The Effects of Sentiment Evolution in Financial Texts: A Word Embedding Approach

Jiexin Zheng[a], Ka Chung Ng[b], Rong Zheng[a] and Kar Yan Tam[a]*

[a] *Department of Information Systems, Business Statistics and Operations Management, School of Business and Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong*

[b] *Department of Management and Marketing, Faculty of Business, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong*

Contact:

jzhengas@connect.ust.hk, kc-boris.ng@polyu.edu.hk, rzheng@ust.hk, kytam@ust.hk

* Corresponding author

Bios of Authors

Jiexin Zheng
Jiexin Zheng is a Ph.D. student in the Department of Information Systems, Business Statistics, and Operations Management at the Hong Kong University of Science and Technology. His research interests include financial text analysis and the economics of AI. His research has appeared at several international conferences or workshops, including *International Conference on Information Systems* and *Workshop on Information Technologies and Systems*.

Prof. Ka Chung Ng
Prof. Ka Chung Ng is an Assistant Professor and Presidential Young Scholar in the Department of Management and Marketing, Faculty of Business, Hong Kong Polytechnic University. He received his Ph.D. in Information Systems from the Hong Kong University of Science and Technology. His research interests lie in fake news, business analytics, and fintech. He has published in *Journal of Management Information Systems*, *Production and Operations Management*, *ACM Transactions on MIS*, and leading IS conference proceedings.

Prof. Rong Zheng
Prof. Rong Zheng is an Associate Professor of information systems at the HKUST Business School. He earned his doctoral degree in information systems from the Stern School of Business at New York University. His general research interest is about realizing business value with AI. More recently, his research examines how the use of AI methods in information processing can change the information environment of financial market. He is currently an associate editor at Business & Information Systems Engineering and was the associate editors for special issues at ISR and MISQ. His work has been published in such leading outlets as *Information Systems Research*, *Management Science*, *The Accounting Review*, and *Communications of the ACM*.

Prof. Kar Yan Tam
Prof. Kar Yan Tam is currently Dean of the Business School and Chair Professor of Information Systems, Business Statistics, and Operations Management at the Hong Kong University of Science and Technology (HKUST). He received his Ph.D. from Purdue University and is a founding member of the HKUST Business School. He has published in *Journal of Management Information Systems*, *MIS Quarterly*, *Information Systems Research*, *Management Science*, and other journals. His current research interests lie in fintech, business analytics, and sustainable and green finance. Prof. Tam is currently serving on the Board of AACSB and EFMD and the editorial boards of a number of IS journals.

The Effects of Sentiment Evolution in Financial Texts: A Word Embedding Approach

Abstract

We examine the evolutionary effects of sentiment words in financial text and their implications for various business outcomes. We propose an algorithm called **Wo**rd **L**ist **Ve**ctor for **S**entiment (WOLVES) that leverages both a human-defined sentiment word list and the word embedding approach to quantify text sentiment over time. We then apply WOLVES to investigate the evolutionary effects of the most popular financial word list, Loughran and McDonald (LM) dictionary, in annual reports, conference calls, and financial news. We find that LM negative words become less negative over time in annual reports compared to conference calls and financial news, while LM positive words remain qualitatively unchanged. This finding reconciles with existing evidence that negative words are more subject to managers' strategic communication. We also provide practical implications of WOLVES by correlating the sentiment evolution of LM negative words in annual reports with market reaction, earnings performance, and accounting fraud.

Keywords: word embedding, word list, sentiment evolution, textual analysis, strategic communication

Introduction

There are increasing IS applications that are built on natural language processing (NLP) techniques to collect intelligence from voluminous text. Many are financial applications (e.g., Refinitiv news analytics, Sentieo) developed to monitor market sentiments derived from unstructured news content [11,51,68]. Currently, these systems reveal sentiment signals using predefined dictionaries and corpora, which are assumed to be fixed and remain semantically constant over time. However, human language changes over time. For example, the vocabulary of Shakespearean times differs greatly from that of today. Even if the vocabulary remains the same, the meanings of words can change drastically as a result of being used to fit communication needs [2].

Although textual analysis has become increasingly important in business research, little is known about how language in financial text evolves and its impact on the effectiveness of content analysis in general. Studies show that financial text like annual reports, earnings conference calls, and financial news, contain information (e.g., sentiment words) that triggers market reactions [41,51,61]. Their findings provide evidence for the business value of financial text, especially words reflecting sentiment. However, being aware of these effects, communicators may choose to change their language to manage audiences' perceptions [5,12,17,41,46]. Therefore, an investigation of the evolutionary pattern of sentiment words and its effects on various business outcomes is of great interest.

We apply the word embedding approach [45] to a predefined sentiment word list to quantify word sentiment evolution patterns in financial text. Word embedding is a machine learning method that captures both the local and global semantic meanings of words with a large amount of training text [19,52,66]. Unlike the conventional NLP method, the word embedding method has the merit of using contextual (neighboring) words to represent the semantic meaning of a word using a high-dimensional vector representation ("word vector," "word embedding," or "embedding" in short). Accordingly, words with similar semantic meanings will have similar vector representations [45]. When the sentiment words in a predefined word list are represented in this way, their dynamic meanings, i.e., their strength in expressing sentiment, can be measured more accurately and are more adaptive to contextual change.

To illustrate the effectiveness of the word embedding method, we calculate the embedding vectors for the extensively used financial sentiment word lists proposed by Loughran and McDonald [41] ("the LM dictionary") and project the embedding vectors into a two-dimensional space using principal component analysis (PCA), as shown in **Figure 1**. The distinguishable positive (blue) and negative (pink) clusters indicate that these two opposing sentiments, represented by embedding vectors, indeed have different semantic meanings. However, it is interesting to note that some sentiment words are located far away from their own center, or even in the wrong clusters, indicating that semantically, they are not fully consistent with their predefined sentiment labels. **Figure 1** provides initial evidence that the word embedding approach can effectively quantify the sentiment using a high-dimensional representation of the LM dictionary and reveals the subtle differences between words in the dictionary in terms of their strength in expressing sentiment.

[Insert **Figure 1** here]

In this study, we propose an algorithm called **Wo**rd **L**ist **Ve**ctor for **S**entiment (WOLVES) that can capture the evolutionary effect of word meanings when conducting sentiment analysis. This algorithm is significantly different from the existing approaches to enhance the word list-based sentiment analysis, as WOLVES incorporates word embedding to dynamically update sentiment word lists (e.g., the LM dictionary in financial context) with their changing semantic meaning and sentiment strength. Toward this goal, WOLVES generates a sentiment intensity score for each word in a word list by using the geometrical distance between sentiment word vectors in the high-dimensional embedding space to quantify its sentiment strength. Given that the embedding is updated with the changing context of a particular sentiment word, this algorithm can capture the dynamic sentiment strength of that word. In this regard, we no longer assume that sentiment words always carry the same level of sentiment regardless of their context.

The sentiment intensity score generated by WOLVES makes it possible to quantitatively trace the level of sentiment delivered in words over time.

The evolutionary sentiment intensity might potentially explain the inconsistent results in the literature about the role of sentiment words in financial text using panel data [17,25]. For example, Loughran and McDonald [35] consider annual reports filed between 1994 and 2008 and indicate that disclosure sentiment can be quantified by applying predefined word lists. However, Frankel et al. [17] find that the dictionary method cannot capture sentiment in 10-K filings when the sample covers a much longer period (between 1996 and 2019), likely because of changes in disclosure language. In this study, we aim to reconcile these inconsistent findings using a refined sentiment measure that considers temporal changes and a large sample of financial text data.

We apply WOLVES to three important types of financial text (annual reports, conference calls, and financial news) to investigate the evolutionary effect of the sentiment words defined in the LM dictionary. We find that after accounting for factors that are not the primary focus of our study or have been documented with effects on word evolution, the average sentiment intensity of both LM positive and negative words is relatively stable over time in all three types of financial texts, except for annual reports, where the intensity of the negative words decreases. This diminishing effect of LM negative words in annual reports is consistent with the view that managers choose their words carefully when preparing annual reports so that they can manage the sentiments communicated [5,17]. However, it is much harder for managers to do the same in extemporaneous settings such as the Q&A sections of conference calls [17,37]. This finding also indicates that news editors may be less likely to practice strategic language manipulation because they do not have the same motivations as managers.

To further validate the results, we consider the implementation of the Sarbanes-Oxley Act (SOX) Section 404 (internal controls)[1] as an exogenous shock, as this requires firms to state any potential risks in their annual reports. We find that the sentiment intensity of LM negative words significantly changes after the SOX is implemented. We also test the effect of the release of the LM dictionary in 2011[2] and find that LM negative words changed significantly after this event. Taken together, this additional evidence is consistent with the view that managers' strategic communication might be a plausible explanation for word evolutional patterns [5,17].

Next, we use WOLVES to construct new sentiment measures to examine the predictive value of sentiments in the MD&A sections of annual reports in terms of *market reaction*, *earnings performance*, and *accounting fraud*. We find that only negative words with high sentiment intensity scores are statistically correlated with these business outcomes. In terms of economic significance, a one-standard-deviation increase in the ratio of high-intensity negative words yields about a 0.160% decrease in excess returns, a 5.27% decrease in future earnings of its standard deviation, and a 20.3% increase in the odds of fraud. Our findings suggest that WOLVES effectively accounts for the effect of word evolution that can significantly improve the power of dictionary-based sentiment measures to predict important business outcomes.

This paper makes three main contributions to literature. First, we contribute a new method for extracting sentiment from financial documents. Our proposed WOLVES combines the human domain knowledge in creating a domain-specific sentiment word list and the computational power implemented by the AI-based approach to capture the evolutionary effect of the sentiment words. Human knowledge, as represented by the LM dictionary, can suffer from limitations such as low generalizability, bias in word selection, and the inability to capture

[1] https://www.sarbanes-oxley-101.com/SOX-404.htm

[2] The sentiment word list became publicly available after Loughran and McDonald published their influential paper in 2011.

dynamic language changes [15,59], while machine learning alone is not always preferable because of its low level of interpretability, which is important in managerial decision-making [54]. Many machine learning methods also require the manual labeling of text data, which can introduce human bias [28]. By leveraging the benefits of both human domain knowledge and the computational power of machine learning methods, our method can not only retain the simplicity and explainability power of the financial lexicon but also capture the evolutionary pattern in language to construct more effective sentiment measures.

Second, our study reconciles to some extent the inconsistent findings in the literature about the information content of financial text. Research applying the dictionary method to capture disclosure sentiment reports inconsistent findings [17,25,41]. Compared with the literature, our proposed sentiment intensity score generated from WOLVES can improve the sentiment measures, as it considers contextual information, the semantic relationships between words, and the temporal changes in word meaning (see **Appendix A**). As a result, the sentiment in the annual report extracted by our method can consistently predict market reaction, future corporate earnings, and accounting fraud.

Finally, our study contributes to the emerging literature concerning how managers behave strategically in their corporate disclosures. Recent research has suggested that managers may attempt to mitigate negative investor perceptions by strategically adjusting their language in business communication or company disclosure [5,12,33]. Thus, analyzing managers' language with a dynamic view is valuable when assessing the information in company disclosures and investors' responses. Our novel findings of a significant evolutionary pattern for LM negative words may reflect the adjustments that managers make to the language of corporate filings in response to the public availability of the LM dictionary.

Related Work

Computational Models of Word Evolution

Research into the semantic changes of words has recently increased because of source data availability and developments in computational linguistics. Earlier studies generally examine the frequency of word usage to assess word evolution [38]. Turney et al. [65] extend this stream of literature by proposing a distributional word representation that captures the temporal semantic meanings of words. Pioneered by Mikolov et al. [45], word embedding models based on deep learning have gradually become popular and effective approaches to NLP problems. For example, Kim et al. [29] examine the semantic changes of words through word embeddings using the Google Books Ngram corpus, which enables them to identify a list of words (e.g., *cell* and *gay*) that evolved significantly between 1900 and 2009. Hamilton et al. [24] empirically compare word evolution patterns identified through embedding methods with the ground truth of known word semantic changes. They confirm that the word embedding method is effective in this NLP task. Schlechtweg et al. [57] conduct a systematic comparison of various computational linguistic models and confirm that the word embedding approach can effectively detect semantic changes automatically. The evolution of words identified through embedding methods has recently been considered when examining various social phenomena, such as gender and ethnic stereotypes [19], social class evolution [32], drug overdose [67], and gender inequality [34]. We extend this stream of literature by quantifying word sentiment evolution in financial text and assessing its implications for various business outcomes. The focus of our method on the temporal changes at the word level is driven by the fact that word list-based sentiment analysis heavily relies on each word's sentiment intensity, which changes over time. Our method contributes to the sentiment analysis literature by capturing such changes.

Word Embedding with the Dictionary-Based Method

The dictionary-based method is commonly applied to the sentiment analyses of financial text [18,41,42,61,62]. This method relies on a predefined set of words to measure the levels of positive or negative sentiment in text [41,61]. Dictionary-based sentiment measures have been applied to predict various business outcomes. For example, García [18] and Tetlock [60] find that the proportion of negative sentiment words in financial articles can predict stock returns. Frankel et al. [17] and Loughran and McDonald [41] report that LM negative words are associated with corporate earnings. Loughran and McDonald [41] find that LM negative words are linked to 10b-5 fraud lawsuits and self-reported material weakness in terms of internal controls. Larcker and Zakolyukina [33] show that deceptive narratives communicated by executives are associated with emotional words.

The dictionary-based approach is regarded as an appropriate tool for analyzing financial text because of its simplicity, ease of implementation, and strong explainability power. However, its efficacy has been questioned because a predefined dictionary is likely to be subjective, domain-specific, and inflexible, and thus unable to adapt to new contexts over time [17]. The deep learning approach can address the limitations of these predefined word lists, and it improves the effectiveness of sentiment extraction from text. For example, Tsai and Wang [64] apply the continuous bag-of-words model [45] to identify additional sentiment words that are semantically similar to those in predefined dictionaries. Theil et al. [63] use a word embedding model to expand an uncertainty word lexicon to include industry-specific terms. Yang et al. [69] illustrate that combining word embeddings and domain-specific dictionaries (i.e., the LM dictionary) significantly improves predictions of stock return volatility. Li et al. [37] demonstrate that the word embedding approach can be used to automatically construct a dictionary containing informative words and phrases related to corporate culture.

Although these methods improve the adaptability of the dictionary-based method, few studies consider temporal changes in language and accordingly adjust the weights of words in the dictionary to better measure sentiment [17]. We extend the prior work in this stream to incorporate temporal language changes, as otherwise, the constructs developed from this method will gradually lose their validity, thus affecting any conclusions and implications drawn. To address this issue, this paper applies the word embedding approach to construct a sentiment intensity score that can be leveraged to improve sentiment construct validity.

Informativeness of Financial Text

The recent financial text analysis literature suggests that company disclosures, such as annual reports and earnings conference calls, are informative to the market. Li [35] shows that firms with lower earnings are associated with poor annual report readability. Larcker and Zakolyukina [33] report that deceptive CEO narratives contain more extremely positive words and fewer anxiety words, while deceptive CFO narratives use more negation and extremely negative words. Huang et al. [27] document that an abnormal positive tone in earnings press release is correlated with poor subsequent earnings and cash flows for up to three years after the initial release. Allee and Deangelis [1] find that firms manage tone in earnings conference calls to temper extreme performance expectations. Dzieliiski et al. [12] find that investors react less to earnings news when managers use more uncertain words in their presentations.

However, as important channels of company disclosure, annual reports and earnings conference calls differ fundamentally [17]. First, earnings conference calls allow participants outside of the company, such as analysts, investors, and media, to interact with managers and discuss firms' financial results during Q&A sessions. Thus, the language is more spontaneous because of the real-time and conversational nature of conference calls. Second, firms typically host conference calls immediately after the press release of earnings announcements, which gives managers limited time to prepare. Unlike these calls, firms have no time pressure when preparing annual reports. Thus, they can craft the language in the reports to achieve their intended goals, particularly in the management's discussion and analysis (MD&A) sections, which receive much attention from investors [37]. In their recent paper, Cao et al. [5] show that managers may avoid LM negative words because they are aware that the disclosure sentiment, measured by the LM dictionary, can be used by investors as the basis of firm valuation or investing decisions. If this task is automated and streamlined using computer programs, it can trigger a stronger effect on the market. Taken together, we are particularly interested in the word sentiment evolution patterns in annual reports using conference calls as a benchmark.

Financial news is another important information source for investors in the financial market, even though it is not company disclosure. Many studies document the informativeness of media content in financial news and its association with investors' reactions [18,61,62]. Unlike annual reports and earnings conference calls, in which managers discuss firms' financial performance and operations, financial news articles are often written by professional reporters from an outsider's perspective [70]. Thus, we expect the word evolution patterns in financial disclosures (i.e., annual reports and earnings conference calls) to differ from those of financial news [4,5,35].

An Embedding-Based Algorithm for Analyzing Word Sentiment Evolution

We propose an embedding-based sentiment extraction algorithm called WOLVES to capture the dynamic sentiment patterns. **Figure 2** outlines the algorithmic procedure of WOLVES. The inputs of the algorithm include a textual corpus split into years and predefined positive and negative sentiment word lists. The output is a vector of the yearly updated sentiment intensity scores for each word in the word list, reflecting the evolutionary sentiment patterns. The following sections explain the details of each step of the WOLVES algorithm.

[Insert **Figure 2** here]

Training Dynamic Word2vec

Word embedding is a numeric vector representation of words to capture their semantic meaning. These vector representations are learned by statistical or machine learning methods (also called language models in general) from a corpus of documents. The learning objective is that words with similar semantic meanings will have similar vector values. Hence, one can measure the semantic similarity between words using vector distance, such as Cosine similarity.[3] In this study, we apply the word2vec model [45] to generate embedding vectors of words for its superior performance in learning the optimal word embedding. **Appendix B** provides a detailed description of our word embedding model.

To capture the word evolution patterns, we separately train word2vec models using corpus from different time periods. We can thus capture a word's temporal geometric properties of its word embedding vectors. In this way, word embedding vectors will be updated to reflect more accurate semantic meanings over time.

Preprocessing and Implementation of Word2vec

As an advantage, the word embedding model requires minimal text preprocessing effort. We simply remove punctuation and convert all characters to lowercase. Traditional preprocessing methods, such as word lemmatization, word stemming, and name entity recognition, are

[3] Cosine similarity between vectors is defined as: $cos(w_1, w_2) = \frac{(w_1 \cdot w_2)}{\|w_1\| \; \|w_2\|} = \frac{\sum_{i=1}^{n} w_{1_i} w_{2_i}}{\sqrt{\sum_{i=1}^{n} {w_{1_i}}^2} \sqrt{\sum_{i=1}^{n} {w_{2_i}}^2}}$, where w_{1_i}, w_{2_i} are components of vector w_1, w_2 respectively.

unnecessary for training word embedding. The minimal preprocessing effort makes our approach easier to reproduce and scale up for other applications. We apply Python's *Gensim* library to train the word2vec models. As suggested by the literature [37,45], we set the vector dimensionality to 300 and use 5 neighboring words as the context. We also omit words that appear fewer than five times in the corpus.

Generating Sentiment Intensity Score

We consider a word as a good sentiment marker if it is close to the center of its own sentiment group and far from the center of the opposite sentiment group. Formally, we define a word's sentiment intensity as the difference between this word's within-group (to its own centroid) and across-group (to the centroid of the opposite group) distance. The two centroids are updated periodically to reflect the temporal sentiment changes. Also, at each update, we exclude the ambiguous sentiment words (i.e., LM words with negative intensity scores) in order to generate more accurate centroids. By doing this, the ground truth of the sentiment is defined with a subset of predefined sentiment word list, rather than a fixed one. Compared to existing embedding-based sentiment analysis methods, such as the one proposed by Garg et al. [19], our method is more flexible in capturing the sentiment strength with a predefined word list.

We outline the steps for calculating sentiment intensity scores as follows. For each year t, we first update the positive and negative word vector lists to remove word vectors with the ambiguous sentiment as

$$V_t' = \left[v_{i,t} \text{ for } v_{i,t} \text{ in } V_t \text{ if} \cos\left(v_{i,t}, average\left(V_{same,t}\right)\right) - \cos\left(v_{i,t}, average\left(V_{opposite,t}\right)\right) > \eta\right],$$

where $cos(\cdot\,,\ \cdot)$ is the cosine similarity function, $average(\cdot)$ is the component-wise mean function, $v_{i,t}$ is the word embedding vector of focal word i in year t, V_t is the list of word embedding vectors, $V_{same,t}$ ($V_{opposite,t}$) is the word embedding vectors of the same (opposite) sentiment as the focal word i, and the embedding vectors are from the word2vec model in year t. The hyperparameter η, ranging between -2 and 2, defines the proportion of the original sentiment word list to be used when constructing the new centroids of the two sentiment groups. We set it to 0 throughout the paper.[4] With the updated sentiment centroids ($\bar{V}_t'$), we then measure the sentiment intensity score of each LM word using the cosine similarity between its word vector to its own centroid, minus the cosine similarity to the centroid of the opposite sentiment. Formally, the sentiment intensity score for a given LM word i in year t is calculated as

$$Intensity_{i,t} = cos\left(v_{i,t}, \bar{V}_{same,t}'\right) - cos\left(v_{i,t}, \bar{V}_{opposite,t}'\right),$$

where $\bar{V}_{same,t}'$ ($\bar{V}_{opposite,t}'$) is the centroid of the same (opposite) sentiment as the focal word i, and the embedding vectors are from the word2vec model in year t.

We then follow the literature in aggregating the sentiment to the document level, which is the ratio of the sentiment words [17,41]. Effectively, WOLVES can improve the overall sentiment measures by assigning higher weights to words with high sentiment intensity or even removing words with low sentiment intensity from the predefined word list in a dynamic way.

Validating Sentiment Intensity Score

To demonstrate the effectiveness and generalizability of WOLVES in a general context, we conduct a validation test using the Amazon Book Review dataset [47]. One merit of Amazon's review data is that the textual review is associated with a quantitative rating, which is a good summative measure of the review's sentiment. In other words, the ground truth of the sentiment in the review text is known to some extent. Therefore, it is an ideal dataset to validate our proposed sentiment measure. The validation procedure is described in **Appendix C**. In brief, our analysis based on this validation dataset shows that our proposed WOLVES algorithm is

[4] Our results are qualitatively consistent for a reasonable range of η.

effective in differentiating sentiment words based on their sentiment intensities and can better explain the summative ratings with a selected subset of words.

Analyses of Word Sentiment Evolution

Data

We apply WOLVES to three types of financial text: annual reports, conference calls, and financial news. We split the three datasets (annual reports, conference calls, and financial news) into multiple folders by year. We thus generate 22 folders for annual reports (1997-2018), 16 for conference calls (2003-2018), and 22 for financial news (1997-2018). Based on these folders, we got 22, 16, and 22 word2vec models in WOLVES for annual reports, conference calls, and financial news, respectively. The details of data collection are presented below.

Annual Report

We download all text from annual report filings between 1997 and 2018 from the EDGAR database.[5] Specifically, we use the dataset prepared and cleaned by the Notre Dame Software Repository for Accounting and Finance (SRAF).[6] In this dataset, each file contains the body text of an annual report and its metadata, including the SEC's Central Index Key (CIK) number, filing date, form type, and file size. We collect 199,176 Form 10-Ks from 34,966 firms identified by their CIK number. As many sections of 10-Ks consist of tables with quantitative data and a boilerplate, we only focus on sections with descriptive text about the firms' business and financial conditions, including Item 1 (Business), Item 1A (Risk Factors), Item 7 (MD&A), and Item 7A (Quantitative and Qualitative Disclosures about Market Risks). We apply regular expression matching operations in Python to extract sections from these 10-K items, resulting in

[5] We exclude 1994-1996 because there are too few 10-Ks in this period.
[6] https://sraf.nd.edu/. The last update was 01/01/2019 when we retrieve the data.

188,244, 95,188, 186,198, and 147,537 per item, respectively. Sections with fewer than 100 words are removed. Then, we use the cleaned text to train the word embedding model.

Conference Calls

We obtain the initial dataset of the earnings conference call transcripts from Thomson Reuters' StreetEvents (SE) database from 2003 to 2018.[7] In the dataset, each file contains the text transcript of the participants' communication during the call and metadata that help us identify the event time, each speaker, and their role (managers, analysts, or investors). We apply regular expression matching operations in Python to match every dialogue with its speaker from the transcript file. We successfully extract 194,928 Q&A sessions with matched speakers and their roles from 265,622 transcript files.[8] We filter out text associated with analysts or investors, as they are not the focus of our study. Our final dataset contains 8,801,959 manager conservations in 194,928 Q&A sessions.

Financial News

We collect financial news articles from 1997 to 2018 via the Factiva Streams API.[9] In the dataset, each news article is associated with a Dow Jones Intelligent Identifier[10] that enables us to extract only relevant financial news. We first filter out irrelevant news types, such as blogs, images, or web pages. We then use the top-level subject codes to extract news articles from the relevant categories, such as commodity/financial market news, corporate/industrial news, and economic

[7] The initial dataset includes transcripts from 2001 to 2018. We exclude 2001-2002 because there are too few transcripts in this period.

[8] We drop transcript files (1) without information about the occupations of the speakers or (2) without dialogue in Q&A sections or (3) if we fail to detect Q&A sections in the full transcript.

[9] https://developer.dowjones.com/site/global/home/index.gsp

[10] A complete list of Factiva fields used for financial news collection is provided here: https://developer.dowjones.com/site/docs/factiva_apis/factiva_analytics_apis/factiva_snapshots_api/index.gsp#product-overview-353.

news.[11] Next, we apply the language and region filters to restrict our data to the U.S. region. Finally, we remove news articles without body or company identifiers. In total, we obtain 6,372,781 news articles for model training and subsequent analyses.

Preliminary Analysis

We apply WOLVES to examine the word sentiment evolution phenomenon within three types of financial text. We compare the yearly change of sentiment intensity for LM words in annual reports, a non-extemporaneous setting in which managers can actively modify the language, with the Q&A section of conference calls and financial news, which are less formal and prepared.

After removing infrequent words from the original LM sentiment dictionary,[12] we identify 308 positive and 1,526 negative words for annual reports, 264 positive and 750 negative words for conference calls, and 278 positive and 1,348 negative words for financial news.

The variable definitions are summarized in **Table 1,** and their descriptive statistics are summarized in **Appendix D**. The mean intensity scores of both LM positive words ($Intensity^{POS}$) and negative words ($Intensity^{NEG}$) are positive for all three types of texts, suggesting that LM words are generally effective in conveying sentiment. The average term frequency of positive words ($TermFreq^{POS}$) is much higher than that of negative words ($TermFreq^{NEG}$) in each dataset. The term frequency of sentiment words is also much higher in annual reports compared to the other two text types, as annual reports are, on average, longer. The geometrical distance between the centers of positive and negative words ($GroupDist$), which is defined as $1 - cos(\bar{V}_{positive}, \bar{V}_{negative})$,[13] is above 0.5 on average in all three types of text, suggesting that positive and negative words are generally semantically different. Finally,

[11] A complete list of top-level subject codes is provided here: http://www.factiva.com/en/cp/content/indexing/DJID_FeaturesBenefits_2017.pdf

[12] A word is removed if its embedding vector appears fewer than five times in a data folder.

[13] $\bar{V}_{positive}$ ($\bar{V}_{negative}$) is the representative group vector of positive (negative) words, calculated as the average of the word embedding vectors of all positive (negative) words defined by the LM dictionary.

$Harvard$ is a dummy variable that indicates whether an LM word is also included in the H4 dictionary. On average, about 30% of sentiment words in the three text types are defined in both LM and H4 dictionaries.

[Insert **Table 1** here]

Figure 3 illustrates the trends of the average sentiment intensity scores of LM words across years and the three text types. Given that we are interested in the individual word's sentiment evolution, we control for the overall changes at the aggregated level, i.e., the similarity between the positive and negative embedding centroids. Also, we control other factors documented in the literature that can influence the calculation of word embedding, including domain-specific information [23] and the frequency of word usage [48]. The scores for the positive and negative words in annual reports clearly demonstrate a different trend, while the polarities in financial news and conference call are relatively stable. Thus, we obtain initial evidence that negative sentiment words in annual reports evolve differently from those in conference calls and financial news, likely because of managers' strategic reporting behavior in such an important disclosure channel.

[Insert **Figure 3** here]

Note that our approach does not measure sentiment by counting the usage of sentiment words. Rather, the word embedding measures reflect the semantic context. Hence, the decreasing intensity scores for negative words in annual reports indicate that the context of using negative words becomes less negative in general. There are two possibilities for this to happen. One is the context for using sentiment words does not really convey negative information. The second is the obfuscation of information by using both positive and negative words in the vicinity of text. To investigate these potential explanations, we report the ratio of LM negative words (Negative word ratio) and the ratio of the sentences containing both LM positive and negative words (Pos-Neg word mixture ratio) in **Figure 4**. Interestingly, we find that managers used more negative words in annual reports over time, resulting in an increase of 85% in the ratio of negative words in 2018 compared to that in 1997, that is, with a 2.97% yearly increase rate. This result is different from a related study by Cao et al. [5] where the authors find that firms avoid negative words in disclosure when facing high machine downloads. Given that their finding is based on cross-sectional evidence, ours is a temporal pattern revealing an increasing trend. This can be explained by firms' increasing litigation concerns [3,43,56]. Nonetheless, we find that negative and positive words are used together more frequently, as the mixed ratio increased by 103% in 2018 compared to that in 1997, that is, with a 3.43% yearly increase rate. This finding suggests that managers may use positive words to manage investors' perceptions when reporting negative news to obfuscate disclosure information [35,39,60]. This language pattern has not been documented by the existing literature, and therefore our study contributes to explaining why negative words lose their expected sentiment intensity over time.

[Insert **Figure 4** here]

Regression Analysis

Next, we apply an econometric analysis to formally investigate the sentiment evolution patterns across financial texts. We focus on the sentiment evolution of LM words in annual reports, and our benchmark consists of the two other types of financial text.[14] We consider the following regression model:

$$\begin{aligned} Intensity_{i,t,d} = {} & \beta_0 + \beta_1\, YearNum + \beta_2\, YearNum \times Report + \beta_3\, TermFreq_{i,t,d} \\ & + \beta_4\, GroupDist_{t,d} + \delta_{i,d} + \varepsilon_{i,t}, \end{aligned} \tag{1}$$

[14] We only include words with valid word embedding vectors across two domains each year in each sample set to ensure a fair comparison. We thus end up with 260 positive words and 740 negative words in the *annual report – conference call* sample set and 269 positive words and 1,163 negative words in the *annual report – financial news* sample set.

where $Intensity_{i,y,d}$ is the sentiment intensity score of an LM word i in a specific year y from a domain d from either the annual reports or the benchmark text types, $YearNum$ is a consecutive year sequence capturing the year trend, and $Report$ is an indicator variable taking a value of 1 if an LM word is from an annual report, and 0 otherwise. We control for the average term frequency of LM words in a specific year and text type ($TermFreq_{i,t,d}$). Note that the sentiment intensity score can be affected by the overall similarity between positive and negative words, so we control for the cosine distance between the positive and negative centroids in annual reports and benchmarks ($GroupDist_{t,d}$). We also control for the heterogeneous effects of words across text types by including the word-domain pair fixed effect $\delta_{i,d}$. We run the regression model separately for positive and negative words on the sample sets of *annual report – conference call* and *annual report – financial news,* respectively, with robust standard errors clustered at the word-domain and year level. We report the results in **Table 2**.

[Insert **Table 2** here]

Columns (1) and (3) of **Table 2** show that the coefficients on $YearNum \times Report$ are not significant for positive words in the LM dictionary. This indicates no significant difference between the sentiment intensity trend of positive words in the annual reports and in the benchmark texts (conference calls or financial news). However, the data shown in Columns (2) and (4) in **Table 2** indicate that the coefficients of $YearNum \times Report$ are negative and significant at the 1% level for LM negative words, suggesting that the negative words in annual reports gradually become relatively less negative over time compared to those in conference calls or financial news.

Additional Analyses

We next examine how the implementation of the SOX 404 controls influences the sentiment intensity of LM dictionary words. The SOX 404 controls implemented in 2002 (see **Figure 3**) required firms to disclose any potential risks and internal vulnerabilities in their annual reports. Thus, managers may be more cautious about using negative words when reporting bad news. We replace the year trend variable $YearNum$ in Equation (1) with $AfterSOX$, an indicator that takes a value of 1 for years after 2002 and 0 otherwise. Since we do not have data for conference calls before 2003, we only analyze the *annual report – financial news* sample set. The results are reported in **Table 3**. As expected, the coefficient on $AfterSOX \times Report$ for negative words is negative and significant at the 1% level, meaning that LM negative words evolve significantly after the SOX 404 controls are implemented.

[Insert **Table 3** here]

We further validate our results by examining the awareness effect of business-context lexicons. We assess the sentiment intensity before and after the publication year of the LM dictionary, which enables us to identify whether the word sentiment evolution in annual reports is driven by managers strategically adjusting their language because of the availability of the LM dictionary. The rationale behind this is that managers may believe that many investors or even automated investing programs will use the LM dictionary to quantify the sentiment level, which will help form their investing decisions. Thus, managers are incentivized to avoid using LM negative words in their reports. We replace the year trend variable $YearNum$ in Equation (1) with the indicator $AfterLM,$ which takes a value of 1 for years after 2011, the official publication year of the LM dictionary, and 0 otherwise. The results are reported in **Table 4**. Consistent with our expectations, we find a negative and significant coefficient for the interaction term $AfterLM \times Report$ for negative words in both sample sets, suggesting that the publication of the LM dictionary influences the evolutionary pattern of sentiment words in annual reports.

[Insert **Table 4** here]

An alternative approach to verify whether awareness of a sentiment dictionary plays a role in the identified word evolution is to apply the H4 dictionary (introduced in 1996), another popular dictionary researchers used to quantify the sentiment of documents [62] before the publication of the LM dictionary. We expect that sentiment words defined by both the LM and H4 dictionaries will be of relatively greater intensity, and thus managers are likely to be more cautious when using these words to communicate with investors. Therefore, we extend the previous model specification in Equation (2):

$$\begin{aligned} Intensity_{i,t,d} &= \beta_0 + \beta_1\, YearNum + \beta_2\, YearNum \times Report \\ &+ \beta_3\, YearNum \times Harvard + \beta_4\, YearNum \times Report \times Harvard \\ &+ \beta_5\, TermFreq_{i,t,d} + \beta_6\, GroupDist_{t,d} + \delta_{i,d} + \varepsilon_{i,t}, \end{aligned} \tag{2}$$

where $Harvard$ is an indicator that takes a value of 1 if an LM word is also defined by the H4 dictionary, and 0 otherwise. We expect that the lexicon awareness effect, captured by the three-way interaction term $YearNum \times Report \times Harvard$, will be negative and significant. Columns (2) and (4) in **Table 5** show that the coefficients associated with $YearNum \times Report \times Harvard$ are negative and significant, indicating that the decrease in the sentiment intensity of LM negative words in annual reports is relatively greater if the words are also included in the H4 dictionary. Together with the previous analysis, we conclude that LM negative words in annual reports change differently from those in conference calls and financial news, which can be explained by managers' strategic reporting behavior.

[Insert **Table 5** here]

Business Impact of Word Sentiment Evolution

In the previous section, we provide empirical evidence that the sentiment intensity of LM negative words in annual reports weakens over time. In this section, we consider the business implications of word sentiment evolution and assess whether our refined sentiment measures generated from WOLVES can provide insights into market reactions, firms' fundamental performance, and accounting fraud.

Sample and Variables Construction

The previous section indicates that word sentiment evolution in annual reports follows a unique pattern. We thus focus on the MD&A section, one of the most important sections of annual reports where firms summarize their current financial performance and provide insights into their prospects [7]. The literature has heavily relied on MD&A to establish the association between disclosure tone and financial indicators like market reaction, trading volume, and return volatility [8,9,16,58]. Besides, we focus our analysis on negative words because of the pattern observed in our earlier analysis and because the previous study suggests that, in terms of annual reports, only negative words are informative [41].

Our initial sample contains 192,916 text filings for 10-Ks and 10-K405s from EDGAR during 1997-2018. We follow Loughran and McDonald [41] and apply several filtering rules to construct our final sample, which contains 63,508 observations from 9,404 firms. The data generation process, the summary statistics of variables, and their definitions are summarized in **Appendix E**. The event period excess return relative to the 10-K filing date is represented by $CAR_{[0,3]}$, measured as the firm's buy-and-hold stock return minus the CRSP value-weighted buy-

and-hold market index return over the 4-day event window (days 0-3) [41].[15] $Earnings_{t+1}$ is the earnings of the next fiscal year, normalized by total assets. Finally, $Fraud_{t+1}$ is an indicator taking the value of 1 if any quarterly financial statement in the next fiscal year is deemed fraudulent and 0 otherwise. We extract firms' fraudulent earnings statements from the Audit Analytics restatements database via the Wharton Research Data Services (WRDS).[16] We identify 188 firm-year observations related to fraudulent earnings statements issued by 119 unique firms in the next fiscal year.

We include a set of control variables that are commonly used in previous studies [17,41]: (1) $PreFFAlpha$, calculated as the intercept from a Fama-French three-factor model estimated prior to the filing date; (2) $InstOwnership$, calculated as the percent of institutional ownership reported in the CDA/Spectrum database for the most recent quarter before the filing date. (3) $Log(Size)$, calculated as the natural logarithm of market capitalization, based on data from COMPUSTAT at the end of the fiscal year; (4) $Log(Book-to-Market)$, calculated as the natural logarithm of the book-to-market ratio, based on data from COMPUSTAT at the end of the fiscal year; (5) $NASDAQ$, a dummy variable with value 1 if the firm traded on NASDAQ and 0 otherwise; (6) $Log(ShareTurnover)$, calculated as the natural logarithm of share turnover. When regressing for future earnings performance ($Earnings_{t+1}$), we follow Li [35] to additionally control current earnings performance ($Earnings_t$), the standard deviation of earnings in the previous five years ($EarningsVol$), and the pre-filing return's root-mean-squared error ($PreFFRMSE$). For the accounting fraud regression, following Richardson et al. [55], we further extend the control variables in the future earnings performance regression to include total

[15] Loughran and McDonald [41] leverage the abnormal stock return during the trading window [0, +3] to validate the informativeness of their proposed LM dictionary. We thus follow their setting to use the trading window [0, +3]. Our results are qualitatively consistent if using another commonly used trading window [0, +1] [17].

[16] We accessed the database in October 2021.

accruals ($Accruals$), measured as the net operating profit minus the cash flow from operations, scaled by total assets.

We assign LM negative words into three evenly distributed groups based on their rankings in terms of sentiment intensity score for each year. We measure the ratio of words with sentiment intensity scores in the top third to the total number of words appearing in MD&A as $NegTone^{TOP}$ for each firm-year pair. Similarly, sentiment variables with words in the middle and bottom thirds are labeled $NegTone^{MID}$ and $NegTone^{BOT}$, respectively. Doing so can directly validate what kind of negative words are more informative. We expect that words with higher sentiment intensity scores are more informative. We also follow the literature [41] to construct a common variable $NegTone^{LM}$ using all LM negative words as the benchmark. This comparison allows us to see whether the proposed sentiment variable performs better than the existing one used by the literature. We only report the main findings here and provide the full regression estimation results in **Appendix E**.

Market Reaction

We first investigate how investors react to negative words with differing levels of sentiment intensity. We expect high-intensity negative words to trigger stronger market reactions. We thus regress $CAR_{[0,3]}$ on the four negative sentiment variables ($NegTone^{TOP}$, $NegTone^{MID}$, $NegTone^{BOT}$, and $NegTone^{LM}$) and all of the relevant control variables used by Loughran and McDonald [41].

Table 6 reports the results of our regression estimation with firm and year fixed effects.[17] First, we find that the relationship between $NegTone^{LM}$ and $CAR_{[0,3]}$ is not significant. Columns (2)-(4) of **Table 6** indicate that only the estimate of $NegTone^{TOP}$ is negative and significant at

[17] The results remain qualitatively consistent when we follow the regression model of Fama and MacBeth [14], in which a standard error with one lag is applied.

the 5% level. Economically, a one-standard-deviation increase in the ratio of high-intensity negative words is associated with a 0.160% decrease in excess returns. This negative and significant result still holds when we include all $NegTone^{TOP}$, $NegTone^{MID}$, and $NegTone^{BOT}$ in the same regression (Column (5)). These results confirm that negative words with higher sentiment intensity are more informative, as they are associated with stronger market reactions.

Existing findings on the efficacy of the dictionary-based approach in explaining market reactions are mixed. For example, Loughran and McDonald [41] find that dictionary-based sentiment measures can explain market reaction, but Frankel et al. [17] find that when they extend their sample period from 1996-2008 to 1996-2019, the measures lose their explanatory power, suggesting that the informativeness of sentiment words changes over time. Our results support this suggestion as the sentiment measure from Loughran and McDonald [41] ($NegTone^{LM}$) has no impact on market reaction in our analysis. By incorporating the effect of word sentiment evolution, we may contribute to the literature by reconciling the inconsistent findings when using the dictionary-based approach to explain market reaction [17,25,41].

[insert **Table 6** here]

Earnings Performance

We next examine whether LM negative words in the MD&A sections of annual reports can help predict the fundamental performance of firms. As indicated in the previous analysis, negative words with higher sentiment intensity have a stronger effect, so we expect them to be better able to predict future performance. To test this hypothesis, we regress future earnings on the three negative sentiment level variables. We include relevant control variables, following Henry and Leone [25].

We use firm and year fixed effects in our regression estimation to strengthen our analysis. The results in **Table 7** show that $NegTone^{LM}$ is negatively and significantly associated with $Earnings_{t+1}$, which is consistent with the literature [25].[18] By comparing the three variables, we find that only negative words with higher sentiment intensity exhibit a strong association with poor future performance. $NegTone^{TOP}$ is significantly associated with $Earnings_{t+1}$ at the 1% level, with a one-standard-deviation increase in the ratio of high-intensity negative words leads to a decrease in $Earnings_{t+1}$ that is 5.27% (0.00538/0.102) of its standard deviation. Column (5) in **Table 7** shows that only $NegTone^{TOP}$ is negative and significant when all three negative sentiment variables are included in the regression model. In summary, we find that negative words with higher sentiment intensity are able to predict future corporate performance to a greater degree than those with lower sentiment intensity, indicating that sentiment words can change over time and that stronger sentiment words have a more consistent effect.

[insert **Table 7** here]

Accounting Fraud

We next investigate whether negative words with higher sentiment intensity are more informative in predicting future accounting fraud, in addition to market reaction and fundamental performance. We define a firm is with future accounting fraud if any of its quarterly financial statements in the next fiscal year is deemed fraudulent afterward. We apply a logistic regression model to regress $Fraud_{t+1}$ on the three proposed negative sentiment variables ($NegTone^{TOP}$, $NegTone^{MID}$, $NegTone^{BOT}$) and the benchmark ($NegTone^{LM}$), together with the same set of control variables from **Table 7**. We also include an industry dummy, following Fama and French [13] and a year dummy in each regression.

[18] Li [35] and Henry and Leone [24] only focus on forward-looking statements in MD&A. In contrast, we analyze the whole MD&A section.

Table 8 reports the regression results. Consistent with the finding of Loughran and McDonald [41], $NegTone^{LM}$ is not significantly associated with fraudulent earnings statements. Interestingly, columns (2)-(4) of **Table 8** show that $Negative^{TOP}$ is positive and significant at the 5% level, while $NegTone^{MID}$ and $NegTone^{BOT}$ are not significant in predicting $Fraud_{t+1}$. Specifically, the coefficient of $NegTone^{TOP}$ (0.185 in Column (2)) suggests that a one-standard-deviation increase in the ratio of high-intensity negative words leads to a 20.3% increase in the odds of fraud. The results from the last regression (Column (5)), which includes all three negative sentiment variables, are consistent with the other regressions. The predictive value of disclosure sentiment for public firms' future risk, including stock return volatility and bankruptcy, has been established in prior research [10,44]. We add to this body of literature by showing that a more effective sentiment measure will associate stronger with firms' future fraud activities.

[insert **Table 8** here]

Robustness Checks

We conduct a battery of robustness checks to strengthen our results, which are summarized in **Appendix F**. First, to further support that the new sentiment measure constructed by WOLVES can reconcile the inconsistent findings in the literature, we replicate our regression analyses of market reactions using two sample periods: (1) the LM sample period that follows Loughran and McDonald [41] to cover firm-year observations only before 2008 and (2) the full sample period that covers all firm-year observations. To align with the setting in Frankel et al. [17], we use the whole annual report sample to construct sentiment measures and investigate their associations with market reaction. The results of the LM sample period reveal that both $NegTone^{LM}$ and $NegTone^{TOP}$ are negatively and significantly associated with the market reaction. However, the association between $NegTone^{LM}$ and market reaction substantially weakens and becomes insignificant in the full sample period. In contrast, the association between $NegTone^{TOP}$ and market reaction remains statistically significant at the 10% level. These findings imply that our proposed WOLVES algorithm can construct more effective sentiment measures.

Second, the current regression analyses are based on the MD&A section of annual reports, which may have a generalizability issue. We thus replicate the regression analyses using the whole annual report text. The results are consistent with our main findings. We also replicate the analyses using the conference call dataset. The results are consistent with our main argument that words with a larger sentiment intensity score are more informative in predicting various important financial indicators. Note that we do not replicate the analyses using financial news as it is not generally considered as company disclosure, and it is extremely noisy to match the news data to the firm's financial data.

Third, we run a series of sensitivity tests to check how WOLVES perform according to different sets of hyperparameters. We vary the vector dimensionality (from 300 to 250 and 350) and the window size for choosing neighboring words (from 5 to 4 and 6) when constructing word2vec. Our results still hold under different hyperparameter settings.

Discussion and Implications

In this study, we propose WOLVES which combines the word embedding approach and a predefined word list to quantify word sentiment evolution in financial text. With WOLVES, we find that the sentiment intensity of LM negative words decreases over time in annual reports but not in conference calls or financial news. We attribute this phenomenon to managers adjusting their language in an attempt to avoid or mitigate negative perceptions from investors. We also find that the sentiment measure generated by WOLVES in annual reports can consistently predict market reaction, firm future performance, and fraudulent statements. These findings have several implications for academic researchers and practitioners.

First, our study contributes to the emerging research into human-AI interactions which focuses on either the effect of AI on human behavior in various tasks [5,40,49] or on how human input and feedback influence AI performance [20,53]. Our study provides insights into human-AI interactions in the complex financial market context. Our evidence that managers leverage business-context lexicons to adjust their use of negative words in an attempt to control investors' negative perceptions reveals how the development of sentiment lexicons can influence managers' behavior. We demonstrate that the sentiment intensity of LM negative words significantly drops over time: (1) after the implementation year (2002) of the SOX 404 controls, (2) after the publication year (2011) of the LM dictionary, or (3) when the words also appear in the H4 dictionary. On the other hand, we demonstrate how the strategic disclosure behavior of managers alters the effectiveness of lexicons and their ability to capture investors' sentiment. We also propose WOLVES, a new algorithm for adjusting AI strategies in response to this behavior. Our empirical analysis shows that only negative words with a sustained sentiment intensity can effectively predict market reaction, corporate performance, or accounting fraud. Consequently, intelligent machines based on lexicons will become less useful in the sentiment analysis of financial text if they do not account for word evolution. Our study thus makes a unique contribution to the understanding of human-AI interaction in financial text analysis.

Second, we contribute to quantitative examinations of word evolution and human language change in the financial context. Other research has related language evolution and semantic changes in words to factors such as word frequency [24], word nature [24], cultural change [30,50], and grammatical rules [38]. Our study suggests that word evolution can be driven by managers’ adversarial behavior, as they struggle to resist the effects of automatically extracting sentiment via computer programs. Other studies suggest that managers have incentives to withhold bad news [31], speculate on buzzwords [6], and manage tone [1,27]. As the text-based sections of disclosures are less precise in nature than the quantitative parts, managers are more able to manipulate them. We identify this effect by examining annual reports, in which authors (managers) have adequate time and strong incentives to adjust word usage, but not in conference calls or financial news, in which managers or authors do not have the time or incentives to manipulate the text content. We provide evidence that the sentiment intensity of LM negative words in annual reports, relative to conference calls and financial news, diminishes over time. Our results suggest that an adversarial approach may have a role in language evolution, which is, to our knowledge, a novel idea in linguistic literature.

Finally, our study provides important implications for practitioners applying textual analysis to financial text analysis. We extend earlier studies that apply deep learning and advanced NLP techniques to reveal how financial text can be used for sentiment signaling. Although some studies have explored word embedding for enhancing sentiment word lists, their approaches mainly rely on a static word embedding model, ignoring the importance of the evolutionary effect of word meanings (see **Appendix A**). We show that WOLVES can effectively quantify the semantic meanings of words, which can inform how LM words change their intended sentiment over time. We can then apply WOLVES to dynamically revise and update lexicons to ensure the validity of the dictionary-based textual analysis. This data-driven approach keeps human effort to a minimum [26,37], which is highly desirable when handling big data in finance research [21]. We further illustrate the practical value of our proposed method by showing that the sentiment intensity score can serve as a powerful yet simple weighting scheme for constructing a more reliable sentiment measure. Unlike studies that consider word frequency [41] or market information [28], our proposed weighting scheme considers the dynamic semantic change of words and is, therefore, novel, practical, and conceptually sound. Our study also extends existing research that captures word importance via their discriminative power of classifying a document's sentiment, which requires a human label for each document [71]. By using a human-generated word list and high-dimensional word embedding vector representation, our method does not rely on a training sample where sentiment labels are required.

Limitations and Future Directions

Our study has several limitations, which can inform future research. First, we train the word embedding models independently for each calendar year. We thus implicitly assume that all word embedding spaces are independent, but this does not consider temporal dependence, which may be crucial to understanding word evolution. Extending the word embedding model to account for temporal-correlated text data can be beneficial in future research. Second, we only study word evolution in financial text from the sentiment perspective. Investigating this evolution in terms of other linguistic dimensions, such as litigious, uncertain, and strong modals, can be valuable and is feasible through the application of various predefined business-context dictionaries [4,41]. Finally, we limit our focus to two specific types of corporate disclosure: annual reports and earnings conference calls. Different types of financial narratives (e.g., analysts' reports and word-of-mouth through social media for financial services) can also be examined to reveal other underlying mechanisms of the word evolution. We can further investigate the word sentiment evolution across different press media to uncover any heterogeneous effect among different news outlets and news editors.

Conclusion

In conclusion, this paper presents a novel algorithm, **Wo**rd **L**ist **Ve**ctor for **S**entiment (WOLVES), which drastically enhances traditional sentiment analysis techniques by leveraging the dynamic nature of language. By incorporating word embedding techniques to dynamically update sentiment word lists according to evolving semantic meaning and sentiment strength, WOLVES addresses the limitations of static sentiment analysis models. Our findings reveal a subtler understanding of sentiment in financial texts over time, reflect strategic communication efforts by managers, and reconcile inconsistencies in prior research. Ultimately, this study underscores the potential of machine learning methods to improve the extraction of valuable business intelligence from financial texts, providing a foundation for future research to further refine these techniques.

Acknowledgements

The authors thank the anonymous reviewers and the Editor-in-Chief for their constructive comments on this paper. The authors have benefited from discussions with Yanzhen Chen, Carlos Fernández-Loría, Ohchan Kwon, Kai-Lung Hui, Chao Jin, Jing Wang, Juanting Wang, Zhitao Yin, and Ruichao Zhu and comments and suggestions from participants in HKUST Business School Ph.D. Student Conference and PACIS 2023 Doctoral Consortium.

Funding

This project was funded by the Theme-based Research Grant on Fintech (T31-604/18-N) of the Research Grant Council of Hong Kong.

References

1. Allee, K.D. and Deangelis, M.D. The structure of voluntary disclosure narratives: Evidence from tone dispersion. *Journal of Accounting Research*, *53*, 2 (2015), 241–274.
2. Birner, B. *Is English Changing?* Linguistic Society of America, Washington DC, 1999.

3. Blankespoor, E. The impact of information processing costs on firm disclosure choice: Evidence from the XBRL mandate. *Journal of Accounting Research*, *57*, 4 (2019), 919–967.
4. Bodnaruk, A., Loughran, T., and McDonald, B. Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, *50*, 4 (2015), 623–646.
5. Cao, S.S., Jiang, W., Yang, B., and Zhang, A. How to talk when a machine is listening: Corporate disclosure in the age of AI. *Review of Financial Studies*, forthcoming (2022).
6. Cheng, S.F., De Franco, G., Jiang, H., and Lin, P. Riding the blockchain mania: Public firms' speculative 8-k disclosures. *Management Science*, *65*, 12 (2019), 5901–5913.
7. Clarkson, P.M., Kao, J.L., and Richardson, G.D. Evidence that management discussion and analysis (MD&A) is a part of a firm's overall disclosure package. *Contemporary Accounting Research*, *16*, 1 (1999), 111–134.
8. D'Augusta, C. and DeAngelis, M.D. Does accounting conservatism discipline qualitative disclosure? Evidence from tone management in the MD&A. *Contemporary Accounting Research*, *37*, 4 (2020), 2287–2318.
9. Davis, A.K. and Tama-Sweet, I. Managers' use of language across alternative disclosure outlets: Earnings press releases versus MD&A. *Contemporary Accounting Research*, *29*, 3 (2012), 804–837.
10. Deveikyte, J., Geman, H., Piccari, C., and Provetti, A. A sentiment analysis approach to the prediction of market volatility. *Frontiers in Artificial Intelligence*, *5*, (2022), 186.
11. Dong, W., Liao, S., and Zhang, Z. Leveraging financial social media data for corporate fraud detection. *Journal of Management Information Systems*, *35*, 2 (2018), 461–487.
12. Dzieliiski, M., Wagner, A.F., and Zeckhauser, R.J. Straight talkers and vague talkers: The effects of managerial style in earnings conference calls. *SSRN Electronic Journal*, (2017).
13. Fama, E.F. and French, K.R. Industry costs of equity. *Journal of Financial Economics*, *43*, 2 (1997), 153–193.
14. Fama, E.F. and MacBeth, J.D. Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, *81*, 3 (1973), 607–636.
15. Feldman, R. Techniques and applications for sentiment analysis: The main applications and challenges of one of the hottest research areas in computer science. *Communications of the ACM 56*, 2013, 82–89.
16. Feldman, R., Govindaraj, S., Livnat, J., and Segal, B. Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, *15*, 4 (2010), 915–953.
17. Frankel, R., Jennings, J., and Lee, J. Disclosure sentiment: Machine learning vs dictionary methods. *Management Science*, *68*, 7 (2022), 5514–5532.
18. García, D. Sentiment during recessions. *Journal of Finance*, *68*, 3 (2013), 1267–1300.
19. Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 16 (2018), E3635–E3644.
20. Ge, R., Zheng, Z., Tian, X., and Liao, L. Human-robot interaction: When investors adjust the usage of robo-advisors in peer-to-peer lending. *Information Systems Research*, *32*, 3 (2021), 774–785.
21. Goldstein, I., Spatt, C.S., and Ye, M. Big data in finance. *Review of Financial Studies*, *34*, 7 (2021), 3213–3225.

22. Goodall, C. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*, 2 (1991), 285-321.
23. Hamilton, W.L., Clark, K., Leskovec, J., and Jurafsky, D. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. Association for Computational Linguistics (ACL), Austin, Texas, 2016, pp. 595–605.
24. Hamilton, W.L., Leskovec, J., and Jurafsky, D. Diachronie word embeddings reveal statistical laws of semantic change. In *54th Annual Meeting of the Association for Computational Linguistics*. 2016, pp. 1489–1501.
25. Henry, E. and Leone, J.A. Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. *Accounting Review*, *91*, 1 (2016), 153–178.
26. Huang, A.H., Lehavy, R., Zang, A.Y., and Zheng, R. Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science*, *64*, 6 (2018), 2833–2855.
27. Huang, X., Teoh, S.H., and Zhang, Y. Tone management. *Accounting Review*, *89*, 3 (2014), 1083–1113.
28. Jegadeesh, N. and Wu, D. Word power: A new approach for content analysis. *Journal of Financial Economics*, *110*, 3 (2013), 712–729.
29. Kim, Y., Chiu, Y.-I., Hanaki, K., Hegde, D., and Petrov, S. Temporal analysis of language through neural language models. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, (2014), 61–65.
30. Kirby, S., Dowman, M., and Griffiths, T.L. Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 12 (2007), 5241–5245.
31. Kothari, S.P., Shu, S., and Wysocki, P.D. Do managers withhold bad news. *Journal of Accounting Research*, *47*, 1 (2009), 241–276.
32. Kozlowski, A.C., Taddy, M., and Evans, J.A. The geometry of culture: Analyzing the meanings of class through word embeddings: *American Sociological Review*, *84*, 5 (2019), 905–949.
33. Larcker, D.F. and Zakolyukina, A.A. Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, *50*, 2 (2012), 495–540.
34. Lawson, M.A., Martin, A.E., Huda, I., Matz, S.C., and Berger, J. Hiring women into senior leadership positions is associated with a reduction in gender stereotypes in organizational language. *Proceedings of the National Academy of Sciences*, *119*, 9 (2022).
35. Li, F. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, *45*, 2–3 (2008), 221–247.
36. Li, F. Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature*, *29*, (2010), 143.
37. Li, K., Mai, F., Shen, R., and Yan, X. Measuring corporate culture using machine learning. *The Review of Financial Studies*, *34*, 7 (2021), 3265–3315.
38. Lieberman, E., Michel, J.B., Jackson, J., Tang, T., and Nowak, M.A. Quantifying the evolutionary dynamics of language. *Nature*, *449*, 7163 (2007), 713–716.
39. Lo, K., Ramos, F., and Rogo, R. Earnings management and annual report readability. *Journal of Accounting and Economics*, *63*, 1 (2017), 1–25.

40. Lou, B. and Wu, L. Artificial intelligence and drug innovation: A large scale examination of the pharmaceutical industry. *MIS Quarterly*, *45*, 3 (2021), 1451–1482.
41. Loughran, T. and McDonald, B. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, *66*, 1 (2011), 35–65.
42. Loughran, T. and McDonald, B. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*, 4 (2016), 1187–1230.
43. Marquis, C., Toffel, M.W., and Zhou, Y. Scrutiny, norms, and selective disclosure: A global study of greenwashing. *Organization Science*, *27*, 2 (2016), 483–504.
44. Mayew, W.J., Sethuraman, M., and Venkatachalam, M. MD&A disclosure and the firm's ability to continue as a going concern. *The Accounting Review*, *90*, 4 (2015), 1621–1651.
45. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger, eds., *Advances in Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 2013, pp. 3111–3119.
46. Newberry, M.G., Ahern, C.A., Clark, R., and Plotkin, J.B. Detecting evolutionary forces in language change. *Nature*, *551*, 7679 (2017), 223–226.
47. Ni, J., Li, J., and McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, (2019), 188–197.
48. Pagel, M., Atkinson, Q.D., and Meade, A. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, *449*, 7163 (2007), 717–720.
49. Park, E.H., Werder, K., Cao, L., and Ramesh, B. Why do family members reject AI in health care? Competing effects of emotions. *Journal of Management Information Systems*, *39*, 3 (2022), 765–792.
50. Pechenick, E.A., Danforth, C.M., and Dodds, P.S. Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE*, *10*, 10 (2015), e0137041.
51. Peng, J., Zhang, J., and Gopal, R. The good, the bad, and the social media: Financial implications of social media reactions to firm-related news. *Journal of Management Information Systems*, *39*, 3 (2022), 706–732.
52. Pennington, J., Socher, R., and Manning, C.D. GloVe: Global vectors for word representation. In A. Moschitti, B. Pang and W. Daelemans, eds., *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543.
53. Rahwan, I., Cebrian, M., Obradovich, N., et al. Machine behaviour. *Nature 568*, 2019, 477–486.
54. Rai, A. Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science 48*, 2020, 137–141.
55. Richardson, S. and Tuna, İ. Predicting earnings management : The case of earnings. *Social Science Research Network Working Paper Series*, (2002).
56. Rogers, J.L., Van Buskirk, A., and Zechman, S.L.C. Disclosure tone and shareholder litigation. *The Accounting Review*, *86*, 6 (2011), 2155–2183.
57. Schlechtweg, D., Hätty, A., del Tredici, M., and Walde, S.S. im. A wind of change: Detecting and evaluating lexical semantic change across times and domains. *ACL 2019 -*

57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, (2019), 732–746.

58. Sun, Y. Do MD&A disclosures help users interpret disproportionate inventory increases? *The Accounting Review*, *85*, 4 (2010), 1411–1440.
59. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*, 2 (2011), 267–307.
60. Tan, H., Wang, E.Y., and Zhou, B. When the use of positive language backfires: The joint effect of tone, readability, and investor sophistication on earnings judgments. *Journal of Accounting Research*, *52*, 1 (2014), 273–302.
61. Tetlock, P.C. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, *62*, 3 (2007), 1139–1168.
62. Tetlock, P.C., Saar-Tsechansky, M., and Macskassy, S. More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance*, *63*, 3 (2008), 1437–1467.
63. Theil, C.K., Štajner, S., and Stuckenschmidt, H. Explaining financial uncertainty through specialized word embeddings. *ACM/IMS Transactions on Data Science*, (March 2020), 1–19.
64. Tsai, M.F. and Wang, C.J. Financial keyword expansion via continuous word vector representations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2014, pp. 1453–1458.
65. Turney, P.D. and Pantel, P. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, (2010), 141–188.
66. Wang, Z., Jiang, C., Zhao, H., and Ding, Y. Mining semantic soft factors for credit risk evaluation in peer-to-peer lending. *Journal of Management Information Systems*, *37*, 1 (2020), 282–308.
67. Wright, A.P., Jones, C.M., Chau, D.H., Matthew Gladden, R., and Sumner, S.A. Detection of emerging drugs involved in overdose via diachronic word embeddings of substances discussed on social media. *Journal of Biomedical Informatics*, *119*, (2021), 103824.
68. Xie, P., Chen, H., and Hu, Y.J. Signal or noise in social media discussions: The role of network cohesion in predicting the bitcoin market. *Journal of Management Information Systems*, *37*, 4 (2020), 933–956.
69. Yang, Y., Zhang, K., and Fan, Y. Analyzing firm reports for volatility prediction: A knowledge-driven text-embedding approach. *INFORMS Journal on Computing*, (2021), 1–19.
70. Zhu, Y., Wu, Z., Zhang, H., and Yu, J. Media sentiment, institutional investors and probability of stock price crash: evidence from Chinese stock markets. *Accounting and Finance*, *57*, 5 (2017), 1635–1670.
71. Zou, Y., Gui, T., Zhang, Q., and Huang, X.J. A lexicon-based supervised attention model for neural sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 868–877.

List of Tables

Table 1. Variable Definitions

Variable Name	Definition
Intensity	The cosine similarity of an LM word vector to the representative group vector of the same sentiment label minus the cosine similarity to the representative group vector of the opposite sentiment label.
YearNum	A consecutive year sequence capturing the year trend.
AfterLM	An indicator that equals 1 for years after 2011 (the publication year of the LM dictionary), and 0 otherwise.
AfterSOX	An indicator that equals 1 for years after 2002 (the implementation of SOX 404 controls), and 0 otherwise.
Report	An indicator equals 1 if an LM word comes from the annual report domain, and 0 otherwise.
Harvard	An indicator equals 1 if an LM word is also defined by the H4 dictionary, and 0 otherwise.
TermFreq	The term frequency of an LM word. It is calculated as the average number of times the word appears in a document.
GroupDist	One minus the cosine similarity between the groups of positive and negative word vectors.

Table 2. Word Sentiment Evolution in Annual Reports

	Annual Reports vs. Conference Calls		Annual Reports vs. Financial News	
	(1)	(2)	(3)	(4)
Variables	$Intensity^{POS}$	$Intensity^{NEG}$	$Intensity^{POS}$	$Intensity^{NEG}$
Constant	0.00631	-0.146	-0.0442	-0.132**
	(0.162)	(-2.011)	(-1.227)	(-4.212)
YearNum	7.85e-05	0.000809	0.00182**	0.000900*
	(0.209)	(1.896)	(4.048)	(2.481)
YearNum × *Report*	0.000559	-0.00236**	-0.000788	-0.00202**
	(1.153)	(-4.770)	(-1.460)	(-6.570)
TermFreq	0.00233	-0.00194	-0.00319	-0.00445*
	(0.882)	(-1.081)	(-1.163)	(-2.338)
GroupDist	0.245**	0.415**	0.315**	0.425**
	(4.715)	(4.377)	(6.258)	(9.422)
Word-Domain FE	Yes	Yes	Yes	Yes
# of Observations	8,320	23,680	11,836	51,172
Adjusted R^2	0.897	0.858	0.852	0.851

Notes. Robust t-statistics based on standard errors clustered by word-domain and year are shown in parentheses below the coefficients. ** and * indicate significance at the 0.01 and 0.05 levels, respectively.

Table 3. Effect of the SOX 404 Controls on the Word Sentiment Evolution in Annual Reports

	Annual Reports vs. Financial News	
	(1)	(2)
Variables	$Intensity^{POS}$	$Intensity^{NEG}$
Constant	-0.171**	-0.200**
	(-4.019)	(-9.447)
AfterSOX	0.00275	-0.00152
	(0.623)	(-0.464)
AfterSOX × *Report*	0.00490	-0.0155**
	(0.852)	(-3.454)
TermFreq	-0.00181	-0.00605**
	(-0.720)	(-2.887)
GroupDist	0.499**	0.524**
	(8.616)	(17.91)
Word-Domain FE	Yes	Yes
# of Observations	11,836	51,172
Adjusted R^2	0.851	0.851

Notes. Robust t-statistics based on standard errors clustered by word-domain and year are shown in parentheses below the coefficients. ** indicates significance at the 0.01 level.

Table 4. Effect of the LM Dictionary on the Word Sentiment Evolution in Annual Reports

	Annual Reports vs. Conference Calls		Annual Reports vs. Financial News	
	(1)	(2)	(3)	(4)
Variables	$Intensity^{POS}$	$Intensity^{NEG}$	$Intensity^{POS}$	$Intensity^{NEG}$
Constant	-0.0240	-0.138	-0.0784	-0.0511
	(-0.662)	(-1.294)	(-1.938)	(-1.597)
AfterLM	-0.00193	0.00412	0.0201**	0.0244**
	(-0.610)	(0.903)	(2.930)	(4.559)
AfterLM × *Report*	0.00626	-0.0142*	-0.0118	-0.0318**
	(1.575)	(-2.396)	(-1.571)	(-5.263)
TermFreq	0.00287	-0.00370	-0.00112	-0.00595*
	(1.153)	(-1.921)	(-0.452)	(-2.762)
GroupDist	0.288**	0.401**	0.374**	0.313**
	(6.218)	(2.959)	(6.747)	(7.150)
Word-Domain FE	Yes	Yes	Yes	Yes
# of Observations	8,320	23,680	11,836	51,172
Adjusted R^2	0.897	0.857	0.853	0.846

Notes. Robust t-statistics based on standard errors clustered by word-domain and year are shown in parentheses below the coefficients. ** and * indicate significance at the 0.01 and 0.05 levels, respectively.

Table 5. Effect of the H4 Dictionary on Word Sentiment Evolution in Annual Reports

	Annual Reports vs. Conference Calls		Annual Reports vs. Financial News	
	(1)	(2)	(3)	(4)
Variables	$Intensity^{POS}$	$Intensity^{NEG}$	$Intensity^{POS}$	$Intensity^{NEG}$
Constant	0.00596	-0.146	-0.0441	-0.132**
	(0.152)	(-2.012)	(-1.224)	(-4.215)
YearNum	0.000340	0.000498	0.00199**	0.000699
	(0.697)	(1.123)	(3.612)	(1.964)
YearNum × *Report*	-0.000134	-0.00174**	-0.00114	-0.00150**
	(-0.237)	(-3.257)	(-1.677)	(-4.584)
YearNum × *Havard*	-0.000609	0.00102**	-0.000398	0.000688*
	(-1.111)	(4.092)	(-0.494)	(2.327)
YearNum × *Report* × *Harvard*	0.00168	-0.00209**	0.000869	-0.00183**
	(2.042)	(-4.360)	(0.915)	(-4.453)
TermFreq	0.00127	-0.00137	-0.00375	-0.00378
	(0.454)	(-0.763)	(-1.334)	(-1.991)
GroupDist	0.246**	0.415**	0.315**	0.425**
	(4.709)	(4.379)	(6.255)	(9.427)
Word-Domain FE	Yes	Yes	Yes	Yes
# of Observations	8,320	23,680	11,836	51,172
Adjusted R^2	0.897	0.858	0.853	0.852

Notes. Robust *t*-statistics based on standard errors clustered by word-domain and year are shown in parentheses below the coefficients. ** and * indicate significance at the 0.01 and 0.05 levels, respectively.

Table 6. Filing Period Excess Return Regression

	(1)	(2)	(3)	(4)	(5)
Variables	$CAR_{[0,3]}$	$CAR_{[0,3]}$	$CAR_{[0,3]}$	$CAR_{[0,3]}$	$CAR_{[0,3]}$
$NegTone^{LM}$	-0.00148				
	(-1.765)				
$NegTone^{TOP}$		-0.00160*			-0.00145*
		(-2.370)			(-2.403)
$NegTone^{MID}$			-0.000900		-0.000611
			(-1.322)		(-0.942)
$NegTone^{BOT}$				-0.000182	0.000181
				(-0.257)	(0.260)
Controls and Constant	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
# of Observations	53,182	53,182	53,182	53,182	53,182
Adjusted R^2	0.0553	0.0553	0.0552	0.0551	0.0553

Notes. For easy comparison, we apply z-score normalization to all sentiment measures. Robust *t*-statistics based on standard errors clustered by firm and year are shown in parentheses below the coefficients. ** and * indicate significance at the 0.01 and 0.05 levels, respectively.

Table 7. Earnings Performance Regression

Variables	(1) $Earnings_{t+1}$	(2) $Earnings_{t+1}$	(3) $Earnings_{t+1}$	(4) $Earnings_{t+1}$	(5) $Earnings_{t+1}$
$NegTone^{LM}$	-0.00508** (-3.193)				
$NegTone^{TOP}$		-0.00538** (-3.094)			-0.00475* (-2.709)
$NegTone^{MID}$			-0.00252 (-1.959)		-0.00128 (-1.003)
$NegTone^{BOT}$				-0.00162 (-1.452)	-0.000591 (-0.493)
Controls and Constant	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
# of Observations	43,510	43,510	43,510	43,510	43,510
Adjusted R^2	0.617	0.617	0.617	0.617	0.617

Notes. For easy comparison, we apply z-score normalization to all sentiment measures. Robust *t*-statistics based on standard errors clustered by firm and year are shown in parentheses below the coefficients. ** and * indicate significance at the 0.01 and 0.05 levels, respectively.

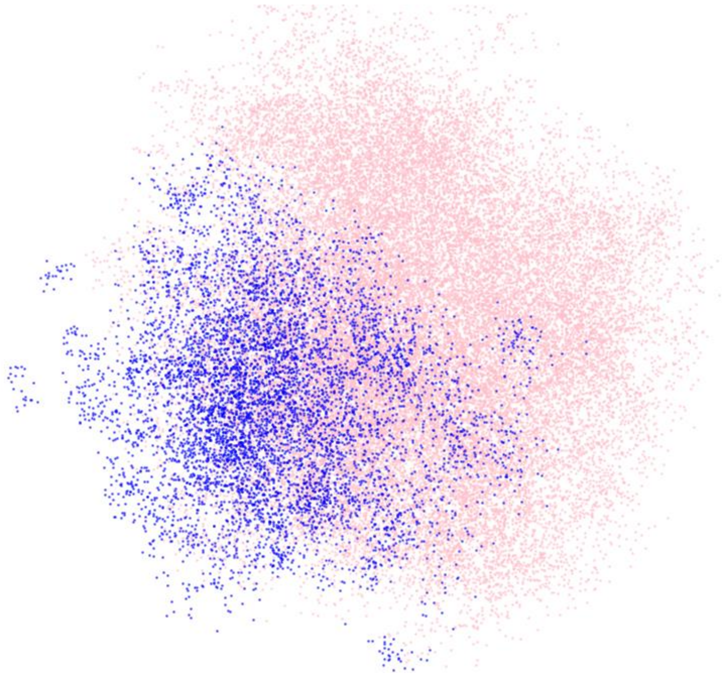
Table 8. Accounting Fraud Regression

Variables	(1) $Fraud_{t+1}$	(2) $Fraud_{t+1}$	(3) $Fraud_{t+1}$	(4) $Fraud_{t+1}$	(5) $Fraud_{t+1}$
$NegTone^{LM}$	0.116 (1.323)				
$NegTone^{TOP}$		0.185* (2.184)			0.210* (2.296)
$NegTone^{MID}$			0.0418 (0.486)		-0.0110 (-0.115)
$NegTone^{BOT}$				-0.0441 (-0.492)	-0.0914 (-0.961)
Controls and Constant	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
# of Observations	39,594	39,594	39,594	39,594	39,594
Log likelihood	-852.8	-851.5	-853.6	-853.6	-851
Pseudo R^2	0.0717	0.0732	0.0709	0.0709	0.0737

Notes. For easy comparison, we apply z-score normalization to all sentiment measures. ** and * indicate significance at the 0.01 and 0.05 levels, respectively

List of Figures

Figure 1. PCA Analysis of LM Word Embedding in Annual Reports



Notes. The following steps are used to generate this figure. First, for each year, we train a word embedding model using all annual reports from that year. Second, we apply the trained yearly model to calculate the word embedding vectors of all LM words. Third, we align the word embedding vector across years by using the Procrustes transformation [22]. Finally, PCA is applied to project the high-dimensional vectors into two-dimensional vectors, with a blue (pink) point representing a positive (negative) word.

For visualization purpose, we apply PCA to reduce the dimension of embedding vectors from 300 to 2. Such significant data compression will influence the geometrical distance between embeddings and thus affects the visualized clusters. To better interpret the results, we provide some statistics below to show that the embedding vectors of the sentiment words are generally closer to their own centroid than to the opposite centroid:

(a) Average cosine similarity between positive words to the positive centroid: 0.2671
(b) Average cosine similarity between negative words to the negative centroid: 0.2426
(c) Average cosine similarity between positive words to the negative centroid: 0.0657
(d) Average cosine similarity between negative words to the positive centroid: 0.0597

Figure 2. The WOLVES Algorithm

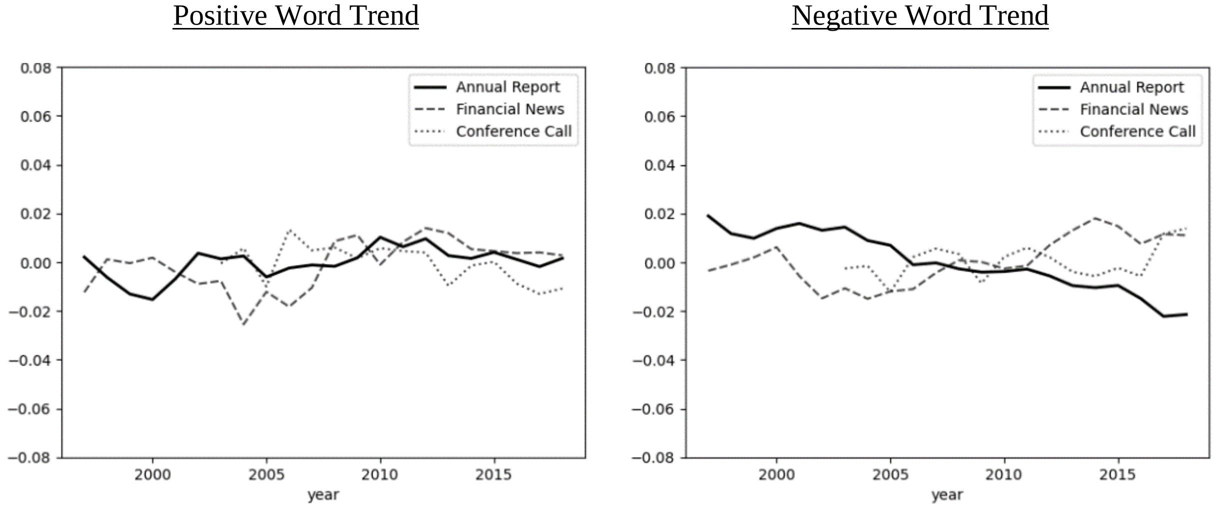
```
Input:
    (1) Corpus split in year $C = [c_1, \dots, c_T]$ where $T$ is the number of years in the corpus,
    (2) A positive word list $WL_p = \left[w_{p,1},\ \dots, w_{p,N_p}\right]$ where $N_p$ is the number of elements in $WL_p$, and
    (3) A negative word list $WL_n = \left[w_{n,1},\ \dots, w_{n,N_n}\right]$ where $N_n$is the number of elements in $WL_n$.
Output:
    Sentiment intensity scores for each word in the positive word list for each year:
        $Intensity_p = \left[Intensity_{1,1}, \dots, Intensity_{N_p,T}\right]$
    Sentiment intensity scores for each word in the negative word list for each year:
        $Intensity_n = \left[Intensity_{1,1}, \dots, Intensity_{N_n,T}\right]$

For $t$=1 to $T$
    Select corpus $c_t$ from $C$
    Conduct preprocessing on $c_t$
    Train a word2vec model $m_t$ using $c_t$
    Use model $m_t$ to generate embedding vectors for all words in $WL_p$: $V_{p,t} = \left[v_{p,1,t} \cdots, v_{p,N_p,t}\right]$
    Use model $m_t$ to generate embedding vectors for all words in $WL_n$: $V_{n,t} = \left[v_{n,1,t} \cdots, v_{n,N_n,t}\right]$

    Update $V_{p,t}$ and $V_{n,t}$ to remove word vectors with the ambiguous sentiment:
      $V'_{p,t} = \left[v_{p,i,t} \text{ for } v_{p,i,t} \text{ in } V_{p,t} \text{ if} \cos\left(v_{p,i,t}, average\left(V_{p,t}\right)\right) - \cos\left(v_{p,i,t}, average\left(V_{n,t}\right)\right) > \eta\right]$ and
        $V'_{n,t} = \left[v_{n,i,t} \text{ for } v_{n,i,t} \text{ in } V_{n,t} \text{ if} \cos\left(v_{n,i,t}, average\left(V_{n,t}\right)\right) - \cos\left(v_{n,i,t}, average\left(V_{p,t}\right)\right) > \eta\right]$,
    where $cos(\cdot\,,\ \cdot)$ is the cosine similarity function, $average(\cdot)$ is the component-wise mean function,
    and $\eta$ is a hyperparameter controlling the proportion of sentiment words in calculating sentiment
    centroids.

    Calculate the centroid based on positive word vectors in $V'_{p,t}$: $\bar{V}'_{p,t} = average\left(V'_{p,t}\right)$
    Calculate the centroid based on negative word vectors in $V'_{n,t}$: $\bar{V}'_{n,t} = average\left(V'_{n,t}\right)$
    For $i$=1 to $N_p$
        Calculate the sentiment intensity score of $w_{p,i}$ in $WL_p$ :
            $Intensity_{i,t} \ = cos\left(v_{p,i,t}, \bar{V}'_{p,t}\right) - cos\left(v_{p,i,t}, \bar{V}'_{n,t}\right)$
    Endfor
    For $j$=1 to $N_n$
        Calculate the sentiment intensity score of $w_{n,j}$ in $WL_n$ :
            $Intensity_{j,t} \ = cos\left(v_{n,j,t}, \bar{V}'_{n,t}\right) - cos\left(v_{n,j,t}, \bar{V}'_{p,t}\right)$
    Endfor
Endfor
```

Figure 3. Sentiment Intensity Trends of LM Words in the Three Financial Text Types

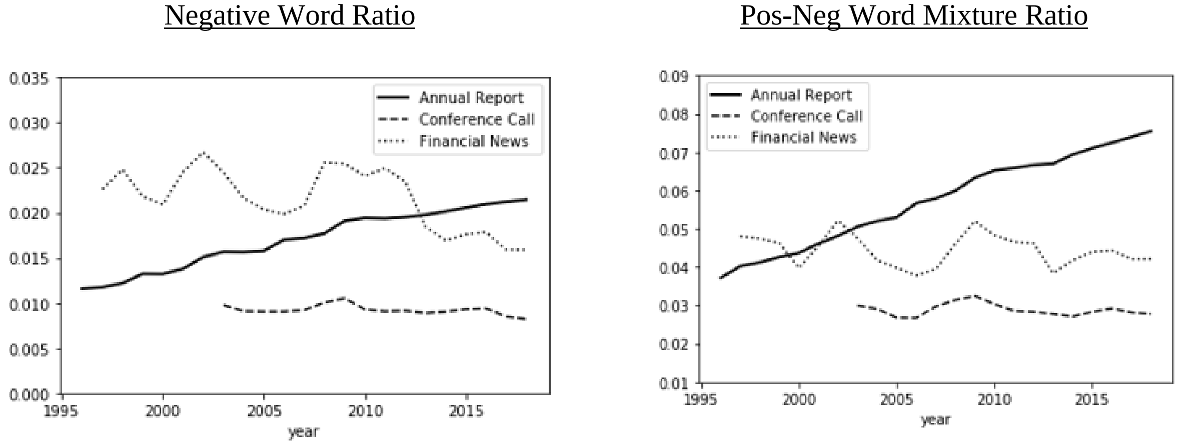


Positive Word Trend

Negative Word Trend

Notes. This figure provides a model-free analysis of the trends of the average sentiment intensity score of LM words across years and the three text types. For each year, we calculate the average sentiment intensity score of all positive and negative words in the annual reports, conference calls, and financial news, respectively. To get rid of the influence of other factors, we run the following regression model for each LM word i in a specific year y from a domain d: $Intensity_{i,t,d} = \beta_0 + \beta_1\ GroupDist_{t,d} + \beta_2\ TermFreq_{i,t,d} + \delta_d + \varepsilon_{i,t}$, where δ_d is a dummy variable representing the domain. Afterward, we obtain $Intensity'_{i,t,d} = Intensity_{i,t,d} - \left(\beta_0 + \beta_1\ GroupDist_{t,d} + \beta_2\ TermFreq_{i,t,d} + \delta_d\right)$ and visualize the trends according to this revised sentiment intensity score. The left (right) panel refers to the average revised score of positive (negative) words.

Figure 4. Different Yearly Ratio of LM Words in the Three Financial Text Types



Negative Word Ratio

Pos-Neg Word Mixture Ratio

Notes. This figure provides a model-free analysis of managers' strategic communication across years and the three text types. The left panel presents the negative word ratio, defined as the number of negative words divided by the total number of words in the corpus per year. The right panel presents the positive-negative word mixture ratio, defined as the number of sentences containing both positive and negative words divided by the number of sentences in the corpus per year.