms often exhibit a negative tone about these topics in the MD&A. The book-to-market ratio has mixed effects on the topic sentiment. Firms with a high book-to-market ratio are more pessimistic about their sales and costs, but optimistic about their employment, taxes, and the perspective on the regulatory environment. Finally, the filing time has no clear effect on the sentiment expressions of firm managers in the MD&As.

7 Conclusion

This paper introduces a topic model with applications in the MD&A section in 10-K filings. To this end, we apply the concept of *anchor words*, which carry the prior insights about the MD&A contents, and utilize the cosine similarity to search for the similar words in the word vector space. The word vectors are derived by the state-of-the-art *Word2vec* model, which are shown to preserve the semantics of words in the vocabulary. To achieve the topic loading, which is a measure of how much a topic is discussed in a MD&A document, we apply the Singular Value Decomposition as a tool to summarize the information compacted in the topic-wise document-term matrices. The topic sentiment is

implemented in the topic-specific manner using the Loughran-McDonald dictionary.

The topics produced by our proposed model exhibit an interpretation that is close to human sense in terms of financial context. With the complication of the language used in the financial applications, our model succesfully categorizes words with high uncertainty in the correct topics (e.g. "income_tax" belongs to topic *Regulation/Tax* and "income_earn" belongs to topic *Profit/Loss*). The time series of topic loadings and sentiment estimated by our model provide some understandings about time variation of major economic events. We find the MD&A contents and tones are determined by firm fundamentals, current performances (sales, incomes), accruals, debt ratio, book-to-market ratio, firm size and firm age. Importantly, well-performed firms tend to express an optimistic tone in their MD&A, while firms with a high accrual-on-asset ratio tend to be pessimistic about their sale and accountings topics. These findings suggest that firm managers tend to report truthfully firm conditions in their textual releases.

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.

Bellstam, G., Bhagat, S. and Cookson, J. A. (2019). A text-based analysis of corporate innovation, *Available at SSRN 2803232* .

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.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2003). A neural probabilistic language model, *The Journal of Machine Learning Research* **3**(Feb): 1137–1155.

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.

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.

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.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning, *Proceedings of the 25th International Conference on Machine learning*, pp. 160–167.

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. and Blei, D. M. (2020). Topic modeling in embedding spaces, *Transactions of the Association for Computational Linguistics* **8**: 439–453.

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.

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

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

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.

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. D., Blei, D. M., Wang, C. and Paisley, J. (2013). Stochastic variational inference, *The Journal of Machine Learning Research* **14**(1): 1303–1347.

Jegadeesh, N. and Wu, D. A. (2017). Deciphering Fedspeak: The information content of FOMC meetings, *Available at SSRN 2939937* .

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. (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.

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.

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

Mitchell, T. M. et al. (1997). Machine learning. 1997, *Burr Ridge, IL: McGraw Hill* **45**(37): 870–877.

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.

Oh, H. I. and Penman, S. H. (2020). The accruals-cash flow relation and the evaluation of accrual accounting, *Available at SSRN 3616871* .

Price, S. M., Doran, J. S., Peterson, D. R. and Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone, *Journal of Banking & Finance* **36**(4): 992–1011.

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.

Schwenk, H. (2007). Continuous space language models, *Computer Speech & Language* **21**(3): 492–518.

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.

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.

Watts, R. L. and Zimmerman, J. L. (1986). Positive accounting theory.

A Building blocks

A.1 The phrase-learning model

Single words cannot fully deliver the intentions of speakers or writers. This statement is especially popular in domain-specific tasks in general and in the economic and financial world in particular when many terminologies have to be expressed by compound words (i.e. words are constructed by more than one single word). For example, the phrase "consumer demand" will not be fully interpreted and replaced by any single words, "consumer" and "demand" alone can not describe comprehensively that word. Therefore, learning phrases[17] in economic and financial textual applications is crucial before thinking about further complicated steps.

To automatically detect phrases in the MD&A corpus of documents, we use a data-driven method described in (Mikolov, Sutskever, Chen, Corrado and Dean, 2013). The main purpose of this method is to compare the frequency of a pair of words when they appear together and when they appear separately in the corpus. More formally, the model computes the following score of two words i and j,

$$s_{ij} = \frac{(\#(\text{word}_i, \text{word}_j) - m))|W|}{\#(\text{word}_i)\#(\text{word}_j)}$$

in which, $\#(\text{word}_i, \text{word}_j)$ is the count that two words appear adjacently in the corpus; $\#(\text{word}_i)$ is the count of word i in the corpus; W is the vocabulary created from the corpus, then $|W|$ is the number of words in the vocabulary; m is the minimum counts of words that we consider to include in the vocabulary, words appear less than this threshold will be excluded out of the vocabulary, this parameter plays a role as a hyperparameter and will be defined before the training phase. Finally, s_{ij}, along with a threshold s, is the quantity of interest from which we determine whether a bigram[18] is a phrase or not. In particular, the threshold s is a hyperparameter such that a pair of two words i and j is considered as a phrase if $s_{ij} > s$. The higher (lower) the threshold is, the less (more) likely a bigram is considered as a phrase, but the more (less) likely high-frequency words get in the place, adding more bias to model results.

[17] To mention about "phrase", we mean a compound word

[18] n-gram is the collection of n words that appear adjacently, a bigram is an n-gram with $n = 2$.

A.2 The Word2vec model

The mission of NLP is to understand the meanings of words and to represent words in a way that computers can understand. Conventional NLP methods try to encode words by their occurrences in the corpus. This way is the so-called "count-based" method and suffers many limitations. Firstly, this way of word representation leads to a very high-dimensional but sparse and noisy representation. Secondly and more importantly, the huge limitation of this approach is that, with the way it treats words as independent tokens, it overlooks the similarity and correlation among words by the context.[19] Moreover, by using n-gram structure of a text, one mitigates the lack of semantic interpretation of count-based approaches, yet also increases the dimensionality of word representation, thus hindering its scalability and efficiency compared to a neural-network-based model (Bengio et al., 2003; Schwenk, 2007).

Recently, a neural-network-based model named *Word2vec*[20], in which words are embedded into a dense and low-dimensional space, is introduced and completed by various authors (Bengio et al., 2003; Collobert and Weston, 2008; Mikolov, Sutskever, Chen, Corrado and Dean, 2013; Mikolov, Chen, Corrado and Dean, 2013). To be more concrete, each word is a d-dimensional vector of real numbers, in which two vectors lie close to each other in the vector space are semantically similar. In addition to significant shrinkage of the number of dimensions of word representation, this method is exhibited empirically to achieve superior performances in various linguistic tasks compared to count-based methods. The result from this model is a matrix that consists of word vectors in its rows. The main idea of this model is that it maximizes the similarity of two words that appear in the context of a certain window (e.g, 5 words on both sides of a word), and at the same time, minimizes the similarity of words that do not appear together. As a result, word vectors produced by this model will capture the similarity of words that usually appear in the same context. It is worth noticing that (Mikolov, Chen, Corrado and Dean, 2013) propose several extensions of this model to increase the training speed and quality of the word vectors via *subsampling* and *negative sampling*, allowing the algorithm to train a vocabulary of billions of distinct words.

Due to its success in capturing word semantics, we decide to adopt this neural-network-based method to learn the word vector representation used for later purposes. To embed the financial context of words, we train the Word2vec model directly on the MD&A corpus after several preprocessing steps. Both Word2vec model and the phrase-learning

[19] This is the concept of the distributional hypothesis that words with similar distributions deliver similar meanings.

[20] Indeed, Word2vec is the name of a class of models that turn words into dense vectors. However, in this paper, we will refer to Word2vec as the model presented in (Mikolov, Sutskever, Chen, Corrado and Dean, 2013).

model are trained using the package *gensim*, an open-source Python library with C++ backend, developed by (Řehůřek and Sojka, 2010) that allows users to train modern NLP models.

B Construction of the anchor word lists via the phrase-learning model

This section elaborates the process that the anchor word lists are constructed from the original word lists of Li (2010) and the phrase-learning model. We first tokenize all terms in each topic in Table 2.[21] Subsequently, a set of thresholds is proposed as $s \in \{0.01, 0.05, 1, 2, 5\}$, and the threshold that allows the model to recover most terms in Table 2 will be chosen. Indeed, we run the phrase-learning model in our corpus twice to detect trigrams and quadrigrams in the corpus. For our experiments, the optimal threshold is $s = 0.01$ in both running times since it reproduces most words from the tokenized anchor word lists. With the optimal threshold, the phrase-learning model is applied to the entire corpus of MD&A documents to detect phrases, and to the word lists in the table 2 to reproduce them. It is worth noting that there are several phrases in the initial anchor word lists which the model fails to detect. In this case, we decide which terms to be kept in the lists. For example, the model is unable to detect the phrase "market position", and two words "market" and "position" alone cannot fully deliver the meaning of the topic about *sale*. Consequently, we decide to exclude both of the individual words from the word list. On the other hand, for the phrase "reserve for contingent liability", the model merely detects the phrase "contingent liability" but not the whole phrase. In this case, the phrase "contingent liability" still delivers the context of the topic about *cost*. Therefore, we decide to keep that phrase in the word list of topic 2. Other phrases that are treated similarly 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 final results of this step are (i) the recovered anchor word lists described in Table 6, and (ii) the new corpus that contains potential words and phrases (instead of the tokenized corpus with all single words).

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

Topic 1: Sale	Topic 2: Cost	Topic 3: Profit	Topic 4: Operation	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 optimal threshold of 0.01. The process to reproduce these lists of anchor words is described in detail in the section 2.1.

C Topic coherence-coverage trade-off

The data-driven determination of the optimal cluster size lies on two aspects, (i) topic coherence[22] and (ii) 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 represented words. Because a lot of machine-learning-based topic models provide no guarantees about 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). Newman et al. (2010) investigate the performance of several topic coherence measures and conclude that the *Pointwise Mutual Information* (PMI)[23] is the best-performing approach to estimate topic coherence in the sense that this measure has the closest Spearman correlation to the human-judged measurement. To represent the coherence of a topic, Newman et al. (2010) aggregate PMI scores of all pairs of two words in the topic word list by mean or median. Experiments suggest that the mean and the median aggregation give insignificantly different results. In our paper, the Normalized-PMI (NPMI)[24] is used instead of PMI to obtain the similar scale to the second criterion, the topic coverage. It is easily observed that a large (smaller) cluster size allows more (fewer) words to be considered as similar to the anchor words, thus leading to longer (shorter) topic word lists. Additionally, a longer topic word list will have a lower topic coherence due to more variant words being included (Lau and Baldwin, 2016). Therefore, a larger (smaller) cluster size results in lower (higher) topic coherence. The topic coherence, in our model, is measured by the averaged NPMI scores of topic word pairs,

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

in which Coh_j^k is the topic coherence of topic j; T_j^k denotes the set of words in topic j and $|T_j^k|$ is the number of topic words in that topic. The superscript k denotes that the quantity depends on the cluster size k.

The latter, topic coverage, measures the probability of a topic given a corpus. Intuitively, high topic coverage indicates a richer content of that topic given the corpus. In LDA, the topic coverage is estimated by the ratio between the total topic word count to the total word count of all topics modified by a set of smoothing coefficients. These coefficients are the parameters of the Dirichlet prior distribution (Hoffman et al., 2013). However, with

[22] In philosophy, one of the theoretical definitions about *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).

[23] The PMI score of two words w_m, w_n is computed as $\mathrm{PMI}(w_m, w_n) = \log \frac{p(w_m, w_n)}{p(w_m)p(w_n)}$, where $p(w_m, w_n)$ is the probability that two words appear together, $p(w_m)$ is the probability of word w_m.

[24] The NPMI score of two words w_m, w_n is computed as $\mathrm{NPMI}(w_m, w_n) = \frac{\mathrm{PMI}(w_m, w_n)}{-\log p(w_m, w_n)}$.

this measurement, one can only measure the ratios between considered topics, which only makes sense under an assumption that a document is covered entirely by considered topics but not noise.[25] Meanwhile, our proposed model assumes that a document comprises the considered topics and noise. Therefore, using the topic coverage measure from the LDA model for our purpose is inappropriate. Consequently, we suggest computing the topic coverage measure as the ratio between total topic word count and total word count. Therefore, a list of only a few topic words, which is more likely to have lower topic word counts, will not be able to cover a topic comprehensively. Formally,

$$Cov_j^k = \frac{C_j^k}{C}$$

in which, 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; C is the total word count of the entire MD&A corpus; T indicates the set of topics, thus $|T|$ is the number of pre-defined topics, which is 11 in our case. With this topic coverage measure, longer (shorter) topic word lists will have wider topic coverage, because more words are taken into account.

As a result of the above reasonings, a larger (smaller) cluster size leads to more (fewer) words being included in the topic word lists, resulting in higher (lower) topic coherence and lower (higher) topic coverage. Therefore, by choosing the optimal cluster size, we aim at balancing the topic coherence/topic coverage trade-off. Because our main target in this step is to form topics that are highly coherent and cover as much information of the corpus as possible, the optimal cluster size is expected to maximize the following quantity,

$$G^k = \sqrt{\sum_{j=1}^{|T|} Coh_j^k \times \sum_{j=1}^{|T|} Cov_j^k} \tag{4}$$

With each k, we proceed to formulate the optimal topic word lists by searching for the similar words of the anchor words in each topic. The topic coherence, topic coverage measures as well as 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 tuning this hyperparameter. Figure 5 showcases the trade-off between topic coherence and topic coverage with varying values of the cluster size. The optimal value of this hyperparameter is 10 which

[25]The coverage of topic j in document d in the LDA model is measure by $Cov_{j,d} = \frac{C_{j,d}+\delta_j}{\sum_{t=1}^{|T|} C_{t,d}+\delta_t}$, where δ_t is the smoothing coefficient corresponding to topic t. Therefore, the topic coverage of a document is always summed up to 1, meaning that a document is always covered entirely by consider topics but not noise.

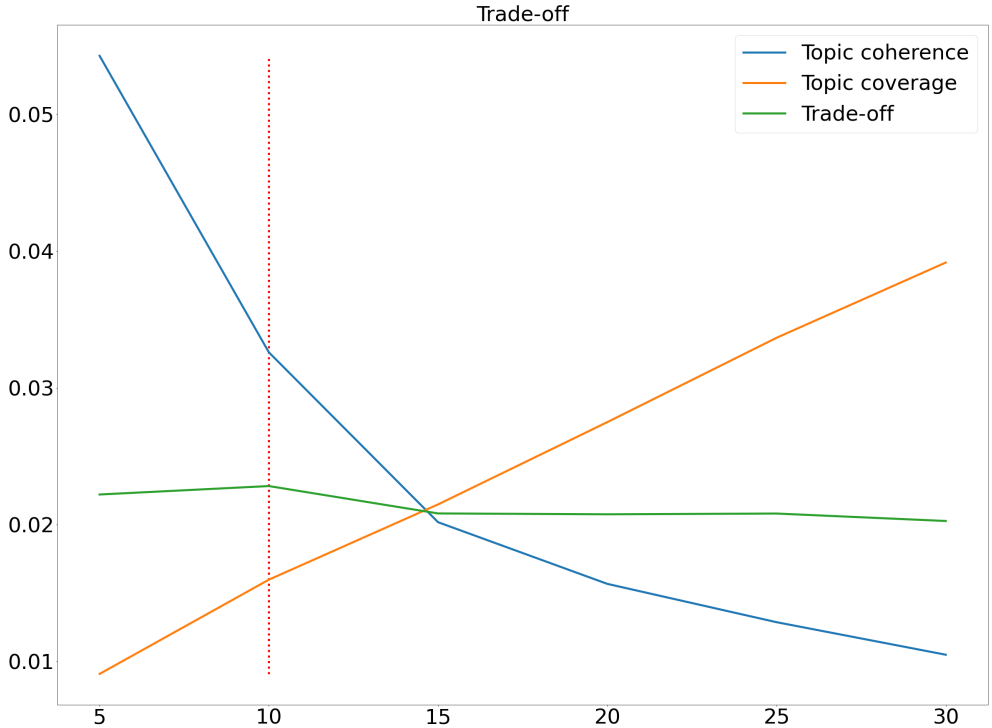

Figure 5: The behavior of topic coverage, topic coherence measures and the trade-off quantity, G_k, with different cluster sizes $k \in \{5, 10, 15, 20, 25, 30\}$. The blue and orange lines correspond to the arithmetic averages of topic coverage and topic coherence over all topics. The value of G_k is depicted 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 $k \in \{5, 10, 15, 20, 25, 30\}$.

means that for each anchor word, we choose the top 10 closest words in the vocabulary to that anchor word. The closeness is determined by the cosine similarity measure. To ensure the separability of the topics, which is a considered property in topic modeling (Arora et al., 2012; Cong et al., 2019), for words that appear in more than one topic, we choose the topic that is closest to these overlapping words. The full topic word lists are reported in Table 7, in Appendix D.

D Topic word lists

	Topic words
Sale/Revenue	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	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	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
Operation	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	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	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	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	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	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	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	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.

E Topic loading estimation by Singular Value Decomposition

Technically, define for each topic j a $|D| \times |T_j|$ document-term matrix A_j with $j = 1, ..., |T|$ ($|T|$ is 11 in our case). This matrix contains word counts of all topic words of the topic j for all MD&A documents. Consequently, each row of the matrix is a vector in the $|T_j|$-dimensional space. By applying the SVD to this document-term matrix, we receive,

$$A_j = U_j \Sigma_j V_j^\top$$

in which U_j and V_j have orthonormal columns of singular vectors, and Σ_j is a diagonal matrix with positive real entries of singular values. According to Theorem 3.1 in Blum et al. (2020), the first k columns of the matrix V_j formulate the best-fit[26] k-dimensional subspace of the matrix N_j. Subsequently, the first right singular vector, $v_j^{(1)}$, which is the first column of the matrix V_j, will be the optimal 1-dimensional subspace that captures most information of the initial document-term matrix N_j. Therefore, this first right singular vector, $v_j^{(1)}$, could capture the importance of topic words in a topic, j.

Due to SEC regulations that firms have to provide more details in the MD&A disclosures, the length of MD&A documents increases over time (Brown and Tucker, 2011). To account for the heterogeneity in document length, we, therefore, normalize the document-term matrix A_j with the length of each segment, i.e. divide each row of A_j by the total word count in that document to obtain matrix $\tilde{A}_j$. Eventually, the optimal summary of information from the initial matrix $\tilde{A}_j$ in a 1-dimensional space could be calculated by the length of the projection of the initial document-term matrix on that optimal subspace, $v_j^{(1)}$, which is formally expressed as,

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

in which, F_j is an $|S|$-dimensional vector whose entries are the loadings of topic j on all MD&A documents in the corpus; $X^{|\cdot|}$ is the element-wise absolute-value operator of the matrix X. Because the matrix $\tilde{A}_j$ is the document-term matrix, thus its entries are non-negative. As a result, according to Perron–Frobenius theorem[27], all elements of $v_t^{(1)}$ have the same sign and the absolute norm is added to convert them into positive numbers. We

[26] The best-fit line is the best line to fit a set of data points by minimizing the L_2 norm of the errors.

[27] The Perron–Frobenius theorem states that a square matrix with positive real entries has a unique largest real eigenvalue and the corresponding eigenvector can always be chosen to have positive entries. In our case, because matrix V_j contains eigenvectors of the matrix $\tilde{A}_j$, which always has non-negative entries, and the first singular value $v_j^{(1)}$ corresponds to the largest eigenvalue, entries of $v_j^{(1)}$ always have identical signs.

proceed with this process overall topics. These vectors F_j with $j = 1, ..., 11$ are used for later analyses.

F Text processing

Firstly, we implement two main textual normalization steps which include, (i) replacing contractions, e.g., converting "don't" to "do not", etc., (ii) removing noise, i.e., single letters, numbers, special characters, punctuations[28], multiple whitespaces, and breaklines. Secondly, we also handle negations carefully. In particular, words in a document that appear in the LM sentiment dictionary are identified. We only take care of these words because they are considered in the tone estimation step. After that, we further determine whether there are negation terms appearing around these sentimentally charged words within a certain window. If sentimentally charged words appear together with a negation term within the considered window, the word is added the "not_" prefix. 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, however with the negation "not" locates in its front, the sentence is transformed to "The Company has **not_defaulted** on any of its obligations since its founding in 1990". Accordingly, the new term **not_defaulted** is added into the positive word list of the LM dictionary. The negation terms which are considered in our implementation are "not", "no", "none", "neither", "nor", and "never". The length of the window around the sentimentally charged words is 5 words on either sides of the considered word.

After handling negation, we further discard stopwords that appear dominantly in the texts but have negligible meanings. We use the stopword lists provided by Loughran-McDonald (LM stopwords) for this process. These lists differ from the stopword list provided in the *nltk* Python library (Bird et al., 2009) in the sense that the LM stopword lists are specifically designed for financial applications and more detailed. It should be noted that in the LM stopwords about names, there are words "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.", "Ms." out of our vocabulary. Finally, texts are lemmatized to remove the ineffectual endings of words. We did not stem words to preserve the meanings of words within a word family.

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

After the above-mentioned normalization steps, the MD&A documents are employed by the phrase-learning model and the Word2vec model. The documents are split into sentences before inputting them into the two models to prevent information of a sentence from spilling over to nearby sentences when training these two models.[29] For training the phrase-learning and Word2vec models, we discard words that appear in less than 15 documents (sentences). Consequently, the vocabulary built from our corpus has $1,884,140$ words. The negative sampling parameter is 15, the context window is 5, and 30 iterations are used for the Stochastic Gradient Descent optimizer.

G Descriptive statistics

Variables	Count	Mean	Std	Min	25%	Median	75%	Max
Sentiment								
Overall	49619	-0.0042	0.00495	-0.05044	-0.00700	-0.00381	-0.00112	0.04018
Sale/Revenue	49619	-0.00030	0.00096	-0.01980	-0.00070	0.0	0.0	0.01539
Cost/Expense	49619	-0.00075	0.00106	-0.01626	-0.00126	-0.00056	0.0	0.01111
Profit/Loss	49619	-0.00116	0.00210	-0.02326	-0.00211	-0.00088	0.0	0.04000
Operation	49619	-0.00033	0.00089	-0.01980	-0.0006	0.0	0.0	0.00971
Liquidity	49619	-1.8e-05	0.00022	-0.00592	0.0	0.0	0.0	0.00761
Investment	49619	-0.00022	0.00070	-0.01887	-0.00039	0.0	0.0	0.01333
Financing	49619	-0.00012	0.00055	-0.01887	-0.00029	0.0	0.0	0.01471
Litigation	49619	-0.00027	0.00067	-0.02518	-0.00033	0.0	0.0	0.00395
Employment	49619	-3.7e-05	0.00034	-0.01070	0.0	0.0	0.0	0.01205
Regulation/Tax	49619	-2.5e-05	0.00052	-0.00658	0.0	0.0	0.0	0.01613
Accounting	49619	-0.00014	0.00049	-0.01197	-0.00027	0.0	0.0	0.01333
Loadings								
Overall	49619	0.00297	0.00164	0.0	0.00182	0.00272	0.00381	0.01941
Sale/Revenue	49619	0.00288	0.00149	0.0	0.00186	0.00264	0.00362	0.01988
Cost/Expense	49619	0.00491	0.00269	0.0	0.00300	0.00420	0.00620	0.03329
Profit/Loss	49619	0.00256	0.00166	0.0	0.00142	0.00223	0.00331	0.02702
Operation	49619	0.00050	0.00071	0.0	0.0	0.00027	0.00070	0.01415
Liquidity	49619	0.00185	0.00120	0.0	0.00100	0.00168	0.00250	0.02364
Investment	49619	0.00137	0.00088	0.0	0.00077	0.00121	0.00176	0.01904
Financing	49619	0.00019	0.00034	0.0	0.0	2e-05	0.00026	0.00808
Litigation	49619	0.00038	0.00045	0.0	6.5e-05	0.00027	0.00054	0.00995

[29]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 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 begining of the second sentence, which is improperly handled. As a result, we first split segments into sentences before training the phrase-learning and Word2vec models, then each sentence now serves as a separate document.

Variables	Count	Mean	Std	Min	25%	Median	75%	Max
Employment	49619	0.00183	0.00136	0.0	0.00087	0.00162	0.00252	0.02334
Regulation/Tax	49619	0.00118	0.00198	0.0	0.00010	0.00021	0.00150	0.04648
Accounting	49619	0.00328	0.00162	0.0	0.00213	0.00305	0.00414	0.01840
Firm fundamentals								
AT_TURN	49619	0.97607	0.94943	0.0	0.36900	0.80200	1.31400	43.7420
ROA	49619	0.06230	0.20504	-4.33800	0.02100	0.09300	0.15800	3.68700
ACC	49619	0.06513	0.23339	-3.01500	0.00700	0.04600	0.09800	28.9860
CAPITAL_RATIO	49619	0.26933	0.27243	-4.27700	0.00500	0.21500	0.45450	4.87100
FIRM_SIZE	49619	5.95531	1.97936	0.45882	4.50988	5.90920	7.25765	13.6284
B/M	49619	0.90413	32.7176	0.0	0.29500	0.52900	0.85700	5152.55
FIRM_AGE	49619	21.6485	13.7870	0.44110	11.06027	19.2110	28.5178	58.9562
Q1	49619	0.71410	0.45185	0.0	0.0	1.0	1.0	1.0
Q2	49619	0.09539	0.29375	0.0	0.0	0.0	0.0	1.0
Q3	49619	0.08847	0.28399	0.0	0.0	0.0	0.0	1.0

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 a cluster size of 10.