Swiss Journal of Economics and Statistics

RESEARCH

Open Access

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

Tri Minh Phan[1*]

Abstract

This paper introduces a novel method to extract the sentiment embedded in the Management's Discussion and Analysis (MD &A) section of 10-K filings. The proposed 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. By contrast, if traditional sentiment extraction methods are used, the MD &A sentiment exhibits no predictive ability for stock markets. Additionally, the MD &A sentiment is associated with dividend-related macroeconomic channels regarding future stock return prediction.

Keywords Knowledge distillation, MD& A, Stock return predictability, Word2Vec

JEL classification J53, G12, G17

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 10-K/Q filings (Bochkay & Levine, 2019; Brown & Tucker, 2011; Cohen, Malloy, and Nguyen, 2020; Davis & Tama-Sweet, 2012; Feldman, Govindaraj, Livnat, and Segal, 2010; Li, 2010a; Loughran & McDonald, 2011; Tavcar, 1998).[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. 75,059). In this scenario, it is natural to expect the MD &A section to encapsulate insights that may influence stock market dynamics. Surprisingly, there are only a few studies that explore the power of the MD &A section to predict future stock returns. In this paper, we explore stock return predictability using solely the sentiment of the MD &A section, which we term *management sentiment*.[2] In particular, we investigate a behavioral implication of management sentiment in asset pricing: the hypothesis is that information in a corporate disclosure with misleading sentiment is absorbed by investors, leading to an overvaluation in the stock price. When the true stock fundamentals are gradually disclosed to the public, the price reverses, implying that 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 are a mixture of informative statements and boilerplate content (Li, 2010b). 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.

*Correspondence:
Tri Minh Phan
triminh.phan@unisg.ch
[1] School of Economics and Political Science, Department of Economics, University of St. Gallen, Bodanstrasse 6, 9000 St., Gallen, Switzerland

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

[2] We refer to the polarity of tone (i.e., negativity (pessimism), neutrality, or positivity (optimism)) as "sentiment," although we are aware of a strand of the literature that prefers to use the word "tone."

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 that integrates both word sentiment and semantics into a pre-defined set of word representations (i.e., vectors). This method results in another set that reflects 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, Vinyals, and Dean, 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 word vectors we obtain inherit word semantics from a pre-trained word representation model and, simultaneously, absorb nuanced sentiment information from the labeled dataset.

First, our proposed approach successfully obtains a new set of word vectors that captures both word sentiment and semantics; henceforth, these vectors are referred to as *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, which we term *semantic-only word vectors*, in clustering words into sentiment categories. Furthermore, our sentiment-semantic word vectors demonstrate a superior capability in document sentiment classification, outperforming competing methods, including semantic-only word vectors and the Loughran–McDonald dictionary (Loughran & McDonald, 2011). In particular, the sentiment-semantic word vectors achieve an F1 score of 0.68 in a sentiment classification task using the Financial Phrasebank dataset (Malo, Sinha, Korhonen, Wallenius, and Takala, 2014). Meanwhile, the corresponding scores for the Loughran–McDonald dictionary (which ignores word semantics) and the semantic-only word vectors (which ignore word sentiment) are 0.58 and 0.64. These findings underscore the importance of integrating sentiment and semantic information into word vectors for accurate sentiment analysis.

Second, the variations of our management sentiment index constructed from the sentiment-semantic word vectors reflect business cycles and historical events unlike the indexes built by the semantic-only word vectors. In concrete terms, our sentiment index reflects the fact that firm managers express pessimism during recessions; the index based on semantic-only word vectors strongly exhibits seasonal patterns without clear associations to historical economic regimes. Our findings are in line with those of Jiang et al. (2019) who also document a downward trend in management sentiment during the 2008 financial crisis. We furthermore suggest that the dot-com crisis hurt 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 twofold. First, with the same MD &A corpus, the management sentiment extracted by the sentiment-semantic word vectors encompasses predictive information beyond that derived from the method based on the Loughran–McDonald dictionary. Importantly, this result holds in both in-sample and out-of-sample settings and is robust to the choice of stock market index. Additionally, we find that our management sentiment index outperforms the powerful historical average model (Campbell & Thompson, 2008) in predicting out-of-sample future stock returns. Second, our measurement of management sentiment, unlike that of Jiang et al. (2019), merely relies on the MD &A section of 10-K filings. Despite the difference in the input data, the two studies arrive at similar conclusions. This similarity may suggest 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 sentiment-semantic word vectors, our work highlights the importance of both word sentiment and semantics in achieving an accurate sentiment estimation of a document. The utilization of sentiment-semantic word vectors unlocks valuable sentiment insights within the MD &A section of 10-K filings that strongly predict future stock returns. These valuable pieces of information may be overlooked by methods that ignore either the sentiment or the semantics of words.

1.1 Related literature and contributions

The past two decades have witnessed a blooming in research on the economic implications of corporate disclosures and the connection of these disclosures to the equity markets (Dyer, Lang, and Stice-Lawrence, 2017; Frankel, Jennings, and Lee, 2022; Henry, 2008; Jegadeesh & Wu, 2013; Jiang, Lee, Martin, and Zhou, 2019; Li, 2010a; Loughran & McDonald, 2011; Price, Doran, Peterson, and Bliss, 2012). Henry (2008) was among the first authors to analyze press releases on earnings using a word-count method. By introducing lists of positive and negative words, Henry (2008) discovers a relationship between the sentiment of earning press releases and investors' reactions. In a similar vein, Loughran and McDonald (2011) introduce a comprehensive sentiment

lexicon tailored for financial context, hereafter referred to as 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 weighting. 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 those words 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 reaction 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 change in the MD &A sentiment. Deviating from sentiment, Brown and Tucker (2011) find that changes in the MD &A content are positively correlated with the magnitude of stock market reactions. Another line of studies on the MD &A section documents its connection to firm characteristics (Bochkay & Levine, 2019; Fengler & Phan, 2023; Li, 2010a; Mayew, Sethuraman, and Venkatachalam, 2015).

So far, studies have documented a link between the MD &A sentiment and contemporaneous market reactions. This is partially in line 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). However, another important part of the SEC's intention, regarding the future implications of the MD &A section, has not been fully explored by the current literature despite the potential for the MD &A section to predict stock market (Feldman, Govindaraj, Livnat, and Segal, 2010). Attempting to fill this gap, our work contributes to the extant literature by providing predictive analyses of the MD &A section in the 10-K filings regarding future stock returns.

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

We seek to overcome this downside of the word-count methods by using the Word2Vec model (Mikolov, Chen, Corrado, and Dean, 2013). Word2Vec, a method based on neural networks, represents words in the form of semantic numerical vectors in which two synonyms tend to be located adjacently in the vector space. Although Word2Vec captures word semantics, its ability to represent sentiment connotations is still questioned. To enhance the adaptability and proficiency of the Word2Vec model in sentiment analysis, we propose an additional component embedded in the modeling process that functions as sentiment guidance for the model. The inclusion of additional components to the likelihood function to capture sentiment is the core idea of many techniques for learning word sentiment representation (Maas et al., 2011; Labutov & Lipson, 2013; Tang, Wei, Qin, Zhou, and Liu, 2014). However, these techniques are constrained by the need for large datasets (Maas et al., 2011; Tang, Wei, Qin, Zhou, and Liu, 2014) or are limited to binary classifications (Labutov & Lipson, 2013). Our approach not only extends these methods to multi-label classification but also demonstrates that it is effective with small sentiment datasets. By enhancing the capabilities of sentiment classification, our proposed technique provides deeper insights into MD &A documents, surpassing the limitations of current dictionary-based methods. This novel adaptation serves as our main methodological contribution, and full details are given in the next section.

2 Methodology

The ultimate goal of our proposed method is to obtain a set of word vectors capturing both the sentiment and the semantic meanings of words. It is thus expected to enhance the sentiment extraction from a document. To this end, our method relies on three building blocks: (i) word vectors that are derived using the Word2Vec model (Mikolov, Chen, Corrado, and Dean, 2013), (ii) a technique that distills the knowledge of a large model into a smaller model, known as knowledge distillation (Hinton, Vinyals, and Dean, 2015), and (iii) the Financial Phrasebank dataset (Malo, Sinha, Korhonen, Wallenius, and Takala, 2014), which serve as sentiment guidance for the Word2Vec model. The first building block functions as an initial model by representing the general semantics of

Table 1 This table reports the top ten most similar words to the word "bad" based on Google pre-trained word vectors, the word vectors trained on our MD &A corpus, and FinText

Google pre-trained	Trained on MD &A	FinText
Good	Not-bad	Good
Terrible	Uncollectible	Problem
Horrible	Troubled	Actually
lousy	Extinguishment	Really
Crummy	Doubtful	Probably

The similarity between two words is measured by the cosine similarity between their corresponding representative vectors

words by numerical vectors, with synonyms tending 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.

2.1 The Word2Vec model

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

2.2 Knowledge distillation

However, leveraging Word2Vec for comprehensive semantic representation while incorporating sentiment meanings is challenging for the following reasons. 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 the datasets are typically small. This is because expensive and time-consuming human annotation is required (Lutz, Pröllochs, and Neumann, 2020). On the other hand, training a Word2Vec model from scratch requires massive data in order to capture the word semantics sufficiently well (Rodriguez & Spirling, 2022). To resolve this paradoxical situation, we need a technique to construct a model that (i) inherits the knowledge of word semantics from a pre-trained Word2Vec model and, at the same time, (ii) integrates this knowledge with the sentiment information carried by a small labeled dataset. Consequently, knowledge distillation (Hinton, Vinyals, and Dean, 2015) appears to be a suitable technique. Specifically, this technique allows us to obtain a model that internalizes the knowledge of a pre-trained model while being encouraged to acquire the supervised information in a labeled dataset autonomously.

For our problem, a new set of word vectors, denoted by W^{SS}, which captures both the semantics and the sentiment of words, is wanted. The pre-trained model in our case is the set of FinText word vectors,[4] denoted by W^{Fin}, because this set is trained with a massive dataset which contains news stories in Dow Jones Newswires Text News Feed (2, 733, 035 unique tokens) covering various economic and financial topics (Rahimikia, Zohren, and Poon, 2021).[5] Finally, we resort to the Financial Phrasebank dataset as the sentiment guidance for W^{SS}. The particular knowledge distillation technique applied particularly to our problem seeks to maximize the following log-likelihood function:

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

[3] We follow the suggestion of Mukherjee et al. (2021), carefully handling negations 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 whether, within a certain window, a negation term, meaning "not," "no," "none," "neither," "nor," and "never," appears within a five-adjacent-word window around the sentiment word. If this is the case, the "not-" prefix is added to the sentiment word. This explains why the word "not-bad" appears in Table 1.

[4] Available at: https://fintext.ai/.

[5] We have also run analyses with the Google pre-trained word vectors as the pre-trained Word2Vec model and confirm that the main findings of this paper are unchanged. The results with the Google pre-trained Word2Vec model are provided upon request.

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; $\Delta(W^{SS}, W^{Fin})$ is the average distance between the vectors corresponding to W^{SS} and W^{Fin} for the same word; θ and W^{SS} are trainable parameters to maximize the log-likelihood function; and W^{Fin} remains fixed during the training process.

The first term, $L(W^{SS}, \theta|X)$, which functions as a document sentiment classifier, integrates the sentiment information encoded by s_i into the word vectors W^{SS}. The second term imposes a semantic penalty when W^{SS} deviates from the FinText pre-trained word vectors W^{Fin} with rich information on word semantics. These competing terms create a trade-off between the amounts 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 accuracy of the sentiment classification on a validation set.

2.3 The Financial Phrasebank dataset and the parameterization of the likelihood function

The first question that arises is how to choose the information set (X_i) and the corresponding sentiment labels s_i. Various methods in the literature address this problem. Li (2010a) utilizes corporate disclosures as the information set. Subsequently, he obtains sentiment labels through human annotations. Similarly, Huang et al. (2023) rely on manual labeling of analysts' reports to determine (X_i) and s_i. A significant disadvantage of this approach is the high cost and time-consuming nature of manual labeling. Another line of studies uses corporate disclosures or newspapers as the information set and employs the associated stock returns as proxies for the sentiment labels (Frankel, Jennings, and Lee, 2022; Lutz, Pröllochs, and Neumann, 2020; Jegadeesh & Wu, 2013). While this method addresses the high cost of human annotation, it potentially introduces noisy sentiment labels since stock returns can be influenced by many non-text factors (Huang, Wang, and Yang, 2023).

Consequently, we use the Financial Phrasebank dataset to acquire s_i and X_i, inspired by Chen et al. (2022). This approach addresses the drawbacks of the above-mentioned methods, as the dataset is publicly available, so the labeling costs are zero, and it is labeled by financial experts, ensuring accurate sentiment labels. There are three sentiment classes in the dataset, *negative* (1), *neutral* (2), and *positive* (3). Accordingly, s_i is assumed to follow a 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; and θ_m is column m of θ with $m = 1, 2, 3$. The matrix multiplication $X_i^\top W^{SS}$ serves as an aggregation of the vectors of the words appearing in document i into a single vector representing the document.

It is worth noting that although we use a linear model, $X_i^\top W^{SS}\theta_m$, to parameterize the likelihood function, the proposed method can be extended to more advanced approaches, including state-of-the-art language models. Specifically, W^{SS} serves as the word embedding layer of the language model, X_i represents the set of tokens encoded by the language models, and θ_m denotes the language model parameters.

2.4 Textual feature extraction and 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*, which stands for *term frequency–inverse document frequency* (Manning & Schutze, 1999). Despite a lack of theoretical justification, Manning and Schutze (1999) suggest that the *tf.idf* representation is useful in document retrieval applications. Technically, we define V as the vocabulary of the FinText model, and $|V|$ as the number of distinct words in V. X_i can now 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 the documents containing word w_j (df_{ij}). We follow the computation specified by Schütze et al. (2008) with a unit smoothing factor for df_{ij} to avoid division by zero. 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 cosine similarity as the distance measure between W^{SS} and W^{Fin}.[6] This choice is motivated by the fact that Word2Vec learns words that are adjacent to each other in terms of cosine similarity (Levy & Goldberg, 2014). Technically,

$$\begin{aligned}\Delta(W^{SS}, W^{Fin}) &= \frac{1}{|V|}\sum_{j=1}^{|V|}\Delta(\mathbf{w}_j^{SS}, \mathbf{w}_j^{Fin}) \\ &= \frac{1}{|V|}\sum_{j=1}^{|V|}\left[1 - \frac{\langle \mathbf{w}_j^{SS}, \mathbf{w}_j^{Fin}\rangle}{\|\mathbf{w}_j^{SS}\|\|\mathbf{w}_j^{Fin}\|}\right]\end{aligned} \tag{5}$$

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

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

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

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

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

3 Data

We estimate our sentiment-semantic word vectors by utilizing the Financial Phrasebank dataset (Malo, Sinha, Korhonen, Wallenius, and Takala, 2014).[7] This dataset was 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 was manually annotated by 16 people with financial expertise who categorized the documents into three sentiment classes: *negative*, *neutral*, and *positive*. The Financial Phrasebank dataset features a high imbalance in the distribution of documents by labeled sentiment (with 604 negative, 2, 879 neutral, and 1, 363 positive documents). In line with Malo et al. (2014), we adopt the F1 score as the evaluation metric for our approach, to accommodate the lack of balance in the dataset.

After obtaining the sentiment-semantic word vectors from 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).[8] 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 in text-file format with HTML tags having been removed. We construct our management sentiment index based only on 10-K filings because it is acknowledged that their information is more significant than that of 10-Q filings (Griffin, 2003). We extract the MD &A section from 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.[9] When compared with the extraction rate of Loughran and McDonald (2011), which is roughly 50%, our rate is reasonable. We further discard 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.[10]

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

- the Standard and Poor' (S &P) 500 and the value-weighted CRSP indexes; both include dividends and are queried from the Wharton Research Data Service (WRDS); and
- the one-month US Treasury bill rate used as the risk-free rate, available from Kenneth R. French's data collection.[11]

[6] 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 the two vectors; and $\|x\|$ is the Euclidean norm of a vector x. To use this as the distance measure between x and y, people usually subtract the cosine similarity from one, i.e., $1 - \text{sim}(x,y)$.

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

[8] Available at: https://sraf.nd.edu/.

[9] Extracting the MD &A section from a 10-K file is not trivial, although it may seem as straightforward as searching for "Item 7. Management Discussion and Analysis." The phrase can appear in the Table of Contents or other items, complicating the search. Even when the phrase is correctly navigated, identifying the subsequent item remains challenging, as it could be "Item 7A" or "Item 8." Another hurdle in this process is that the MD &A does not always appear as "Item 7." These issues often result in the incomplete or inaccurate extraction of the MD &A section from 10-K filings. For a detailed discussion of these problems, refer to Section 6 of Loughran and McDonald (2016).

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

[11] Available at: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

The predictive power of the management sentiment in relation to stock returns could be rooted in the reflection of firm managers concerning the business cycle or macroeconomic conditions. To delve into the macroeconomic channels associated with the stock return predictability based on management sentiment, we leverage the monthly macroeconomic dataset provided by Welch and Goyal (2008).[12] As it is expected to connect directly with macroeconomic fundamentals, this dataset has gained popularity in stock return forecasting literature that uses macroeconomic variables (Chen, Pelger, and Zhu, 2023; Cochrane, 2011; Gu, Kelly, and Xiu, 2020; D. Huang, Jiang, Tu, and Zhou, 2015; Jiang, Lee, Martin, and Zhou, 2019). In particular, the dataset includes 14 monthly macroeconomic variables: log dividend-price ratio (*DP*), log 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 Sect. 2, semantic-only word vectors, that is, W^{Fin}, tend to group together 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, by clustering words with similar sentiments together. Despite that, we also show that to a large extent W^{SS} retains word semantics in the financial context. Moreover, we show that W^{SS}, which excels in sentiment and semantic encapsulation, classifies document sentiment more accurately than W^{Fin} and the Loughran–McDonald dictionary-based method.

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

We first assess the proficiency of W^{SS} at capturing both the sentiment and the semantics of words, through a comparison with W^{Fin}. To this end, we implement a sentiment classification at the word level. This classification relies on the presumption that if word vectors are more proficient at capturing sentiment, positive (negative) words will tend to be surrounded by more words delivering positive (negative) sentiment.[13]

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 is just used in this study to validate 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^{Fin}. To determine neighboring words, we combine two criteria: (i) the top n most similar words, denoted by n, and (ii) a pre-defined similarity threshold above which two words are determined to be similar, denoted by τ. Because W^{SS} is expected to capture sentiment meanings more effectively than W^{Fin}, we anticipate that more positive (negative) words and fewer negative (positive) words will be found in the neighborhood of positive (negative) words with W^{SS} compared to W^{Fin}. 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 entry of 2.68 means that, for every positive word defined by the Loughran–McDonald dictionary, W^{SS} yields on average 2.68 other positive words exhibiting a cosine similarity above 0.2 within the top 10 words that are most similar to the given word. Consequently, the classification sees this number as the *true positive* of W^{SS} under the corresponding set of criteria values. Similarly, based on W^{SS}, there are 0.26 negative words found within the neighborhood of positive words given the same set of criteria. Therefore, this is 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 3.45 and 0.40, respectively.

At first glance, W^{SS} outperforms W^{Fin} in allocating words into the correct sentiment categories. In particular, W^{SS} has higher true positive/negative and lower false positive/negative than W^{Fin}. Put differently, W^{SS} clusters words into the corresponding sentiment more accurately than W^{Fin}, 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.

[12] Available at: https://github.com/powder197/Goyal-and-Welch-2008-/tree/master.

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

Table 2 This table reports the confusion matrix of the word-level sentiment classification using W^{SS} and W^{Fin} with different values of the top n most similar words and similarity thresholds τ

		(True) Positive				(True) Negative			
n		**10**	**15**	**20**	**30**	**10**	**15**	**20**	**30**
Panel A: $\tau = 0.2$									
Positive	W^{SS}	**2.68**	**3.66**	**4.48**	**5.88**	**0.40**	**0.51**	**0.60**	**0.91**
	W^{Fin}	2.58	3.62	4.37	5.56	0.44	0.52	0.70	1.00
Negative	W^{SS}	**0.26**	**0.32**	**0.45**	**0.66**	**3.45**	**4.68**	**5.74**	**7.81**
	W^{Fin}	0.26	0.33	0.48	0.75	3.32	4.57	5.61	7.61
Panel B: $\tau = 0.4$									
Positive	W^{SS}	**2.66**	**3.64**	**4.42**	**5.72**	**0.40**	**0.51**	**0.58**	**0.89**
	W^{Fin}	2.56	3.60	4.32	5.48	0.44	0.52	0.67	0.95
Negative	W^{SS}	**0.24**	**0.30**	**0.40**	**0.56**	**3.39**	**4.57**	**5.51**	**7.18**
	W^{Fin}	0.25	0.30	0.42	0.62	3.27	4.45	5.38	6.95
Panel C: $\tau = 0.6$									
Positive	W^{SS}	**0.93**	**0.96**	**0.97**	**0.97**	**0.12**	**0.12**	**0.12**	**0.12**
	W^{Fin}	0.86	0.89	0.89	0.89	0.12	0.12	0.12	0.12
Negative	W^{SS}	**0.08**	**0.08**	**0.08**	**0.08**	**0.77**	**0.77**	**0.77**	**0.77**
	W^{Fin}	0.08	0.08	0.08	0.08	0.74	0.74	0.74	0.74

The positive and negative words are determined by the Loughran–McDonald dictionary. The bold numbers indicate the word vectors among W^{SS} and W^{Fin} that are more proficient in capturing sentiment, measured by their classification accuracy

So far, W^{SS} has demonstrated its superior proficiency in capturing word sentiment compared to W^{SS}. However, the 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 for each given word, and then, 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 whose meaning in the financial context is different from their casual meaning (e.g., "bond," "capital," "share"). This assessment, although prone to some subjectivity, is widely used in word representation research to evaluate the quality of word vectors regarding semantics (Das, Donini, Zafar, He, and Kenthapadi, 2022; Dieng, Ruiz, and Blei, 2020; Li, Mai, Shen, and Yan, 2021; Mikolov, Chen, Corrado, and Dean, 2013).

Table 3 presents the top ten most similar words to "bank," "bond," "capital," "cash," "debt," "inflation," "interest," "liability," "share," and "yield" based on cosine similarity and retrieved using W^{SS} and W^{Fin}. Overall, these words are surrounded by words with strong economic and financial meanings, even for those such as "bond," "capital," and "share" whose meanings depend on the context. This serves as compelling evidence that W^{SS} efficiently preserves the word semantics in the financial context.

Comparing the top similar words generated by our method and using the FinText word vectors underlines the preservation of word semantics by our method. Specifically, we find that the top ten most similar words for these chosen words remain consistent between W^{SS} and W^{Fin}, albeit with minor differences in word order. These findings extend to the other benchmark words, indicating the reliability of our method in maintaining semantic coherence.

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

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

W^{SS} has demonstrated greater proficiency than W^{Fin} in capturing both word sentiment and semantics. However, how it performs in sentiment classification remains unanswered. In conjunction with the findings in Sect. 4.1, a superior performance of W^{SS} compared to W^{Fin} in sentiment classification will add robustness to our approach to calibrating word vectors for effective sentiment and semantic representation. Indeed, many studies validate their proposed models by sentiment classification

Table 3 This table reports the top ten most similar words to the corresponding words based on the sentiment-semantic word vectors (W^{SS}) and FinText (W^{Fin})

Bank		**Bond**		**Capital**		**Cash**		**Debt**	
W^{SS}	W^{Fin}	W^{SS}	W^{Fin}	W^{SS}	W^{Fin}	W^{SS}	W^{Fin}	W^{SS}	W^{Fin}
Banking	Banking	Debt	Debt	Tlcom	Tlcom	Liquidity	Liquidity	Liquidity	Liquidity
Lender	Lender	Treasury	Treasury	Funding	Funding	Debt	Debt	Indebtedness	Indebtedness
Central	Central	Yield	Yield	Investment	Equity	Paid	Paid	Financing	Bond
Deutsche	Deutsche	Issuance	Issuance	Equity	Investment	Payment	Payment	Borrowing	Borrowing
Lending	Citigroup	Eurobond	Eurobond	Financing	Financing	Pay	Pay	Bond	Financing
Citigroup	Lending	Fund	Fund	Liquidity	Liquidity	dividend	dividend	Funding	Funding
Ubs	Ubs	Municipal	Municipal	Fund	Expenditure	Financing	Financing	Issuance	Issuance
Institution	Institution	Mortgage	Mortgage	Expenditure	Fund	Quarterly	Consideration	Repayment	Repayment
Mortgage	Mortgage	Schuldschein	Schuldschein	Venture	Venture	Consideration	Quarterly	Cash	Cash
Finance	Finance	Frn	Equity	Cap	Financial	Hand	Hand	Equity	Burden
Inflation		**Interest**		**Liability**		**Share**		**Yield**	
GDP	GDP	Rate	Rate	Damage	Damage	Per	Per	Treasury	Bond
Economy	Economy	Open	Open	responsibility	Responsibility	Common	Common	Bond	Treasury
Economic	Economic	Income	Income	Exposure	Exposure	Stock	Stock	Spread	Spread
Wage	Wage	Ownership	Ownership	Claim	Claim	Unit	Million	Curve	Benchmark
Headline	Headline	Raise	Expense	Environmental	Environmental	Million	Unit	Benchmark	Curve
Output	Growth	Equity	Raise	Compensation	Compensation	Purchase	Purchase	Two year	Two year
Growth	Output	Debt	Equity	Legal	Legal	Dividend	Billion	Five year	Five year
Consumption	Persistently	Margin	Debt	Warranty	Warranty	Price	Earnings	Percentage	Percentage
Persistently	Consumption	Expense	Margin	Expense	Expense	Billion	Dividend	Return	Bps
Rate	Rate	Payment	Increase	Adjustment	Adjustment	Countervalue	Price	Bps	Return

Similar words are chosen by cosine similarity

(Huang, Wang, and Yang, 2023; Li, 2010a; Lutz, Pröllochs, and Neumann, 2020), 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,

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

where ϕ^k is a d-dimensional vector of model parameters associated with the word vectors W^k; the other notations are defined in Sect. 2. It should be 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 as $\hat{p}(s_i = m|\hat{\phi}^k_m, W^k, X_i)$.

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^{Fin}. 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,

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

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^{Fin}, and (iii) the word vector approach based on W^{SS}

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^{Fin}	0.13	0.80	0.46	0.66	0.46
C - W^{SS}	**0.45**	**0.80**	**0.60**	**0.71**	**0.61**

The bold numbers indicate the most superior method in classifying document sentiment in terms of each evaluation metric

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, as shown. The F1 scores showcased in this table are calculated using the testing part of the Financial Phrasebank dataset

Besides W^{SS} and W^{Fin}, we implement a sentiment classification model using the Loughran–McDonald dictionary. While comparing the classification capability of W^{SS} with that of W^{Fin} makes manifest the importance of capturing sentiment when classifying sentiment using word vectors, the comparison of the word vectors (i.e., W^{SS} and W^{Fin}) and the Loughran–McDonald dictionary demonstrates the significance of word semantics in sentiment classification. Following the convention used in many studies (Henry, 2008; Jiang, Lee, Martin, and Zhou, 2019; Loughran & McDonald, 2011), 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 *negative*, *neutral*, and *positive* sentiments; and $\#(\text{pos})_i$ and $\#(\text{neg})_i$ are, respectively, the number of positive and number of negative words defined by the Loughran–McDonald dictionary that appear in document *i*.

Like Malo et al. (2014), we opt for the F1 score as the evaluation metric for this classification. We present both class-wise and global F1 scores across sentiment categories for a more comprehensive assessment. Since the F1 score is traditionally used in binary classification, one needs to adapt it for multi-class classification problems via aggregation. Specifically, we apply two types of aggregation: micro- and macro- averages.[14]

Table 4 reports a wide range of 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^{Fin} and W^{SS}. In general, the word vector approach using W^{SS} outperforms its competitors in classifying sentiment. Together with the results of Sect. 4.1, this reinforces the success of our approach in injecting the sentiment meanings into semantic-only word vectors, and subsequently in reflecting a more accurate document sentiment classification.

A more thorough examination of Table 4 reveals deeper insights into the sentiment classification capabilities of the methods under consideration. Surprisingly, with a high F1 score in the *negative* class relative to its competitors, the Loughran–McDonald dictionary performs fairly well in identifying pessimistic documents. 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^{Fin} performs the worst in classifying pessimism, even in comparison with the dictionary-based approach, by a large margin: its F1 score is only 0.13 compared to the scores of the other approaches of 0.36 and 0.45. Combined with the lowest macro-average F1 score of W^{Fin}, this result suggests that relying exclusively on word semantics is inadequate for gauging nuanced sentiment expressions precisely.

Third, the superior performance of W^{SS} in most cases suggests that, in order to measure sentiments accurately,

[14] Technically, for the micro-averaged F1 score, we calculate the true positives, false positives, and false negatives across all sentiment categories, thereby accounting for the imbalanced labels. The global F1 score under this aggregation is then derived from these aggregated counts. In contrast, the macro-averaged F1 score is computed by first determining the F1 score for each sentiment class individually. The global F1 score under this aggregation is then obtained as the unweighted average of these individual F1 scores, treating each class equally regardless of its size. Consequently, we prioritize the micro-averaged F1 score as our main evaluation metric, because of the imbalanced nature of the Financial Phrasebank dataset. We refer readers to Grandini et al. (2020) and Takahashi et al. (2023) for further technical details of F1 scores.

both word sentiment and semantics are required. Comparing W^{SS} with the Loughran–McDonald approach reveals that word semantics, when standing alone, may not be powerful but are still crucial in classifying sentiment precisely. Our findings correspond with many criticisms of bag-of-word methods because of their treatment of words as independent units; see Huang et al. (2023), Mikolov et al. (2013), Li et al. (2021) among others.

Associated with the empirical results shown in Sect. 4.1, two conclusions can be drawn. 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 sentiment. Ultimately, the next question is how an accurate sentiment measurement can be applied to explore economic values or answer financial puzzles. Subsequent sections will delve into this intriguing topic.

5 Appendix A: Construction of the management sentiment index

To examine the predictive effects of management sentiment on stock markets, it appears natural to construct an index that conveys firm managers' sentiment through corporate disclosures. Subsequently, the connection between this index and the series of future stock returns is investigated. Attempts of this nature are commonly based on bag-of-words approaches, in which the sentiment of a document is a projection of a pre-defined lexicon (Feldman, Govindaraj, Livnat, and Segal, 2010; Henry, 2008; X. Huang, Teoh, and Zhang, 2014; Jiang, Lee, Martin, and Zhou, 2019; Loughran & McDonald, 2011; Price, Doran, Peterson, and Bliss, 2012). Other studies obtain sentiment labels by human annotation (Li, 2010a) or by the associated market reactions (Frankel, Jennings, and Lee, 2022; Jegadeesh & Wu, 2013).

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.[15] 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. The moving average is implemented retrospectively, utilizing data from the past four months to prevent look-ahead bias. 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^{Fin}. The first sentiment index is constructed based on a word-count approach using the Loughran–McDonald dictionary (Loughran & McDonald, 2011). The second sentiment index is formed in a similar way to S^{SS}, but using the FinText word vectors, W^{Fin}, instead of W^{SS}. The two sentiment indexes are both derived from the MD &A corpus and are aggregated in a similar way to S^{SS} as described in the previous paragraph. As the final step, the three management sentiment indexes are standardized to have zero means and unit variances to eliminate the effects of scale difference.[16] As shown by the empirical results in Sect. 4, using S^{SS} in presenting the management sentiment index is more advantageous than using S^{Fin} or S^{LM} because of the more effective sentiment representation of W^{SS} over W^{Fin} and the Loughran–McDonald approach.

Figure 1 presents the variations of the three sentiment indexes over time. At first glance, S^{LM} and S^{Fin}, the sentiment indexes built by the Loughran–McDonald dictionary and the FinText word vectors, respectively, exhibit strong seasonality throughout most of the data sample. While S^{LM} only shows a decline during the dot-com crisis, S^{Fin} displays a local decreasing trend during the financial crisis and remains steady during the dot-com recession. This implies that S^{LM} and S^{Fin} do not adequately capture explanatory information about business cycles or historical events. This observation is expected because the Loughran–McDonald dictionary and W^{Fin}, which are the core of S^{LM} and S^{Fin}, capture only one aspect of the sentiment-semantic trade-off.

Unlike S^{LM} and S^{Fin}, S^{SS} aligns well with business states, especially the essence of the two recessions. In particular, the management sentiment based on S^{SS} starts low but gradually increases until before the dot-com crisis. During the dot-com crisis, the management sentiment

[15] 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 unbalanced. Consequently, biases caused by a distributional shift may occur if weights are not applied.

[16] The standardization is implemented for the whole time series for the visualization and in-sample regression studies in Sect. 6.1. For the out-of-sample studies, index standardization is executed in a recursive-window manner to avoid look-ahead bias; see Sect. 6.2 for further details of the study design.

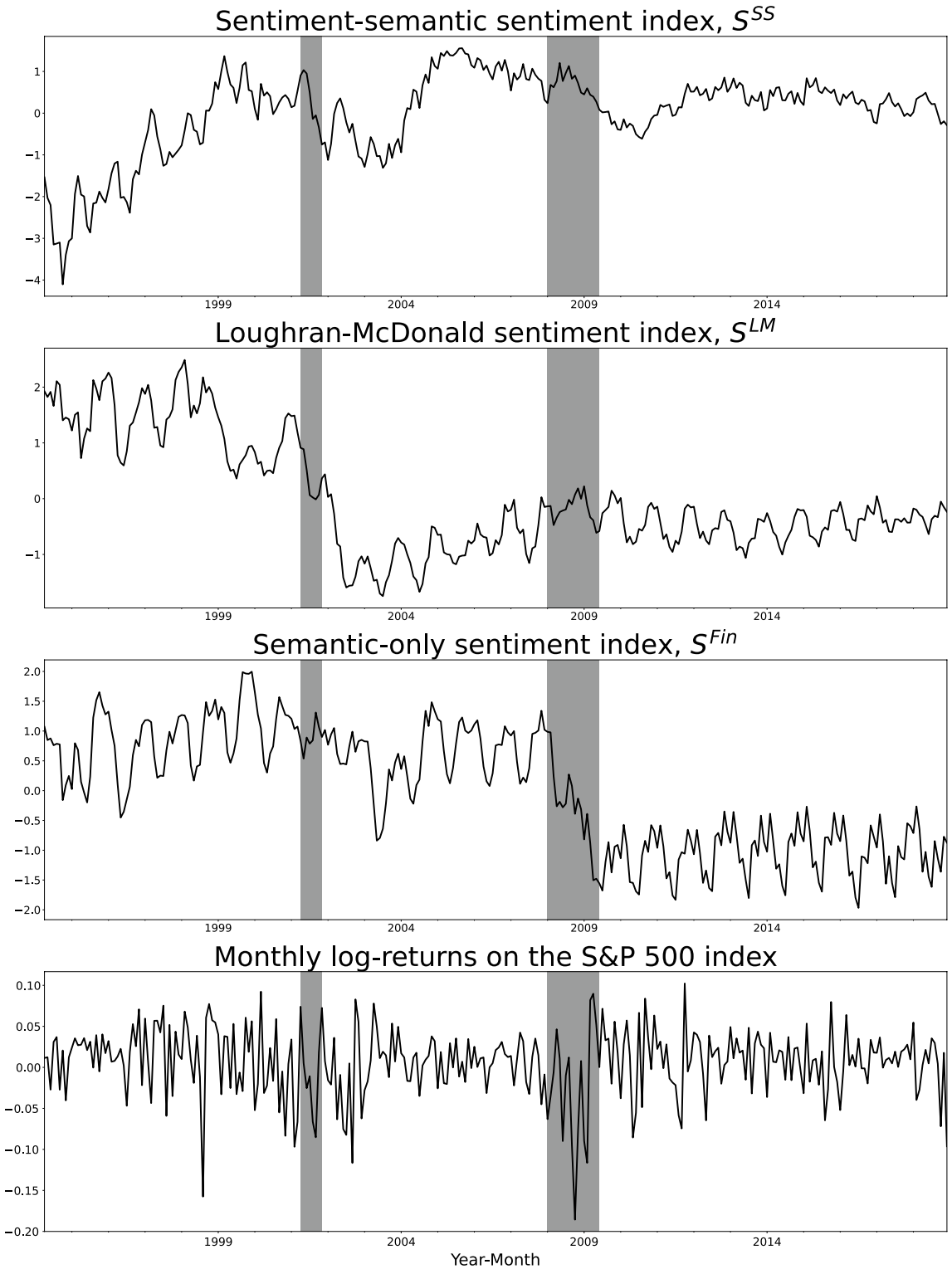

Fig. 1 The market-level management sentiment indexes extracted from the MD &A section of 10-K filings. The first plot depicts the management sentiment index S^{SS} constructed using the sentiment-semantic word vectors trained on the Financial Phrasebank dataset. The second plot is of the sentiment index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

drops, and it then remains low until around 2003. This period coincides with several high-profile accounting fraud cases (e.g., Enron and Worldcom) coming to light, which may have driven down management sentiment. Following this period, S^{SS} rises and then remains high in value until before the 2008 financial crisis, implying that firm managers tended to use an optimistic tone in their MD &A during this time. During the financial crisis, S^{SS} again displays a decrease in management sentiment. As opposed to S^{LM} and S^{Fin}, S^{SS} does not exhibit noticeable seasonality across the data sample.

6 Predictive regression analysis

In this section, we provide empirical evidence regarding the 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}. We focus our comparative analyses on S^{LM} instead of S^{Fin} because the Loughran–McDonald dictionary is widely used to extract sentiment in the current literature (Loughran & McDonald, 2011; Jiang, Lee, Martin, and Zhou, 2019; Sautner, Van Lent, Vilkov, and Zhang, 2023), thus serving as a strong benchmark. We implement the analyses in both an in-sample and an out-of-sample manner to guarantee the robustness of our findings.

6.1 In-sample market return predictability

We first examine the market return predictability for 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,

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

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 $t+h$); and *Recession* is a dummy variable indicating the economic recessions defined by the National Bureau of Economic Research.

Our experiment is inspired by the paper by Jiang et al. (2019) yet possesses two important differences. First, in addition to the S &P 500 index, we also, like Jegadeesh and Wu (2013), use the value-weighted CRSP index to compute the market returns. This additional index is expected to enhance the robustness of the test results. Second, we control for recession fixed effects and monthly fixed effects in all equations to capture the potential variations caused by seasonality and business cycles. As shown in Fig. 1, the recessions negatively affect both management sentiment and the S &P 500 index.[17] Therefore, omitting the recession control may lead to inconsistent estimates. While the seasonality is not visible with S^{SS}, it is pronounced in S^{LM}. As a result, we decide to include monthly dummies in all equations to obtain fair comparisons.

Table 5 presents the regression results of equation 10 over h-month horizons with $h = 1, 3, 6, 9, 12$,[18] First, for the univariate regressions, the coefficients on S^{SS} are negative and significant at the 5% level for the semi-annual to one-year horizons. We do not, however, observe any significant coefficients on S^{LM}. The level of significance is stronger with the S &P 500 index compared with the value-weighted CRSP index, evidenced by significant coefficients across all horizons. Intuitively, the sentiment index, which integrates word sentiment and semantics, is negatively associated with future cumulative excess market returns. In contrast, the index based solely 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 with S^{LM}. In particular, adding S^{LM} to the models with only S^{SS} results in limited changes to the sign and the significance of the coefficients for S^{SS} in all regressions. Moreover, the R^2s of the bivariate models are similar to those 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 for the two models is 16.8% and 16.4%, respectively). With a substantial correlation of -0.499 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 semi-annual horizon, an increase of one standard deviation in the management sentiment is associated with a decrease of 2.3% in the cumulative returns with the value-weighted CRSP index and a decrease of 2.7% with the S &P 500 index. Furthermore, the estimated coefficient on S^{SS} increases in absolute value as h increases. This result implies that S^{SS} consistently and significantly predicts the cumulative excess

[17] In particular, the S &P 500 index suffered negative returns during most of the 2008 financial crisis.

[18] It is well known that the return predictive regressions usually suffer from various econometric issues: spurious inference results due to persistent independent variables (Ferson, Sarkissian, and Simin, 2003) small-sample bias (Stambaugh, 1999), and potential biased standard error estimation (Hodrick, 1992). We use the Newey–West heteroscedasticity- and autocorrelation-robust t-statistics with small-sample adjustment for consistent covariance matrix estimation to cope with the above-mentioned issues.

Table 5 This table reports the results of the OLS regressions of equation 10 over *h*-month horizons with $h = 1, 3, 6, 9, 12$

	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}	−0.003		−0.003	−0.010		−0.011	−0.023**		−0.029*	−0.033***		−0.043**	−0.044***			−0.056**
	(−1.231)		(−0.845)	(−1.575)		(−1.151)	(−2.286)		(−1.829)	(−2.778)		(−2.280)	(−3.186)			(−2.578)
S^{LM}		0.002	0.000		0.003	−0.003		0.003	−0.012		0.003	−0.019		0.007		−0.022
		(0.552)	(0.095)		(0.398)	(−0.234)		(0.252)	(−0.624)		(0.189)	(−0.805)		(0.298)		(−0.796)
R^2	0.100	0.099	0.100	0.127	0.117	0.127	0.162	0.132	0.168	0.180	0.136	0.190	0.184	0.127		0.195
Panel B: S &P 500 index																
S^{SS}	−0.005*		−0.004	−0.014**		−0.015*	−0.027***		−0.032**	−0.038***		−0.045***	−0.050***			−0.058***
	(−1.669)		(−0.902)	(−2.491)		(−1.726)	(−2.784)		(−2.170)	(−3.311)		(−2.605)	(−3.395)			(−2.842)
S^{LM}		0.003	0.001		0.006	−0.002		0.007	−0.010		0.010	−0.013		0.015		−0.016
		(0.189)	(−0.383)		(0.868)	(−0.192)		(0.540)	(−0.553)		(0.568)	(−0.605)		(0.634)		(−0.584)
R^2	0.103	0.100	0.103	0.133	0.114	0.133	0.164	0.121	0.168	0.197	0.140	0.202	0.200	0.130		0.205

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, respectively, the sentiment-semantic word vectors and the Loughran–McDonald dictionary (Loughran & McDonald, 2011). A constant term (α) and a recession dummy (*Recession*) are also included in each regression equation. The coefficients, Newey–West heteroscedasticity- 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

market returns in the long run. In addition, the predictive power of 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 10.0% to 18.4% with the value-weighted CRSP index, and from 10.3% to 20.0% with the S &P 500 index. This means that S^{SS} is a factor that can explain large in-sample variations of future excess market returns. Moreover, the out-of-sample tests presented in Sect. 6.2 show that this result is maintained out-of-sample.

In conclusion, these results contribute to those of 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 of the stock market, provided that the sentiment is precisely measured.

6.2 Out-of-sample market return predictability

In numerous predictive analyses, researchers discover substantial predictive evidence with in-sample data, yet struggle to obtain significant predictive power with out-of-sample data (Inoue & Kilian, 2005). Additionally, out-of-sample analyses tend to be more resilient to the econometric issues described in Sect. 6.1 (Busetti & Marcucci, 2013). Therefore, to provide a robust validation of 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 great interest. Moreover, as demonstrated in Sect. 6.1, S^{SS} encompasses predictive information additional to that of S^{LM} for 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 when compared to S^{LM}. Accordingly, we compare the market return predictive power of: (i) each of the sentiment indexes S^{SS} and S^{LM} with that of the historical average benchmark; and (ii) the model combining 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: $CER_{t+1\to t+h} = \alpha_{2A} + \beta_{2A} S_t^k + \epsilon_{t+1\to t+h}$, $k \in \{SS, LM\}$

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. As in the in-sample studies, 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.

Unlike the in-sample studies, the recession dummy is not included in all models because recessions are typically determined ex-post using macroeconomic variables, which often suffer from delays and revisions in their publication (Clements, 2019). By excluding the recession dummy, the out-of-sample prediction resembles a real-time prediction. We also exclude monthly dummies from the models in Test A for two reasons. First, Model 1A without monthly dummies serves as the robust historical average benchmark described by Welch and Goyal (2008). Second, including monthly dummies only in Model 2A would skew the comparison of the forecasting power between the sentiment indexes and the benchmark, as the forecasting power of Model 2A would be partly impacted by the monthly dummies. For Test B, the results remain robust to the inclusion of monthly dummies; see Appendix C for further details.

The tests are implemented in a recursive-window manner (West & McCracken, 1998), in which the data from 1994:01 to 1999:12 is the initial training set, and the data period from 2000:01 to 2018:12 is used as the evaluation period.[19] It is worth noting that the sentiment indexes are also standardized recursively to avoid look-ahead bias. In particular, in one window, we execute the following steps in order: (i) standardize the index; (ii) estimate the model; (iii) standardize the index values in the prediction part by the statistics of the index in the training part; and (iv) make predictions of returns.

We define the mean squared prediction error (MSPE) as a measure of prediction accuracy.[20] In both tests A and B, we want to test the null hypothesis that the MSPE of the parsimonious model is smaller than or equal to that of the nested models against the alternative hypothesis that the nested model has a smaller MSPE than the parsimonious model. To this end, we use the Campbell and Thompson (2008) out-of-sample R^2 statistic (R^2_{OS}) which is defined as follows:

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

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

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 $t+h$

		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}	0.129	3.055*	2.356	5.555	6.544
		(0.755)	(1.599)	(1.063)	(1.115)	(1.010)
	S^{LM}	−1.916	−1.966	−2.311	−2.460	−1.218
		(0.099)	(−0.081)	(0.631)	(0.623)	(0.380)
Test B		−1.736	2.567*	1.047	3.500	6.639
		(0.231)	(1.553)	(1.114)	(1.178)	(1.204)
Panel B: S &P 500 index						
Test A	S^{SS}	0.263	3.476*	7.365**	8.693**	10.19**
		(0.674)	(1.521)	(1.731)	(1.709)	(1.656)
	S^{LM}	−0.020	−0.015	−0.011	−0.020	−0.022
		(0.827)	(1.002)	(0.936)	(0.980)	(1.130)
Test B		−2.163	1.325*	7.315**	8.416**	9.384**
		(0.803)	(1.291)	(1.692)	(1.703)	(1.667)

S^{SS} and S^{LM} are the management sentiment indexes extracted from the MD &A section of 10-K filings using, respectively, the sentiment-semantic word vectors and the Loughran–McDonald dictionary (Loughran & McDonald, 2011). 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 MSPE of the parsimonious models (i.e., the historical average model in test A, and the S^{LM}-only model in test B) is smaller than or equal to the MSPE of the nested models. The tests are implemented in a recursive-window manner (West & McCracken, 1998), in which the data from 1994:01 to 1999:12 is the initial training set, and the data period from 2000:01 to 2018:12 is used as the evaluation period. * and ** denote significance at the 10% and 5% levels, respectively

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

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

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^2_{OS}. 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 500 indexes. For the value-weighted CRSP index, we observe that S^{SS} significantly (at the 10% level) improves the historical average in predicting the cumulative excess market returns at the three-month horizon. In contrast, S^{LM} exhibits no predictive power for the cumulative market returns. With significant R^2_{OS} at the three-month, semi-annual, nine-month, and one-year horizons in test B, we see that S^{SS} adds significant predictive 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 for future market returns than it does for the value-weighted 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 the time horizons considered 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 any model, suggesting that the sentiment of the MD &A documents, as extracted by the Loughran–McDonald

dictionary, does not contain more predictive information for future market returns than the historical average model. Test B conducted on the S &P 500 index further corroborates the findings regarding the enhanced predictive capability of S^{SS} over 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, which were previously shown to lack predictive power regarding future market returns, now exhibit enhanced predictive ability because of the inclusion of S^{SS}.

Jiang et al. (2019) find that management sentiment in the current month t contains predictive information beyond that of the historical average benchmark for predicting the market return in the next month $t+1$, but our results suggest the opposite. 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, p. 129). Consequently, their management sentiment index can exploit earlier effects of the sentiment on the stock markets. However, with our results, we show that exploiting the sentiment exclusively from the MD &A section of 10-K filings is capable of capturing mispricing information in stock prices. Our findings contribute to the literature on stock return predictability and corporate disclosures by showing that the mispricing information contained in the management sentiment embedded in the 10-K filings may to some extent be concentrated in the MD &A section.

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 on this finding, using the lens of macroeconomic channels.

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 the 14 macroeconomic variables described in Sect. 3. Unlike the model in Sect. 6.1, we exclude monthly dummies here because neither S^{SS} nor the macroeconomic variables show seasonal patterns.

Table 7 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 all the macroeconomic variables except the dividend-price ratio (*DP*) and the dividend yield (*DY*). These results imply that the management sentiment index S^{SS} may capture information relating to the dividend payments 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*) and the default return spread (*DFR*).

The out-of-sample test results, which are detailed in Table 8, reinforce these findings. We find that S^{SS} makes a limited contribution to the predictive ability of the dividend-related variables (the dividend-price ratio (*DP*), dividend yield (*DY*), and dividend-payout ratio (*DE*)). As 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.

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 Sect. 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.[21]

We conjecture that this result is rooted to some extent in the discussions of firm managers in the MD &A section regarding the dividend payment plans. 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..."

[21] For reasons of space, we present the mutual correlations of the 14 macroeconomic variables and S^{SS} for reference in Table 9 in Appendix B.

Table 7 This table reports the in-sample OLS regression results of equations 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	0.027	0.202***	0.427***	0.024	−0.134	0.181***	−0.988	0.400***	−1.297
	(1.129)	(3.420)	(4.890)	(0.879)	(−0.420)	(2.725)	(−1.065)	(4.038)	(−0.865)
DY	0.064***	0.234***	0.457***	0.067***	0.157	0.219***	-0.710	0.436***	−0.999
	(2.817)	(3.825)	(5.350)	(2.611)	(0.487)	(3.288)	(−0.744)	(4.666)	(−0.660)
DE	0.023	0.109*	0.186***	0.021	−0.164	0.097*	−1.599**	0.165***	−2.941***
	(1.190)	(1.963)	(2.866)	(1.097)	(−0.736)	(1.932)	(−2.477)	(2.652)	(−2.970)
EP	−0.014	−0.036	−0.027	−0.014	−0.300	−0.038	−2.200***	−0.030	−3.939***
	(−0.651)	(−0.409)	(−0.235)	(−0.604)	(−1.315)	(−0.434)	(−3.097)	(−0.275)	(−3.470)
SVAR	−4.224***	−1.217	3.807	−4.198***	−0.231	−0.974	−2.163***	4.255*	−3.981***
	(−8.497)	(−0.702)	(1.608)	(−5.214)	(−1.012)	(−0.553)	(−3.061)	(1.847)	(−3.541)
B/M	−0.012	0.393**	0.866**	−0.006	−0.287	0.448**	−2.486***	0.967***	−4.587***
	(−0.170)	(2.157)	(2.434)	(−0.088)	(−1.296)	(2.482)	(−2.998)	(2.795)	(−3.584)
NTIS	0.149	1.043	1.551	0.074	−0.246	0.466	−1.895***	0.439	−3.655***
	(0.637)	(1.381)	(0.962)	(0.255)	(−0.970)	(0.554)	(−3.113)	(0.269)	(−3.917)
TBL	−0.130	−0.620	−1.639	−0.187	−0.377	−1.024	−2.645***	−2.405**	−5.021***
	(−0.649)	(−0.941)	(−1.234)	(−0.986)	(−1.519)	(−1.603)	(−3.145)	(−2.050)	(−3.547)
TMS	0.186	0.933	3.516*	0.128	−0.263	0.476	−2.072***	2.787	−3.308***
	(0.572)	(0.722)	(1.690)	(0.312)	(−1.134)	(0.363)	(−3.116)	(1.368)	(−3.225)
DFY	−0.572	4.909	13.49*	−0.406	−0.272	6.424	−2.484***	16.36**	−4.704***
	(−0.377)	(0.842)	(1.831)	(−0.240)	(−1.196)	(1.074)	(−3.050)	(2.310)	(−3.7199)
DFR	1.221***	1.446***	1.632***	1.218***	−0.260	1.418***	−2.141***	1.582***	−3.880***
	(10.19)	(3.420)	(2.974)	(6.087)	(−1.366)	(3.472)	(−3.176)	(3.125)	(−3.612)
INFL	0.790	−3.288	−8.113**	0.833	−0.300	−2.984	−2.144***	−7.569**	−3.837***
	(0.547)	(−1.323)	(−2.332)	(0.627)	(−1.342)	(−1.266)	(−3.110)	(−2.206)	(−3.536)

in which the complementary predictive power of S^{SS} in addition to the 14 macroeconomic variables (Welch & 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 Sect. 3. A constant α and a recession dummy are also included in each regression equation. The coefficients, Newey–West heteroscedastic- 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

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 the 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 the dividend-related information located in the MD &A section is a cause of the predictive power of the management sentiment. The causal effects are left for future studies.

8 Conclusion

This paper sheds light on the ability of the sentiment contained in the MD &A section of 10-K filings from January 1994 to December 2018 to predict future returns. Unlike most existing studies, we introduce a novel method for

Table 8 This table reports the out-of-sample stock return predictability of the management sentiment index

	R^2 (%) and adjusted MSPE (in parentheses)				
h **(months)**	**1**	**3**	**6**	**9**	**12**
DP	−4.182	−1.694	2.142	2.495	3.571
	−0.175	0.382	0.632	0.850	0.938
DY	−3.178	−1.032	1.866	2.223	3.246
	−0.149	0.349	0.605	0.845	0.977
DE	−0.529	1.097	3.642	5.661	8.071*
	0.206	0.701	0.948	1.124	1.302
EP	−3.773	0.397	6.794*	8.253*	9.882**
	−0.570	0.528	1.552	1.560	1.685
SVAR	0.134	1.903**	6.685**	9.895**	12.20**
	1.073	1.795	2.080	2.109	2.270
B/M	−2.015	2.916*	10.60**	14.48**	18.89**
	−0.429	1.474	2.181	2.141	2.208
NTIS	−0.259	0.944	2.806*	3.290*	4.261*
	−0.380	1.220	1.596	1.527	1.620
TBL	0.361	4.138**	9.758***	12.43***	16.12**
	1.040	2.318	2.355	2.340	2.281
TMS	0.340	3.522	7.087*	8.212*	9.163*
	0.580	1.128	1.355	1.344	1.393
DFY	0.085	3.195*	8.408**	10.70**	13.49**
	0.622	1.603	1.760	1.907	2.283
DFR	0.340	3.499*	7.482**	8.910**	10.43**
	0.821	1.607	1.792	1.751	1.675
INFL	0.315	3.772*	7.474**	8.889**	10.51**
	0.666	1.470	1.785	1.792	1.723

S^{SS}, in addition to the 14 macroeconomic variables (Welch & 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 Sect. 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 & 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

accurately gauging 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, whose variations match well, conceptually, with different economic episodes. The index based on the semantic-only approach is, however, unable to produce meaningful interpretations of the states of the economy. This observation once again reaffirms the importance of sentiment nuances captured by word vectors in exploring the economic implications of MD &A documents.

Finally, our proposed management sentiment index is a strong negative predictor of future stock returns. Moreover, we show that it embraces predictive insights concerning future stock returns beyond the dictionary-based sentiment index. These findings hold in both in-sample and out-of-sample setups. Based on these results, three conclusions are drawn concerning the 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 information regarding firm conditions that may lead to stock mispricing. 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 is because a word is encoded by a single numerical vector regardless of the surrounding context in a sentence or paragraph. This limitation suggests an extension of the current work with language models like FinBERT (Huang, Wang, and Yang, 2023), associated with our proposed method. The dynamical contextualization of language models is anticipated to uncover more insights into corporate disclosures.

Appendix A: 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 $X^{MDA}_{i,t}$. Follow the instructions in Sect. 4.2, the predicted probability of each sentiment class m, with $m = 1, 2, 3$, conditioning on W^{SS} is $\hat{p}(s_i = m|\hat{\phi}^{SS}_m, W^{SS}, X^{MDA}_{i,t})$. It is worth noting that $\hat{\phi}^{SS}_m$ 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,

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

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.

Appendix B: B Correlation of S^{SS} and macroeconomic variables

See Tables 9 and 10

Table 9 This table reports the correlations for the management sentiment index S^{SS} and the 14 macroeconomic variables

	S^{SS}	*DP*	*DY*	*DE*	*EP*	*SVAR*	*B/M*	*NTIS*	*TBL*	*TMS*	*DFY*	*DFR*	*INFL*
S^{SS}	1.000												
DP	−0.327	1.000											
DY	−0.340	0.982	1.000										
DE	−0.088	0.446	0.438	1.000									
EP	−0.096	0.098	0.097	−0.847	1.000								
SVAR	0.109	0.190	0.115	0.366	−0.294	1.000							
B/M	0.105	0.656	0.636	−0.019	0.410	0.054	1.000						
NTIS	−0.451	−0.307	−0.284	−0.278	0.127	−0.251	−0.216	1.000					
TBL	−0.277	−0.345	−0.338	−0.247	0.070	−0.104	−0.614	0.356	1.000				
TMS	−0.181	0.319	0.314	0.354	−0.204	0.154	0.353	0.137	−0.688	1.000			
DFY	0.231	0.414	0.382	0.696	−0.528	0.597	0.310	−0.498	−0.446	0.366	1.000		
DFR	−0.021	0.004	0.086	0.168	−0.185	−0.248	-0.028	0.023	−0.086	0.114	0.108	1.000	
INFL	0.032	−0.119	−0.110	−0.092	0.032	−0.326	−0.063	0.077	0.123	−0.039	−0.221	−0.019	1.000

The definitions of the 14 macroeconomic variables are given in Sect. 3. The data sample spans the period from 1994:01 to 2018:12

Table 10 This table reports the out-of-sample Test B in Sect. 6.2 in case monthly dummies are included. R^2_{OS} is the out-of-sample Campbell and Thompson (2008) R^2

	R^2_{OS}				
	1 month	**3 months**	**6 months**	**9 months**	**1 year**
Value-weighted CRSP	0.400	3.472*	3.359	6.296	8.128
	(−0.166)	(1.578)	(1.213)	(1.261)	(1.264)
S &P 500	0.545	3.656*	7.977**	10.43**	12.20**
	(0.980)	(1.410)	(1.670)	(1.732)	(1.674)

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 model, i.e., the S^{LM}-only model, has smaller or equal MSPE than the nested model. The tests are implemented in a recursive-window manner (West & 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

Out-of-sample return forecasting with monthly dummy inclusion

Acknowledgements
I would like to thank the editor, Daniel Kaufmann, and two anonymous referees for their valuable feedback. I also thank Bruno Jaeger, Erik-Jan Senn, Joshua Traut, Matthias R. Fengler, and the participants of the 2023 Annual Congress of the Swiss Society of Economics and Statistics for their helpful comments. Finally, I would like to thank Barry Meehan for his help in editing the manuscript.

Author Contributions
I fully contributed to my paper. The author read and approved the final manuscript.

Funding
Not applicable

Declarations

Conflict of interest
The author declares that they have no conflict of interest.

Ethical approval
Not applicable

Received: 23 February 2024 Accepted: 12 July 2024
Published online: 13 August 2024

References

Bochkay, K., & Levine, C. B. (2019). Using MD &A to improve earnings forecasts. *Journal of Accounting, Auditing & Finance, 34*(3), 458–482.
Brown, S. V., & 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., & Marcucci, J. (2013). Comparing forecast accuracy: A Monte Carlo investigation. *International Journal of Forecasting, 29*(1), 13–27.
Campbell, J. Y., & 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., & Liu, Y. (2022). Media-expressed tone, option characteristics, and stock return predictability. *Journal of Economic Dynamics and Control, 134*, 104290.
Chen, L., Pelger, M., & Zhu, J. (2023). Deep learning in asset pricing. *Management Science, 72*(2), 714–750.
Clark, T.E., & 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*, 1127–1148,
Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics, 138*(1), 291–311.
Clements, M. P. (2019). Do forecasters target first or later releases of national accounts data? *International Journal of Forecasting, 35*(4), 1240–1249.
Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance, 66*(4), 1047–1108.
Cohen, L., Malloy, C., & Nguyen, Q. (2020). Lazy prices. The. *Journal of Finance, 75*(3), 1371–1415.
Das, S. R., Donini, M., Zafar, M. B., He, J., & 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., & 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., & 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., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics, 8*, 439–453.
Dyer, T., Lang, M., & 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., & Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies, 15*(4), 915–953.
Fengler, M., & Phan, M. T. (2023). *A topic model for 10-K management disclosures Tech*. Gallen, School of Economics and Political Science: Rep. University of St.
Ferson, W. E., Sarkissian, S., & Simin, T. T. (2003). Spurious regressions in financial economics? *The Journal of Finance, 58*(4), 1393–1413.
Frankel, R., Jennings, J., & Lee, J. (2022). Disclosure sentiment: Machine learning vs. dictionary methods. *Management Science, 68*(7), 5514–5532.
Grandini, M., Bagli, E., 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., & 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., & 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., 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., & 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., & 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., & 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., & Zhang, Y. (2014). Tone management. *The Accounting Review, 89*(3), 1083–1113.
Inoue, A., & 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., & 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., & Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics, 132*(1), 126–149.
Labutov, I., & Lipson, H. (2013). Re-embedding words. *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 489–493).
Levy, O., & 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. (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., & Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies, 34*(7), 3265–3315.
Loughran, T., & 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., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *The Journal of Accounting Research, 54*(4), 1187–1230.
Lutz, B., Pröllochs, N., & 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., & 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., 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* (pp. 142–150).
Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology, 65*(4), 782–796.
Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
Mayew, W. J., Sethuraman, M., & 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., Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*arXiv:1301.3781,
Miranda-Belmonte, H. U., Muñiz-Sánchez, V., & 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., & 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., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance, 36*(4), 992–1011.
Rahimikia, E., Zohren, S., Poon, S.-H. (2021). Realised volatility forecasting: Machine learning via financial word embedding. *arXiv preprint*arXiv:2108.00480,
Rodriguez, P. L., & 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.
Sautner, Z., Van Lent, L., Vilkov, G., & Zhang, R. (2023). Firm-level climate change exposure. *The Journal of Finance, 78*(3), 1449–1498.
Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge University Press Cambridge.
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., & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics, 20*(2), 147–162.
Stock, J. H., & Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature, 41*(3), 788–829.
Stock, J. H., & 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., & Koyama, T. (2023). Hypothesis testing procedure for binary and multi-class F1-scores in the paired design. *Statistics in Medicine, 42*(23), 4177–4192.
Tang, D., Wei, F., Qin, B., Zhou, M., Liu, T. (2014). Building large-scale twitter-specific sentiment lexicon: A representation learning approach. *Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 172–182).
Tavcar, L. R. (1998). Make the MD &A more readable. *The CPA Journal, 68*(1), 10.
Welch, I., & 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., & McCracken, M. W. (1998). Regression-based tests of predictive ability. *International Economic Review, 39*(4), 817–840.

Publisher's Note