Explaining Financial Uncertainty through Specialized Word Embeddings

CHRISTOPH KILIAN THEIL, University of Mannheim, Germany
SANJA ŠTAJNER, Symanto Research, Germany
HEINER STUCKENSCHMIDT, University of Mannheim, Germany

The detection of vague, speculative, or otherwise uncertain language has been performed in the encyclopedic, political, and scientific domains yet left relatively untouched in finance. However, the latter benefits from public sources of big financial data that can be linked with extracted measures of linguistic uncertainty as a mean of extrinsic model validation. Doing so further helps in understanding how the linguistic uncertainty of financial disclosures might induce financial uncertainty to the market. To explore this field, we use term weighting methods to detect linguistic uncertainty in a large dataset of financial disclosures. As a baseline, we use an existing dictionary of financial uncertainty triggers; furthermore, we retrieve related terms in specialized word embedding models to automatically expand this dictionary. Apart from an industry-agnostic expansion, we create expansions incorporating industry-specific jargon. In a set of cross-sectional event study regressions, we show that the such enriched dictionary explains a significantly larger share of future volatility, a common financial uncertainty measure, than before. Furthermore, we show that—different to the plain dictionary—our embedding models are well suited to explain future analyst forecast uncertainty. Notably, our results indicate that enriching the dictionary with industry-specific vocabulary explains a significantly larger share of financial uncertainty than an industry-agnostic expansion.

CCS Concepts: • **Computing methodologies** → **Information extraction**; • **Applied computing** → *Economics*;

Additional Key Words and Phrases: Linguistic uncertainty, financial uncertainty, word embeddings, financial disclosures, text mining

ACM Reference format:
Christoph Kilian Theil, Sanja Štajner, and Heiner Stuckenschmidt. 2020. Explaining Financial Uncertainty through Specialized Word Embeddings. *ACM/IMS Trans. Data Sci.* 1, 1, Article 6 (February 2020), 19 pages.
https://doi.org/10.1145/3343039

1 INTRODUCTION

The financial markets are driven by large sources of both structured and unstructured data. The recently growing area of financial text mining has shown that linking both of these sources is a

This work was conducted during Sanja Štajner's affiliation with the University of Mannheim, Germany.
Authors' addresses: C. K. Theil and H. Stuckenschmidt, University of Mannheim, Data and Web Science Group, B6, 26, Mannheim 68159, Germany; emails: {christoph, heiner}@informatik.uni-mannheim.de; S. Štajner, Symanto Research, Pretzfelder Str. 15, Nürnberg 90425, Germany; email: sanja.stajner@symanto.net.

2577-3224/2020/02-ART6 $15.00
https://doi.org/10.1145/3343039

NEW BUSINESS VENTURE

The securities being offered by the Company are **subject to the risks** inherent in any new business venture. Although the Company has operated as a contract research firm since 1986, it have limited experience and a short history of operations with respect to marketing and selling susceptibility tests or therapeutics. The Company has had only minimal revenues related to the sale of its genetic susceptibility testing services. With the exception of its periodontal susceptibility test, the genetic susceptibility tests **anticipated to be sold** by the Company have not yet been finally designed, developed, tested or marketed. Therefore, **there can be no assurance** that the Company will be able to complete these genetic susceptibility tests, that those tests will be accepted in the marketplace, or that the tests can be sold at a profit. The Company's business **may also be affected** significantly by economic and market conditions **over which the Company has no control**. Consequently, **an investment in the Company's Common Stock is highly speculative**. The Company **does not guarantee** any return on an investment in its Common Stock.

Fig. 1. An excerpt of a 10-K filed by a pharmaceuticals company in spring 1998. Parts classifiable as uncertain according to our methodology[1] are in bold.

fruitful avenue for research: Retrieving textual information from financial disclosures can help to explain and anticipate market movements.

In this article, we analyze relationships between the uncertainty of disclosure language and several financial uncertainty measures. For this purpose, we collected a large dataset of 10-Ks (see Section 1.1 for a definition), extracted a dictionary-based measure of linguistic uncertainty, and regressed market- and analyst-based measures of financial uncertainty on it. We thus show how analyzing the uncertainty of disclosed information helps to explain subsequent financial uncertainty.

1.1 Disclosure Type: 10-Ks

According to the *Securities Exchange Act of 1934*, all publicly traded U.S. companies with above 10 million dollars in assets and more than 500 shareholders are required to file 10-Ks by the Securities and Exchange Commission (SEC). A 10-K is an annual report giving a comprehensive summary of a company's activities throughout the preceding year. Figure 1 presents a 10-K excerpt exemplifying typical characteristics of these documents and how they can convey uncertainty. This 10-K belongs to a relatively young and small company with "limited experience," thus explaining the large share of uncertain wording in its business description.

Past literature has shown that 10-Ks are important informative disclosures for investors and analysts [15, 16]. Furthermore, 10-Ks have been analyzed to explain uncertainty of the information environment [17, 19]. Yet, past studies have predominantly relied on the independent variable of *readability* or *complexity* instead of assessing the linguistic uncertainty as determinant of real-worldly uncertainty. As of now, the scientific community lacks a systematic study examining how linguistic uncertainty of 10-Ks helps in explaining financial uncertainty measures apart from volatility.

A 10-K can contain up to 15 sections, with Sections 1 and 2, "Business and Property Description," Section 7, "Management's Discussion and Analysis" (MD&A), and Section 8, "Financial Statements," being the most predominantly included [6]. While some past studies have argued to focus on specific sections they deem most informative to stakeholders (typically the MD&A), Loughran

[1]Following Loughran and McDonald [18, p. 45], we define words and phrases as *uncertain* if they are (1) imprecise or (2) referring to risk.

and McDonald [19] argue to use the full document, as they have shown that using only the MD&A section "does not provide more powerful statistical tests" while introducing the probability of parsing errors. The findings of Dyer et al. [6] hint that using only Section 1a, "Risk Factors," might be of interest to be exploited. However, since this section became only mandatory to be included after 2005, it is usually less available.[2] Hence, we follow Loughran and McDonald [19] and analyze the documents in their entirety.

1.2 Independent Variable: Linguistic Uncertainty

Linguistic uncertainty can be inherent to any form of communication, may it be written or spoken, formal or informal. On one hand, users of uncertain communication may engage in it unintentionally due to lacking knowledge: An early study in educational research found that students exposed to less informative lectures adopt vaguer language than others when asked to summarize what they have learned [12]. On the other hand, people may also use uncertain language as a strategic tool to shape the opinions of their audience. This social phenomenon has especially been noticed in settings of asymmetric information like politics [21] or business [29]. Likewise, company executives might use uncertain language due to their ignorance about the current or future economic situation of their business, or they might follow a specific agenda and use it to intentionally obfuscate information from stakeholders.

In spite of the theoretical appeal of uncertainty, the variable is comparably under-explored in the financial text-mining community. Indeed, a comprehensive survey of the state-of-the-art methods by Loughran and McDonald [20] criticizes that "[m]any textual analysis studies have focused on the simple positive/negative dichotomy of sentiment analysis" in spite of its "low power" [20, p. 1224]. Instead, the authors specifically propose to investigate the more promising concept of uncertainty.

Another significant share of financial literature has been denoted to studying *readability* or *complexity* as measures of disclosure understandability [20, pp. 1193–1198]. The theory holds that a low readability of disclosures causes the documents to be less informative for stakeholders and hence induces uncertainty to the market. It is unlikely that educated and skilled individuals like banking analysts or investors have trouble to correctly assess the performance of a business due to, e.g., high shares of polysyllabic words or subclauses appearing in its annual report.[3] Therefore, it should predominantly be deliberate over-complication of information as form of obfuscation that affects the markets [3]. However, Loughran and McDonald [20] conclude in their survey on textual complexity in financial disclosures that it is problematic to disentangle whether (1) managers use complex language due to the nature of their business (2) or whether they intentionally over-complicate information to mitigate negative reactions to bad news. Thus, we are additionally motivated to explore linguistic uncertainty instead of the already well-explored variables of sentiment polarity and text complexity.

1.3 Contributions

We are the first to develop specialized word embedding models accounting for the industry-specific vocabulary of different business sectors. We use these models effectively to expand a dictionary of uncertainty triggers. Doing so, we provide a fine-grained analysis of how the choice of an industry

[2]Indeed, during parsing the disclosures (see Section 3.1.1), we noticed that such a restriction would reduce our sample size of 76,991 instances by about 50% (only 37,438 instances contained that section).

[3]Notably, Loughran and McDonald [19] have shown that the share of polysyllabic words, a common readability measure and one of the components of the Gunning Fog index [11], is actually a misspecified variable for measuring uncertainty of the information environment.

classification scheme (distinguishing between 5 to 49 industries) and altering the number of added similarity candidates k impact the dictionary expansions. We evaluate the effectiveness of our embedding models and expansions by providing: (1) an intrinsic error analysis exploring the suitability of our expansions qualitatively and (2) an extrinsic analysis measuring to which regard the expanded uncertainty dictionary explains drifts of overall market uncertainty with cross-sectional regression analyses. In summary, we contribute to the scientific community by:

- Developing the first word embedding models accounting for industry-specific vocabulary;
- Exploring different levels of granularity (i.e., the number of industries) according to which we train these industry-specific models;
- Exploring different values for the number of added candidate terms in our dictionary expansions;
- Performing a rigorous error analysis assessing whether the expanded dictionary indeed contains relevant terms according to human perception;
- Successfully using the expanded dictionary to statistically explain both drifts in stock return volatility and analyst uncertainty—the latter of which neither the plain dictionary nor an industry-agnostic expansion were capable of.

2 RELATED WORK

2.1 Financial Literature

Loughran and McDonald [18] were the first to introduce a set of sentiment dictionaries containing vocabulary specific to the financial domain. Based on a large sample of 10-Ks from 1994 to 2008, they manually developed dictionaries[4] spanning the categories of *positive*, *negative*, *litigious*, *strong modal*, *weak modal*, and—most important for our task—*uncertain* words. While purpose and content of the other dictionaries are rather self-explanatory, *litigious* "categorizes words reflecting a propensity for legal contest" and the *uncertain* dictionary was developed "with emphasis on the general notion of imprecision rather than exclusively focusing on risk" [18, p. 45]. Perhaps not surprisingly, the authors found that the cumulative tf-idf of *uncertain* words shares a positive and highly significant relationship with stock return volatility measured in the period after the filing date.

Following up and using a slightly expanded dataset, the same authors switched their scope of attention to the measure of *readability*, which they define as the "effective communication of valuation-relevant information" by companies [19, p. 1643]. They found that a simple file-size-based measure is better suited for explaining volatility, analyst forecast error, and analyst forecast dispersion than a traditional readability formula, the Gunning Fog Index [11]. Following their approach, we also perform event studies to quantify the impact of 10-K content on the previously mentioned financial uncertainty measures; yet, for reasons outlined in Section 1.2, we focus on the independent variable *uncertainty* instead of *readability*. We hypothesize that enriching the dictionary of uncertainty with industry-specific vocabulary should also reflect in more decisive regression results.

2.2 Natural Language Processing Literature

Tsai and Wang [32] automatically expanded Loughran and McDonald's [18] six dictionaries by training word embedding models on a corpus of 10-Ks from 1994 to 2006 and adding the 20 most cosine similar terms to each original dictionary entry. They found that doing so both improves the performance of a Ranking Support Vector Machine (SVM^{rank}) as well as a Support Vector

[4] https://www3.nd.edu/~mcdonald/Word_Lists.html.

Regression (SVR) model with bag-of-word vectors as features and stock return volatility in the year after the filing date as label. Following up, Tsai et al. [33] show that such an expanded dictionary can effectively be used to not only predict return volatility, but also post-event volatility (estimated with the Fama–French 3-factor model [7]) in the following year. Although the authors acknowledge that the regression on post-event volatility is sensitive with regard to the number of added candidates k [33, cf. Figure 3], they keep k equal to 20 due to a "diminishing return between prediction performance" for increasing numbers of k [33, p. 14]. While our findings confirm that generally, larger numbers of k benefit regressions of short-term post-event volatility, we will show that this not the case for the analyst-based measures that we study (see Section 4.2).

More recently, Rekabsaz et al. [26] refined this approach by including additional financial features and contrasting different methods for term weighting and feature fusion. They expanded the financial sentiment dictionaries in a similar fashion to Tsai and Wang [32] yet focusing on the *positive*, *negative*, and *uncertain* dictionaries and a set of 10-Ks from 2006 to 2015. Apart from bag-of-word vectors, they used the current volatility, a Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model [1], and a sector variable as features in their SVR. They found that limiting the data to a subset of the most recent 10-Ks (ca. 4,000 instances) and stacking the market features upon the text features further improves the performance of the predictive task. Yet they did not observe noticeable performance increases when splitting the set into 11 industry-specific subsets and re-running the experiments, which they attributed to the scarcity of training data (on average ca. 300 10-Ks per industry).

Theil et al. [31] perform a dictionary expansion following Tsai and Wang [32], yet with the aim of comparing it to (1) a general-domain expansion using a pre-trained embedding model based on the Google News dataset[5] and (2) manually filtered versions of such expansions. While the effect of manual filtering seems negligible, the results show that for a volatility regression task, a domain-specific model is better suited than a generic model. The authors further show that such an expansion can effectively be used for a binary classification task of sentences into *uncertain* or *certain* classes. Different to this article, we focus on training industry-specific embedding models. In addition, we provide a holistic view of financial uncertainty by showing that our industry-specific models can explain subsequent drifts in not only volatility but also in analyst forecast error and analyst dispersion.

2.3 Locating the Current Work

We argue that industry-specific knowledge should already be induced within the training process of the embedding models. Depending on the industry of the training data, the similar candidates of a term might change substantially (see Table 2). Furthermore, 20 as the number of adding closest neighboring terms appears arbitrary to us, since in related domains, the common threshold is 10 [10, 24]. Therefore, we were motivated in both comparing different values for the number of added candidates k and the number of industries according to which we allocate our data for the subsequent analyses and for training the embedding models.

However, in contrast to previous work [26, 32], our approach is not concerned with optimizing a predictive model of volatility but rather measuring the effect strength of linguistic uncertainty on the overall market in a cross-sectional study. Following Loughran and McDonald's [19] extensive financial methodology, we use the financial data as an extrinsic validation for the quality of our method. Apart from volatility, we deploy the analyst-based measures and an extended number of control variables introduced by the aforementioned authors. We thus demonstrate the robustness

[5] https://code.google.com/archive/p/word2vec/.

of our approach and make a case for training industry-specific instead of industry-agnostic "all-purpose" embedding models explored in previous work.

3 METHODOLOGY

To address our task, we apply the following pipeline: First, we collect a large dataset of 10-Ks (Section 3.1). Next, we train word embedding models and expand an existing dictionary of uncertainty triggers (Section 3.2). Afterward, we conduct an extensive error analysis of the expansions (Section 4.1). Finally, we perform a set of event study regressions as an extrinsic validation (Section 3.3).

3.1 Collecting a Large Dataset of 10-Ks

3.1.1 Parsing Procedure. We download all 220,565 10-Ks during 1994 to 2015 from the SEC's public filing database EDGAR.[6] First, we retrieve all text appearing between section headings with term matching heuristics. Afterward, we remove exhibits, graphics, HTML tags, and other anomalies leaving plain paragraph text. Last, following Loughran and McDonald [18, p. 40], we only keep sections with at least 250 words, as others are usually only "incorporated by reference." All texts are tokenized, stripped of punctuation and numbers, and lowercased with the exception of proper nouns, which are identified through part-of-speech tagging.[7] Hence, we ensure that, e.g., the modal verb "may" can be distinguished from the month "May." Our preprocessed text data can be found online.[8]

3.1.2 Data Screens. Following Loughran and McDonald [18, 19], we then perform a set of data screens: We remove duplicates (dropping 3,301) and instances with a filing date less than 180 days from the prior filing (dropping 653). Next, we require a match with the financial database CRSP[9] (dropping 113,396), the stock to be ordinary common equity (dropping 4,466), a stock price of greater than \$3 (dropping 15,281), a positive book-to-market ratio (dropping 3,384), as well as stock return data available for trading day windows t_{-252} to t_{-6} before, t_0 to t_1 around, and t_6 to t_{28} after the filing date; for the window prior to the filing, we consider instances with at least 60 days of return data available and for the window after the filing, we consider instances with at least 10 days of data (dropping 409). Last, we remove reports in which we could not identify at least one complete section (dropping 2,684). This leaves us with 76,991 reports for financial regression analyses. The residual of 121,235 files (excluding 22,339 duplicates and documents with less than one section) is used to train the embedding models. All data needed to replicate our regressions can be found in our Online Appendix.[10]

3.1.3 Choice of an Industry Classification Scheme. To divide our corpus into industry-specific sub-corpora, an industry classification scheme had to be deployed. Perhaps the two most popular of such schemes among professionals [34] are the Global Industry Classification Standard (GICS)[11] and the Industrial Classification Benchmark (ICB).[12] In financial research, however, the industry classification scheme developed by Fama and French [8] can be described as "default choice" [4, p. 57]. As it was also used by Loughran and McDonald [18, 19], we decided to use it for our purposes,

[6] https://www.sec.gov/edgar.shtml.
[7] We use NLTK 3.2.1 for all of these steps.
[8] http://data.dws.informatik.uni-mannheim.de/theil/10k.zip.
[9] http://www.crsp.com.
[10] http://data.dws.informatik.uni-mannheim.de/theil/acm-tds-19.zip.
[11] https://www.msci.com/gics.
[12] http://www.ftserussell.com/financial-data/industry-classification-benchmark-icb.

too. Another advantage of this scheme is that, depending on the preferred level of granularity, one can distinguish between {5, 10, 12, 17, 30, 38, 48, 49} industries (from now on: FF5 to FF49).

3.2 Training Word Embeddings and Expanding a Dictionary of Uncertainty Triggers

We address our task by performing a grid search over the parameters *number of added candidates (k)* and *number of industries*, thus investigating $20 \cdot 9 = 180$ possible parameter combinations.

3.2.1 Industry-Agnostic Model. We begin by training an industry-agnostic word embedding model on the full training data (126,330 10-Ks with approximately 2.3 billion words). For this purpose, we use word2vec [22] with standard parameters. We then use this model to automatically expand Loughran and McDonald's [18] list of 297 financial uncertainty triggers such as "anomalous," "predict," or "volatility." While Tsai and Wang [32] use a top-k approach with $k = 20$, past work by Rekabsaz et al. [27, 28] suggests that filtering related terms based on a cosine similarity (S_C) threshold might be worth exploring.[13] Cosine similarity is defined as follows:

$$S_C = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}. \tag{1}$$

Here A_i and B_i represent components of word vectors A and B, respectively.

To find out whether a top-k approach or a threshold based on S_C is more suitable for our task, we explored expansions in both fashions with the parameters suggested by the previous literature ($k = 20$, $S_C = 0.7$) and compared the regression results (see Section 3.3 for the general experimental setup). We found that an expansion retrieving the top-20 candidates leads to more decisive regression results across all three explored independent variables: For the volatility regressions, eight out of nine top-k models (compared to three based on S_C) yielded a significant correlation; eight (one) for the analyst forecast error regressions; and seven (none) for the analyst dispersion regressions. An explanatory hypothesis for this apparent performance difference can be found in the different distributions of S_C across the sets of similar candidates: We found that a similarity-based filtering leads to largely different numbers of added candidates given that some seed terms have more candidates with large S_C values than others; this would imply a model bias toward such terms.

For reasons outlined in Section 2.3, we were interested in how lowering k could help to retain more relevant terms and even stronger regression results. Hence, we view $k = 20$ as a maximum and explore values $\forall k \in \mathbb{N}, 1 \leq k \leq 20$ in our experiments (see Sections 4.1 and 4.2).

3.2.2 Industry-Specific Models. Apart from an industry-agnostic model, we train word embedding models according to each of the eight industry schemes (FF5 to FF49): For example, according to FF49, we train embedding models for the precious metals industry, the computer hardware industry, and 47 others. As summarized in Table 1, the allocation of training data per industry differs substantially depending on the specific scheme. As can be seen, even at the most granular level (FF49), the average number of documents per industry is still substantially higher than the one of Rekabsaz et al. [26], approximately 2,500 vs. 300. This difference is even more noticeable when selecting a scheme with a comparable granularity to theirs (11 industries): Both FF10 and FF12 assign close to 10,000 documents per industry. Therefore, we are confident to have overcome aforementioned authors' hurdle of data scarcity. We furthermore hypothesize that the relatively low skewness and standard deviation of FF12, FF48, and FF49 compared to their neighbors should also reflect in more favorable results due to a more even allocation of training data.

[13] More specifically, they propose the general threshold $S_C = 0.7$ for retaining similar candidates.

Table 1. Descriptive Statistics for the Number of Documents per Industry According to Each Industry Classification Scheme

	FF5	FF10	FF12	FF17	FF30	FF38	FF48	FF49
Mean	21,055	11,485	9,718	7,018	4,075	3,335	2,578	2,527
Std. Dev.	21,881	17,609	10,707	12,984	6,342	7,225	3,375	3,090
Skew.	1.9	3.1	2.1	3.2	3.3	4.3	2.3	2.0
Min.	967	967	742	493	105	105	105	105
Max.	63,496	63,496	39,265	53,374	31,939	40,346	14,649	12,735

3.3 Using Event Study Regressions as Extrinsic Validation

For each dictionary and document, we calculate the cumulative tf-idf of uncertain terms to gauge linguistic uncertainty. Apart from the plain uncertainty dictionary [18], we also evaluate our 180 expansions that we created as outlined in Section 3.2. These expansions include 20 industry-agnostic expansions with $k = [1, 20]$ and 160 industry-specific expansions along the eight industry schemes (FF5 to FF49) and $k = [1, 20]$. The documents are our 76,991 10-K instances (Section 3.1). The cumulative tf-idf score for a respective dictionary and document D is calculated by summing up the tf-idf scores of the dictionary's terms $w_1, w_2, \ldots, w_n$ occurring in D. For each dictionary–document combination, this procedure yields a continuous measure of linguistic uncertainty (in the following: "Uncertainty").

To compare which dictionary provides the best assessment of Uncertainty, for each dictionary and document, we follow Loughran and McDonald [19] by performing regressions of the financial variables volatility, analyst forecast error ("Error"), and analyst dispersion ("Dispersion"). Note that different to them, our main explanatory measure is Uncertainty and not readability, however. The sampling procedure and calculation of these three financial variables is outlined in the following Sections 3.3.1 and 3.3.2. The regressions adhere to the following formulae (Equations (2)–(4)):

$$\text{Volatility}_i = \alpha_i + \beta_i \cdot \text{Uncertainty}_i + \delta_i, \tag{2}$$

$$\text{Error}_i = \alpha_i + \beta_i \cdot \text{Uncertainty}_i + \delta_i, \tag{3}$$

$$\text{Dispersion}_i = \alpha_i + \beta_i \cdot \text{Uncertainty}_i + \delta_i. \tag{4}$$

In these equations, α_i is the estimated regression intercept (*bias term* in Machine Learning terminology), β_i is the estimated slope coefficient (*weight*) for the independent variable Uncertainty, and δ_i is a vector of control variables that are calculated as outlined in Section 3.3.3. The slope coefficient β denotes the number of standard deviations that the dependent variable changes for each standard deviation increase in the predictor variable; it can be interpreted as the effect strength of our linguistic uncertainty measure to a given financial uncertainty measure.

3.3.1 Volatility. We consider the filing date of a 10-K as event date after which we try to measure return fluctuation (i.e., volatility) attributable to the 10-K disclosure. We follow Loughran and McDonald [19] and calculate subsequent volatility as the RMSE of a post-filing market model [30] using trading days t_6 to t_{28} (approximately a month) relative to the 10-K filing date. To control for historic volatility, we additionally estimate a pre-filing market model using trading days t_{-252} to t_{-6} (approximately a year).

A market model is routinely estimated by regressing a respective company's returns (r_i) on the return of the overall market (r_m) in said windows. Return data are obtained from the financial

database CRSP. As proxy for r_m, we use the CRSP value-weighted index,[14] which is calculated as outlined in Reference [9, p. 7]. The market model regressions adhere to the following formula:

$$r_i = \alpha_i + \beta_i \cdot r_m. \tag{5}$$

These regressions yield intercepts (bias terms) α_i and slope coefficients (weights) β_i. We use these two variables to estimate expected returns in the given window. We further calculate the volatility, our main independent variable, as root mean square error (RMSE) of the market models. Calculating volatility in such a manner as opposed to simply using the standard deviation of returns is a common procedure [2, 19, *inter alia*] to obtain a measure of *idiosyncratic*, i.e., unsystematic *risk*. Using market model return estimates (called *expected returns*) and quantifying the differences toward actual returns yields residuals that cannot be explained through fluctuations of the overall market alone. These residuals (called *unexpected returns*) reflect gains or losses attributable to unforeseen events.

3.3.2 Analyst-based Measures. In addition to this market-based measure, using data from the financial database I/B/E/S,[15] we deploy two common measures of information uncertainty based on analyst forecasts: *analyst forecast error* and *analyst dispersion*. These measures focus on the key figure *earnings per share*, which indicates the proportion of company profit allocated to each outstanding share. Due to the lower availability of data, the sample size gets reduced to 32,799 for the analyst forecast error and to 21,166 for the analyst dispersion regressions.

We follow Loughran and McDonald's definitions of Reference [19] analyst forecast error and analyst dispersion: Analyst forecast error is calculcated as the absolute value of the standardized unexpected earnings, which are defined as (actual earnings − average expected earnings)/stock price. The *actual earnings* are the earnings per share as published in the earnings announcement. The *average expected earnings* are calculated as the mean of all earnings forecasts issued by banking analysts between the 10-K filing date and the date of the earnings announcement. Both figures are obtained from the I/B/E/S unadjusted data files. For analysts with more than one forecast reported between the 10-K filing and the earnings announcement, we retain only the forecast closest to the filing date. Finally, the variable is winsorized at the 1% level.[16] Analyst dispersion is calculated as standard deviation of analyst forecasts appearing in the forecast error estimate divided by stock price. We retain only firms with at least two analyst forecasts and again winsorize the variable at the 1% level.

3.3.3 Control Variables. Beyond these three independent variables, following Loughran and McDonald [19], we use the following set of control variables within our regressions:

- The intercepts α and the RMSE from market model regressions with trading days t_{-252} to t_{-6} as indicators of *historic performance* and *historic volatility* (see Section 3.3.1 for details).
- The *filing period abnormal return* as absolute value of the buy-and-hold return[17] in trading days t_0 to t_1 minus the buy-and-hold return of the market index.

[14] A *value-weighted index* is a market index that is obtained by weighting the returns of its components (a sample of stocks deemed representative of the overall market) by their market value. Individual market values are obtained by multiplying a stock price by the total number of outstanding shares.

[15] https://financial.thomsonreuters.com/en/products/data-analytics/company-data/ibes-estimates.html.

[16] *Winsorization* is a transformation method to handle outliers in series of continuous data. Winsorizing at a 1% limit means that all data points below the 1st percentile are set to the value of the 1st percentile and all data points above the 100 − 1 = 99th percentile are set to the value of the 99th percentile, respectively.

[17] The *buy-and-hold return* is the return of a security bought on a specific date $t - \tau$ and held for τ days up until a specific date t.

- The log-transformed *firm size* calculated as current stock price multiplied by the number of outstanding shares.
- The log-transformed *book-to-market* ratio, calculated as the book value of equity according to COMPUSTAT[18] divided by the market value of equity according to CRSP.[19] Here, we only considered firms with a positive book value and winsorized at the 1% level.
- A *NASDAQ dummy* variable set to one if the firm is listed on the NASDAQ at the time of the 10-K filing, otherwise zero.

In addition to these variables, we control for year- and industry-specific effects by adding the filing year and the assigned industry according to the respective Fama and French [8] industry scheme as one-hot-encoded categorical features. In the Error and Dispersion regressions, we additionally include the number of analyst forecasts appearing in the analyst forecast error calculation as control.

4 RESULTS AND DISCUSSION

The discussion of our results is twofold: Section 4.1 provides an in-depth error analysis exploring how our embedding models are suitable to capture uncertainty. Following up, Section 4.2 provides a purely data-driven, quantitative analysis assessing how well our models are suited to explain financial uncertainty in terms of volatility, analyst forecast error, and analyst dispersion.

4.1 Intrinsic Validation: Error Analysis of the Expansions

4.1.1 Industry-Agnostic Model. As a starting point, we retrieved the 20 closest terms to each original dictionary term within an industry-agnostic embedding model. Afterward, we asked two annotators of linguistic and financial domain knowledge[20] to co-annotate all retrieved candidates that occurred at least once in the dataset used for our regressions (2,710 terms) as either relevant or not. For this task, the IAA measured in terms of Cohen's κ [5] amounted 0.80, which can be deemed a "substantial" agreement [14]. Considering only the terms with perfect agreement, this analysis showed that 9.8% out of all candidates were indeed relevant. We thus hypothesized that lowering the relevance criterion k would retain more cosine similar and thus probably relevant terms. Yet, an in-depth analysis of the candidates revealed that certain candidate sets—despite their relatively high cosine similarity—were comprised of terms specific to a certain domain yet irrelevant for general uncertainty detection. For example, the top three candidates of the term "anomalous" are "outcrop" (S_C = 0.76), "gold-quartz" (0.76), and "silicified" (0.75). Of these terms, none could be classified as uncertain. Apparently, in specific industries (in this case: mining), certain key terms are discussed in contexts differing substantially from the general domain. In such a context, they are used to describe highly technical processes, thus diluting the overall results.

4.1.2 Industry-Specific Models. Therefore, in the next step, we explored how training industry-specific embedding models according to FF5–FF49 might mitigate said issue. It could be expected that creating an embedding of the word "anomalous" in, e.g., the pharmaceuticals industry should yield substantially different results than in the mining industry. Indeed, we noticed that an increase of industry granularity is accompanied by an increase in domain specificity of the employed

[18] http://www.crsp.com/products/research-products/crspcompustat-merged-database.

[19] The *book value* reflects the value of a firm as recorded on its balance sheet. The *market value* denotes firm value according to the stock market forces of supply and demand; it is calculated by multiplying the current stock price by the number of outstanding shares. The ratio of book over market value is commonly interpreted as degree of over- or undervaluation.

[20] One of these annotators graduated with a major in linguistics, the other with a major in finance; both have practical experience in the financial services sector. We chose these annotators to get a holistic view of the problem, which we expected to be dependent on both linguistic and financial domain knowledge.

Table 2. Top Five Similarity Candidates for the Word "Anomalous" with Their Cosine Similarity in Exemplary Industries According to the FF49 Scheme

Hardware		Lab Equipment		Precious Metals		Software		Utilities	
suspicious	0.69	overheating	0.71	pathfinder	0.79	intruder	0.62	gaming	0.63
exploits	0.67	cysts	0.70	anomalies	0.75	unfortunate	0.60	deceptive	0.62
disk-to-disk	0.63	false-positives	0.70	elevated	0.75	ever-growing	0.57	intentional	0.61
denial-of-service	0.62	out-of	0.68	arsenic	0.74	oversupply	0.56	omissions	0.59
alerts	0.62	Smallpox	0.68	trace	0.74	re-send	0.56	unethical	0.58

Table 3. Share of Relevant Terms (Left Side) and IAA (Right Side) for Varying Levels of k and Different Embedding Models

		Agn.	FF12	FF49
% rel.	k_1	16.0%	7.4%	5.2%
	k_{20}	10.0%	6.1%	3.4%

		Agn.	FF12	FF49
IAA (κ)	k_1	0.84	0.80	0.78
	k_{20}	0.81	0.77	0.70

candidates: Recurring to our previous example of the term "anomalous," Table 2 provides an overview of its similarity candidates according to some of our FF49 embedding models. As can be seen, the new candidates are recruited from a more industry-specific vocabulary, which makes them more applicable for varying domain-specific needs.

4.1.3 Manual Analysis. To quantify the beneficial effect of lower k values and industry-specific embedding models, we deepened our analysis. Since there are 297 original dictionary terms and to each term, we add up to 20 candidates according to nine different schemes spanning up to 49 different embedding models, the total number of candidate terms (ca. 60K distinct types) would be too high to make a complete manual evaluation feasible. Hence, we focused our analysis on the industry-agnostic and the FF12 and FF49 expansions, which yielded the most distinctive results in the event study regressions (see Section 4.2). According to Zipf's Law [35], the frequency of a word within a corpus is approximately inversely proportional to its rank in the frequency table; i.e., intuitively speaking, a small set of words dominates the frequency counts. In the case of the industry-agnostic model, indeed only 100 terms account for approximately 25% of the cumulative tf-idf count. FF12 and FF49 follow highly similar patterns.

Hence, we retrieved the 100 terms with the highest cumulative tf-idf at both $k = 1$ and $k = 20$ for the industry-agnostic model and all industry-specific models according to FF12 and FF49. This yielded a list of 4,820 distinct top-scoring candidate terms. Then, we let our two annotators co-annotate all of these terms regarding their relevancy for uncertainty detection; we only considered terms as relevant if both annotators agreed on a "relevant" label. Intuitively, one would expect that the share of relevant to total candidate terms should be higher for $k = 1$ (i.e., the least inclusive relevance threshold) than for $k = 20$ (most inclusive threshold). Our results show that this is the case across all models (see Table 3, left side): The share of relevant candidates drops from 16% to 10% for the industry-agnostic model, from 7.8% to 6.1% for FF12, and from 5.2% to 3.4% for FF49. In general, we can also observe that the share of relevant terms decreases consistently with increasing industry granularity. This is explainable due to an increasing number of industries leading to both a higher specificity of the candidates (see Section 4.1.2) and a larger number of candidates. Hence, it can be expected that an expansion based on, e.g., FF49 contains more noise than an industry-agnostic expansion.

Interestingly, we can observe that this trend is also accompanied by a consistent decrease of IAA in terms of Cohen's κ for models with a higher industry granularity and increased levels of k (see Table 3, right side). This analysis suggests that while candidate specificity increases, so do both the subjectivity and the difficulty of the task due to more required industry-specific knowledge (reconsider e.g., Table 2). In summary, our expectation that a lower level of k leads to a higher share of meaningful terms could be confirmed across all analyzed models. However, aforementioned numbers can only be taken as indications, since a complete manual evaluation of all 60K candidate terms would be unfeasible.

4.1.4 Task-Specific Shortcomings of Word2Vec. Our analysis further suggests that word2vec might have limitations for the given task. Slightly simplified, word2vec deems a word A as similar to a word B if (1) A could be used interchangeably for B or (2) A appears in a similar context as B. Especially the latter case is problematic for our endeavor. To a large degree, the *uncertain* sentiment dictionary consists of word classes such as modal verbs ("may") and adverbials of degree ("somewhat") or probability ("maybe"). Within a sentence, these rather functional word classes, also known as *hedges*, evoke semantic slots. That is, instead of being particularly meaningful on their own, they only limit the truth value of other entities and actions [13] by assigning a probability through fuzzy quantification—and a quantifier is meaningless without the expression that it quantifies. Hence, these words usually share a context with objects, people, places or verbs. Not surprisingly, therefore, seemingly irrelevant similarity candidates such as names of brands ("Monsanto") and organizations ("SEC"), units of quantity ("ft") or even other functional verbs ("be") appear within our expansions.

However, this previously discussed tendency of word2vec to suggest co-contextual words as similarity candidates could also be useful for future research: Another way to improve the dictionary would be removing instead of adding terms. Our analyses suggest that some of the dictionary terms might indeed be misspecified. For example, the original dictionary terms with the root "random" predominantly neighbor candidates such as "sample," "trial," or "researcher" in our embedding models. This suggests that such terms are mostly used in the context of experimental trials—something that neither fits the dictionary creators' goal of capturing an imprecise nor a risk-related choice of words.

A conventional solution for this issue would be removing said terms either (1) based on a knowledge-poor approach relying on part-of-speech tagging or named entity recognition or (2) based on a knowledge-rich, dictionary-based filtering. As a matter of fact, in an earlier step of our research, we tried how the filtering of proper nouns (identified through part-of-speech tagging) from the candidate terms might influence the results. However, we noticed that doing so actually worsened the results of the regression task (Section 4.2) substantially, while not making place for more relevant candidates as evaluated through our qualitative analysis—hence, we dismissed this idea.

4.1.5 "To Be" or Not "to Be"? Concerning Stopwords. In addition, we considered removing stopwords by using Porter's [25] relatively comprehensive list of 153 functional terms such as "at," "should," or "be." However, we noticed that for the industry-agnostic model, even at $k = 20$ (i.e., the most inclusive level), only 16 of these terms were among the approximately 3,500 candidates. By definition, such stopwords are usually also among the most frequently appearing, which is why they usually get assigned an inverse document frequency of approximately zero. Indeed, in our case, out of previously mentioned 16 terms, 14 appear in the lower 10 percentile of all average tf-idf scores. The remaining two are the pronouns "yours" and "himself" (lower 30 and 40 percentile, respectively). Since the industry-specific expansions follow highly similar patterns, the effect of such functional terms is barely noticeable for our endeavor. Furthermore, modal words

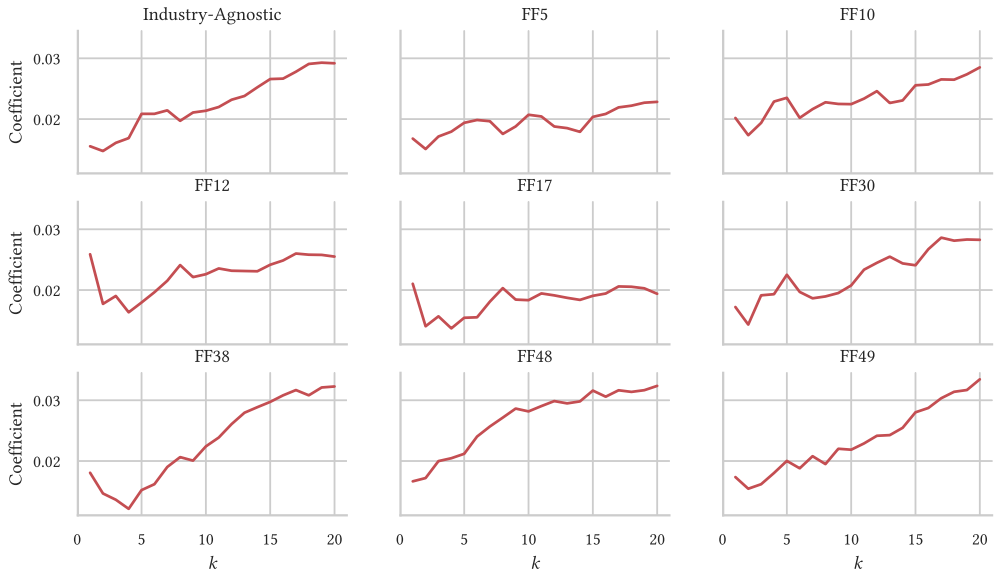

Fig. 2. Influence of industry scheme and number of added candidates k on slope coefficients β in regressions of volatility on uncertainty. Coefficients are standardized with a mean of zero and a standard deviation of one. All regressions include 76,991 observations.

such as "should" or "can," which are considered to be stopwords according to Porter's [25] list, are indeed of high relevance for our analyses. This shows that the distinction of stopwords vs. non-stopwords, at least for the purpose of uncertainty detection in the financial domain, is somewhat arbitrary. For these reasons, we decided not to filter out stopwords. As a novel idea of avoiding irrelevantly added candidates due to the functional word class of the original terms, we propose to introduce filtering methods based on grammatical properties in future work. One way to do so would be giving priority to similarity candidates sharing the same word class as the original term.

4.2 Extrinsic Validation: Regressions of Financial Uncertainty Measures

How useful are our expansions in a real-world application? In this section, we analyze the beneficial effect of inducing industry-specific jargon in regressions of the financial uncertainty measures volatility, analyst forecast error, and analyst dispersion. The results of these event study regressions are summarized in Figures 2–4 and additionally provided in greater detail in Appendix A.

4.2.1 Volatility Regressions. Consistent with previous research, linguistic uncertainty and volatility are positively related (see Figure 2); i.e., an annual report with a higher uncertainty of its content is usually also followed by a higher stock return volatility than its peers. Furthermore, for all volatility regressions, the control variables behave similarly (see Appendix A, Table 4): Firms with a high pre-filing performance and market value are subject to less post-filing volatility. Firms with a low book-to-market ratio, with a higher pre-filing volatility, with larger unexpected returns around the filing date, and NASDAQ-listed firms experience a higher volatility.

In most cases, the relationship between uncertainty and volatility is considerably stronger using the expanded dictionaries than using the plain dictionary (coefficient β of 0.014, significant at the 5% level). Among all industry schemes, FF49 yields the most decisive results (see Figure 2). The highest coefficient (0.033) is significant at the 0.1% level and achieved for $k = 20$. This value is 14% higher than the leading industry-agnostic model ($\beta = 0.029$, significant at the 1% level) and about

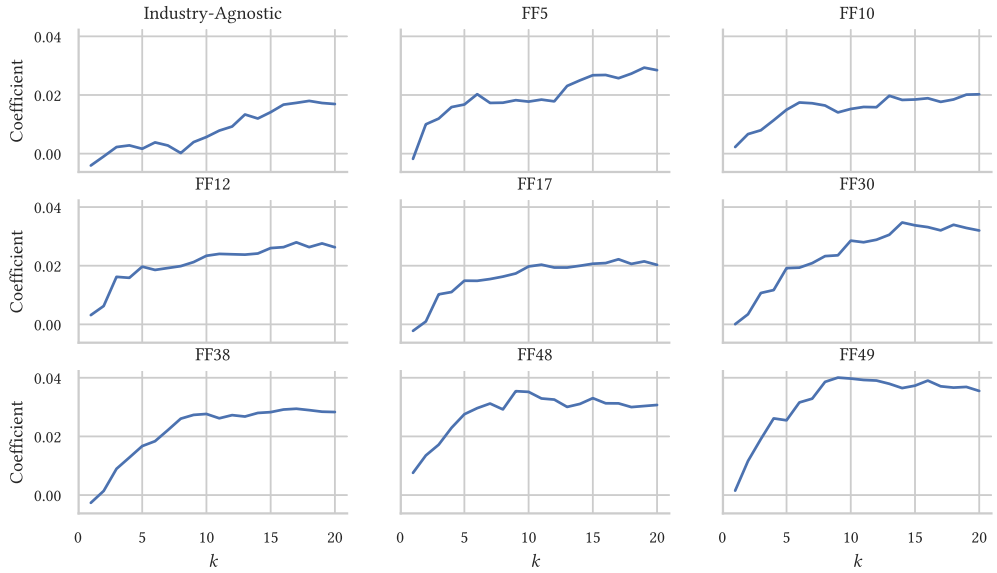

Fig. 3. Influence of industry scheme and number of added candidates k on slope coefficients β in regressions of analyst forecast error on uncertainty. Coefficients are standardized with a mean of zero and a standard deviation of one. All regressions include 32,799 observations.

135% higher than the one using the plain unexpanded dictionary (see Appendix A, Table 4). While there does not seem to be a clear relationship between the height of the coefficient and increased industry granularity, the strength of the association between uncertainty and volatility generally seems to increase with higher values of k.

4.2.2 Regressions of Analyst-based Measures. Figures 3 and 4 indicate that linguistic uncertainty and analyst uncertainty in terms of forecast error and dispersion are positively related. This means that uncertain 10-Ks are associated with both more erroneous and more dispersed analyst forecasts. For both sets of regressions, the control variables follow similar patterns again (see Appendix A, Tables 5 and 6): Firms with a higher analyst uncertainty tend to be smaller, not listed on the NASDAQ, and are subject to both lower performance, larger volatility before the filing, and a higher book-to-market ratio. The only control variable behaving differently is the number of analysts, which is negatively related with forecast error and positively related with dispersion.

For the regressions of analyst forecast error (see Figure 3), again FF49 obtains the highest coefficient (0.040). This value, which is achieved through $k = 9$, is significant at the 0.1% level, considerably higher than the insignificant value of the plain dictionary (−0.004), and 2.2 times as high as the leading industry-agnostic approach ($\beta = 0.018$, insignificant). In contrast to Figure 2, the beneficial effect of a higher and thus more inclusive relevance criterion k seems to be saturated around quadrant two ($5 \leq k \leq 10$) or three ($10 \leq k \leq 15$).

For the regressions of analyst dispersion (see Figure 4), FF12 obtains the highest coefficient (0.039) for $k = 17$. This value is significant at the 1% level and again considerably higher than the value of the plain dictionary, which is insignificant and indistinguishable from zero. It furthermore is 8% higher than the leading industry-agnostic model ($\beta = 0.036$, significant at the 5% level). While FF5–FF38 generally seem to benefit from an increased k (with FF12–FF38 experiencing slight downturns in quadrant four), FF48 and FF49 again experience visible performance drops in the last quadrants.

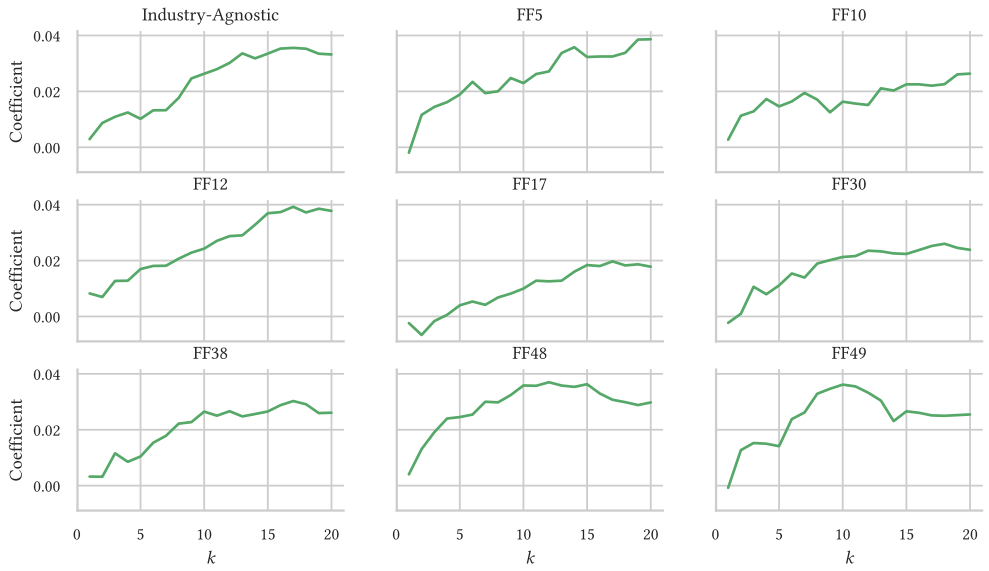

Fig. 4. Influence of industry scheme and number of added candidates k on slope coefficients β in regressions of analyst dispersion on uncertainty. Coefficients are standardized with a mean of zero and a standard deviation of one. All regressions include 21,166 observations.

4.2.3 Discussion. We proceed with discussing the economic magnitude of the association between linguistic uncertainty and financial uncertainty measures. The regression results imply that an increase of one standard deviation in uncertainty (according to the optimal industry-specific expansions) leads to an increase of 3.3% to 4% of financial uncertainty's standard deviation (see A, Tables 4–6). While these coefficients might appear small, they are well in line with recent research: For example, Bonsall et al. [2] find that their proposed plain English measure explains 3.5% of subsequent volatility's standard deviation. Furthermore, in their study on textual analysis in accounting and finance, Loughran and McDonald conclude that the "economic magnitude of the soft information [i.e., text] is somewhat limited" [20, p. 1202].

In summary, all regressions of financial uncertainty on linguistic uncertainty benefit substantially from incorporating industry-specific vocabulary. This beneficial effect is most profound for the regressions of analyst forecast error. Interestingly, as already hypothesized in Section 3.1, FF12, FF48, and FF49 are consistently among the most decisively performing models, which can probably at least partly be attributed to the relatively even allocation of training data per industry.

5 CONCLUSION

In this article, we addressed the automatic detection of linguistic uncertainty in financial disclosures by extracting dictionary-based features from a large corpus of 10-Ks. We created a set of automatically expanded dictionaries by training industry-specific word embedding models. Furthermore, we provided an in-depth error analysis revealing how inducing industry-specific vocabulary leads to more specific candidate terms. As expected, our results indicate that this increase in specificity is also accompanied by more noise being introduced to the expansions. Finally, we have shown how dictionary expansions incorporating industry-specific jargon lead to more decisive regression results for the financial uncertainty measures (volatility, analyst forecast error, and analyst dispersion) than both the plain dictionary as well as an industry-agnostic expansion.

Given the novelty of the area, future work could refine the methodology in several ways. First, relevance of the similarity candidates could be increased by prioritizing candidates sharing word

classes with the original dictionary terms. Furthermore, since prior and the current research focus on unigrams, the dictionary as well as the word embedding models could be expanded to cover *n*-grams. Moreover, the problem could be approached from the other side by removing existing instead off adding new dictionary terms: Our analysis suggests that dictionary terms like "random" might indeed be misspecified. Last, as our error analysis points out, word2vec's limitations for the given task could be overcome by using a different embedding model—possibly refining the context-based restrictions with topical criteria such as in lda2vec [23].

A DETAILED REGRESSION RESULTS

Table 4. Detailed Results for the Regressions of Volatility

	Volatility	Volatility	Volatility
Independent Variable:			
Plain dictionary	0.014*		
	(2.419)		
Optimal industry-agnostic expansion		0.029**	
		(3.121)	
Optimal industry-specific expansion			0.033***
			(5.168)
Control Variables:			
Historical performance	−0.085**	−0.085**	−0.084**
	(−3.501)	(−3.476)	(−3.471)
Historical volatility	0.465***	0.464***	0.462***
	(12.170)	(12.023)	(12.121)
Filing date abnormal return	0.100***	0.100***	0.100***
	(12.655)	(12.609)	(12.551)
Firm size	−0.107***	−0.110***	−0.110***
	(−5.830)	(−5.889)	(−6.102)
Book-to-market	−0.064**	−0.065***	−0.064***
	(−3.616)	(−3.691)	(−3.653)
NASDAQ dummy	0.054**	0.054**	0.055**
	(3.444)	(3.433)	(3.490)
Adjusted R^2	47.93%	47.97%	47.98%
N	76,991	76,991	76,991

All regressions include intercepts, calendar year dummies, and Fama and French industry dummies. Coefficients are standardized with a mean of zero and a standard deviation of one. *t*-statistics are in parentheses with standard errors clustered by year and industry.
$^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

Table 5. Detailed Results for the Regressions of Analyst Forecast Error

	Error	Error	Error
Independent Variable:			
Plain dictionary	−0.003		
	(−0.352)		
Optimal industry-agnostic expansion		0.018	
		(1.996)	
Optimal industry-specific expansion			0.040**
			(3.052)

(Continued)

Table 5. Continued

	Error	Error	Error
Control Variables:			
Historical performance	−0.085***	−0.124***	−0.123***
	(−5.302)	(−5.285)	(−5.290)
Historical volatility	0.275***	0.275***	0.272***
	(5.547)	(5.487)	(5.503)
Filing date abnormal return	0.057***	0.057**	0.057***
	(4.855)	(4.848)	(4.841)
Firm size	−0.169***	−0.169***	−0.171***
	(−8.476)	(−8.479)	(−8.645)
Book-to-market	0.132***	0.132***	0.131***
	(6.524)	(6.549)	(6.589)
NASDAQ dummy	−0.073***	−0.073***	−0.073***
	(−4.733)	(−4.795)	(−4.826)
Number of analysts	−0.037**	−0.038**	−0.037**
	(−3.369)	(−3.333)	(−3.425)
Adjusted R^2	19.52%	19.53%	19.61%
N	32,799	32,799	32,799

All regressions include intercepts, calendar year dummies, and Fama and French industry dummies. Coefficients are standardized with a mean of zero and a standard deviation of one. *t*-statistics are in parentheses with standard errors clustered by year and industry.
$^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

Table 6. Detailed Results for the Regressions of Analyst Dispersion

	Dispersion	Dispersion	Dispersion
Independent Variable:			
Plain dictionary	0.000		
	(0.022)		
Optimal industry-agnostic expansion		0.036*	
		(1.996)	
Optimal industry-specific expansion			0.039**
			(4.017)
Control Variables:			
Historical performance	−0.139***	−0.140***	−0.136***
	(−5.660)	(−5.663)	(−6.965)
Historical volatility	0.305***	0.303***	0.297***
	(5.630)	(5.584)	(6.693)
Filing date abnormal return	0.054***	0.054***	0.055***
	(4.057)	(4.047)	(5.246)
Firm size	−0.136***	−0.138***	−0.135***
	(−6.155)	(−6.249)	(−9.539)
Book-to-market	0.138***	0.136***	0.143***
	(7.327)	(7.354)	(5.616)
NASDAQ dummy	−0.079***	−0.080***	−0.094**
	(−4.026)	(−4.082)	(−3.884)
Number of analysts	0.024	0.022	0.030
	(1.104)	(1.010)	(1.645)
Adjusted R^2	21.23%	21.29%	19.74%
N	21,166	21,166	21,166

All regressions include intercepts, calendar year dummies, and Fama and French industry dummies. Coefficients are standardized with a mean of zero and a standard deviation of one. *t*-statistics are in parentheses with standard errors clustered by year and industry.
$^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful comments.

REFERENCES

[1] Tim Bollerslev. 1986. Generalized autoregressive conditional heteroskedasticity. *J. Econometr.* 3, 31 (1986), 307–327.
[2] Samuel B. Bonsall, Andrew J. Leone, Brian P. Miller, and Kristina Rennekamp. 2017. A plain english measure of financial reporting readability. *J. Account. Econ.* 63, 2 (2017), 329–357.
[3] Brian J. Bushee, Ian D. Gow, and Daniel J. Taylor. 2018. Linguistic complexity in firm disclosures: Obfuscation or information? *J. Account. Res.* 56, 1 (2018), 85–121.
[4] Louis K. C. Chan, Josef Lakonishok, and Bhaskaran Swaminathan. 2007. Industry classifications and return comovement. *Financ. Anal. J.* 63, 6 (2007), 56–70.
[5] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 1 (1960), 41–48.
[6] Travis Dyer, Mark Lang, and Lorien Stice-Lawrence. 2017. The evolution of 10-k textual disclosure: Evidence from latent dirichlet allocation. *J. Account. Econ.* 64, 2 (2017), 221–245.
[7] Eugene F. Fama and Kenneth R. French. 1993. Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* 33, 1 (1993), 3–56.
[8] Eugene F. Fama and Kenneth R. French. 1997. Industry costs of equity. *J. Financ. Econ.* 43, 2 (1997), 153–193.
[9] Center for Research in Security Prices (CRSP). 2018. CRSP Indexes: Methodology Guide.
[10] Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the Conference of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL–IJCNLP'15).* Association for Computational Linguistics, Stroudsburg, PA, 63–68.
[11] Robert Gunning. 1952. *The Technique of Clear Writing.* McGraw–Hill, New York, NY.
[12] Jack H. Hiller. 1971. Verbal response indicators of conceptual vagueness. *Am. Educ. Res. J.* 8, 1 (1971), 151–161.
[13] George Lakoff. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *J. Philos. Logic* 2, 4 (1973), 458–508.
[14] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.
[15] Reuven Lehavy, Feng Li, and Kenneth Merkley. 2011. The effect of annual report readability on analyst following and the properties of their earnings forecasts. *Account. Rev.* 86, 3 (2011), 1087–1115.
[16] Feng Li. 2008. Annual report readability, current earnings, and earnings persistence. *J. Account. Econ.* 45, 2–3 (2008), 221–247.
[17] Jun Li and Xiaofei Zhao. 2015. Complexity and Information Content of Financial Disclosures: Evidence from Evolution of Uncertainty Following 10-K Filings. (2015).
[18] Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Financ.* 66, 1 (2011), 35–65.
[19] Tim Loughran and Bill McDonald. 2014. Measuring readability in financial disclosures. *J. Financ.* 69, 4 (2014), 1643–1671.
[20] Tim Loughran and Bill McDonald. 2016. Textual analysis in accounting and finance: A survey. *J. Account. Res.* 54, 4 (2016), 1187–1230.
[21] Adam Meirowitz. 2005. Informational party primaries and strategic ambiguity. *J. Theor. Pol.* 17, 1 (2005), 107–136.
[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Arxiv E-prints 1301.3781* (2013).
[23] Christoph E. Moody. 2016. Mixing dirichlet topic models and word embeddings to make lda2vec. *Arxiv E-prints 1605.02019* (2016).
[24] Gustavo Henrique Paetzold and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the AAAI Conference on Artificial Intelligence.* Association for the Advancements of Artificial Intelligence, Palo Alto, CA, 3761–3767.
[25] Martin F. Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (1980), 130–137.
[26] Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Allan Hanbury, Alexander Duer, and Linda Anderson. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based IR models. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL'17).* Association for Computational Linguistics, Stroudsburg, PA, 1712–1721.
[27] Navid Rekabsaz, Mihai Lupu, and Allan Hanbury. 2016. Uncertainty in neural network word embedding: Exploration of threshold for similarity. In *Proceedings of the Neu-IR Workshop at the ACM SIGIR Conference on Research and Development in Information Retrieval.* Association for Computing Machinery, NY.

[28] Navid Rekabsaz, Mihai Lupu, and Allan Hanbury. 2017. Exploration of a threshold for similarity based on uncertainty in word embedding. In *Proceedings of the European Conference on Information Retrieval (ECIR)*. Springer, Cham, Switzerland.
[29] Jonathan L. Rogers. 2008. Disclosure quality and management trading incentives. *J. Account. Res.* 46, 5 (2008), 1265–1296.
[30] William F. Sharpe. 1963. A simplified model for portfolio analysis. *Manage. Sci.* 9, 2 (1963), 277–293.
[31] Christoph Kilian Theil, Sanja Štajner, and Heiner Stuckenschmidt. 2018. Word embeddings-based uncertainty detection in financial disclosures. In *Proceedings of the ACL Workshop on Economics and Natural Language Processing (ECONLP'18)*. Association for Computational Linguistics, Stroudsburg, PA, 32–37.
[32] Ming-Feng Tsai and Chuan-Ju Wang. 2014. Financial keyword expansion via continuous word vector representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. Association for Computational Linguistics, Stroudsburg, PA, 1453–1458.
[33] Ming-Feng Tsai, Chuan-Ju Wang, and Po-Chuan Chien. 2016. Discovering finance keywords via continuous-space language models. *ACM Trans. Manage. Inf. Syst.* 7, 3 (2016), 1–17.
[34] Maximilian A. M. Vermorken. 2011. GICS or ICB, How different is similar? *J. Asset Manage.* 12, 1 (2011), 30–44.
[35] George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Boston, MA (USA).

Received August 2018; revised March 2019; accepted June 2019