Word Embeddings-Based Uncertainty Detection in Financial Disclosures

Christoph Kilian Theil, **Sanja Štajner** and **Heiner Stuckenschmidt**
Data and Web Science Group
University of Mannheim, Germany
{christoph, sanja, heiner}@informatik.uni-mannheim.de

Abstract

In this paper, we use NLP techniques to detect linguistic uncertainty in financial disclosures. Leveraging general-domain and domain-specific word embedding models, we automatically expand an existing dictionary of uncertainty triggers. We furthermore examine how an expert filtering affects the quality of such an expansion. We show that the dictionary expansions significantly improve regressions on stock return volatility. Lastly, we prove that the expansions significantly boost the automatic detection of uncertain sentences.

1 Introduction

Despite its real world impact in tasks like volatility prediction, the automatic detection of linguistic uncertainty has been left relatively untouched in finance. Motivated by this research gap, we created the first classifier capable of detecting uncertain sentences in so-called *10-Ks*. These annual reports are required by the U.S. Securities and Exchange Commission (SEC) and give a comprehensive overview of a company's business activities. We selected this disclosure type since it has to be filed by all public companies in the U.S., thus ensuring a large sample size. Furthermore, it is the only disclosure type for which a tailored dictionary resource exists.

1.1 Loughran & McDonald's Dictionary

As basis for our experiments, we took an existing financial domain dictionary containing 297 uncertain terms. This dictionary has been shown to possess explanatory power of future stock return volatility (Loughran and McDonald, 2011) and is the only of its kind specifically designed for 10-Ks. Its creators developed it "with emphasis on the general notion of imprecision rather than exclusively focusing on risk" (Loughran and McDonald, 2011, p. 45). As this quote indicates, on one hand, the dictionary contains terms marking imprecision (e.g. "could", "may", "probably", "somewhat"). On the other hand, it contains terms referring to real-worldly risk and uncertainty (e.g. "anomaly", "risk", "uncertainty", "volatility").

1.2 Contributions

We automatically expanded Loughran and McDonald's (2011) uncertainty dictionary by adding semantically close candidate terms according to word embeddings. Apart from training our own domain-specific embedding model, we compared such an expansion to one using a general-domain embedding model. Moreover, we investigated whether manual filtering of candidate terms by a domain expert can further improve the results. We evaluated the quality of our expansions in both a set of regressions on stock return volatility and a binary sentence classification task by posing two research questions:

- **RQ1** How do a general-domain and a domain-specific expansion compare?
- **RQ2** How do an automatic and a semi-automatic, expert-filtered expansion compare?

We show that our unfiltered domain-specific expansion significantly increases the explanatory power of regressions on stock return volatility over the plain dictionary. We furthermore introduce a dataset of annual reports newly annotated for this study and train a binary classifier distinguishing uncertain from certain sentences. Again, the domain-specific expansion significantly improves the classification performance over the plain dictionary. In this case, however, the expert-filtering

Proceedings of the First Workshop on Economics and Natural Language Processing, pages 32–37
Melbourne, Australia, July 20, 2018.

provides a small performance increase over the fully automatic expansion.

2 Related Work

Loughran and McDonald (2011) introduced financial dictionaries spanning the categories of *positive*, *negative*, *litigious*, *strong modal*, *weak modal*, and—most important for us—*uncertain* words. Perhaps not surprisingly, they find that the cumulative tf–idf of *uncertain* terms in a set of 10-Ks shares a positive and highly significant relation with future stock return volatility. To quantify the improvement of our new expansions over this dictionary, we use a regression setup similar to their subsequent paper (Loughran and McDonald, 2014).

Tsai and Wang (2014) automatically expanded said dictionaries by training word embeddings and adding the 20 most cosine similar terms to each original dictionary term. Using a dataset of 10-Ks, they show that this expansion improves a prediction of future stock return volatility. In contrast to them, we provide a systematic analysis how a domain-specific vs. a general-domain (RQ1) and an automatic vs. a semi-automatic expansion (RQ2) perform in a set of regressions. Furthermore, for the first time in the community, we perform a binary sentence classification task on 10-Ks to assess directly whether our models are indeed suitable to detect linguistic uncertainty.

Theil et al. (2017) created the first classifier capable to detect uncertain sentences in the financial domain. Yet, they sample their sentences from earnings call transcripts, a largely different disclosure type than 10-Ks. Apart from typical characteristics of spoken language such as less structure and more spontaneity, these disclosures are voluntary and thus usually less available. As previous studies have hinted that analyzing the language of 10-Ks can help to explain uncertainty of the information environment (Loughran and McDonald, 2011, 2014), we were further motivated to create the first sentence classifier for 10-Ks.

3 Data

We downloaded Loughran and McDonald's (2011) dictionary[1] of 297 financial uncertainty triggers such as "may", "probably", or "volatility". From now on, we refer to this dictionary as *Unc*. We further downloaded all 220,565 10-Ks during 1994 to 2015 from the SEC's database EDGAR[2]. We removed duplicates and filings shorter than 250 words, thus leaving 203,321 files. We divided this set into three non-overlapping subsets: First, using word2vec (Mikolov et al., 2013) with standard parameters, we deployed 124,830 10-Ks (approximately 2.3 billion words) to train a domain-specific embedding model. As benchmark, we also retrieved Google's generic word2vec embedding model,[3] which was trained on approximately 100 billion words from the Google News dataset.

Second, we used 76,991 10-Ks in our regressions. For each instance, we retrieved stock pricing data from the databases CRSP[4] and CRSP/Compustat Merged. To facilitate replication, our data screening and parsing procedures are described in greater detail in our Online Appendix.[5] It further contains all textual and financial data needed to replicate our regressions.

Third, we used a random sample of 1,500 10-Ks for the classification task. Out of these, we randomly sampled 100 sentences and let two annotators of financial and linguistic knowledge coannotate them as either *certain* or *uncertain*. The guidelines which we gave to our annotators can be found in the Online Appendix.

It has to be noted that the task of evaluating uncertainty as an inherently subjective semantic concept—especially in such a specialized domain as finance—is of particular intricacy. First, consider the following sentence, which both annotators labeled *uncertain*:

Example 3.1. "These factors raise substantial doubt regarding the Company's ability to continue as a going concern."

In contrast, consider the following sentence, on which the annotators disagreed; words and phrases considered to be uncertainty triggers by the annotator proposing an *uncertain* label are underlined:

Example 3.2. "Fidelity is subject to interest rate risk to the degree that its interest-bearing liabilities, primarily deposits with short and medium term maturities, mature or reprice at different rates than its interest-earning assets."

This sentence references "risk" and contains

[1] https://sraf.nd.edu/textual-analysis/resources

[2] https://www.sec.gov/edgar.shtml

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

[4] http://www.crsp.com

[5] http://dws.informatik.uni-mannheim.de/en/people/researchers/christoph-kilian-theil/

additional imprecisions, which speaks in favor of an *uncertain* label. Yet, the referenced risk is nonopaque and said imprecisions could be attributed to legal requirements as inherent to any regulated corporate disclosure; hence, a case could also be made for an *certain* label.

Nevertheless, the IAA measured as κ (Cohen, 1960) was 0.73, which can be considered "substantial" (Landis and Koch, 1977). Notably, Ganter and Strube (2009) report an even lower pairwise IAA with $0.45 \leq \kappa \leq 0.80, \bar{x}_\kappa = 0.56$ for an annotation of Wikipedia sentences as *certain* or *uncertain*. Despite making use of highly trained domain experts, Štajner et al. (2017) also obtained a lower IAA with $0.47 \leq \kappa \leq 0.70, \bar{x}_\kappa = 0.61$ for a comparable annotation task. They sampled their sentences from transcribed debates held by the U.S. central bank's monetary policy committee (FOMC).

Given our comparably high IAA, we were confident of our annotation quality and let the first annotator annotate an additional 900 sentences, thus forming our newly created dataset REPORTS. Out of its 1,000 sentences, 870 were labeled *certain* and 130 were labeled *uncertain*. This new dataset can also be found in our Online Appendix as useful resource for others to advance the field.

4 Methodology

4.1 Expanding the Dictionary

To answer RQ1, we first determined the 20 most cosine similar terms according to the generic embedding model for each of the 297 terms of *Unc*. We chose 20 as the number of added terms since this is the value suggested by Tsai and Wang (2014). After lowercasing, we removed 28 anomalous tokens (e.g. "##.million"), 1,657 *n*-grams, and 2,139 duplicates. We excluded *n*-grams, since *Unc* contains only unigrams and we wanted to keep its expansions comparable. We added the remaining 2,036 terms to *Unc* and thus created *UncGen* with 2,333 terms.

For our domain-specific model, we derived a list of 5,820 candidate terms and removed 1,947 duplicates. We did not lowercase in this case, as this was already part of our preprocessing. We again added the remaining 3,873 terms to *Unc* and thus created *UncSpec* with 4,170 terms. Remarkably, *UncGen* and *UncSpec* share an overlap of 458 (23% and 12%) of the newly added terms, which indicates that they employ a largely different vocabulary. An exemplary overview of added terms according to both models can be found in our Online Appendix.

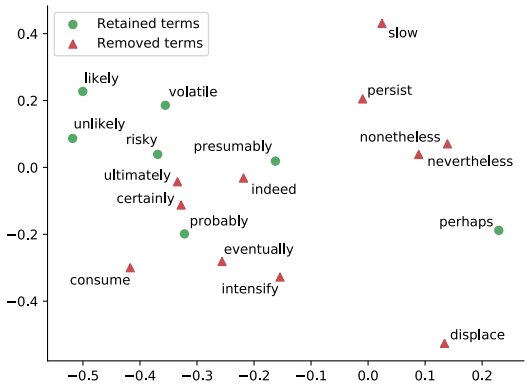

Figure 1: Candidate terms for "probably" in our domain-specific embedding model. Terms retained/removed by the annotator are marked by dots/triangles. Dimensionality is reduced through t-SNE (van der Maaten and Hinton, 2008).

We found that antonyms—despite their opposite meaning—were frequently embedded in similar semantic spaces. Coalescing relations of synonymy and antonymy is a well-known and often undesired property of distributional models (Mohammed et al., 2008). Hence, it can be explained why both *UncGen* as well as *UncSpec* contain the token "certainly" as cosine similar term to "probably" (similarity of 0.68 and 0.45, respectively). In addition, other irrelevant terms (e.g. "event", "significance") appeared frequently in close proximity to uncertain terms.

This motivated us to explore how manual filtering could improve the expanded dictionaries (RQ2). Therefore, we let an annotator of both financial and linguistic domain knowledge evaluate and remove such terms he deemed inappropriate to cover uncertainty. Figure 1 provides an exemplary visualization of this procedure. Thus, we created the dictionaries $UncGen_{exp}$ with 538 and $UncSpec_{exp}$ with 475 terms. Notably, the filtering caused the vocabulary of both lists to converge, as they now shared an overlap of 241 (45% and 51%) of the added terms. Finally, we created the combinations *UncGen*+*UncSpec* and $UncGen_{exp}$+$UncSpec_{exp}$. An overview of all dictionaries can be found in Table 1.

Table 1: Number of uncertainty triggers per dictionary.

Dictionary	# of Triggers
Unc	297
Automatic Expansions:	
UncGen	2,333
UncSpec	4,170
UncGen+UncSpec	5,748
Expert-Aided Expansions:	
$UncGen_{exp}$	538
$UncSpec_{exp}$	475
$UncGen_{exp}$+$UncSpec_{exp}$	652

Table 2: Results of the regression on volatility. Standard errors are clustered by year and industry. Coefficients are standardized with a mean of zero and a standard deviation of one.

Dictionary	Coefficient	t	R^2
Unc	0.016^{*}	2.28	47.91%
UncGen	0.014^{*}	2.20	47.90%
$UncGen_{exp}$	0.017^{*}	2.56	47.91%
UncSpec	0.034^{***}	3.90	47.96%
$UncSpec_{exp}$	0.020^{*}	2.68	47.91%
UncGen+UncSpec	0.032^{***}	3.67	47.95%
$UncGen_{exp}$+$UncSpec_{exp}$	0.017^{*}	2.59	47.91%

$^{*}p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$

4.2 Regressing Uncertainty on Volatility

To assess the real world impact of our problem, we performed event studies measuring the association of linguistic uncertainty in our set of 76,991 10-Ks with *volatility*, a common measure of financial uncertainty. To be comparable with previous work (Loughran and McDonald, 2011, 2014), we measure volatility as the deviation between the expected and the actual returns after the report's filing date in terms of root mean square error (RMSE). We calculate the expected returns estimating market models (Sharpe, 1963) using trading days $[6, 28]$ relative to the filing date.

In addition, following Loughran and McDonald (2014), we used an extensive set of control variables: the intercepts α and the RMSE from market models using trading days $[-252, -6]$ as indicators of *historic performance* and *historic volatility*. The *filing period abnormal return* as absolute value of the buy-and-hold return in trading days $[0, 1]$ minus the buy-and-hold return of the market index. The *firm size* as current stock price multiplied by the number of outstanding shares. The *book-to-market* ratio, a valuation measure, calculated as the book value of equity divided by the market value of equity. Here, we only considered firms with a positive book value and winsorized at the 1% level. Lastly, we used 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.

We calculated the cumulative tf–idf of all uncertain terms according to the dictionary *Unc* and its six expansions (see Table 1). Then, we conducted seven regressions with each containing uncertainty gauged via the respective dictionary as main independent variable, the control variables, and post-filing volatility as dependent variable. This setup allows us to quantify financial uncertainty likely induced by the filing event. All regressions include an intercept, calendar year dummies, and Fama and French (1997) 48-industry dummies for time- and industry-fixed effects. For brevity, we only report the key statistics for our main independent variables—a more detailed overview with all control variables can be found in our Online Appendix.

4.3 Creating a Classifier of Uncertainty

We used the seven dictionaries (Table 1) as feature sets in the binary classification task. As feature representation, we tried both relative term frequency (tf) as well as term frequency–inverse document frequency (tf–idf). Next, using Weka Experimenter (Hall et al., 2009), we applied six machine algorithms in a 10-fold cross-validation setup with ten repetitions: Logistic Regression (le Cessie and van Houwelingen, 1992); SMO, the Weka implementation of Support Vector Machine (Platt, 1999); JRip, the Weka implementation of RIPPER (Cohen, 1995); J48, the Weka implementation of C4.5 decision tree (Quinlan, 1993); Random Forest (Breiman, 2001); and a Convolutional Neural Network (Amtén, 2014). As a JRip classifier outperformed the other five, we only report its performance—for the same reason, we only report the results for tf–idf weighting. The full results according to all algorithms and both feature representations can be found in our Online Appendix.

5 Results and Discussion

5.1 Regression

The results of our regressions are depicted in Table 2. In all regressions, uncertainty and post-filing volatility have a positive statistical relation. This relation is significant ($p \leq 0.05$) for

Table 3: Results of the classification task for the *uncertain* class of REPORTS. The best results are boldfaced and significant performance increases ($\alpha = 0.05$) over *Unc* are marked with asterisks.

Dictionary	P	R	F	A
Unc	0.65	0.49	0.54	89.66%
UncGen	0.66	0.51	0.56	89.86%
$UncGen_{exp}$	0.67	0.52	0.56	90.05%
UncSpec	**0.69***	0.54*	0.58*	90.31%
$\boldsymbol{UncSpec_{exp}}$	0.67	**0.56***	**0.59***	**90.37%**
UncGen+UncSpec	**0.69***	0.52	0.57	90.30%
$UncGen_{exp}$+$UncSpec_{exp}$	0.66	0.54	0.58	90.07%
Majority Class (certain)	0.00	0.00	0.00	87.00%

Unc, *UncGen*, $UncGen_{exp}$, $UncSpec_{exp}$, as well as $UncGen_{exp}$+$UncSpec_{exp}$, and highly significant ($p \leq 0.001$) for *UncSpec* and *UncGen+UncSpec*.

The strength of this association is also the highest for both *UncSpec* and *UncGen+UncSpec* (0.034 and 0.032), twice as high than that of *Unc* (0.016). This shows that these expansions have a decisively higher explanatory power of volatility. Furthermore, concerning RQ1, the regressions seem to benefit most from a specific instead of a generic dictionary expansion. Additionally, with regard to RQ2, the expert filtering does not improve the results—in some cases, it even worsens them. As shown in Table 1, our expert annotator retained a relatively small subset of the candidate terms (23% of *UncGen* and 11% of *UncSpec*). Such a rigid filtering causes a smaller coverage of the expansions and furthermore carries the danger of false negative errors. We hypothesize that the effect of erroneously added terms is already mitigated through tf–idf weighting, thus rendering manual work unnecessary.

Above discussed coefficients might appear small, as a one standard deviation increase of *UncSpec* explains only a 3.4% of such an increase in volatility. However, this is in line with previous research attesting that the "economic magnitude of the soft information is somewhat limited" (Loughran and McDonald, 2016, p. 1202).

5.2 Classification

Table 3 shows the results of the classification task on the newly created dataset REPORTS. Performance is evaluated in terms of precision (P), recall (R), F_1 score (F) on the *uncertain* class, and in terms of overall accuracy (A). Additionally, significant performance increases over *Unc* were determined through paired *t*-tests with $\alpha = 0.05$.

The highest precision (0.69) is obtained through *UncSpec* and *UncGen+UncSpec*, which significantly outperform *Unc*. $UncSpec_{exp}$ scores highest in terms of recall (0.56), which again is significantly higher than that of *Unc*. This value in combination with a relatively high precision (0.67) makes the former feature set the strongest overall in terms of an F_1 score of 0.59, thus significantly exceeding *Unc* and $UncGen_{exp}$.

Overall, the high precision of *UncSpec* and *UncGen+UncSpec* coincides with the regressions (see Section 5.1), where both already yielded the highest coefficients. Another similarity are the implications for our research questions: Again, the domain-specific expansion performs best (RQ1), while the expert filtering does not provide visible improvements (RQ2).

Out of the newly added terms of $UncGen_{exp}$, 24 are contained in the 130 sentences labeled as *uncertain*. This stands in contrast to 29 matches with the terms of $UncSpec_{exp}$, which again indicates an advantage of the domain-specific model. The domain-dependent and legalese language of the latter reflects in matching terms such as "uninsured", "more-likely-than-not", and "interpretive".

In summary, our results show that for the given task, training own domain-specific word embedding models gives an advantage over relying on generic, off-the-shelf solutions. Lastly, the results reveal that the manual filtering of candidate terms has only a negligible impact on performance.

6 Conclusion

In this paper, we expanded a dictionary of financial uncertainty triggers through both a generic and a domain-specific embedding model. In a set of financial regressions, we showed that our domain-specific expansion shares a two times greater and highly significant association with subsequent volatility than the plain dictionary. Furthermore, we presented a newly annotated dataset of annual reports and showed that the dictionary expansions significantly boost the performance in a binary classification task of uncertain sentences.

Acknowledgments

This work was supported by the SFB 884 on the Political Economy of Reforms at the University of Mannheim (project C4), funded by the German Research Foundation (DFG).

References

Johannes Amtén. 2014. Neural network plugin for Weka. https://github.com/amten/NeuralNetwork. Accessed on January 16, 2018.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Saskia le Cessie and Hans C. van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):41–48.

William W. Cohen. 1995. Fast effective rule induction. In *Proceedings of the ICML*, pages 115–123.

Eugene F. Fama and Kenneth R. French. 1997. Industry costs of equity. *Journal of Financial Economics*, 43:153–193.

Viola Ganter and Michael Strube. 2009. Finding hedges by chasing weasels: Hedge detection using Wikipedia tags and shallow linguistic features. In *Proceedings of the ACL–IJCNLP: Short Papers*, pages 173–176.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explorations Newsletter*, 11:10–18.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Tim Loughran and Bill McDonald. 2014. Measuring readability in financial disclosures. *The Journal of Finance*, 69(4):1643–1671.

Tim Loughran and Bill McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Laurens J. P. van der Maaten and Geoffrey E. Hinton. 2008. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *ArXiv e-prints*.

Saif Mohammed, Bonnie Dorr, and Graeme Hirst. 2008. Computing word-pair antonymy. *Proceedings of the EMNLP*, pages 982–991.

John C. Platt. 1999. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods—Support Vector Learning*.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Francisco.

William Sharpe. 1963. A simplified model for portfolio analysis. *Management Science*, 9:277–293.

Christoph Kilian Theil, Sanja Štajner, Heiner Stuckenschmidt, and Simone Paolo Ponzetto. 2017. Automatic detection of uncertain statements in the financial domain. In *Lecture Notes in Computer Science: Proceedings of the CICLing*, Berlin. Springer. In press.

Ming-Feng Tsai and Chuan-Ju Wang. 2014. Financial keyword expansion via continuous word vector representations. *Proceedings of the EMNLP*, pages 1453–1458.

Sanja Štajner, Goran Glavaš, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Domain adaptation for automatic detection of speculative sentences. In *Proceedings of the IEEE ICSC*, pages 164–171.