Sentiment-semantic word vectors: A new method to estimate management sentiment

Tri Minh Phan*

Version: February 15, 2024

Abstract

This paper investigates the stock return predictability in relation to the Management's Discussion and Analysis (MD&A) section of 10-K filings of US firms from January 1994 to December 2018. Based on innovations in Natural Language Processing and statistical learning, we introduce a novel method to extract the sentiment embedded in the MD&A section. We find that our method outperforms traditional approaches in terms of sentiment classification accuracy. Utilizing this method, the MD&A sentiment is found to be a strong negative predictor of future stock returns, demonstrating consistency in both in-sample and out-of-sample settings. Notably, with traditional sentiment extraction methods, the MD&A sentiment exhibits no predictive ability to stock markets. This finding underlines the MD&A section as an important source for stock return prediction, providing an accurate sentiment analysis method. Additionally, we examine the stock return predictability of the MD&A sentiment in conjunction with macroeconomic variables. This examination reveals that the MD&A sentiment is associated with dividend-related macroeconomic channels regarding future stock return prediction.

Keywords: knowledge distillation, MD&A, stock returns, Word2Vec

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

1 Introduction

Serving as the main focus of numerous studies, the Management's Discussion and Analysis (MD&A) section is undoubtedly one of the most important parts of the 10-K/Q filings (Tavcar, 1998; Li, 2010; Feldman et al., 2010; Brown and Tucker, 2011; Loughran and McDonald, 2011; Davis and Tama-Sweet, 2012; Bochkay and Levine, 2019; Cohen et al., 2020).[1] It purports to "...provide investors and other users with material information that is necessary to an understanding of the company's financial condition and operating performance, as well as its prospects for the future" (SEC, 2003, Chapter III.B, p. 75059). According to this scenario, it is natural to expect the MD&A section to encapsulate insights that potentially influence the stock market dynamics. Surprisingly, few studies are exploring the predictive power of the MD&A section to future stock returns.

In this paper, we explore stock return predictability using sentiment derived solely from the MD&A section, voiced as *management sentiment*.[2] In particular, we investigate a behavioral implication of management sentiment in asset pricing which hypothesizes that misleading sentimental information in corporate disclosures is absorbed by investors, leading to overvaluation in stock prices. When the true stock fundamentals are gradually disclosed to the public, the prices reverse, hence implying management sentiment negatively predicts future stock returns in the long run. This hypothesis is theoretically modeled by De Long et al. (1990) and empirically confirmed by Jiang et al. (2019) using 10-K/Q filings and conference calls to represent management sentiment. However, 10-K/Q filings mix with informative and boilerplate contents (Li et al., 2010). As a valuable part of 10-K/Q filings (Tavcar, 1998), whether the stand-alone sentiment in the MD&A section is predictive of future stock returns remains an open question.

We construct a management sentiment index from the MD&A section of 10-K filings using a word-representing model with novel adaptations. Specifically, we introduce a method integrating both word sentiment and semantics into a pre-defined set of word representations (i.e., vectors), resulting in another set absorbing both sentiment and semantic connotations. To achieve this target, our method relies on three components: (i) a word representation model embracing rich word semantics, which is pre-trained with a massive dataset; (ii) a Knowledge Distillation technique (Hinton et al., 2015); and (iii) a dataset with sentiment labels. The first component acts as a "semantic anchor" for the word vectors, while the second component seeks to infuse the sentiment meanings, carried by the third component, into these vectors. Intuitively, the obtained word vectors inherit word semantics from a pre-trained word representation model and, simultaneously, absorb nuanced sentiment information from a labeled dataset.

[1] "10-K/Q filings" in the context of this paper means "10-K and 10-Q filings".

[2] We refer to "sentiment" as the polarity of tone, i.e., negativity (pessimism), neutrality, and positivity (optimism) with the awareness of a strand of literature that prefers using the word "tone".

First, our proposed approach successfully obtains a new set of word vectors that captures both word sentiment and semantics, which is henceforth referred to as the *sentiment-semantic word vectors*. By a word-level sentiment classification, the sentiment-semantic word vectors outperform another set of word vectors carrying only word semantics, termed as *semantic-only word vectors*, in clustering words into sentiment categories. Furthermore, our sentiment-semantic word vectors demonstrate a superior capability in document sentiment classification over the competing methods, including the semantic-only word vectors and the Loughran-McDonald dictionary (Loughran and McDonald, 2011). In particular, the sentiment-semantic word vectors achieve a 0.68 F1 score in a sentiment classification task using the Financial Phrasebank dataset (Malo et al., 2014). Meanwhile, the corresponding quantities of the Loughran-McDonald dictionary, (ignoring word semantics) and the semantic-only word vectors (ignoring word sentiment) are 0.58 and 0.64. These findings underscore the importance of integrating both sentiment and semantic information into word vectors for accurate sentiment analysis.

Second, the variations of our management sentiment index, constructed by the sentiment-semantic word vectors, reflect business cycles and historical events, as opposed to that built by the semantic-only word vectors. For concreteness, our sentiment index reflects the fact that firm managers express pessimism during recessions, meanwhile, the index based on the semantic-only word vectors only exhibits seasonal patterns without associations with historical economic regimes. Our finding is in line with Jiang et al. (2019) who also document a downward trend of the management sentiment during the 2008 financial crisis. We furthermore suggest that the dot-com crisis also hurts the management sentiment. These findings align with the nature of economic recessions.

Third, we find that our management sentiment index serves as a strong predictor of future stock returns, directly confirming the above-mentioned behavioral hypothesis. This result is two-fold. First, with the same set of MD&A corpus, the management sentiment extracted by the sentiment-semantic word vectors encompasses predictive information beyond that derived from the Loughran-McDonald dictionary-based method. Importantly, this result holds in both in-sample and out-of-sample settings and is robust to the choice of the stock market indexes. Additionally, we find that our management sentiment index outperforms the powerful historical average model (Campbell and Thompson, 2008) in predicting out-of-sample future stock returns. Second, our management sentiment merely relies on the MD&A section of 10-K filings, as opposed to Jiang et al. (2019). Despite the difference in the input data, both studies arrive at similar conclusions. This similarity potentially suggests that the MD&A section contains useful sentiment signals for future stock return prediction, providing accurate sentiment measurement. We further find that the predictive power of our management sentiment index relates to the information provided by firm managers regarding dividend payment plans in the MD&A section.

In conclusion, by introducing the sentiment-semantic word vectors, our work highlights the importance of both word sentiment and semantics in achieving accurate estimation of document sentiments. The utilization of sentiment-semantic word vectors unlocks the valuable sentiment insights within the MD&A section of 10-K filings that strongly predict future stock returns. These valuable pieces of information are possibly overlooked by methods ignoring either the sentiment or semantics of words.

Related literature and contributions

The past two decades have witnessed a blooming amount of research on the economic implications of corporate disclosures and their connection to the equity markets (Henry, 2008; Li, 2010; Loughran and McDonald, 2011; Price et al., 2012; Jegadeesh and Wu, 2013; Dyer et al., 2017; Jiang et al., 2019; Frankel et al., 2022). Henry (2008) is among the earliest attempts to analyze earning press releases using a word-count method. By introducing two lists of positive and negative words, they discover a relationship between earning press sentiment and investor's reactions. In a similar vein, Loughran and McDonald (2011) introduce a comprehensive sentiment lexicon specifically in the financial context, hereafter termed the *Loughran-McDonald* dictionary. They find that only negative words within 10-K filings are associated with contemporaneous stock returns. Jegadeesh and Wu (2013) argue that words in the Loughran-McDonald dictionary should be subject to some weights. Accordingly, they develop a market-dependent scheme of word weighting and show that stock returns are influenced by both positive and negative words in 10-K filings as long as they are appropriately weighted. Using the Loughran-McDonald dictionary, Jiang et al. (2019) show that management sentiment extracted from 10-K/Q filings and conference calls is predictive of future stock returns. Frankel et al. (2022) compare the information contained in corporate disclosures using machine learning and dictionary-based methods.

Studies linking the sentiment of the MD&A section and the market reactions are surprisingly infrequent. Loughran and McDonald (2011), besides 10-K filings, provide evidence of a significant relationship between the MD&A section and the stock returns via negative words. By using the Loughran and McDonald (2011) lexicon, Feldman et al. (2010) detect a significant association between short-window market reactions around the 10-K filing dates and the change of the MD&A sentiment. Deviating from the sentiment, Brown and Tucker (2011) find that changes in MD&A contents positively correlate with the magnitude of stock market reactions. Another line of studies on the MD&A section documents its connection to firm characteristics (Li, 2010; Mayew et al., 2015; Bochkay and Levine, 2019).

So far, studies have well-documented a link between the MD&A sentiment and contemporaneous market reactions. This is in line partially with the intention of the US Securities and Exchange Commission (SEC) that the MD&A section should provide explanatory

information to investors regarding current firm conditions (SEC, 2003). Yet another important part of the SEC's intention regarding future implications of MD&A has not been fully explored by current literature despite the potential stock market predictability of the MD&A section (Feldman et al., 2010). Attempting to fill this gap, our work contributes to the extant literature by providing predictive analyses of the MD&A section located in 10-K filings regarding future stock returns.

We also contribute to the burgeoning literature on techniques used in economic and financial sentiment analysis. The current state of the literature on this area is dominated by lexicon-based methods due to their simplicity (Henry, 2008; Feldman et al., 2010; Loughran and McDonald, 2011; Jiang et al., 2019). Although several efforts have been made to deviate from the reliance on a pre-defined sentiment lexicon (Li, 2010; Jegadeesh and Wu, 2013; Chen et al., 2022; Frankel et al., 2022), the underlying techniques for textual feature extraction are still word-count-based. However, due to the ignorance of word semantics, the current methods may overlook potential sentiment resulting from word interactions (Huang et al., 2023).

We seek to overcome this downside of the word-count methods by using the Word2Vec model (Mikolov, Chen, Corrado and Dean, 2013). Word2Vec is a neural-network-based method that represents words in the form of semantic numerical vectors in which two synonyms tend to locate adjacently in the vector space. Despite word semantics capturing, the ability of Word2Vec to represent sentiment connotations is still questioned. To enhance the adaptability and proficiency of a Word2Vec model in sentiment analysis, we propose an additional component embedded in the modeling process that functions as sentiment guidance to the model. We show that our approach enhances sentiment classification, thereby allowing us to unlock more insights into the MD&A documents beyond the reach of current dictionary-based methods. This novel adaptation serves as our main methodological contribution which is elaborated in the next section.

2 Methodology

The ultimate goal of our proposed method is to obtain a set of word vectors capturing both sentiment and semantic meanings of words, thus expectedly enhancing the sentiment extraction from a document. To this end, our method relies on three building blocks: (i) word vectors that are derived by the Word2Vec model (Mikolov, Chen, Corrado and Dean, 2013), (ii) a technique that distills the knowledge of a large model into a smaller model, i.e., Knowledge Distillation (Hinton et al., 2015), and (iii) the Financial Phrasebank dataset (Malo et al., 2014), serving as sentiment guidance to the Word2Vec model. The first building block functions as an initial model to represent the general semantics of

words by numerical vectors, in which synonyms tend to be represented adjacently in the vector space. The second aims to inject the financial context and the sentiment meanings, which are extracted from the third building block, into the word vectors while preserving the general semantics captured by the initial model.

The Word2Vec model

Since introduced by Mikolov, Chen, Corrado and Dean (2013), studies in economics and finance that adopt Word2Vec to explore financial documents have gained popularity; see Li et al. (2021), Das et al. (2022), Ma et al. (2023), Miranda-Belmonte et al. (2023), among others. The ability to capture the immediate contexts in representing words is the key feature that sets Word2Vec apart from the count-based word-representation methods, which have been widely used in economic research using textual data (Loughran and McDonald, 2011; Jegadeesh and Wu, 2013; Huang, Zang and Zheng, 2014; Henry and Leone, 2016; Jiang et al., 2019). However, despite Word2Vec's success in capturing word semantics, word sentiment representation is still beyond its capability. For concreteness, Table 1[3] illustrates this downside of the vanilla Word2Vec model by presenting the top ten most similar words to the word "bad" based on the Google pre-trained Word2Vec and the Word2Vec model trained on the MD&A corpus. At first glance, the most similar words to "bad" are "good" and "not-bad". While this result seems logical in the sense of semantic similarity, it appears counterintuitive when considering their polarized sentiments. Intuitively, these Word2Vec models tend to group words with opposite sentiments. An effective Word2Vec model for sentiment representation is anticipated to ensure words with similar sentiments are clustered together.

Knowledge Distillation

Nevertheless, leveraging Word2Vec for comprehensive semantic representation while incorporating sentiment meanings is challenging due to the following circumstances. On the one hand, to integrate sentiment meanings into a Word2Vec model, data with sentiment labels are required (Maas et al., 2011). Labeled data are, however, scarce in economics and finance and typically small. This is because of their requirement for expensive and time-consuming human annotation (Lutz et al., 2020). On the other hand, training a Word2Vec model from scratch requires a massive amount of data to sufficiently capture the word semantics (Rodriguez and Spirling, 2022). To resolve this paradoxical situation, we need a technique constructing a model that (i) inherits the knowledge of word semantics from a pre-trained Word2Vec model, and at the same time, (ii) integrates it with the

[3]To take negation into account as suggested by Mukherjee et al. (2021), negations are carefully handled before proceeding to the sentiment analysis. In particular, we first locate the sentiment words defined by the Loughran-McDonald sentiment dictionary in the MD&A documents. After that, we determine, within a certain window, if negation terms, which are "not", "no', "none", "neither", "nor", and "never" appear within a 5-adjacent-word window around the sentiment words. If that is the case, the "not-" prefix is added to the sentiment words. This explains why the word "not-bad" appears in Table 1.

Google pre-trained Word2Vec	Trained on the MD&A corpus
good	not-bad
terrible	uncollectible
horrible	troubled
lousy	extinguishment
crummy	doubtful
horrid	forgiveness
awful	uncollectable
dreadful	extinquishment
horrendous	restructurings
nasty	trouble

Table 1: This table reports the top ten most similar words to the word "bad" based on Google pre-trained word vectors and the word vectors trained on our MD&A corpus. The similarity between two words is measured by the cosine similarity between their corresponding representative vectors.

sentiment information carried by a small labeled dataset. Consequently, Knowledge Distillation (Hinton et al., 2015) appears as a suitable technique. Specifically, it allows us to obtain a model that internalizes the knowledge of a pre-trained model while encouraging the new model to acquire the supervised information in a labeled dataset autonomously.

Particularly to our problem, a new set of word vectors, denoted by W^{SS}, that captures both the semantics and sentiment of words is wanted. The pre-trained model in our case is the Google word vectors,[4] denoted by W^{GG}, because it is trained with a massive dataset containing approximately 100 billion words in total and three million distinct words in the vocabulary (Mikolov, Sutskever, Chen, Corrado and Dean, 2013). Finally, we resort to the Financial Phrasebank dataset as the sentiment guidance for W^{SS}. The knowledge distillation technique applied particularly to our problem seeks to maximize the following log-likelihood function,

$$L(W^{SS}, \theta | X) = \sum_{i=1}^{N} \log p(s_i | \theta, W^{SS}, X_i) - \lambda \Delta(W^{SS}, W^{GG}), \tag{1}$$

in which, s_i is the sentiment label of document i; X is the information set, $\{X_1, X_2, ..., X_N\}$, of N documents and X_i is the set of features extracted from document i; $\Delta(W^{SS}, W^{GG})$ is the average distance between the vectors corresponding to W^{SS} and W^{GG} of a same word; θ and W^{SS} are trainable parameters to maximize the log-likelihood function; W^{GG} remains fixed during the training process.

The first term of $L(W^{SS}, \theta | X)$, functioning as a document sentiment classifier, integrates the sentiment information encoded by s_i into the word vectors W^{SS}. The second term

[4] Available at: https://code.google.com/archive/p/word2vec/.

imposes a semantic penalty when W^{SS} deviates from the Google pre-trained word vectors W^{GG} with rich information on word semantics. These competing terms create a trade-off between the amount of sentiment and semantic information captured by W^{SS} during the training process. The trade-off is controlled by λ which is optimally chosen by the sentiment classification accuracy on a validation set.

The Financial Phrasebank dataset and the parameterization of the likelihood function

We use the Financial Phrasebank dataset to acquire s_i and X_i. There are three sentiment classes in the dataset, i.e., *negative* (1), *neutral* (2), and *positive* (3). Accordingly, s_i is assumed to follow the multinomial distribution with $M = 3$ levels. The conditional likelihood function becomes,

$$p(s_i|\theta, W^{SS}, X_i) \propto \prod_{m=1}^{M} \pi_{i,m}^{s_{i,m}} \tag{2}$$

in which,

$$\pi_{i,m} = \frac{\exp(X_i^\top W^{SS}\theta_m)}{\sum_{n=1}^{M}\exp(X_i^\top W^{SS}\theta_n)}. \tag{3}$$

Technically, W^{SS} is a $|V| \times d$ matrix; θ is a $d \times M$ matrix; θ_m is column m of θ with $m = 1, 2, 3$.

Textual feature extraction and the choice of the distance measure

Two problems remain: (i) how to extract X_i from document i; and (ii) how to choose the distance measure, Δ. For the first problem, inspired by Jegadeesh and Wu (2013), we rely on a method called *tf.idf* standing for *term frequency - inverse document frequency* (Manning and Schutze, 1999). Despite the lack of theoretical justification, Manning and Schutze (1999) suggest that the *tf.idf* representation is useful in document retrieval applications. Technically, define V as the vocabulary of the Google pre-trained Word2Vec model, and $|V|$ as the number of distinct words in V. X_i now can be represented by a $|V|$-dimensional vector, $(X_{i1}, X_{i2}, ..., X_{i|V|})^\top$. Each element of this vector, X_{ij}, is calculated as the ratio between the occurrences of word w_j in document i (tf_{ij}) and the transformed count of documents containing word w_j (df_{ij}). Formally,

$$X_{ij} = tf_{ij} \times \log\frac{N}{(df_{ij}+1)}, \tag{4}$$

in which, N is the number of documents in the Financial Phrasebank training set.

To address the second problem, we choose the cosine similarity as the distance measure

between W^{SS} and W^{GG}.[5] This choice is motivated by the fact that Word2Vec learns words that are adjacent to each other in terms of cosine similarity (Levy and Goldberg, 2014). Technically,

$$\Delta(W^{SS}, W^{GG}) = \frac{1}{|V|}\sum_{j=1}^{|V|}\Delta(\vec{w}_j^{SS}, \vec{w}_j^{GG}) = \frac{1}{|V|}\sum_{j=1}^{|V|}\left[1 - \frac{\langle \vec{w}_j^{SS}, \vec{w}_j^{GG}\rangle}{\|\vec{w}_j^{SS}\|\|\vec{w}_j^{GG}\|}\right] \tag{5}$$

in which, $\vec{w}_j^k$ is the vector representation of word w_j based on W^k with $k \in \{SS, GG\}$; technically, $\vec{w}_j^k$ is row j of matrix W^k.

Putting all components together gives us the following log-likelihood function,

$$L(W^{SS}, \theta | X) = \sum_{i=1}^{N} \log p(s_i | \theta, W^{SS}, X_i) + \frac{\lambda}{|V|}\sum_{j=1}^{|V|}\frac{\langle \vec{w}_j^{SS}, \vec{w}_j^{GG}\rangle}{\|\vec{w}_j^{SS}\|\|\vec{w}_j^{GG}\|}, \tag{6}$$

where p is parameterized by equations 2 and 3; X_i is defined by equation 4.

To prevent overfitting, we randomly split the Financial Phrasebank dataset into three parts: training, validation, and testing parts. The training part serves to solve for W^{SS} and θ by maximizing the log-likelihood function 6. The validation part is used to optimally choose the trade-off hyperparameter λ. We use the testing part to compare the sentiment classification power between W^{SS} and W^{GG}. We provide a comprehensive discussion of this comparison in Section 4.2.

3 Data

We estimate our sentiment-semantic word vectors by utilizing the Financial Phrasebank dataset (Malo et al., 2014).[6] The dataset is constructed to address the scarcity of high-quality labeled data specifically for financial sentiment analysis. It consists of English news articles centered around Finnish firms listed on the Nasdaq Helsinki stock exchange and comprises 4,846 documents. The dataset is manually annotated by 16 people with financial expertise who categorize the documents into three sentiment classes: *negative*, *neutral*, and *positive*. The Financial Phrasebank dataset features a high imbalance in the labeled sentiment distribution (with 604 negative, 2,879 neutral, and 1,363 positive doc-

[5] The cosine similarity between two vectors x and y is defined as $\text{sim}(x, y) = \frac{\langle x, y\rangle}{\|x\|\|y\|}$, where $\langle x, y\rangle$ is the inner product of two vectors x, y; and $\|x\|$ is the Euclidean norm of a vector x. To use it as the distance measure between x and y, people usually subtract the cosine similarity from one, i.e., $1 - \text{sim}(x, y)$.

[6] Available at: https://huggingface.co/datasets/financial_phrasebank.

uments). In line with Malo et al. (2014), we adopt the F1 score as the evaluation metric for our approach to accommodate the imbalanced characteristic of the dataset.

After obtaining the sentiment-semantic word vectors by the Financial Phrasebank dataset, we construct the management sentiment index using the corpus of the Management's Discussion and Analysis (MD&A) section of 10-K filings of US firms from 1994 to 2018. The 10-K filings can be downloaded from The Notre Dame Software Repository for Accounting and Finance (SRAF).[7] The SRAF page also provides additional resources for textual data analysis, such as stopword lists and the Loughran-McDonald dictionary. The SRAF data consists of both 10-K and 10-Q filings which are in the text-file format with HTML tags having been removed. We construct our management sentiment index based on 10-K filings only because their embedded information is acknowledged to be more significant than that of 10-Q filings (Griffin, 2003). We then extract the MD&A section out of each 10-K file following the advice of Loughran and McDonald (2016) and manage to extract 68% of all the 10-K files in the corpus. In comparison with Loughran and McDonald (2011) who achieve roughly 50% of the successful MD&A extraction, our rate is reasonable. We further discard the MD&A documents that have fewer than 250 words. After these purges, we retain 124, 133 MD&A documents spanning the period 1994:01 to 2018:12.[8]

To supplement our regression analyses in Section 6, we further employ multiple sources of numerical data, including:

- the Standard and Poor's (S&P) 500 and the value-weighted CRSP indexes; both include dividends and are queried from the Wharton Research Data Service (WRDS);
- the one-month US Treasury bill rate used as the risk-free rate, available on Kenneth R. French's data.[9]

The predictive power of the management sentiment in relation to stock returns could be rooted in the reflection of firm managers concerning the business cycles or macroeconomic conditions. To delve to macroeconomic channels associated with the stock return predictability based on management sentiment, we leverage the monthly macroeconomic dataset provided by Welch and Goyal (2008). As expected to directly connect with macroeconomic fundamentals, this dataset has gained popularity in stock return forecasting literature using macroeconomic variables (Cochrane, 2011; Huang et al., 2015; Jiang et al., 2019; Gu et al., 2020; Chen et al., 2023). In particular, the dataset includes 14 monthly macroeconomic variables which are: the log dividend-price ratio (DP), log

[7] Available at: https://sraf.nd.edu/.
[8] From now on, we use the format *yyyy:mm* to indicate the time of month *mm* in year *yyyy*.
[9] Available at: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

dividend yield (DY), log earnings-price ratio (EP), log dividend-payout ratio (DE), stock return variance ($SVAR$), book-to-market ratio (B/M), net equity expansion ($NTIS$), Treasury bill rate (TBL), long-term bond yield (LTY), long-term bond return (LTR), term spread (TMS), default yield spread (DFY), default return spread (DFR), and inflation rate ($INFL$). Detailed definitions of these variables are given in Section 2.2 of Jiang et al. (2019).

4 Empirical results

This section provides empirical evidence about the effectiveness of the sentiment-semantic word vectors, W^{SS}, in sentiment analyses. As mentioned in Section 2, semantic-only word vectors, i.e., W^{GG}, tend to group words with semantic similarity regardless of sentiment. Therefore, we expect that W^{SS}, which captures more sentiment meanings via our proposed method, can mitigate this problem, i.e., clustering words with similar sentiments together. Despite that, we further show that W^{SS} retains a large extent of word semantics in the financial context. Moreover, we show that W^{SS}, which excels in sentiment and semantic encapsulation, classifies document sentiment more accurately in comparison with W^{GG} and the Loughran-McDonald dictionary-based method.

4.1 How well does W^{SS} cluster words sentimentally?

We first assess the proficiency of W^{SS} in capturing both the sentiment and semantics of words through a comparison with W^{GG}. To this end, we implement a sentiment classification at the word level. This classification relies on the presumption that, given word vectors with more proficiency in capturing sentiment, positive (negative) words tend to be surrounded by more words delivering positive (negative) sentiment.[10]

We employ the Loughran-McDonald dictionary to determine the set of positive and negative words. The choice of the Loughran-McDonald dictionary guarantees the relevance of these sentiment words in the financial context. It is worth noting that the Loughran-McDonald dictionary solely participates in validating the word vectors, thus ensuring our approach is fully data-driven. After that, for each sentiment word, we examine the sentiment types of its neighboring words using W^{SS} and W^{GG}. To determine neighboring words, we combine two criteria: (i) the top n most similar words, denoted by n, and (ii) a predefined similarity threshold above which two words are determined to be similar,

[10] By *positive* (*negative*) words, we imply words that deliver optimism (pessimism) based on several rules, e.g., sentiment dictionaries.

		(True) Positive				(True) Negative			
n		10	15	20	30	10	15	20	30
Panel A: $\tau = 0.2$									
Positive	W^{SS}	**3.28**	**4.35**	**5.51**	**7.35**	**0.42**	**0.68**	**0.89**	**1.18**
	W^{GG}	3.15	4.13	5.17	7.03	0.51	0.74	0.92	1.27
Negative	W^{SS}	0.33	**0.40**	**0.51**	**0.76**	**2.73**	**3.58**	**4.47**	**6.05**
	W^{GG}	**0.32**	0.43	0.55	0.83	2.65	3.45	4.32	5.71
Panel B: $\tau = 0.3$									
Positive	W^{SS}	**3.28**	**4.28**	**5.35**	**6.78**	**0.41**	**0.64**	**0.77**	**0.98**
	W^{GG}	3.14	4.04	4.99	6.42	0.50	0.70	0.83	1.08
Negative	W^{SS}	0.33	**0.40**	**0.51**	**0.74**	**2.68**	**3.48**	**4.15**	**5.14**
	W^{GG}	**0.31**	0.43	0.56	0.79	2.62	3.33	4.02	4.82
Panel C: $\tau = 0.4$									
Positive	W^{SS}	**2.53**	**2.88**	**3.03**	**3.13**	**0.29**	**0.32**	**0.33**	**0.33**
	W^{GG}	2.44	2.74	2.85	2.93	0.35	0.36	0.39	0.39
Negative	W^{SS}	**0.25**	**0.26**	**0.28**	**0.29**	**1.79**	**1.88**	**1.94**	**1.94**
	W^{GG}	0.26	0.29	0.35	0.36	1.74	1.85	1.88	1.88

Table 2: This table reports the confusion matrix of the word-level sentiment classification using W^{SS} and W^{GG} with different values of the top n most similar words and similarity thresholds τ. The positive and negative words are determined by the Loughran-McDonald dictionary. The bold numbers indicate the word vectors among W^{SS} and W^{GG} that are more proficient in capturing sentiment, measured by their classification accuracy.

denoted by τ. Because W^{SS} is expected to capture sentiment meanings more effectively than W^{GG}, we anticipate that more positive (negative) words and less negative (positive) words are found in the neighborhood of positive (negative) words with W^{SS} compared to W^{GG}. To enhance the robustness of the classification, we apply various values of the criteria to choose the neighboring words.

Table 2 reports the confusion matrix of the sentiment classification described above. The table presents the average numbers of correct and incorrect assignments regarding the word classification of two sentiment categories: *negative* and *positive*. The first element of 3.28 means that, for every positive word defined by the Loughran-McDonald dictionary, W^{SS} yields on average 3.28 other positive words exhibiting a cosine similarity above 0.2 within the top 10 most similar words of the given word. Consequently, this number is seen as the *true positive* of W^{SS} in this classification at the top 10 most similar words and the similarity threshold of 0.2. Similarly, based on W^{SS}, there is 0.42 negative word found within the neighborhood of positive words given the same set of criteria. This is,

Words	Top ten most similar words
bank	banking, lender, branch, brokerage, lending, deposit, mortgage, institution, treasury, insurer
bond	debt, fund, unsecured, warrant, refinance, maturity, issuance, municipal, mortgage, equity
capital	liquidity, cash, investment, equity, treasury, financing, debt, fund, infrastructure, finance
cash	money, financing, liquidity, fund, debt, payment, capital, treasury, proceeds, financial
debt	refinance, repayment, liquidity, loan, equity, mortgage, credit, financing, financial, finance
inflation	economy, mediumterm, currency, rate, growth, economic, index, ruble, norminal, sector
interest	interested, concern, intention, commitment, royalty, attention, stake, expectation, confidence, income
liability	insurance, responsibility, risk, compensation, statutory, damage, incur, legal, burden, insurer
share	pershare, stock, earnings, dividend, common, pretax, repurchase, profit, value, gain
yield	benchmark, bps, rate, maturity, bp, throughput, note, harvest, basis, produce

Table 3: This table reports the top ten most similar words to the corresponding words. The similar words are chosen by the cosine similarity based on W^{SS}.

therefore, the *false positive* of W^{SS} in this particular case. With the same interpretation, the *true negative* and *false negative* of W^{SS} with this set of criteria are 2.73 and 0.33, respectively.

At first glance, W^{SS} outperforms W^{GG} in allocating words into the correct sentiment categories. In particular, W^{SS} has higher true positive/negative and lower false positive/negative in comparison with W^{GG}. Put differently, W^{SS} clusters words into the corresponding sentiment more accurately than W^{GG}, thus demonstrating the superiority of W^{SS} in capturing the sentiment meanings of words. Moreover, these results are robust to the varying values of the cluster size n and the similarity threshold τ.

So far, W^{SS} has demonstrated its superior proficiency in capturing word sentiment compared to W^{SS}, yet another question concerning the preservation of word semantics in the financial context of W^{SS} remains. To provide an impression of how well W^{SS} maintains the word semantics in the financial context, we retrieve the top ten most similar words of each given word, and subsequently, qualitatively assess their coherence and relevance in the financial context. For robustness, we combine words with strong financial meanings (e.g., "cash", "debt") and words used in the financial context differently from their casual meanings (e.g., "bond", "capital", "share"). This assessment, although prone to some extent of subjectivity, is widely used in word representation research to evaluate the quality of the word vectors regarding semantics (Mikolov, Chen, Corrado and Dean, 2013; Dieng et al., 2020; Li et al., 2021; Das et al., 2022).

Table 3 presents the top ten most similar words to "bank", "bond", "capital", "cash", "debt", "inflation", "interest", "liability", "share", and "yield" retrieved using W^{SS} and the cosine similarity. Overall, these words are surrounded by words with strong economic and financial meanings, even for those with different meanings depending on the contexts such as "bond", "capital", and "share". For the word "interest", the top similar words tend to have casual meanings however several economics-related words like "expectation" and "income" also appear. This serves as compelling evidence that W^{SS}

efficiently preserves the word semantics in the financial context.

In conclusion, the sentiment-semantic word vectors W^{SS} derived by our approach outperform the semantic-only word vectors W^{GG} in capturing sentiment. Moreover, while proficiently conveying the word sentiment, W^{SS} effectively retains the word semantics inherent in W^{GG}.

4.2 How accurately does W^{SS} classify document sentiment?

W^{SS} has demonstrated greater proficiency in capturing both word sentiment and semantics in comparison with W^{GG}. However, how it performs in sentiment classification remains unanswered. In conjunction with the findings in Section 4.1, a superior performance of W^{SS} compared to W^{GG} in sentiment classification will add robustness to our approach in calibrating word vectors for effective sentiment and semantic representation. Indeed, many studies validate their proposed models by sentiment classification (Li, 2010; Lutz et al., 2020; Huang et al., 2023), demonstrating the efficacy of this validation method in evaluating a novel approach.

Formally, we maximize the following log-likelihood function for the sentiment classification task,

$$\tilde{L}(\phi^k|W^k, X_i) = \sum_{i=1}^{N} \log p(s_i|\phi^k, W^k, X_i), \qquad \text{with} \qquad k \in \{SS, GG\}, \tag{7}$$

where ϕ^k is a d-dimensional vector of model parameters associated with the word vectors W^k; the other notations are defined in Section 2. It is noted that, with this sentiment classification, the word vectors W^k are fixed and are not subject to further training. The predicted probability for each sentiment class m conditioning on the word vectors W^k and document i is calculated by $\hat{p}(s_i = m|\hat{\phi}^k_m, W^k, X_i)$.

To estimate ϕ^k, we use the training part of the Financial Phrasebank dataset that was used to estimate W^{SS}. The testing part is then used to evaluate the classification accuracy of W^{SS} and W^{GG}. For every document i, the predicted sentiment class $\hat{s}^k_i$ based on the word vectors W^k is the one associated with the highest predicted probability. Technically,

$$\hat{s}^k_i = \underset{m}{\text{argmax}}\, \hat{p}(s_i = m|\hat{\phi}^k_m, W^k, X_i). \tag{8}$$

Besides W^{SS} and W^{GG}, we implement a sentiment classification model using the Loughran-McDonald dictionary. While comparing the classification capability of W^{SS} with that of W^{GG} manifests the importance of sentiment-capturing in classifying sentiments using

Approach	Class-wise F1 scores			F1 score micro	F1 score macro
	negative	*neutral*	*positive*		
A - Loughran-McDonald dictionary	**0.36**	0.70	0.40	0.58	0.49
B - W^{GG}	0.08	0.77	0.39	0.64	0.41
C - W^{SS}	0.27	**0.79**	**0.48**	**0.68**	**0.51**

Table 4: This table compares the performances of three approaches for sentiment classification in the Financial Phrasebank dataset: (i) the Loughran-McDonald dictionary-based approach, (ii) the word vector approach based on W^{GG}, and (iii) the word vector approach based on W^{SS}. The first three columns present the component F1 scores of three sentiment classes, i.e., *negative*, *neutral*, and *positive*, respectively. The last two columns exhibit the global F1 scores using the micro and macro averages, correspondingly. The F1 scores showcased in this table are calculated using the testing part of the Financial Phrasebank dataset.

word vectors, the comparison of the word vectors (i.e., W^{SS} and W^{GG}) and the Loughran-McDonald dictionary unveils the significance of word semantics in sentiment classification. Follow the convention used in many studies (Henry, 2008; Loughran and McDonald, 2011; Jiang et al., 2019), the predicted sentiment class of document i using the Loughran-McDonald dictionary is determined as follows,

$$\hat{s}_i^{LM} = \begin{cases} 1 & \text{if } \#(\text{pos})_i < \#(\text{neg})_i \\ 2 & \text{if } \#(\text{pos})_i = \#(\text{neg})_i \\ 3 & \text{if } \#(\text{pos})_i > \#(\text{neg})_i \end{cases} \tag{9}$$

in which, 1, 2, and 3 indicate the *negative*, *neutral*, and *positive* sentiments; $\#(\text{pos})_i$ and $\#(\text{neg})_i$ are the number of positive and negative words defined by the Loughran-McDonald dictionary that appear in document i, respectively.

Similar to Malo et al. (2014), we opt for the F1 score as the evaluation metric for this classification. However, to obtain in-depth analyses across each sentiment class, we also present the class-wise F1 scores besides the global F1 scores with the micro and macro average; see Grandini et al. (2020) and Takahashi et al. (2023) for further technical details of the F1 scores.

Table 4 reports a wide range of the F1 scores for sentiment classification within the Financial Phrasebank dataset. It includes the results from the Loughran-McDonald dictionary-based approach and the word vector approaches using W^{GG} and W^{SS}. In general, the word vector approach using W^{SS} outperforms the competitors in classifying sentiments. Together with the results of Section 4.1, this reinforces the success of our approach in injecting the sentiment meanings into semantic-only word vectors, subsequently, reflecting in more accurate document sentiment classification.

A more thorough examination of Table 4 reveals deeper insights into the sentiment classification capabilities of the considered methods. First, with the highest F1 score in the

negative class, the Loughran-McDonald dictionary is the best in identifying pessimistic documents among the competing approaches. The proficiency in detecting the negative sentiment of the Loughran-McDonald dictionary may explain the findings of Loughran and McDonald (2011), wherein only the pessimism embedded in 10-K filings is associated with stock returns.

Second, the semantic-only word vector W^{GG} surprisingly performs the worst in classifying pessimism, even in comparison with the dictionary-based approach, by a large margin, i.e., its F1 score is only 0.08 compared to 0.27 and 0.36 of the others. Combined with the lowest macro-average F1 score of W^{GG}, this result suggests that relying exclusively on word semantics is inadequate for precisely gauging nuanced sentiment expressions.

Third, the superior performance of W^{SS} in most cases suggests that, in order to measure sentiments accurately, both word sentiment and semantics are required, especially with the *neutral* and *positive* sentiment classes. Comparing only W^{SS} with the Loughran-McDonald approach reveals that word semantics, when standing alone may not be powerful, is still crucial in precisely classifying sentiments. Our findings correspond with many criticisms of bag-of-word methods regarding their treatment of words as independent units; see Mikolov, Chen, Corrado and Dean (2013), Li et al. (2021), Huang et al. (2023), among others.

It is noticed that there is a difference in the model performance order determined by the global F1 scores between the two averaging methods. This difference can be explained by the imbalance of the Financial Phrasebank dataset. In particular, the micro-average F1 score is less prone to class imbalance (Grandini et al., 2020; Takahashi et al., 2023). Despite that, W^{SS} robustly exhibits the best performance in classifying the sentiments globally.

Associated with the empirical results shown in Section 4.1, two conclusions are made. First, our approach successfully obtains a set of word vectors that captures both word sentiment and semantics in the financial context. Furthermore, the small size of the Financial Phrasebank dataset highlights the adaptability of our approach in handling small and domain-specific data. Second, both captured sentiment and semantics play crucial roles in accurately identifying sentiments. Ultimately, how an accurate sentiment measurement is applied to explore economic values or answer financial puzzles arises as the next question. Subsequent sections will delve into this intriguing topic.

5 Construction of the management sentiment index

To examine the predictive effects of management sentiment on stock markets, it appears natural to construct an index conveying firm managers' sentiment through corporate dis-

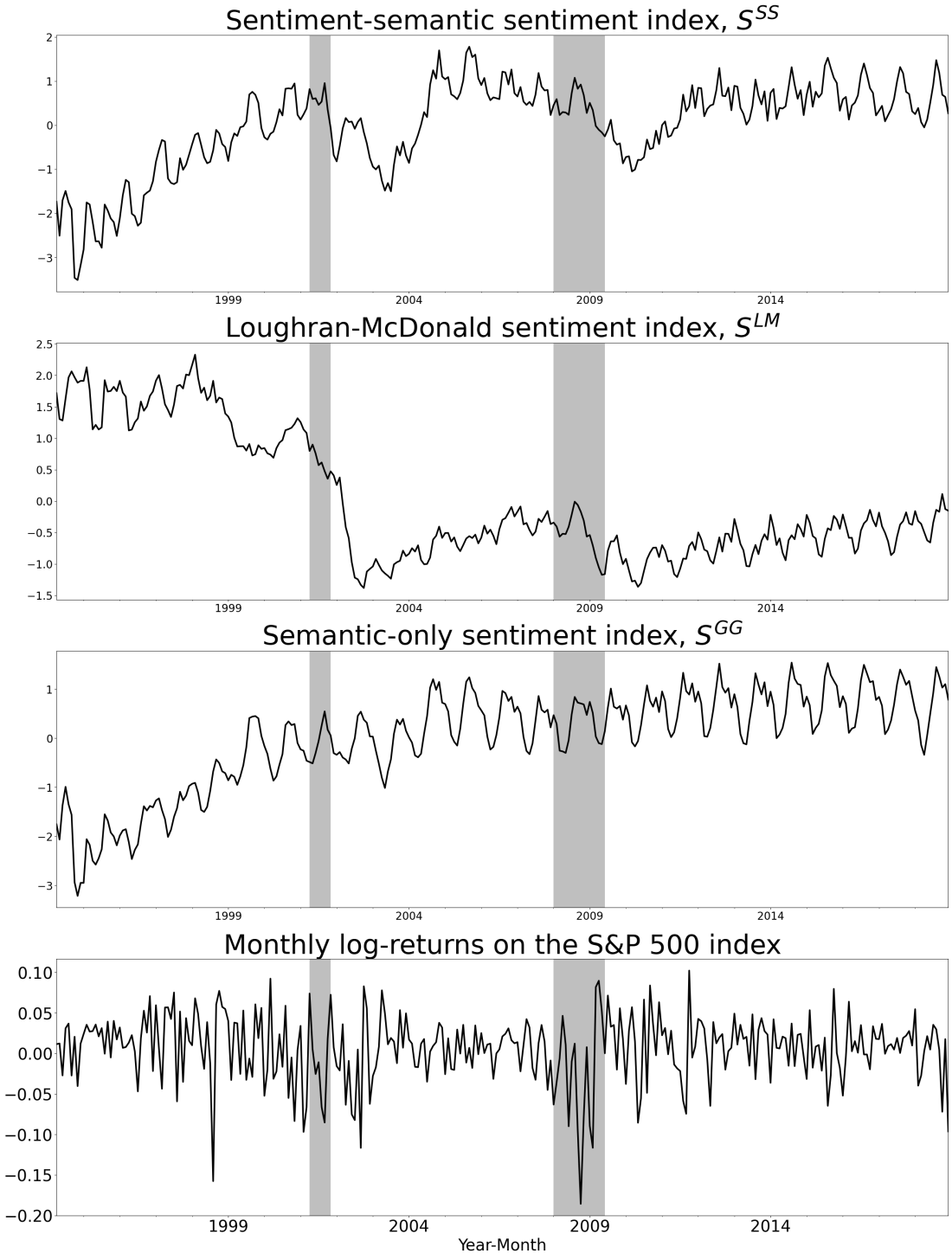

Figure 1: The market-level management sentiment indexes extracted from the MD&A section of 10-K filings. The first plot depicts the management sentiment index S^{SS} constructed using the sentiment-semantic word vectors trained on the Financial Phrasebank dataset. The second plot is sentiment index S^{LM} constructed using the bag-of-words method based on the Loughran-McDonald dictionary (Loughran and McDonald, 2011). The third plot depicts the sentiment index S^{GG} built by the Google pre-trained word vectors. We also present the series of log returns on the S&P 500 index. The vertical grey bars indicate the economic recessions defined by the NBER. The data sample spans the period from 1994:01 to 2018:12.

closures. Subsequently, the connection between this index and the series of future stock returns is investigated. To this end, many attempts are commonly based on bag-of-words approaches, wherein the sentiment of a document is a projection of a pre-defined lexicon (Henry, 2008; Feldman et al., 2010; Loughran and McDonald, 2011; Price et al., 2012; Huang, Teoh and Zhang, 2014; Jiang et al., 2019). Other studies obtain sentiment labels by human annotation (Li, 2010) or by the associated market reactions (Jegadeesh and Wu, 2013; Frankel et al., 2022).

We instead measure the management sentiment using our word vectors W^{SS} and the MD&A corpus. In particular, we apply the sentiment classification model using W^{SS} (i.e., model C in Table 4) to produce the predictions of the sentiment classes (i.e., *negative*, *neutral*, and *positive*) on the MD&A corpus. The sentiment score of each MD&A document is computed as the weighted expected values of its predicted sentiment.[11] We then construct our management sentiment index, which is hereafter denoted as S^{SS}, by a simple average of the management sentiment scores from the MD&A documents released in a given month. Following Jiang et al. (2019), we smooth the management sentiment index by a four-month moving average to mitigate the effects of idiosyncratic noises. We provide the details of the sentiment estimation in Appendix A.

For the sake of methodological comparison, we also construct two other management sentiment indexes similar to S^{SS}: (i) the Loughran-McDonald sentiment index, S^{LM}, and (ii) the semantic-only sentiment index, S^{GG}. The first sentiment index is constructed based on a word-count approach using the Loughran-McDonald dictionary (Loughran and McDonald, 2011). The second sentiment index is formed similarly to S^{SS}, however, using the Google word vectors, W^{GG}, instead of W^{SS}. The two sentiment indexes are both derived from the MD&A corpus and are aggregated similarly to S^{SS} as described in the previous paragraph. As the final step, three management sentiment indexes are standardized to have zero means and unit variances to eliminate the effects of scale difference. As shown by the empirical results in Section 4, using S^{SS} in presenting the management sentiment index is more advantageous compared to S^{GG} and S^{LM} because of the effective sentiment representation of W^{SS} over W^{GG} and the Loughran-McDonald approach.

Figure 1 presents the variations of the three sentiment indexes over time. At first glance, S^{GG}, the sentiment index built by the semantic-only word vectors, presents only seasonality over nearly the entire data sample, thus implying that it contains limited explanatory information about business cycles or historical events. This observation is expected because, while W^{GG} contains rich semantics, it reflects scarce sentiment meanings. Consequently, S^{GG}, which is constructed based on W^{GG} as the core ingredient, captures sen-

[11] The weights are the inverse proportions of the sentiment classes in the Financial Phrasebank dataset. We decide to use the weighted average because the distribution of the sentiment classes in the MD&A corpus may differ from that in the Financial Phrasebank dataset, which is well-known to be imbalanced. Consequently, biases caused by a distributional shift may occur if the weights are not applied.

timent fluctuations limitedly.

Based on S^{SS}, the management sentiment is low initially but increases gradually until the beginning of the dot-com crisis. This period witnesses the disagreement between S^{SS} and S^{LM}. In particular, based on S^{LM}, the sentiment is high and starts to decrease slightly until the time before the dot-com crisis. However, both S^{SS} and S^{LM} acknowledge that during the dot-com crisis the management sentiment drops, with a decreasing trend continuing to around 2003. This period coincides with the exposure of several high-profile accounting fraudulent cases (e.g., Enron and Worldcom), which may drive down the management sentiment. After this period, both S^{SS} and S^{LM} showcase a rise in their values until before the 2008 financial crisis, implying that firm managers tend to express their MD&A with an optimistic voice during this time. However, while S^{LM} witnesses a steady upward trend, S^{SS} exhibits a sharp increase after 2004 and a gradual decrease until the financial crisis. Both S^{SS} and S^{LM} once again agree on a decrease in management sentiment during the 2008 financial crisis. The time after the financial crisis witnesses a strong cyclicity pattern in both S^{SS} and S^{LM}. This seasonal pattern is similar to the management sentiment index built by Jiang et al. (2019) when it also observes a bold seasonal pattern after the financial crisis.

6 Predictive regression analysis

In this section, we provide empirical evidence regarding stock return predictability of our sentiment-semantic management sentiment index, S^{SS}. This goal can be achieved by numerous comparative analyses between our management sentiment index and the index built by the Loughran-McDonald dictionary-based method, S^{LM}. The sentiment index S^{GG} is not considered in this analysis due to its limitation in reflecting informational fluctuations described in Section 5. We implement the analyses in both in-sample and out-of-sample manners to guarantee the robustness of potential findings.

6.1 In-sample market return predictability

We first examine the market return predictability regarding the sentiment indexes, S^{SS} and S^{LM}. To empirically test the market return predictability of the sentiment indexes, we design the following set of equations,

$$\begin{aligned} CER_{t\to t+h} &= \alpha + \beta S_t^k + Recession + \epsilon_{t\to t+h}, \qquad k \in \{SS, LM\} \\ CER_{t\to t+h} &= \alpha + \beta S_t^{SS} + \gamma S_t^{LM} + Recession + \epsilon_{t\to t+h} \end{aligned} \tag{10}$$

	Monthly cumulative excess market returns ($CER_{t\to t+h}$)														
h (months)	1			3			6			9			12		
Panel A: Value-weighted CRSP index															
S^{SS}	-0.008**		-0.010*	-0.015**		-0.019**	-0.028***		-0.038***	-0.035***		-0.050***	-0.047***		-0.067***
	(-2.107)		(-1.907)	(-2.466)		(-2.029)	(-3.178)		(-2.716)	(-3.151)		(-3.111)	(-3.366)		(-3.530)
S^{LM}		0.000	-0.005		0.001	-0.009		-0.002	-0.020		-0.030	-0.038*		-0.009	-0.042
		(0.030)	(-0.779)		(0.062)	(-0.775)		(-0.114)	(-1.166)		(-0.283)	(-1.394)		(-0.341)	(-1.606)
R^2	0.078	0.064	0.081	0.119	0.094	0.126	0.170	0.120	0.190	0.179	0.130	0.208	0.188	0.125	0.230
Panel B: S&P 500 index															
S^{SS}	-0.007**		-0.008*	-0.016***		-0.029**	-0.029***		-0.036***	-0.036***		-0.057***	-0.049***		-0.073***
	(-2.423)		(-1.738)	(-3.229)		(-2.220)	(-3.506)		(-2.898)	(-3.458)		(-3.254)	(-3.549)		(-3.487)
S^{LM}		0.002	-0.002		0.003	-0.006		0.020	-0.017		-0.001	-0.024		-0.002	-0.034
		(0.407)	(-0.383)		(0.406)	(-1.575)		(0.101)	(-1.025)		(-0.033)	(-1.164)		(-0.053)	(-1.276)
R^2	0.067	0.054	0.068	0.118	0.087	0.122	0.157	0.103	0.170	0.182	0.126	0.202	0.194	0.112	0.221

Table 5: This table reports the results of the OLS regressions of equation 10 over h-month horizons with $h = 1, 3, 6, 9, 12$. The dependent variable, $CER_{t\to t+h}$, is the cumulative excess market returns, i.e., the monthly returns on (i) the value-weighted average CRSP index (Panel A) and (ii) the S&P 500 index (Panel B) in excess of the risk-free rate, from month t to month $t+h$. S^{SS} and S^{LM} are the management sentiment indexes extracted from the MD&A section of 10-K filings using the sentiment-semantic word vectors and the Loughran-McDonald dictionary (Loughran and McDonald, 2011), correspondingly. A constant term (α) and a recession dummy (*Recession*) are also included in each regression equation. The coefficients, Newey-West heteroskedasticity- and autocorrelation-robust t-statistics (in parentheses), and R^2 are reported. The data sample spans the period from 1994:01 to 2018:12. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

where $CER_{t \to t+h}$ is the cumulative excess market returns, i.e., the monthly returns on (i) the value-weighted average CRSP index, and (ii) the S&P 500 index, in excess of the risk-free rate from month t to month $t + h$; *Recession* is a dummy variable indicating the economic recessions defined by the National Bureau of Economic Research.

Our experiment is inspired by Jiang et al. (2019) yet possesses two important differences. First, in addition to the S&P 500 index, we also use the value-weighted CRSP index similarly to Jegadeesh and Wu (2013) to compute the market returns. Accordingly, this additional usage is expected to enhance the robustness of the test's results. Second, we further control the recession fixed effects in all equations to capture potential variations caused by seasonality and business cycles. As shown in Figure 1, the recessions negatively affect both management sentiment and the S&P 500 index.[12] Therefore, omitting the recession control possibly leads to inconsistent estimates.

It is well-known that the return predictive regressions usually suffer several econometric issues, i.e., spurious inference results due to persistent independent variables (Ferson et al., 2003), the small-sample bias (Stambaugh, 1999), and potential biased standard error estimation (Hodrick, 1992). We use the heteroskedasticity- and autocorrelation-robust Newey-West t-statistic with small-sample adjustment for consistent covariance matrix estimation to cope with the above-mentioned issues.

Table 5 presents the regression results of equation 10 over h-month horizons with $h = 1, 3, 6, 9, 12$. First, for the univariate regressions, the coefficients on S^{SS} are negative and significant at the 5% level with all considered horizons. We, however, do not observe any significant coefficients on S^{LM}. This result remains robust when either the value-weighted CRSP index or the S&P 500 index is considered as the stock market index. Intuitively, the sentiment index, which integrates word sentiment and semantics, is negatively associated with future cumulative excess market returns. In contrast, one solely based on word sentiment shows no correlation with market returns.

The bivariate regression results reveal deeper insights about the superior predictive capacity of S^{SS} in comparison to S^{LM}. In particular, adding S^{LM} to the models with only S^{SS} does not change the sign and the significance of the coefficients on S^{SS} in all regressions. Moreover, R^2s of the bivariate models are similar to that of the corresponding regressions on S^{SS} alone in most of the cases (e.g., with the S&P 500 index at the semi-annual horizon, R^2 of the two models are 17.0% and 15.7%, respectively). With a substantial correlation of -0.520 between S^{SS} and S^{LM}, these findings suggest that S^{SS} possesses predictive insights regarding future market returns that are beyond those of S^{LM}.

Economically, at the three-month horizon, an increase of one standard deviation in the management sentiment is associated with a decrease of 1.5% in cumulative returns on

[12] Especially, the S&P 500 index suffered negative returns in most of the time of the 2008 financial crisis.

the value-weighted CRSP index and 1.6% with the S&P 500 index. Furthermore, the estimated coefficient on S^{SS} increases in the absolute values as h increases. This result implies that S^{SS} consistently and significantly predicts the cumulative excess market returns in the long run. In addition, the predictive power of the S^{SS} becomes stronger when the horizon gets longer. Across the horizons, the in-sample R^2 of the regressions on S^{SS} ranges from 7.8% to 18.8% with the value-weighted CRSP index, and from 6.7% to 19.4% with the S&P 500 index. This means that S^{SS} is a factor that can explain large in-sample variations of the future excess market returns. Moreover, out-of-sample tests presented in Section 6.2 show that this result is maintained out-of-sample.

In conclusion, contribute to Jiang et al. (2019), who discover a negative correlation between sentiment extracted from 10-K/Q reports and conference calls and future market returns, our finding suggests that the sentiment information derived exclusively from the MD&A section of 10-K filings is also a strong and negative predictor to the stock market, provided that the sentiment is precisely measured.

6.2 Out-of-sample market return predictivity

In numerous predictive analyses, researchers often discover substantial predictive evidence with in-sample data, yet struggle to obtain significant predictive power with out-of-sample data (Inoue and Kilian, 2005). Additionally, out-of-sample analyses tend to be more resilient to the econometric issues described in Section 6.1 (Busetti and Marcucci, 2013). Therefore, to robustly validate the predictive power of our management sentiment index, S^{SS}, we conduct several out-of-sample return predictive analyses at the market level. According to Welch and Goyal (2008), in stock returns forecasting, a historical average of stock returns frequently outperforms regression models of stock returns on economic predictors. Therefore, whether S^{SS} outperforms the historical average benchmark model is of high interest. Moreover, as demonstrated in Section 6.1, S^{SS} encompasses additional predictive information beyond S^{LM} concerning future market returns based on in-sample tests. If this result is preserved out-of-sample, it can be demonstrated that S^{SS} holds significant economic value complementarily to S^{LM}. Accordingly, we compare the market return predictive power of: (i) the sentiment indexes S^{SS} and S^{LM} with that of the historical average benchmark, respectively; and (ii) the model combining both S^{SS} and S^{LM} with that containing only S^{LM}. Technically, we conduct the two following tests,

Test A:

- Model 1A: $CER_{t+1\to t+h} = \alpha_{1A} + \epsilon_{t+1\to t+h}$
- Model 2A: $CER_{t+1\to t+h} = \alpha_{2A} + \beta_{2A} S_t^k + \epsilon_{t+1\to t+h}, \qquad k \in \{SS, LM\}$

Test B:

- Model 1B: $CER_{t+1\to t+h} = \alpha_{1B} + \beta_{1B} S_t^{LM} + \epsilon_{t+1\to t+h}$
- Model 2B: $CER_{t+1\to t+h} = \alpha_{2B} + \beta_{2B} S_t^{SS} + \beta_{3B} S_t^{LM} + \epsilon_{t+1\to t+h}$

in which model 1 in the two tests is a parsimonious model and model 2 nests the corresponding model 1. In all models, we regress $\{CER_{s+1\to s+h}\}_{s=1}^{t+1-h}$ on predictor variables $\{X_s\}_{s=1}^{t+1-h}$, in which X_s is a combination of a constant term, S_s^{SS}, and S_s^{LM} depending on the model. The tests are implemented in a recursive-window manner (West and McCracken, 1998), in which the data from 1994:01 to 1999:12 is the initial training set, and the data from 2000:01 to 2018:12 is used as the evaluation period.[13]

Define the mean squared prediction error (MSPE) as a measure of prediction accuracy,[14] in both tests A and B, we want to test the null hypothesis of smaller or equal MSPE of the parsimonious model compared to that of the nested models against the alternative hypothesis of which the nested model has a smaller MSPE compared to the parsimonious model. To this end, we use the Campbell and Thompson (2008) out-of-sample R^2 statistic (R_{OS}^2) which is defined as follows,

$$R_{OS}^2 = 1 - \frac{\sum_{t=P}^{T}(CER_{t+1\to t+h} - \widehat{CER}_{2,t+1\to t+h})^2}{\sum_{t=P}^{T}(CER_{t+1\to t+h} - \widehat{CER}_{1,t+1\to t+h})^2} = 1 - \frac{MSPE_2}{MSPE_1} \tag{11}$$

in which, P is the starting time point of the evaluation dataset, which is 2000:01 in our case; $\widehat{CER}_{j,t+1\to t+h}$ with $j = 1, 2$ are the out-of-sample forecasts produced by the parsimonious (model 1) and the nested (model 2) models in each test, correspondingly. By definition, R_{OS}^2 has the range of $(-\infty, 1]$. A significantly positive R_{OS}^2 implies that the nested models have better forecasting ability than the parsimonious models. Accordingly, the above-mentioned testing hypotheses turn out to be, $H_0 : R_{OS}^2 \leq 0$ against $H_A : R_{OS}^2 > 0$.

We adopt the adjusted MSPE statistic proposed by Clark and West (2007), which is the difference between the MSPE statistics of models 1 and 2 with a bias adjustment, to test the significance of R_{OS}^2. Clark and West (2007) show that the adjusted MSPE statistic asymptotically follows a standard normal distribution, and the null hypothesis is rejected if the statistic exceeds $+1.282$, $+1.645$, and $+2.323$ for a one-sided test at the 10%, 5%, and 1% significance levels, respectively.

Table 6 reports the results of tests A and B on the value-weighted CRSP and the S&P

[13] Although the initial training data ranges from 1994:01 to 1999:12, the true training data of each model varies depending on h.

[14] This statistic is also used by Stock and Watson (2002, 2003, 2004), Clark and McCracken (2006), to name but a few.

		R^2_{OS}(%) and adjusted MSPE (in parentheses)				
h (months)		1	3	6	9	12
Panel A: Value-weighted CRSP index						
Test A	S^{SS}	-0.458	1.353**	1.326	3.219	2.874
		(0.208)	(1.736)	(1.018)	(1.053)	(0.989)
	S^{LM}	-2.353	-2.137	-3.261	-3.754	-2.299
		(0.136)	(0.693)	(0.754)	(0.659)	(0.484)
Test B		-0.280	2.877**	3.788*	6.345**	5.287*
		(0.079)	(1.757)	(1.493)	(1.699)	(1.584)
Panel B: S&P 500 index						
Test A	S^{SS}	0.623	3.638**	6.182**	6.769*	8.758*
		(1.271)	(1.691)	(1.715)	(1.619)	(1.569)
	S^{LM}	-0.024	-0.025	-0.025	-0.033	-0.044
		(0.894)	(1.107)	(1.146)	(1.185)	(1.246)
Test B		0.517	3.882*	8.366**	9.849**	13.79**
		(0.938)	(1.478)	(1.832)	(1.886)	(2.019)

Table 6: This table reports the out-of-sample performance of the management sentiment indexes, S^{SS} and S^{LM}, in predicting the cumulative excess market returns, i.e., the monthly returns on (i) the value-weighted average CRSP index (Panel A) and (ii) the S&P 500 index (Panel B) in excess of the risk-free rate, from month $t+1$ to month $t+h$. S^{SS} and S^{LM} are the management sentiment indexes extracted from the MD&A section of 10-K filings using the sentiment-semantic word vectors and the Loughran-McDonald dictionary (Loughran and McDonald, 2011), correspondingly. Test A evaluates the predicting performance of S^{SS} and S^{LM} in comparison with the historical average benchmark. Test B evaluates the predicting performance of S^{SS} in addition to S^{LM}. R^2_{OS} is the out-of-sample Campbell and Thompson (2008) R^2. The adjusted MSPE statistic (in parentheses) is the mean squared prediction error statistic introduced by Clark and West (2007) to test the null hypothesis that the parsimonious models (i.e., the historical average model in test A, and the S^{LM}-only model in test B) have smaller or equal MSPE than the nested models. The tests are implemented in a recursive-window manner (West and McCracken, 1998), in which the data from 1994:01 to 1999:12 is the initial training set, and the data from 2000:01 to 2018:12 is used as the evaluation period. * and ** denote significance at the 10% and 5% levels, respectively.

500 indexes. For the value-weighted CRSP index, we observe that S^{SS} outperforms the historical average in predicting the cumulative excess market returns at the three-month horizon. On the opposite, S^{LM} exhibits no predicing power to the cumulative market returns. With significant R^2_{OS} at the three-month, semi-annual, nine-month, and one-year horizons in test B, it is shown that S^{SS} adds significant predicting information to the model with only S^{LM} in the middle and long run.

For the S&P 500 index, S^{SS} possesses even stronger predictive power to future market returns than that regarding the value-weighed CRSP index. In comparison with the historical average benchmark, the model with S^{SS} is capable of producing more precise forecasts of cumulative excess market returns across all considered time horizons, except for the one-month period. We also observe a monotonic increase in R^2_{OS} along the expanding horizons in this case, implying that S^{SS} increasingly predicts the future market returns when the portfolio is held for a longer period.

We still do not observe a significant positive R^2_{OS} with S^{LM} in all models, suggesting that the sentiment of the MD&A documents, as extracted by the Loughran-McDonald dictionary, does not contain predictive information to future market returns. Test B conducted on the S&P 500 index further corroborates the findings regarding the enhanced predictive capability of S^{SS} in contrast to S^{LM}. More concretely, the presence of numerous significant and positive R^2_{OS} statistics highlights the considerable predictive capacity contributed by S^{SS} to models solely reliant on S^{LM}. These models, previously indicated to lack predictive power regarding future market returns, now exhibit enhanced predictive ability due to the inclusion of S^{SS}.

Different from Jiang et al. (2019) who find that management sentiment in the current month t outperforms the historical average benchmark in predicting the market returns in the next month $t+1$, our results suggest that the management sentiment possesses significantly higher predictive power compared to the historical average in longer horizons. We conjecture that the difference is rooted in the inclusion of more abundant data sources in the Jiang et al. (2019) management sentiment index. This inclusion allows them to "...examine manager sentiment on a more timely basis" (Jiang et al., 2019). Consequently, their management sentiment index can exploit more prompt effects of the sentiment on the stock markets. However, with our results, we show that the sentiment exploited exclusively from the MD&A section of 10-K filings is capable of capturing the mispricing information to the stock prices. Our findings contribute to the literature on stock return predictability and corporate disclosures that the mispricing information contained in the management sentiment embedded in the 10-K filings possibly concentrates on the MD&A section to some extent.

7 Management sentiment and macroeconomic channels

So far, the management sentiment index S^{SS} has been found to negatively predict future stock returns. According to Jiang et al. (2019), the negative predictive power of management sentiment may be due to the misjudgment of investors regarding future firm earnings. This section aims to provide another angle of this finding under the lens of macroeconomic channels.

To this end, we examine the complementary predictive power of S^{SS} in addition to the 14 macroeconomic variables provided by Welch and Goyal (2008). In particular, we re-implement the in-sample and out-of-sample predictive regression analyses used in Section 6 with S^{LM} being replaced by each of the 14 macroeconomic variables. It should be noted that, within this section, we use only the S&P 500 index as the market index. This is because several macroeconomic variables are derived from the S&P 500 index. Due to the

limitations in space, we present the result tables of these analyses in Appendix B together with the mutual correlations of the 14 macroeconomic variables and S^{SS} for reference.

We first implement the in-sample analysis regarding the predictive information covered by S^{SS} in relation to the 14 macroeconomic variables, using the following equations,

$$\begin{aligned} CER_{t\to t+h} &= \alpha + \beta X_t + Recession + \epsilon_{t\to t+h}, \\ CER_{t\to t+h} &= \alpha + \beta X_t + \gamma S_t^{SS} + Recession + \epsilon_{t\to t+h} \end{aligned} \tag{12}$$

in which, X_t is one of 14 macroeconomic variables described in Section 3.

Table 8 reports the estimation results for the above regression equations. We observe that S^{SS} exhibits significant correlations to future stock returns at the 5% level when nested with the macroeconomic variables, except with the dividend-price ratio (DP) and dividend yield (DY). These results imply that the management sentiment index S^{SS} possibly captures the information relating to the dividend payment of S&P 500 firms. With the significant and negative coefficients in the regressions other than DP and DY, S^{SS} demonstrates that its predictive information is orthogonal to that of the other macroeconomic variables, even to those with strong stock return predictability such as the book-to-market ratio (B/M) or the default return spread (DFR).

The out-of-sample test results, which are detailed in Table 9, reinforce these findings. We find that S^{SS} limitedly contributes to the predictive ability of the dividend-related variables, i.e., the dividend-price ratio (DP), dividend yield (DY), and dividend-payout ratio (DE). Consistently with the in-sample findings, S^{SS} is found to add significant power to the other macroeconomic variables in predicting out-of-sample future stock returns.

We conjecture that this result is possibly rooted to some extent in discussions made by firm managers regarding the dividend payment plans in the MD&A section. For example, in the MD&A section of Apple Inc.'s 10-K filing in 2015, the company wrote,

"... In April 2014, the Company increased its share repurchase authorization to $90 billion and the quarterly dividend was raised to $0.47 per common share, resulting in an overall increase in its capital return program from $100 billion to over $130 billion. During 2014, the Company utilized $45 billion to repurchase its common stock and paid dividends and dividend equivalents of $11.1 billion...
... The Company currently anticipates the cash used for future dividends, the share repurchase program and debt repayments will come from its current domestic cash, cash generated from ongoing U.S. operating activities and from borrowings..."

Another example can be found in the MD&A in the 2012 10-K filing of Microsoft Corporation, in which the company wrote,

"...Cash used for financing increased $1.0 billion to $9.4 billion due mainly to a $6.0 billion net decrease in proceeds from issuances of debt and a $1.2 billion increase in dividends paid, offset in part by a $6.5 billion decrease in cash used for common stock repurchases...
... We expect existing domestic cash, cash equivalents, short-term investments, and cash flows from operations to continue to be sufficient to fund our domestic operating activities and cash commitments for investing and financing activities, such as regular quarterly dividends, debt repayment schedules, and material capital expenditures, for at least the next 12 months and thereafter for the foreseeable future..."

In general, we provide evidence that the predictive power of the management sentiment index S^{SS} is fully absorbed by the information about dividend payment plans of S&P 500 firms. This absorption can be attributed to discussions regarding dividends made by firm managers in the MD&A section of 10-K filings. Our findings, however, are not equivalent to the assertion that dividend-related information located in the MD&A section is a cause of the predictive power of the management sentiment. Its causal effects are then left for future studies.

8 Conclusion

The paper sheds light on the return predictability of the sentiment contained in the MD&A section of 10-K filings from January 1994 to December 2018. As opposed to most of the existing studies, we introduce a novel method to accurately gauge the MD&A sentiment. In particular, our method relies on three components: (i) the Google pre-trained Word2Vec model to nail word representations to initial semantic information; (ii) the knowledge distillation method; and (iii) a dataset with sentiment labels acting as sentiment guidance. The result of our approach is a set of word vectors capturing both sentiment and semantic meanings.

Our proposed method enhances sentiment classification at both word and document levels. Explicitly, we suggest that omitting either sentiment or semantic meanings leads to inefficient sentiment classification. This result underlines the importance of these two facets in obtaining an accurate sentiment measurement.

By using the sentiment-semantic word vectors, we build a management sentiment index of which their variations conceptually match well with different economic episodes. The index based on the semantic-only approach is, however, unable to produce meaningful interpretations of the economic states. This observation once again reaffirms the importance of sentiment nuances captured by word vectors in exploring economic implications presented in the MD&A documents.

Finally, our proposed management sentiment index is a strong negative predictor of future stock returns. Moreover, our management sentiment index is shown to embrace predictive insights concerning future stock returns beyond the dictionary-based sentiment index. Moreover, these findings hold in both in-sample and out-of-sample setups. Based on these results, three conclusions are withdrawn concerning sentiment analysis of the MD&A documents. First, it is crucial to have an accurate measurement to obtain meaningful sentiment information. Second, the MD&A section of 10-K filings contains mispricing information regarding firm conditions rather than true fundamentals. Third, the predictive power of the management sentiment of the MD&A documents relates to the information about dividend payment plans.

A potential limitation, however, is that our model, although based on semantic word representation, remains statically contextualized. This means that a word is encoded by a single numerical vector regardless of the surrounding context in a sentence or paragraph. This limitation elicits an extension of the current work with the usage of language models, e.g., FinBERT (Huang et al., 2023), associated with our proposed method. As dynamically contextualized, this extension is anticipated to uncover more insights into corporate disclosures.

References

Bochkay, K. and Levine, C. B. (2019). Using MD&A to improve earnings forecasts, *Journal of Accounting, Auditing & Finance* **34**(3): 458–482.

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.

Busetti, F. and Marcucci, J. (2013). Comparing forecast accuracy: A Monte Carlo investigation, *International Journal of Forecasting* **29**(1): 13–27.

Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average?, *The Review of Financial Studies* **21**(4): 1509–1531.

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.

Chen, L., Pelger, M. and Zhu, J. (2023). Deep learning in asset pricing, *Management Science* .

Clark, T. E. and McCracken, M. W. (2006). The predictive content of the output gap for inflation: Resolving in-sample and out-of-sample evidence, *Journal of Money, Credit and Banking* pp. 1127–1148.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models, *Journal of Econometrics* **138**(1): 291–311.

Cochrane, J. H. (2011). Presidential address: Discount rates, *The Journal of Finance* **66**(4): 1047–1108.

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

Das, S. R., Donini, M., Zafar, M. B., He, J. and Kenthapadi, K. (2022). Finlex: An effective use of word embeddings for financial lexicon generation, *The Journal of Finance and Data Science* **8**: 1–11.

Davis, A. K. and Tama-Sweet, I. (2012). Managers' use of language across alternative disclosure outlets: earnings press releases versus md&a, *Contemporary Accounting Research* **29**(3): 804–837.

De Long, J. B., Shleifer, A., Summers, L. H. and Waldmann, R. J. (1990). Noise trader risk in financial markets, *Journal of Political Economy* **98**(4): 703–738.

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

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

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.

Ferson, W. E., Sarkissian, S. and Simin, T. T. (2003). Spurious regressions in financial economics?, *The Journal of Finance* **58**(4): 1393–1413.

Frankel, R., Jennings, J. and Lee, J. (2022). Disclosure sentiment: Machine learning vs. dictionary methods, *Management Science* **68**(7): 5514–5532.

Grandini, M., Bagli, E. and Visani, G. (2020). Metrics for multi-class classification: An overview, *arXiv preprint arXiv:2008.05756* .

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.

Gu, S., Kelly, B. and Xiu, D. (2020). Empirical asset pricing via machine learning, *The Review of Financial Studies* **33**(5): 2223–2273.

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

Henry, E. and Leone, A. J. (2016). Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone, *The Accounting Review* **91**(1): 153–178.

Hinton, G., Vinyals, O. and Dean, J. (2015). Distilling the knowledge in a neural network, *arXiv preprint arXiv:1503.02531* .

Hodrick, R. J. (1992). Dividend yields and expected stock returns: Alternative procedures for inference and measurement, *The Review of Financial Studies* **5**(3): 357–386.

Huang, A. H., Wang, H. and Yang, Y. (2023). Finbert: A large language model for extracting information from financial text, *Contemporary Accounting Research* **40**(2): 806–841.

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

Huang, D., Jiang, F., Tu, J. and Zhou, G. (2015). Investor sentiment aligned: A powerful predictor of stock returns, *The Review of Financial Studies* **28**(3): 791–837.

Huang, X., Teoh, S. H. and Zhang, Y. (2014). Tone management, *The Accounting Review* **89**(3): 1083–1113.

Inoue, A. and Kilian, L. (2005). In-sample or out-of-sample tests of predictability: Which one should we use?, *Econometric Reviews* **23**(4): 371–402.

Jegadeesh, N. and Wu, D. (2013). Word power: A new approach for content analysis, *Journal of Financial Economics* **110**(3): 712–729.

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

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization, *Advances in Neural Information Processing Systems*, pp. 2177–2185.

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

Li, F. et al. (2010). Textual analysis of corporate disclosures: A survey of the literature, *Journal of accounting literature* **29**(1): 143–165.

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

Loughran, T. and McDonald, B. (2011). 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.

Lutz, B., Pröllochs, N. and Neumann, D. (2020). Predicting sentence-level polarity labels of financial news using abnormal stock returns, *Expert Systems with Applications* **148**: 113223.

Ma, Y., Liu, C., Zhang, J. T. and Liu, Y. (2023). Reliability study of stock index forecasting in volatile and trending cities using public sentiment——based on word2vec and LSTM models, *Applied Economics* **55**(43): 5013–5032.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. and Potts, C. (2011). Learning word vectors for sentiment analysis, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-volume 1*, Association for Computational Linguistics, pp. 142–150.

Malo, P., Sinha, A., Korhonen, P., Wallenius, J. and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts, *Journal of the Association for Information Science and Technology* **65**.

Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*, MIT press.

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

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.

Miranda-Belmonte, H. U., Muñiz-Sánchez, V. and Corona, F. (2023). Word embeddings for topic modeling: an application to the estimation of the economic policy uncertainty index, *Expert Systems with Applications* **211**: 118499.

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.

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.

Rodriguez, P. L. and Spirling, A. (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research, *The Journal of Politics* **84**(1): 101–115.

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.

Stambaugh, R. F. (1999). Predictive regressions, *Journal of Financial Economics* **54**(3): 375–421.

Stock, J. H. and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes, *Journal of Business & Economic Statistics* **20**(2): 147–162.

Stock, J. H. and Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices, *Journal of Economic Literature* **41**(3): 788–829.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set, *Journal of Forecasting* **23**(6): 405–430.

Takahashi, K., Yamamoto, K., Kuchiba, A., Shintani, A. and Koyama, T. (2023). Hypothesis testing procedure for binary and multi-class F1-scores in the paired design, *Statistics in Medicine* **42**(23): 4177–4192.

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

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction, *The Review of Financial Studies* **21**(4): 1455–1508.

West, K. D. and McCracken, M. W. (1998). Regression-based tests of predictive ability, *International Economic Review* **39**(4): 817–840.

A Construction of the management sentiment index

Denote the *tf.idf* representation of the MD&A document i that is released in month t as $X_{i,t}^{MDA}$. Follow the instructions in Section 4.2, the predicted probability of each sentiment class m, with $m = 1, 2, 3$, conditioning on W^{SS} is $\hat{p}(s_i = m|\hat{\phi}_m^{SS}, W^{SS}, X_{i,t}^{MDA})$. It is worth noting that $\hat{\phi}_m^{SS}$ is the estimated parameter of model C in Table 4. The sentiment score of this MD&A document based on W^{SS} is computed as,

$$s_{i,t}^{SS} = \frac{\sum_{m=1}^{M} \omega_m.m.\hat{p}(s_i = m|\hat{\phi}_m^{SS}, W^{SS}, X_{i,t}^{MDA})}{\sum_{m=1}^{M} \omega_m}$$

where $\omega_1 = \frac{1}{604}$, $\omega_2 = \frac{1}{2879}$, and $\omega_3 = \frac{1}{1363}$, which are the inverse proportions of the sentiment classes in the Financial Phrasebank dataset.

Further, define N_t as the number of all MD&A documents released in month t. Consequently, the management sentiment index based on W^{SS} is as follows,

$$\tilde{S}_t^{SS} = \frac{1}{N_t} \sum_{i=1}^{N_t} s_{i,t}^{SS}$$

The final sentiment index is smoothed by a four-month moving average following Jiang et al. (2019). Technically,

$$S_t^{SS} = \frac{1}{4} \sum_{p=0}^{3} \tilde{S}_{t-p}^{SS}$$

Replacing $\hat{\phi}_m^{SS}$ by $\hat{\phi}_m^{GG}$ and W^{SS} by W^{GG} in the above steps yields the management sentiment index S^{GG}. For the management sentiment index built by the Loughran-McDonald dictionary, the sentiment score of the MD&A document i in month t is computed similarly to Henry (2008). Particularly,

$$s_{i,t}^{LM} = \frac{\#(\text{pos})_{i,t} - \#(\text{neg})_{i,t}}{\#(\text{pos})_{i,t} + \#(\text{neg})_{i,t}}$$

in which, $\#(\text{pos})_{i,t}$ and $\#(\text{neg})_{i,t}$ denote the number of positive and negative words in the MD&A document i in month t, respectively. Different from Jiang et al. (2019) who put document length in the denominator, this way of calculation shields the sentiment measure from being diluted caused by non-sentiment words. The aggregation and smoothing steps remain unchanged.

B Comparison of S^{SS} and macroeconomic variables

	S^{SS}	DP	DY	EP	DE	$SVAR$	B/M	$NTIS$	TBL	TMS	DFY	DFR	$INFL$
S^{SS}	1.000												
DP	-0.197	1.000											
DY	-0.220	0.981	1.000										
EP	-0.022	0.081	0.080	1.000									
DE	-0.083	0.449	0.441	-0.854	1.000								
$SVAR$	0.081	0.206	0.129	-0.292	0.369	1.000							
B/M	0.179	0.664	0.644	0.409	-0.020	0.055	1.000						
$NTIS$	-0.504	-0.354	-0.331	0.114	-0.287	-0.247	-0.225	1.000					
TBL	-0.354	-0.382	-0.376	0.059	-0.253	-0.099	-0.624	0.346	1.000				
TMS	-0.155	0.309	0.304	-0.213	0.352	0.159	0.352	0.125	-0.706	1.000			
DFY	0.188	0.444	0.412	-0.525	0.702	0.595	0.314	-0.493	-0.440	0.377	1.000		
DFR	-0.034	0.002	0.087	-0.186	0.168	-0.248	-0.028	0.022	-0.087	0.114	0.109	1.000	
$INFL$	-0.057	-0.127	-0.118	0.030	-0.093	-0.325	-0.064	0.074	0.121	-0.041	-0.221	-0.020	1.000

Table 7: This table reports the correlations for the management sentiment index S^{SS} and the 14 macroeconomic variables. The definitions of the 14 macroeconomic variables are given in Section 3. The data sample spans the period from 1994:01 to 2018:12.

	$CER_{t\to t+h} = \alpha + \beta X_t + \epsilon_{t\to t+h}$			$CER_{t\to t+h} = \alpha + \beta X_t + \gamma S_t^{SS} + \epsilon_{t\to t+h}$					
h (months)	1	6	12	1		6		12	
		β		β	γ	β	γ	β	γ
DP	0.029	0.213***	0.439***	0.022	-0.006*	0.192***	-0.019*	0.407***	-0.029*
	(1.172)	(3.625)	(4.984)	(0.960)	(-1.824)	(3.129)	(-1.911)	(4.302)	(-1.665)
DY	0.068***	0.247***	0.470***	0.063***	-0.004	0.227***	-0.017	0.441***	-0.026
	(2.902)	(4.081)	(5.517)	(2.632)	(-1.052)	(3.595)	(-1.582)	(4.815)	(-1.442)
EP	-0.014	-0.037	-0.029	-0.013	-0.007**	-0.033	-0.028***	-0.023	-0.049***
	(-0.638)	(-0.411)	(-0.263)	(-0.604)	(-2.217)	(-0.375)	(-3.540)	(-0.211)	(-3.568)
DE	0.023	0.109*	0.185***	0.020	-0.006**	0.098*	-0.023***	0.165**	-0.039***
	(1.194)	(1.966)	(2.845)	(0.991)	(-2.152)	(1.854)	(-3.072)	(2.568)	(-2.979)
$SVAR$	-4.230***	-1.222	3.902*	-4.190***	-0.007**	-1.051	-0.029***	4.199*	-0.050***
	(-8.496)	(-0.705)	(1.673)	(-9.174)	(-2.391)	(-0.599)	(-3.494)	(1.735)	(-3.489)
B/M	-0.012	0.394**	0.862**	0.008	-0.007**	0.493***	-0.035***	1.039***	-0.063***
	(-0.170)	(2.164)	(2.427)	(0.131)	(-2.438)	(2.994)	(-3.791)	(3.242)	(-4.013)
$NTIS$	0.158	1.072	1.493	-0.070	-0.008**	0.314	-0.026***	0.074	-0.049***
	(0.658)	(1.390)	(0.906)	(-0.228)	(-2.207)	(0.335)	(-2.677)	(0.040)	(-3.155)
TBL	-0.125	-0.631	-1.721	-0.274	-0.009**	-1.244*	-0.038***	-2.860**	-0.070***
	(-0.610)	(-0.942)	(-1.278)	(-1.475)	(-2.585)	(-1.905)	(-3.463)	(-2.450)	(-3.835)
TMS	0.195	0.942	3.468*	0.089	-0.007**	0.528	-0.028***	2.823	-0.043***
	(0.598)	(0.722)	(1.671)	(0.241)	(-2.379)	(0.418)	(-3.749)	(1.406)	(-3.304)
DFY	-0.602	4.945	13.82*	-0.315	-0.007**	6.235	-0.032***	16.14**	-0.057***
	(-0.395)	(0.846)	(1.882)	(-0.199)	(-2.234)	(1.058)	(-3.242)	(2.310)	(-3.651)
DFR	1.227***	1.450***	1.633***	1.216***	-0.007**	1.400***	-0.028***	1.547***	-0.048***
	(10.26)	(3.430)	(2.906)	(6.715)	(-2.573)	(3.445)	(-3.544)	(2.955)	(-3.604)
$INFL$	0.779	-3.271	-8.190**	0.674	-0.007**	-3.706	-0.029***	-8.943**	-0.051***
	(0.537)	(-1.313)	(-2.352)	(0.499)	(-2.494)	(-1.509)	(-3.576)	(-2.499)	(-3.564)

Table 8: This table reports the in-sample OLS regression results of equations 12, in which the complementary predictive power of S^{SS} in addition to the 14 macroeconomic variables (Welch and Goyal, 2008) is examined. The dependent variable, $CER_{t\to t+h}$, is the monthly returns on the S&P 500 index in excess of the risk-free rate, from month t to month $t+h$. The definitions of the 14 macroeconomic variables are given in Section 3. A constant α and a recession dummy are also included in each regression equation. The coefficients, Newey-West heteroskedastic- and autocorrelation-robust t-statistics (in parentheses) are reported. The data sample spans the period from 1994:01 to 2018:12. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	$R^2(\%)$ and adjusted MSPE (in parentheses)				
h (months)	1	3	6	9	12
DP	-2.737	-0.903	3.189	3.289	4.710
	(-0.545)	(0.300)	(0.914)	(0.867)	(0.973)
DY	-2.183	-0.496	2.841	2.918	4.192
	(-0.515)	(0.342)	(0.917)	(0.852)	(0.960)
EP	-1.461	1.517	5.597**	6.149**	8.631**
	(-0.126)	(0.923)	(1.863)	(1.833)	(2.055)
DE	-0.003	1.902	3.777*	5.035	7.822
	(0.812)	(1.062)	(1.290)	(1.242)	(1.269)
$SVAR$	0.291*	1.886**	5.505**	7.88**	1.072**
	(1.577)	(1.944)	(2.039)	(2.041)	(2.161)
B/M	-1.448	3.195**	11.00***	13.82***	18.38***
	(0.523)	(1.784)	(2.543)	(2.756)	(2.809)
$NTIS$	-0.246	0.688	1.444	1.525	3.131*
	(0.520)	(1.063)	(1.183)	(1.195)	(1.353)
TBL	0.963**	4.921**	9.556**	11.56**	16.42**
	(1.693)	(2.016)	(2.030)	(2.078)	(2.102)
TMS	0.775	3.885*	6.502*	6.938*	8.441*
	(1.121)	(1.427)	(1.523)	(1.365)	(1.340)
DFY	0.438	3.619*	7.310**	8.607**	11.81**
	(1.228)	(1.577)	(1.730)	(1.881)	(2.618)
DFR	0.781*	3.863**	6.537**	7.173*	9.197*
	(1.380)	(1.732)	(1.759)	(1.631)	(1.567)
$INFL$	0.607	3.716**	6.846**	7.742**	10.07**
	(1.225)	(1.661)	(1.785)	(1.702)	(1.639)

Table 9: This table reports the out-of-sample stock return predictability of the management sentiment index, S^{SS}, in addition to the 14 macroeconomic variables (Welch and Goyal, 2008). The stock market returns are computed as the monthly returns on the S&P 500 index in excess of the risk-free rate, from month $t+1$ to month $t+h$. The definitions of the 14 macroeconomic variables are given in Section 3. R^2_{OS} is the out-of-sample Campbell and Thompson (2008) R^2. The adjusted MSPE statistic (in parentheses) is the mean squared prediction error statistic introduced by Clark and West (2007) to test the null hypothesis that the parsimonious models (i.e., the model contains exclusively the macroeconomic variables) have smaller or equal MSPE than the nested models. The tests are implemented in a recursive-window manner (West and McCracken, 1998), in which the data from 1994:01 to 1999:12 is the initial training set, and the data from 2000:01 to 2018:12 is used as the evaluation period. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.