Predicting Modality in Financial Dialogue

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

Abstract

In this paper, we perform modality prediction in financial dialogue. To this end, we introduce a new dataset and develop a binary classifier to detect strong or weak modal answers depending on surface, lexical, and semantic representations of the preceding question and financial features. To do so, we contrast different algorithms, feature categories, and fusion methods. Perhaps counter-intuitively, our results indicate that the strongest features for the given task are financial uncertainty measures such as market and individual firm risk.

1 Introduction

In this paper, we predict the modality of answers depending on their preceding question and other features in financial dialogue. Modality is an important concept in principal–agent settings of asymmetric information such as the stock market, since it can be used as a strategic tool by company executives: Using modality markers such as "probably" or "certainly," investor expectations can be managed or the effect of negative news can be mitigated without having to commit to false statements. Loughran & McDonald (2016, p. 1224) suggest to examine the hypothesis that larger shares of modal words in conference calls might worsen stock or operating performance. Subsequently, Dzieliński *et al.* (2019) found that executive modality is indeed predictive of stock price as well as analyst's earnings forecasts and firm valuations (Dzieliński et al., 2019). Although different to past work, we explore causes, not effects, of modality in the financial domain, this shows that modality prediction has potential down-stream uses in return, risk, and analyst forecast prediction. Specifically, modality prediction models could be employed for intra-day return prediction.

1.1 Modality

Linguistic modality, a concept related to politeness (Danescu-Niculescu-Mizil et al., 2013) and hedging (Lakoff, 1973; Hyland, 1998), is most commonly categorized into *dynamic*, *priority*, and *epistemic modality* (Portner, 2009, p. 47). In this work, we focus on epistemic modality, which expresses a speaker's confidence in the truth of their proposition [*ibid.*]: a high epistemic modality (variously expressed through markers such as "certainly," "must") describes a high confidence and a low modality ("probably," "might") stands for a low degree of confidence. While past socio-linguistic research has shown that a manual annotation of modality on a 5-item scale is a comparably hard task for humans (Rubin, 2007), past work in the financial domain indicates that the task seems to be easier for a binary distinction and a broader definition of uncertainty (Theil et al., 2018a). As manual annotation is costly and time-consuming, we were interested in automatically creating a silver standard dataset based on an established lexicon of modality markers (Loughran and McDonald, 2011) in the financial domain. To the best of our knowledge, there is no study investigating the determinants of modality in dialogue using natural language processing.

Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, pages 226–234
Barcelona, Spain (Online), December 12, 2020.

1.2 Earnings Calls

Earnings calls—the textual form we analyze in this paper—are quarterly public teleconferences or webcasts in which companies present the financial results of the ending business quarter. Past literature has examined indirectness (Crawford Camiciottoli, 2009), persuasion (Crawford Camiciottoli, 2011; Crawford Camiciottoli, 2018), and deception (Larcker and Zakolyukina, 2012) in earnings calls. Earnings calls typically consist of two parts: first, the company management (usually the CEO and/or CFO) as well as investor relations representatives hold a scripted presentation which closely follows the accompanying press release. Second, the call is opened to investors and banking analysts, which pose questions to the management in a Q&A session. Together with the information asymmetry, this unscripted kind of interaction makes the Q&A part especially suitable for our modality prediction task. Hence, we were motivated to extract question–answer pairs from the Q&A and to predict the modality of an answer depending on the content of the preceding question.

1.3 Contributions

We provide the following contributions to the community:

- We publish a dataset of 5K question–answer pairs for modality prediction.
- We introduce the first modality classifier including semantic information and learning from heterogeneous features.
- We provide interpretable results by visualizing the importance and effect of the used features.

2 Related Work

In the financial domain, the task of modality or vagueness detection is closely related to risk and return prediction. Loughran & McDonald (2011) handcrafted a set of sentiment lexica based on frequent terms in a sample of 60K 10-Ks. These lexica (from now on: LM) span the categories *positive*, *negative*, *uncertain*, *litigious*, *strong modal*, and *weak modal* and have been shown to possess predictive power of risk. Subsequent work in the NLP community automatically expands said lexica by adding semantically similar terms according to word embedding models for predictions of risk in form of return volatility (Tsai and Wang, 2014; Rekabsaz et al., 2017) or correlations with it (Theil et al., 2018b; Theil et al., 2020).

Štajner *et al.* (2017) perform speculation detection in the monetary policy domain as a binary sentence classification task. They use a list a list of uncertainty triggers extracted from the CoNLL-2010 shared task's training set (Farkas et al., 2010), the LM *uncertain* lexicon, and an own list of speculation triggers tailored to the task. Theil *et al.* (2018a) train a binary sentence classifier predicting the linguistic uncertainty of 1K sentences randomly sampled from a dataset of earnings calls. They use lemmatized bag-of-words (BoW) vectors, part-of-speech tags, a set of handcrafted syntactic rules, the CoNLL-2010 list of uncertainty triggers (Farkas et al., 2010), and the LM *uncertain* lexicon. Their results indicate that BoW vectors and the LM lexicon are the strongest features, which is why we include them in our classifier, too. Note that different to these works, we do not aim to predict the uncertainty of a sentence given its content, but rather the uncertainty of an answer given the content of the preceding question. Furthermore, we explore additional feature categories, such as semantic or financial features.

Using a set of 120K earnings calls, Dzieliński *et al.* (2019) find that the modality of executive utterances is correlated with post-call stock price, analyst's earnings forecasts, and firm valuation. Keith and Stent (2019) gather 12K earnings call transcripts and find that pragmatic and semantic features are moderately predictive of analysts' price forecast targets following the call dates. Their pragmatic feature set contains a dictionary of uni- and n-gram hedges (Prokofieva and Hirschberg, 2014) as well as the LM dictionary (Loughran and McDonald, 2011); however, they find the influence of semantic features (BoW and doc2vec (Le and Mikolov, 2014) vectors) to be stronger. Theil *et al.* (2019) collect a dataset of 90K earnings calls and develop an attention-based neural model to predict financial risk (i.e. return volatility) given the transcripts and several financial features. We include their financial features in our classifier,

types	tokens	sentences	utterances
7.7K	232.1K	15.1K	5.0K

Table 1: Descriptive statistics of our dataset.

as past research suggests a correlation between linguistic modality and financial risk. Different to these works, we do not predict external financial measures based on linguistic features. Instead, we aim to predict a linguistic variable (modality) as we are interested in uncovering its determinants in financial Q&A settings.

3 Methodology

We begin by introducing a new dataset for modality prediction in financial dialogue (cf. Section 3.1), proceed by defining different features sets (cf. Section 3.2), and finally introduce a classifier for our binary classification task (cf. Section 3.3).

3.1 Dataset

We obtain 20K earnings call transcripts from SeekingAlpha[1] and sample all question–answer (Q&A) pairs from them. Numbers are identified with SpaCy's named entity recognizer and replaced with uniform placeholder tokens. We remove Q&A pairs with inaudible parts, audiogaps, or multiple speakers talking at once.

We use the established LM dictionary (Loughran and McDonald, 2011) as a basis to induce the binary modality label of the answers, thus forming a silver standard dataset used in the subsequent classification. To this end, we focus on the two categories *weak* and *strong modality* and extract the answers with the highest share of these words—to avoid ambiguous labels, we require the *weak modal* answers to contain zero *strong modal* words and vice versa:

– The *weak modality* lexicon contains 27 tokens conveying vagueness such as "maybe" and "possibly." We take the 2.5K answers with the highest share of weak modal tokens and assign them a *weak modal* label.

  Example: "Well, the numbers might suggest that."

– The *strong modality* lexicon contains 20 tokens conveying certainty such as "always" and "undoubtedly." We take the 2.5K answers with the highest share of these tokens and assign them a *strong modal* label.

  Example: "It will. That's right, it will."

This yields a balanced dataset of 5K (2.5K *weak* and 2.5K *strong modal*) instances; Table 1 describes this set in terms of surface features. For the subsequent experiments, we apply an $80:20$ training–test split. Both our dataset and code can be found online.[2]

3.2 Features

Since we aim to predict the modality of an answer given the preceding question, all features are extracted from the questions. In total, we evaluate four different feature categories, which are partly motivated by the previous literature (cf. Section 2).

3.2.1 Surface Features

In the SURFACE feature set, we explore the following:

[1] `seekingalpha.com` is a crowd-sourced provider of data and research on financial markets. We comply with their reproduction policy of not quoting more than 400 words of any given transcript.

[2] `https://www.uni-mannheim.de/dws/people/researchers/phd-students/kilian-theil`

– **Length** is once represented by the number of sentences and once by the number of tokens in the respective question.

– **Positivity** and **negativity** are the share of tokens according to the respective LM lexica. These are defined by 354 positive tokens such as "breakthrough" or "win" and 2,355 negative tokens such as "decline" and "worsen."

– **Strong** and **weak modality** of a question could influence the modality of the respective answer. Examples of strong and weak model tokens according to the LM lexicon are given in Section 3.1.

– **Uncertainty** is again measured by the respective LM lexicon which contains 297 tokens referring to linguistic imprecision or risk, e.g. "hypothesis" and "volatility."

3.2.2 Lexical (Semantic) Features

In the LEXICAL category, we compare tf and tfidf vectors, which have been shown to perform strong for an uncertainty detection task (Theil et al., 2018b). To reduce sparsity, we apply singular value decomposition (SVD) and experiment with dimensions $d_{BoW} \in \{100, 200, ..., 1000\}$. Additionally, to expand the LEXICAL feature set with semantic information, we train word embedding models with word2vec (Mikolov et al., 2013) on the entire earnings call corpus (cf. Section 3.1). We evaluate dimensions $d_{w2v} \in \{100, 200, 300\}$ with both the continuous bag-of-words (CBOW) and the skip-gram (SG) architecture. Finally, we represent all questions as embedding centroids. Our results indicate that out of all previously mentioned representations, tfidf vectors with $d_{BoW} = 300$ are optimal for the given task.

3.2.3 Semantic Features

We use the Latent Dirichlet Allocation (LDA) algorithm to obtain topic models forming our SEMANTIC feature set. To find an optimal number of topics n, we evaluate the sensitivity of the log-likelihood l and the perplexity P to $n \in \{5, 10, ..., 45, 50\}$ in a five-fold cross validation setup on our training set. Our results indicate that an optimal l and P are obtained for $n = 5$.[3]

3.2.4 Financial Features

We use the FINANCIAL feature set proposed by Theil et al. (2019) to contrast the predictive power of linguistic features to that of performance measures about the firm or the overall economy:

– **Firm volatility**, measured by the standard deviation of stock returns, is the most important measure for financial risk. We include the volatility in the preceding business quarter as this feature should have an impact on investor and manager confidence.

– **Market volatility** as gauged by the CBOE Volatility Index (VIX),[4] reflects the overall market uncertainty and should have a similar (albeit more global) impact as firm volatility.

– **Firm size** or market value is the number of outstanding shares multiplied by the stock price and is a well-known driver of risk (Fama and French, 1992).

– **Book-to-market** reflects the firm value according to the balance sheet divided by the market value and thus reflects the degree of over- or undervaluation. Similar to the preceding measures, this ratio is considered to be a major risk driver (Fama and French, 1992).

– **Earnings surprise** reflects the deviation from the actual earnings per share figure from the mean of previous analyst forecasts. Negative surprises tend to decrease stock returns (Price et al., 2012) which may lead the executives to manage investor expectations.

– **Industry** dummies are obtained from the established Fama–French 12-industry scheme,[5] which distinguishes between e.g. "energy" or "healthcare."

[3] $l = -145218.44$ and $P = 1782.15$.

[4] http://www.cboe.com/vix

[5] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Features	Weak Modal			Strong Modal			Average		
	P	R	F	P	R	F	P	R	F
SURFACE	0.52	0.55	0.53	0.51	0.48	0.50	0.52	0.52	0.52
LEXICAL	0.57	0.60	0.59	0.56	0.53	0.54	0.57	0.57	0.57
SEMANTIC	0.51	0.52	0.51	0.49	0.47	0.48	0.50	0.50	0.50
FINANCIAL	**0.89**	**0.95**	**0.92**	**0.95**	0.87	**0.91**	**0.92**	**0.91**	**0.91**
ALL$_{early}$	0.86	**0.95**	0.90	0.94	0.85	0.89	0.90	0.90	0.90
ALL$_{late}$	**0.89**	0.85	0.87	0.85	**0.89**	0.87	0.87	0.87	0.87
Random	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50

Table 2: Classification results per class (*weak* and *strong modal*) and on average.

3.3 Classifier

Since we are interested in examining the influence of different features on an answer's modality, we select a set of algorithms with interpretable weights. In sum, we consider: (Gaussian) Naïve Bayes, Logistic Regression, Support Vector Machines (with RBF kernel), Decision Trees, Random Forest, and XGBoost (Chen and Guestrin, 2016). The classifier is implemented and evaluated using `sklearn 0.21.2` and `xgboost 0.90`.

3.3.1 Feature Fusion

To fusion our four feature categories, we use the following methods: (1) Early fusion involves representing all feature categories in the same vector space; (2) late fusion (or "stacking") implies that for each feature category, a separate classifier is trained—the predicted labels of these classifiers are then used as feature inputs for a meta-classifier predicting the final label. Our results show that, when representing all features in one vector space (early fusion), the XGBoost classifier outperforms all other algorithms. We furthermore find that the Gaussian Naïve Bayes algorithm performs best as meta-classifier for the late fusion approach.

3.3.2 Evaluation

We evaluate the performance of our classifiers with precision, recall, and F-score metrics. Furthermore, to quantify relative feature importance in case of the early fusion approaches, we use SHAP (SHapley Additive exPlanations) values, which were introduced by Lundberg and Lee (2017) and subsequently adapted for tree-based learners (Lundberg et al., 2020):

$$\phi_i(f_x) = \sum_{R \in \mathcal{R}} \frac{1}{M!}[f_x(P_i^R \cup i) - f_x(P_i^R)], \tag{1}$$

where ϕ_i is the SHAP value for feature i, f_x is the model output, $\mathcal{R}$ is the set of all feature orderings, P_i^R is the set of all features preceding feature i in ordering R, and M is the total number of features.

4 Results and Discussion

4.1 Feature Performance

Table 2 shows the results of our classification task in terms of precision (P), recall (R), and F-score (F) for both the *strong* and the *weak modal* class as well as on average. The early fusion approach uses an XGBoost classifier trained on a single vector containing all features; the late fusion approach additionally uses a Gaussian Naïve Bayes meta-classifier stacked upon two XGBoost classifiers trained separately on the linguistic and financial features. Since the binary labels are evenly distributed, a useful classifier should exceed a value of 0.50 across all measures. The SURFACE, LEXICAL, SEMANTIC, and FINANCIAL feature sets are defined as outlined in Section 3.2 and the fused features are represented by ALL with separate subscripts for the *early* and the *late* fusion approach.

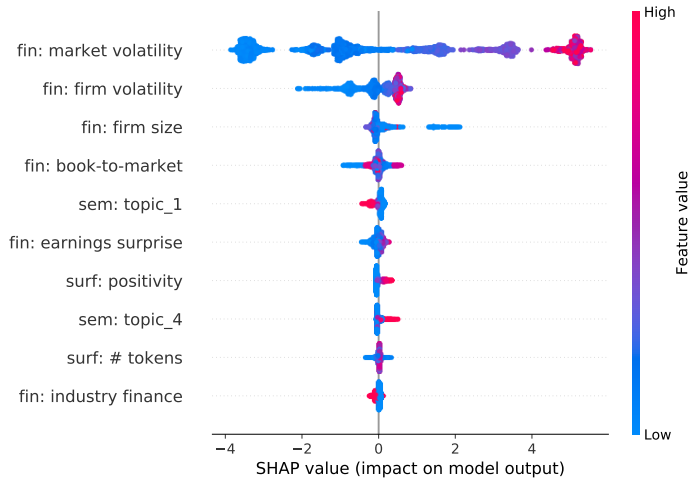

Figure 1: Violin plot of SHAP values for the top-10 features in the binary classification with early fusion.

All feature sets (with the exception of SEMANTIC for the *strong modal* class) improve over a random prediction. Furthermore, although late fusion improves slightly in terms of precision on the *weak modal* class ($P = 0.89$ vs. 0.86) and in terms of recall on the *strong modal* class ($R = 0.89$ vs. 0.85), the overall performance is slightly worse than that of an early fusion approach. When looking at individual features sets, we find that, perhaps counter-intuitively, the financial feature set alone has the strongest performance—even when compared to the more complex fusion approaches. This suggests that e.g. market or firm risk have a comparably larger influence on the modality of executive answers than the content of the preceding question. Therefore, while past literature asserts a comparably small impact of textual information for correlations with financial risk (Loughran and McDonald, 2011; Theil et al., 2018b), the same seems to apply when predicting a linguistic variable such as modality. Furthermore, this motivates to explore whether the effect persists when featuring a larger context window of textual information (perhaps including the earnings call presentation or prior questions and answers) or different methods of textual representation.

4.2 Feature Importance

One advantage of the early fusion approach is its interpretability: since all features are represented in the same vector, we can obtain a notion of relative feature importance quantitatively. To do so, we calculate the SHAP values (cf. Section 3.3.2) for all features and present the results in Figure 1. The intuition behind these values is to compare the contribution of a feature value to the difference between the actual and the mean prediction.

The strongest feature is market volatility, followed by firm volatility, and firm size. Interestingly, a high market and firm volatility positively impact the model output (and vice versa), implying that risky economic conditions may prompt managers to create a sense of security by committing to *strong modal* answers more frequently. Apart from two topical features, the strongest linguistic feature is positivity: Less positive questions tend to decrease the modality of an answer which could be attributed to their unsettling impact on manager confidence.

In addition, we were interested to explore the importance of individual linguistic types for the final prediction. To this end, we ranked the average SHAP values of all components of the purely LEXICAL model. In addition, we ran SpaCy's part-of-speech tagger on the ranked terms to explore the prevalence of different word classes. The vocabulary size of the complete dataset is 7,679 types. Out of these, only 153 terms have an average SHAP value > 0, i.e. are important for the final prediction. We found that the majority of terms are nouns (63), followed by verbs (47), adjectives (21), and adverbs (10). The top-20 terms according to their average SHAP value can be found in Table 3. Questions with numerical content (e.g. containing the token "number" or the placeholder tokens "DATE," "MONEY," "CARDINAL" for

term	SHAP in %	term	SHAP in %
DATE	5.67	share	0.94
okay	4.61	obviously	0.91
really	2.37	CARDINAL	0.89
number	1.85	color	0.88
MONEY	1.71	doing	0.85
does	1.46	tax	0.81
capex	1.44	charge	0.79
opportunity	1.26	performance	0.77
opportunities	1.13	stores	0.69
timing	1.12	given	0.68

Table 3: Average SHAP values for the top-20 terms. Uppercase terms represent placeholder tokens for the respective numerical named entity types identified with SpaCy.

dates, monetary values, and cardinal numbers) appear to influence an answer's modality. Likewise, business jargon terms such as "capex," "share," or "tax" are important for modality prediction.

Lastly, we were motivated to compare the feature distributions of the 434 misclassified instances to the total population of 1K test instances. For example, systematically higher VIX values in the misclassified instances compared to the rest of the population would motivate further experiments with a different weighting/sampling procedure of this feature in the training process. To do so, we checked for significant differences in the SURFACE and FINANCIAL feature sets across both misclassified and test instances using independent t-tests. Although none of the features showed significant differences in mean for $p \in \{0.05, 0.01, 0.001\}$, we found that the p-value for question uncertainty approaches conventional levels for significance ($p = 0.144$). This indicates that, apart from the increased context window mentioned above, future work could deeper explore the measurement of and prediction based on uncertainty for the given task—perhaps building on prior work on modality, hedging, or uncertainty detection presented in Section 2.

5 Conclusion

In this paper, we present a new dataset for modality prediction in financial dialogue and introduce a binary classifier to address this task. In our experiments, we contrast the performance of various algorithms, feature sets, and fusion methods. Interestingly, we reach a counter-intuitive result indicating that financial features (most prominently market and firm risk) possess a higher predictive power for answer modality than linguistic features (such as bags-of-words, topic models, or word embeddings) of the preceding question. In future work, it would be interesting to explore whether this effect persists when using a larger context window for the textual representations.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments.

References

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of ACM SIGKDD*, pages 785–794.

Belinda Crawford Camiciottoli. 2009. "Just Wondering if You Could Comment on That": Indirect Requests for Information in Corporate Earnings Calls. *Text & Talk*, 29(6):661–681.

Belinda Crawford Camiciottoli. 2011. Ethics and Ethos in Financial Reporting: Analyzing Persuasive Language in Earnings Calls. *Business Communication Quarterly*, 74(3):298–312.

Belinda Crawford Camiciottoli. 2018. Persuasion in Earnings Calls: A Diachronic Pragmalinguistic Analysis. *International Journal of Business Communication*, 55(3):275–292.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of ACL*, pages 250–259.

Michał Dzieliński, Alexander Wagner, and Richard J. Zeckhauser. 2019. Straight Talkers and Vague Talkers: The Effects of Managerial Style in Earnings Conference Calls. *Swiss Finance Institute Research Paper Series*, 17(13).

Eugene F. Fama and Kenneth R. French. 1992. The Cross Section of Expected Stock Returns. *Journal of Finance*, 47(2):427–465.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of CoNLL: Shared Task*, pages 1–12.

Ken Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins, Amsterdam/Philadelphia.

Katherine A. Keith and Amanda Stent. 2019. Modeling Financial Analysts' Decision Making via the Pragmatics and Semantics of Earnings Calls. In *Proceedings of ACL*, pages 493–503.

George Lakoff. 1973. Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. *Journal of Philosophical Logic*, 2:458–508.

David F. Larcker and Anastasia A. Zakolyukina. 2012. Detecting Deceptive Discussions in Conference Calls. *Journal of Accounting Research*, 50(2):494–540.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of ICML*, pages 272–280.

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. 2016. Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, 54(4):1187–1230.

Scott M. Lundberg and Su-in Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of NIPS*, pages 1–10.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, 2(1):56–67.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arxiv:1301.3781*.

Paul Portner. 2009. *Modality*. Oxford University Press.

S. McKay Price, James S. Doran, David R. Peterson, and Barbara A. Bliss. 2012. Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone. *Journal of Banking and Finance*, 36(4):992–1011.

Anna Prokofieva and Julia Hirschberg. 2014. Hedging and Speaker Commitment. In *International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data*. LREC.

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 ACL*, pages 1712–1721.

Victoria L. Rubin. 2007. Stating with Certainty or Stating with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements. In *Proceedings of NAACL HLT 2007*, pages 141–144.

Sanja Štajner, Goran Glavaš, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Domain Adaptation for Automatic Detection of Speculative Sentences. In *Proceedings of the International Conference on Semantic Computing*, San Diego.

Christoph Kilian Theil, Sanja Štajner, Heiner Stuckenschmidt, and Simone Paolo Ponzetto. 2018a. Automatic Detection of Uncertain Statements in the Financial Domain. In *Proceedings of CICLing*, pages 642–654. Springer.

Christoph Kilian Theil, Sanja Štajner, and Heiner Stuckenschmidt. 2018b. Word Embeddings-Based Uncertainty Detection in Financial Disclosures. In *Proceedings of the ACL Workshop on Economics and Natural Language Processing (ECONLP)*, pages 32–37.

Christoph Kilian Theil, Samuel Broscheit, and Heiner Stuckenschmidt. 2019. PRoFET: Predicting the Risk of Firms from Event Transcripts. In *Proceedings of IJCAI*, pages 5211–5217.

Christoph Kilian Theil, Sanja Štajner, and Heiner Stuckenschmidt. 2020. Explaining Financial Uncertainty through Specialized Word Embeddings. *ACM/IMS Transactions on Data Science*, 1(1).

Ming-Feng Tsai and Chuan-Ju Wang. 2014. Financial Keyword Expansion via Continuous Word Vector Representations. In *Proceedings of the EMNLP*, pages 1453–1458.