UNCERTAINTY, RISK, AND FINANCIAL DISCLOSURES

Applications of Natural Language Processing in Behavioral Economics

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von
CHRISTOPH KILIAN THEIL
aus Mannheim

Mannheim, 2022

Dekan:	Dr. Bernd Lübcke, Universität Mannheim
Referent:	Prof. Dr. Heiner Stuckenschmidt, Universität Mannheim
Korreferent:	Prof. Dr. Simone Paolo Ponzetto, Universität Mannheim
Korreferent:	Prof. Dr. Dirk Hovy, Università Commerciale Luigi Bocconi

Tag der mündlichen Prüfung: 16. Mai 2022

ABSTRACT

In the last decade, natural language processing (NLP) methods have received increasing attention for applications in behavioral economics. Such methods enable the automatic content analysis of large corpora of financial disclosures, e.g., annual reports or earnings calls. In this setting, a conceptually interesting but underexplored variable is linguistic uncertainty: Due to the unpredictability of the financial market, it is often necessary for corporate management to use hedge expressions such as "likely" or "possible" in their financial communication. On the other hand, management can also use uncertain language to influence investors strategically, for example, through deliberate obfuscation. In this dissertation, we present NLP methods for the automated detection of linguistic uncertainty. Furthermore, we introduce the first experimental study to establish a causal link between linguistic uncertainty and investor behavior. Finally, we propose regression models to explain and predict financial risk. In addition to the independent variable of linguistic uncertainty, we explore a psychometric and an assumption-free model based on Deep Learning.

ZUSAMMENFASSUNG

Methoden des Natural Language Processing (NLP) haben im letzten Jahrzehnt zunehmend Einzug in die Verhaltensökonomie gehalten. Diese Methoden ermöglichen die automatische Inhaltsanalyse von großen Korpora finanzieller Berichterstattung wie Jahresberichten oder Earnings Calls. In diesem Zusammenhang ist die sprachliche Unsicherheit eine konzeptionell spannende aber wenig erforschte Variable: Durch die Unwägbarkeit des Finanzmarktes ist es für das Management von Unternehmen oft erforderlich, in seiner Finanzkommunikation auf Heckenausdrücke wie „wahrscheinlich" oder „möglicherweise" zurückzugreifen. Andererseits kann unsichere Sprache auch zur strategischen Beeinflussung von Anlegenden—etwa durch bewusste Verschleierung—genutzt werden. Im Rahmen dieser Dissertation werden zunächst NLP-Methoden zur automatisierten Erkennung sprachlicher Unsicherheit präsentiert. Um eine kausale Verknüpfung zwischen sprachlicher Unsicherheit und Anlegeverhalten zu etablieren, wird die erste experimentelle Studie zu diesem Thema vorgestellt. Schließlich werden Regressionsmodelle zur Erklärung und Vorhersage von finanziellem Risiko vorgeschlagen. Neben der unabhängigen Variable der sprachlichen Unsicherheit werden auch ein psychometrisches und ein annahmefreies Modell basierend auf Deep Learning erforscht.

ACKNOWLEDGEMENTS

Throughout the past five years, I have received tremendous support from numerous people.

First, I would like to thank my supervisor and referee Heiner Stuckenschmidt. He provided me with academic guidance, granted me creative freedom, always supported my decisions, and demonstrated great foresight in staffing me on exciting projects and steering me in the right directions.

I am also indebted to Simone Paolo Ponzetto, who agreed to co-refer my dissertation and encouraged me to join the weekly meetings of his lab during the dawning times of the pandemic. These fruitful discussions provided me with renewed motivation and fresh ideas in a period where both were scarce for many of us.

Furthermore, I am very thankful to my co-referee Dirk Hovy, with whom I had the pleasure to collaborate from 2020 onward. Moreover, he invited me to spend a research visit at his lab at Bocconi in Milan. Apart from having a great two months with fantastic colleagues and friends, this environment and his input have really inspired me and fostered my personal development.

I would also like to thank Sanja Štajner, who was a mentoress to me in the first two years of my doctorate. Beyond that, thanks belong to my collaborators: Samuel Broscheit, Jens Daube, and Jakob Kappenberger (née Gutmann). Finally, I am grateful to several colleagues at the University of Mannheim's Data and Web Science Group: Kiril Gashteovski, Goran Glavaš, Oliver Lehmberg, Federico Nanni, Daniel Ruffinelli, Jörg Schönfisch, and Cäcilia Zirn.

From our psychology department, I would like to thank Jutta Mata, Moritz Ingendahl, and Sophie Scharf. From the area finance, I would like to thank Kristina Meier and Pavel Lesnevski.

Special thanks go to my family: To my grandparents, Alfred and Hannelore. To my parents, Lisa and Hans. And to my siblings, Mirjam, Markus, and Clara.

Last but not least, I would like to thank my dear friends who were ever-supportive on this journey: Lukas, Dani, Sebastian, Elena, Alex D., Niklas, Max, Simon, Clemens, Stolli, Santanu, Jens, Jakob, Alex K., Anne, Arne, and my girlfriend Cathrin.

Thank you very much to all of you—without you, this work would not have been possible.

OWN PUBLICATIONS

Some ideas, tables, and figures have previously appeared in the following publications and working papers co-authored by me:

Theil, Christoph Kilian, Samuel Broscheit, and Heiner Stuckenschmidt (2019). "PRoFET: Predicting the Risk of Firms from Event Transcripts." In: *Proceedings of IJCAI*, pp. 5211–5217. DOI: 10.24963/ijcai.2019/724.

Theil, Christoph Kilian, Sanja Štajner, and Heiner Stuckenschmidt (2020). "Explaining Financial Uncertainty through Specialized Word Embeddings." In: *ACM/IMS Transactions on Data Science* 1.1, pp. 1–19. DOI: 10.1145/3343039.

Theil, Christoph Kilian, Sanja Štajner, Heiner Stuckenschmidt, and Simone Paolo Ponzetto (2017). "Automatic Detection of Uncertain Statements in the Financial Domain." In: *Proceedings of CICLing*, pp. 642–654. DOI: 10.1007/978-3-319-77116-8_48.

Theil, Kilian, Jens Daube, and Heiner Stuckenschmidt (2022). *Linguistic Uncertainty and Risk Perception in Financial Disclosures*. Working Paper. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4012946.

Theil, Kilian, Dirk Hovy, and Heiner Stuckenschmidt (2023). "Top-Down Influence? Predicting CEO Personality and Risk Impact from Speech Transcripts." In: *Proceedings of ICWSM (forthcoming)*. URL: https://arxiv.org/abs/2201.07670.

Theil, Kilian and Heiner Stuckenschmidt (2020). "Predicting Modality in Financial Dialogue." In: *Proceedings of the COLING Workshop on Financial Narrative Processing (FNP)*, pp. 226–234. URL: https://www.aclweb.org/anthology/2020.fnp-1.35/.

Appendix A contains an overview of the models, code, and data published along with this thesis.

CONTENTS

LIST OF ACRONYMS

BiLSTM	Bidirectional Long Short-Term Memory
BoW	Bag-of-Words
BM25	Okapi Best Matching 25
BTM	Book-to-Market
CAR	Cumulative Abnormal Return
CRSP	Center for Research in Security Prices
CBoW	Continuous Bag-of-Words
CEO	Chief Executive Officer
CFO	Chief Financial Officer
CNN	Convolutional Neural Network
CAR	Cumulative Abnormal Return
CV	Cross-Validation
DL	Deep Learning
EDGAR	Electronic Data Gathering, Analysis, and Retrieval system
EPS	Earnings per Share
FOMC	Federal Open Market Committee
FNN	Feed-Forward Neural Network
GAAP	Generally Accepted Accounting Principles
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GPU	Graphics Processing Unit
IAA	Inter-Annotator Agreement
IBES	Institutional Brokers Estimate System
LDA	Latent Dirichlet Allocation
LASSO	Least Absolute Shrinkage and Selection Operator
LIWC	Linguistic Inquiry and Word Counts
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MBTI	Myers–Briggs Type Indicator
MD&A	Management's Discussion and Analysis
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Squared Error
NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
NLTK	the Natural Language Toolkit

OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PoS	Part-of-Speech
Q&A	Question-and-Answer session
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
RMSE	Root Mean Squared Error
SEC	United States Securities and Exchange Commission
SG	Skip-Gram
SHAP	Shapley Additive Explanations
SVD	Singular Value Decomposition
SL	Sequence Labeling
S&P 500	Standard and Poor's 500
SUE	Standardized Unexpected Earnings
SVM	Support Vector Machine
SVR	Support Vector Regression
tf	term frequency
tf–idf	term frequency–inverse document frequency
US	United States
VIX	Chicago Board Options Exchange Volatility Index
WEKA	the Waikato Environment for Knowledge Analysis
WRDS	Wharton Research Data Services

Part I

PRELIMINARIES

1

INTRODUCTION

The most useful words at the stock exchange are: 'maybe,' 'hopefully,' 'possible,' 'it could,' 'nonetheless,' 'although,' 'indeed,' 'I believe,' 'I mean,' 'however,' 'probably,' 'it seems to me.' Everything you believe and say is conditional.[1]

— André Bartholomew Kostolany

The waning years of the last millennium were one of the most volatile periods in financial history (Turner and Weigel, 1992). When the predictability of the Dow Jones, the oldest and most eminent equity index in the U.S., was at an all-time low (Qian and Rasheed, 2005), it became apparent that traditional econometric models proved insufficient for the ever more interconnected and chaotic financial market. In a spiral of creative destruction (Schumpeter, 1950), this crisis created a fertile ground for scientific innovation: The birth of prospect theory (Kahneman and Tversky, 1979) marked a breakthrough for the aspiring discipline of behavioral economics and would later earn Daniel Kahneman the Nobel Memorial Prize in Economic Sciences.

While the neoclassic notion of economics as a natural rather than a social science previously had led to an elimination of psychology from economic textbooks, Kahneman and his colleagues uncovered a growing number of biases that were unexplainable without this discipline. Specifically, they found that when facing investment decisions, humans are not only influenced by informational content, but also by its form. A key example is the *certainty effect* which states that, incongruent with expected utility theory, investors consistently prefer certain bets over uncertain ones, even when the latter have the same (or, in some cases, even a higher) expected value (Tversky and Kahneman, 1986). This illustrates how uncertainty and perceived risk play a pivotal role in framing financial decision-making problems.

Mirroring the advances of behavioral economics, scientists and practitioners increasingly began to include the vast yet untapped sources of textual information into their financial models. While early approaches of financial Natural Language Processing (NLP) were usually based on word counts and other sparse representations of text (Frazier et al., 1984; Lewis et al., 1986, *inter alia*), today's growing computational resources and innovations in the area of Deep Learning (DL) enable the automatic analysis of large-scale datasets, often including several modalities such as text, speech, and financial data (Qin and Yang, 2019; Sawhney et al., 2021, *inter alia*). The thesis at hand is located at the intersection of behavioral economics and the

1 Quotation translated from German by the author.

state-of-the-art in NLP. In particular, it proposes computational methods to automatically detect expressions of uncertainty and other linguistic phenomena in financial disclosures and uses them to predict reactions of the financial market. The following sheds light on the motivation, key contributions, and outline of this thesis.

1.1 MOTIVATION

Following Kahneman and Tversky (1979), an increasing number of studies have analyzed intra- and inter-subjective perceptions of numerical uncertainty in financial settings (Diecidue and van de Ven, 2008; Holzmeister et al., 2020; Klos et al., 2005; Zeisberger, 2021). However, little is understood about perceptions of linguistic probability or uncertainty, also known as *vagueness*. As hinted in the introductory quote by business magnate André Kostolany, vague expressions such as "probably" or "more-likely-than-not" are frequently used devices of financial communication. An evident reason for this stems from the unpredictability and complexity of the stock market: By toning down their language, managers and other market actors avoid seeming overconfident or untrustworthy in the face of uncertain financial prospects. Furthermore, investor relations communication—and the relationship between management and shareholders in general—is shaped by asymmetric information and, at times, conflicting interests. For example, the management might be interested in keeping its job despite bleak business outlooks; investors, on the other hand, usually have little personal involvement in the company and may therefore favor drastic measures to secure returns (e.g., layoffs or restructuring).

This setting, also called the *agency dilemma* (Eisenhardt, 1989), increases management incentives to selectively share or sugarcoat information with evasive language. As of now, however, systems to automatically detect such language in financial communication are scarce. Apart from stock price prediction, an important downstream application for such systems is deception and fraud detection (Bachenko et al., 2008; Fornaciari et al., 2021; Larcker and Zakolyukina, 2012). Furthermore, analyzing the use of uncertain language in financial Question-and-Answer sessions (Q&As) may allow us to understand more about psychological drivers of vagueness. For example, suggestive or negatively framed questions by banking analysts may prompt firm Chief Executive Officers (CEOs) to "weasel out" of an unpleasant situation without having to commit to false statements. Hence, as part of this thesis, we were interested in developing systems for linguistic **uncertainty detection** as our first task (T_1).

First observational studies show that firms issuing financial disclosures with uncertain language face fluctuating stock prices (Loughran and McDonald, 2011, 2013) and less favorable analyst valuations

(Dzieliński et al., 2021). That is, linguistic uncertainty in financial disclosures seems to increase financial uncertainty on a firm-level. However, the behavioral component of this mechanism is not fully understood. Is linguistic uncertainty a reflection of uncertain business conditions? Or rather, is linguistic uncertainty a *cause* of increased stock price volatility and analyst misjudgment? It is impossible to answer this question with observational regression analyses alone, since those (even when carefully controlling for possible confounders) can merely quantify a correlation as opposed to a causal interaction. Methodologically, a natural candidate for this problem is a laboratory study randomly allocating a stimulus to a treatment and a control group (Floyd and List, 2016). Due to the lack of such studies, we were motivated to conduct the first laboratory experiment addressing our second task (T_2), the **causal modeling of uncertainty and risk**. In doing so, we aim to find out how linguistic uncertainty impacts investor behavior. Furthermore, we wanted to investigate how investor characteristics such as personality, gender, or risk tolerance determine their perception of linguistic uncertainty and financial risk.

Lastly, although prior studies have explored the effect of linguistic uncertainty on stock price fluctuation and other financial risk measures, they usually rely on existing word lists deemed to cover linguistic uncertainty. Such a word list may, e.g., contain the terms "anomaly" and "possibility." A disadvantage of such lists is that they are based on subjective assessments of their creators and that they are usually not exhaustive. For example, a word list aimed to quantify economic uncertainty created before 2016 will likely not contain the words "Brexit" or "coronavirus," despite their undeniable importance in the years onward. Furthermore, novel text-based risk regressors based on DL provide a powerful complement to traditional count-based approaches. Hence, we were motivated to tackle financial **risk regression** as our third and final task (T_3). Apart from a method to automatically enrich an existing uncertainty dictionary with industry- and time-specific uncertainty terms (hence increasing its exhaustiveness while not relying on manual annotation), we experiment with a Transformer-based model to detect the impact of CEO personality on financial risk, and an assumption-free model for risk prediction.

1.2 CONTRIBUTIONS

The novelty of this thesis lies in two factors: It is the first to analyze the entire mechanism of companies issuing uncertain financial disclosures, over their influence on individual investor behavior, down to predictions of market reactions based on these disclosures. In doing so, it employs a holistic toolkit of classic text classification based on sparse features, laboratory experiments for causal inference, large-scale observational studies, and advanced NLP methods based on DL

and Transformers. Along the three tasks outlined above, we provide the following contributions to the academic community:

(T_1) Uncertainty detection: *How can we detect linguistic uncertainty automatically in financial disclosures? What are its economic and linguistic determinants?*

In Chapter 4, we present a new gold standard dataset for a binary classification of linguistic uncertainty in finance. We furthermore introduce the first classifier to address this task. In Chapter 5, we develop a silver standard dataset for uncertainty prediction in a financial Q&A setting. Moreover, we propose the first classifier to detect linguistic uncertainty based on multimodal linguistic and financial features.

(T_2) Causal modeling of uncertainty and risk: *How can we quantify the causal influence of linguistic uncertainty on risk perception and investment behavior? Which personal characteristics of investors play a role in this setting?*

To establish a notion of causality between linguistic uncertainty and investor behavior, we conduct the first laboratory experiment analyzing the effect of uncertain language in financial disclosures on risk perception and investments (Chapter 6). Moreover, we explore the confounding effect of investor characteristics such as age, personality, and financial literacy.

(T_3) Risk regression: *What is the influence of linguistic uncertainty on financial risk measures? What other linguistic phenomena are explanatory or predictive of risk?*

Chapter 7 presents an approach to explain financial risk and analyst uncertainty based on linguistic uncertainty; to measure uncertainty, we combine an automatic dictionary expansion method with a feature selection approach based on multi-task learning. Chapter 8 proposes the first regressor of CEO personality; providing evidence for the *upper echelons theory* (Hambrick and Mason, 1984), we find a robust and significant effect of personality on financial risk. Finally, Chapter 9 introduces the first assumption-free DL regressor of risk based on semantic and financial representations.

Published and forthcoming papers created during this thesis are listed in Own Publications. Links to the publicly available models, code, and data can be found in Appendix A.

1.3 OUTLINE

Following tasks T_1–T_3 outlined above, the structure of this thesis is described on a high level in Figure 1. In the current Part I, the the-

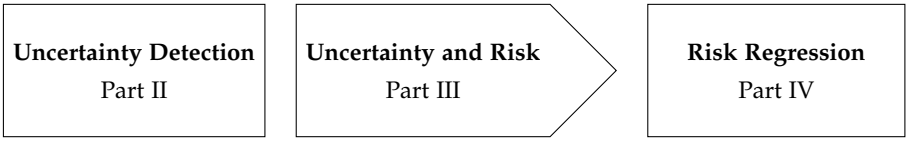


Figure 1: Structure of the core parts of this thesis.

oretical foundations of this research (Chapter 2) and related work (Chapter 3) are discussed.

To address T_1, Part II introduces two models to classify financial disclosure language into either linguistically certain or uncertain style. Chapter 4 introduces a binary classifier for uncertain sentences in earnings calls based on lexico-syntactic features. Afterward, Chapter 5 presents an uncertainty classifier in the earnings call Q&A based on financial and lexico-semantic features.

Concerning T_2, Part III describes an experimental approach to establish a causal link between linguistic uncertainty in financial disclosures and investment behavior (Chapter 6).

Addressing T_3, Part IV deals with regressions of volatility on linguistic uncertainty and other text-based measures: Chapter 7 presents a method to expand an established dictionary of uncertainty automatically; we use the such assessed linguistic uncertainty to explain future volatility and analyst uncertainty. Chapter 8 discusses a regressor of CEO personality and its application for volatility prediction. Then, Chapter 9 introduces an assumption-free regressor of volatility based on latent linguistic variables.

The final Part V presents a summary and conclusion (Chapter 10) together with limitations and directions for future work (Chapter 11).

2

THEORETICAL BACKGROUND

* What is uncertainty, and how does it shape our reality, thinking, and language? Which types of uncertainty affect the financial markets, and how do they interact? This chapter elaborates on the theoretical background of this thesis by defining linguistic and financial uncertainty, their connection to the agency dilemma, and their manifestations in financial disclosures.

2.1 ALEATORY AND EPISTEMIC UNCERTAINTY: TWO SIDES OF THE SAME COIN

In science, *uncertainty* is commonly categorized into an *aleatory* and an *epistemic* component (Hacking, 1975). *Aleatory uncertainty* derives from Latin "alea" which means "dice" and captures the stochastic notion of chance, which is thought to be fundamental and irreducible (Chowdhary and Dupuis, 2011). Instead, through repeated empirical observation, it can be recorded with probability distributions across known action outcomes. An example is the toss of a coin, where in theory, both outcomes are equally probable yet unpredictable ex-ante. *Epistemic uncertainty*, derived from Ancient Greek ἐπιστήμη ("insight" or "knowledge"), is the type of uncertainty that arises due to incomplete or incorrect knowledge about the environment (Chowdhary and Dupuis, 2011). Examples include inaccurate measurements, missing data, or erroneous computational models. Epistemic uncertainty is theoretically reducible since one can repeat measurements with more accurate sensors (*ibid.*), gather additional data, and create more exhaustive models. However, there are practical restrictions due to the limitations of technology and human understanding itself.

In practice, epistemic and aleatory uncertainty are often mingled, and it remains an open question whether any uncertainty might ultimately be due to lacking knowledge. For example, a coin toss might in actuality not be truly random, as the laws of classical mechanics are thought to be deterministic[1] and features such as the coin's mo-

* §2.2 expands on the introductory discussions of linguistic uncertainty phenomena and financial disclosure types in Theil and Stuckenschmidt (2020), Theil, Štajner, and Stuckenschmidt (2020), Theil, Broscheit, and Stuckenschmidt (2019), and Theil, Štajner, Stuckenschmidt, and Ponzetto (2017).

1 There is an ongoing academic debate whether classical mechanics is indeed deterministic (Nikolić, 2006) and, on the other hand, whether true randomness actually exists, e.g., in interpretations of quantum mechanics (Barrett et al., 2014). However, as the focus of this thesis lies on computational social science and since the classification into epistemic and aleatory uncertainties is useful for our use case, these questions will not be discussed in greater detail herein.

mentum vector are to some degree predictive of the toss outcome (Diaconis et al., 2007). Looking at chaotic systems[2] such as the financial market (Hristu-Varsakelis and Kyrtsou, 2008), however, it can be helpful to distinguish between uncertainty due to human error (e.g., imprecise observation or communication) and a component of uncertainty that can not be meaningfully reduced and hence is best approximated with probability distributions (e.g., the return on an investment). Therefore, although it is scientifically impossible to prove whether aleatory uncertainty truly exists as an intrinsic property of our world (Beck, 2009), we will adhere to this proposed dichotomy as a conceptual simplification.

Several scholars have attempted to embed the aleatory and epistemic components of uncertainty into their scientific disciplines. Table 1 summarizes the synonymous conceptual pairs, which we discuss in the following.

In economics, a prominent example is given in Frank Knight's seminal work "Risk, Uncertainty, and Profit," in which he distinguishes between economic *risk*, which can be measured and recorded in the form of probability distributions; and fundamental *uncertainty* (hence also called *Knightian uncertainty*), for which this is impossible (Knight, 1921). Here, risk corresponds to the aleatory component of uncertainty, while fundamental uncertainty refers to the epistemic side. The dichotomy is re-investigated by psychologists Kahneman and Tversky (1982), who call it "external" and "internal" uncertainty. Economist Davidson (1996) calls these two types of uncertainty "ontological" (from Ancient Greek ὄντος + λογία: "science of being") and "epistemological" (derived from the same root as "epistemic"). Finally, in statistics, two kinds of error are called "random" and "systematic" (Taylor, 1999). While random errors reflect the stochastic variability between measurements, systematic errors comprise observational errors, miscalibrated sensors, or environmental interference (*ibid.*).

All of these classifications have in common that they assume one side of uncertainty to be an intrinsic property of our surrounding world and another side to be the result of lacking knowledge or erroneous information. In the following, we will adhere to the conceptual pair of "aleatory" and "epistemic" uncertainty and discuss these types of uncertainty in the context of our object of research, the financial market.

2 Chaotic systems are systems composed of many components or agents interacting with each other (Amaral and Ottino, 2004). These interactions are shaped by properties such as asymmetric causality and emergence (*ibid.*). Therefore, although such systems are theoretically thought to be deterministic, accurate predictions of their global outcomes are a non-trivial task. This property is concisely summarized in a quote ascribed to mathematician and meteorologist Edward Lorenz: "[in chaotic systems,] the present determines the future, but the approximate present does not approximately determine the future" (Danforth, 2015).

Table 1: An overview of conceptual pairs used to dichotomize uncertainty in the literature. The first column contains terms related to the worldly facet of uncertainty, and the second column those related to the mental facet of uncertainty.

Worldly	Mental	defined in
Aleatory	Epistemic	Hacking (1975)
External	Internal	Kahneman and Tversky (1982)
Ontological	Epistemological	Davidson (1996)
Random	Systematic	Taylor (1999)

2.2 UNCERTAINTY IN THE AGENCY DILEMMA

In the financial market, the relationship between investors and the management of companies is shaped by information asymmetry and conflicting interests. An important economic paradigm to conceptualize this relationship is the *agency dilemma* or principal–agent problem. This section introduces this framework and relates it to various sources of uncertainty. Figure 2 provides a draft of the assumed causality. §2.2.1 and §2.2.2 bisect the types of uncertainty occurring in this setting and relate them to the theoretical foundation established in §2.1. §2.2.3, finally, motivates the financial disclosures that are analyzed in this thesis.

On an abstract level, the agency dilemma describes any setting in which a principal commissions an agent to do work for them (Eisenhardt, 1989, p. 58). In this thesis, investors or banking analysts represent the principal, and the company management represents the agent. Due to conflicting interests and information asymmetry, both the compensation and information disclosure must ensure an optimal alignment of principal and agent incentives. Information asymmetry arises as the company management usually has a more complete picture of the current financial situation and might withhold unpleasant information from investors and the public. An example of conflicting interests is risk-taking behavior, as managers and investors usually have different goals and attitudes toward risk: Increased risk-taking usually translates to increased returns, which should satisfy investors. However, past research has shown that, e.g., the fear of job or reputation loss increases managers' risk aversion (Amihud and Baruch, 1981). While managers also have personal involvement in the firm, investors usually own diversified portfolios with many different companies, increasing their risk proneness (Jensen and Meckling, 1976).

Apart from these motivations to selectively share information, committing to definite and reliable statements is particularly challenging in financial disclosures. Frequently, the company management has

to make prognoses about the future of their business; however, the economy is a continuously evolving system whose outcomes are inherently hard to predict (Arthur, 2014). Furthermore, even in the face of challenging business situations, the management may be interested in receiving favorable analyst forecasts, for example, due to the self-fulfilling prophecy character of economic decision-making (Petalas et al., 2017). An inability to answer open questions spontaneously due to incomplete knowledge provides further reasons for using non-committal statements falling under the linguistic terminus of *hedging* (Lakoff, 1973).

This complex environment sets the stage for the theoretical model underlying this thesis: Figure 2 provides an overview of the specific kinds of aleatory and epistemic uncertainty that we focus on in the context of the agency dilemma: financial risk, linguistic uncertainty, and analyst uncertainty. The following introduces these concepts.

2.2.1 *Aleatory Uncertainty in the Agency Dilemma*

A financial key concept located on the aleatory side of uncertainty is *economic risk* coined by Knight (1921). It arises due to uncertainty about economic outcomes. Conceptually, this risk is thought to be quantifiable in the form of a probability distribution across these outcomes. Examples of economic risk include the rate of inflation, swings in exchange rates, or credit risk. Economy- or industry-wide developments in supply and demand[3] constitute another source of economic risk. In Figure 2, we furthermore consider environmental conditions (such as the weather, climate, or natural disasters) to be a driver of economic risk. Uncertain environmental conditions might especially influence fundamental industries close to processing raw materials (e.g., agriculture or mining). Finally, at the bottom left of Figure 2, analysts and investors invest in the economy as a whole or single companies. These investments are represented by $ signs. In an emergent manner, such investments aggregate to trading volume and price of financial instruments, which determines economic and financial risk.

FINANCIAL RISK In the taxonomy established in Figure 2, we define *financial risk* as a specific kind of economic risk capturing the uncertainty concerning financial returns of an investment or business. Most commonly, financial risk is defined as the quantifiable and known probability distribution of possible losses or gains. In the literature, the most important financial risk measure is *volatility*. In the broader sense, volatility describes the "variability of the random (unforeseen) component of a time series" (Andersen et al.,

3 An example for this are the worldwide supply shortages of Graphics Processing Units (GPUs) due to increasing cryptocurrency mining in the early 2020s (Linus, 2021).

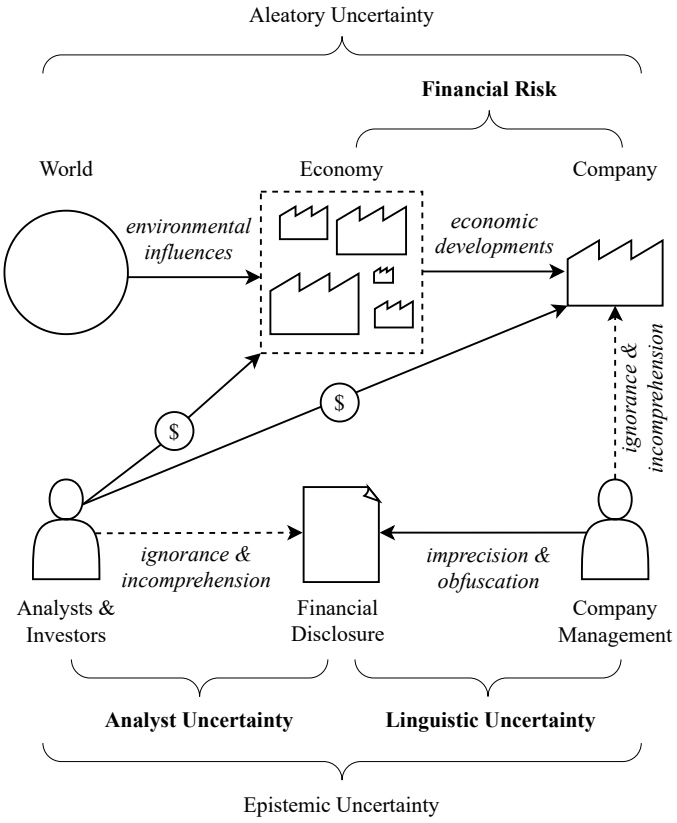


Figure 2: Sources of aleatory and epistemic uncertainty in the agency dilemma; figure created by the author. Uncertainty types studied in this thesis are boldfaced. In the top left corner, environmental influences affect the economy, whose developments in turn affect individual companies. The management's understanding of their companies is hindered by human *ignorance* and *incomprehension*. Based on this limited understanding, financial disclosures, who may contain **linguistic uncertainty** due to *imprecision* or *obfuscation*, are drafted. Analysts and investors read these disclosures, again with a restricted understanding due to mental limitations. If analysts interpret the financial situation of the underlying company differently or wrongly, this disagreement can be captured with measures of **analyst uncertainty**. Based on their understanding, analysts and investors conduct investments (denoted by $) into the economy and individual companies. In an emergent process, these investments determine economic and **financial risk**.

2006, p. 780). As such, the measure is applied in, e.g., presidential approval prediction, weather forecasting, or neuro-muscular activation modeling (*ibid.*, p. 797). Volatility is a crucial measure in financial decision-making (particularly asset or company valuation), thus making volatility prediction an essential task for finance and risk management (*ibid.*, p. 789). Conceptually, volatility proxies for investor uncertainty about the future development of firms (Doshi et al., 2021). This thesis focuses on two measurements of volatility outlined below.

Since Markowitz (1952), a common way of measuring volatility is the sample standard deviation of logarithmic stock returns. This measure indicates the possible spread or fluctuation of returns. Intuitively speaking, if the volatility is low, the underlying stock's price stays relatively constant over time. The second volatility measure investigated in this thesis is the error (uncertainty) of a regression model aiming to explain the movement of future stock returns. This is done by estimating market models (Sharpe, 1963) which forecast the future stock returns of a company in excess of overall market returns. In this case, volatility is defined as the Root Mean Squared Error (RMSE) of such a market model, where a large error indicates low predictability of returns. This measure factors in movements of the overall market and aims to capture the residual uncertainty on a firm-level over market uncertainty.

As shown in Figure 2, financial risk can both be a cause and an effect of linguistic or analyst uncertainty: Taking a top-down view, externalities such as the climate or evolving economic conditions might cause increases in financial risk; on the other hand, the investment behavior of individual market participants (i.e., buying and selling of specific assets) can also increase risk in an emergent manner. Financial disclosures have the function of mirroring these uncertainties and hence are indispensable information channels to market participants. If such disclosures contain obfuscated information or other residual uncertainty, they are less informative to shareholders, leading to more dispersed investment behavior. Successively, this increases return fluctuation and thus financial risk. In this thesis, we present works modeling the assumed causality of risk $\rightarrow$ linguistic uncertainty (Theil and Stuckenschmidt, 2020) or the other way around (Theil et al., 2019, 2020; Theil et al., 2023) to account for the bidirectional relationship between financial risk and linguistic uncertainty.

2.2.2 *Epistemic Uncertainty in the Agency Dilemma*

Sources of epistemic uncertainty are shown at the bottom of Figure 2: Usually, the management has the most complete understanding of the company's financial situation. This understanding, however, is capped by mental limitations and distorted by psychological biases. Furthermore, as described above, conflicting interests might motivate the management to communicate imprecisely or obfuscate disclosure information. Further epistemic uncertainty arises due to the ignorance and incomprehension of analysts and investors. Concerning these epistemic types of uncertainty, this thesis focuses on:

1. the incremental linguistic uncertainty introduced in financial disclosures; according to Figure 2, this can be thought of as a sender-side epistemic uncertainty

2. analyst uncertainty reflecting in more erroneous or dispersed forecasts; this is a receiver-side epistemic uncertainty

We now describe both epistemic uncertainty measures in more detail.

LINGUISTIC UNCERTAINTY Linguistic uncertainty can be divided into ambiguity and vagueness (Klir, 1987). Simply put, ambiguity is a problem of reference and vagueness one of boundaries (*ibid.*): A unit of language with more than one possible interpretation is called *ambiguous*; it is called *vague* if the meaning is clear but not the degree to which it applies. The focus of this thesis lies on the vagueness component of linguistic uncertainty. Vagueness is commonly exemplified with the sorites paradox (from Ancient Greek σωρός: the heap) which, in a shortened form, poses the question "How many grains of sand constitute a heap?" There is no fixed boundary or number of grains after which a non-heap turns into a heap. In natural language, adjectives such as "tall" are deemed vague, as their degree membership is subjective and demands clarification. The use of degree modifiers such as "relatively" or "quite" can induce additional vagueness. Closely related to vagueness, we explore *epistemic modality* and *hedging*. The former describes the degree of confidence a speaker has in the truth of their proposition (Portner, 2009, p. 47), e.g., "it might rain tomorrow." The latter describes the use of linguistic devices to communicate "a lack of complete commitment to the truth value" or cautionary language (Hyland, 1998; Lakoff, 1973), for example, due to ignorance or politeness (Danescu-Niculescu-Mizil et al., 2013).

Large shares of everyday language contain uncertainty (Konstantinova et al., 2012; Light et al., 2004). Motivations for using such language are: The stochasticity of our world (cf. the discussion of aleatory uncertainty above) which motivates the use of fuzzy quantifiers such as "maybe" or "almost certainly." Furthermore, incomplete knowledge or the complexity of the underlying issue might prompt speakers to use vague language (Kahneman and Tversky, 1982). For example, an early study in educational research found that students exposed to less informative lectures adopt vaguer language when asked to summarize what they have learned (Hiller, 1971).

In financial disclosures, forward-looking statements and spontaneous, unprepared answers necessitate the use of disclosure vagueness (Cazier et al., 2019; Dzieliński et al., 2021). Furthermore, uncertain language can be used to obfuscate information deliberately (Serra-Garcia et al., 2011) or to manipulate an audience (Meirowitz, 2005; Rogers, 2008). In line with this hypothesis, findings by Burgoon et al. (2016) show that companies use linguistic uncertainty strategically for deceptive purposes. Moreover, empirical studies suggest that CEOs tend to use vaguer language than CFOs (Burgoon et al., 2016,

p. 138), and male executives tend to use vaguer language than female ones (de Amicis et al., 2020).

Despite the theoretical appeal of linguistic uncertainty, the variable is comparably under-explored in the financial NLP community. Loughran and McDonald (2016) summarize that "[m]any textual analysis studies have focused on the simple positive/negative dichotomy of sentiment analysis" despite its "low power" (p. 1224). As an alternative, the authors propose investigating the construct of epistemic modality or linguistic uncertainty.

ANALYST UNCERTAINTY In Chapter 7, we explore two other epistemic uncertainty measures in a financial context. These measures revolve around analyst uncertainty and, more precisely, the error and dispersion of Earnings per Share (EPS) forecasts. Here, similar to volatility, we measure uncertainty internally as the standard deviation of forecasts, or externally as the error compared to the actual EPS value. The EPS are a key figure for shareholders since they measure the allocated profit per stock. Hence, analysts and investors base their trading decisions on this figure.

The first considered analyst uncertainty measure is Standardized Unexpected Earnings (SUE), measured as the mean absolute error of analyst EPS forecasts. The larger this value is, the less accurate were the projections of analysts on average. The second investigated measure is analyst dispersion, calculated as the standard deviation of analyst EPS forecasts. A large dispersion indicates little consensus among analysts. Conceptually, it could be expected that *ceteris paribus*, heightened linguistic uncertainty in financial disclosures increases the error and dispersion of analyst forecasts.

2.2.3 *Financial Disclosures*

This thesis focuses on regulatory financial disclosures instead of social media or news data. We chose financial disclosures as their content tends to be more truthful and objective.[4] This is due to the Generally Accepted Accounting Principles (GAAP), a set of accounting standards mandated by the United States Securities and Exchange Commission (SEC), which aim to ensure the correctness and completeness of financial disclosures. Hence, we eliminate the possible confounding factor of factuality which could dilute the results of an uncertainty detection or risk regression task.

Past literature has extensively analyzed two forms of financial disclosures: Form 10-K and earnings calls. We will also study these dis-

4 In line with this reasoning, the Canadian Securities Administrators (CSA) reviewed the social media disclosures of 100 public companies and found that almost 80% of them did not have guidelines for online disclosure practices (CSA, 2017). In particular, many of the analyzed disclosures were selective, imbalanced, or biased, and in some cases even untrue.

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.

Figure 3: Excerpt of a 10-K filed by a pharmaceuticals company in spring 1998. Phrases containing vague formulations (epistemic uncertainty) or referring to economic risk factors (aleatory uncertainty) are boldfaced. Figure from Theil et al. (2020).

closures types in this thesis. Presumably due to their easier availability, most research has revolved around 10-Ks. Furthermore, the state-of-the-art financial dictionary (Loughran and McDonald, 2011) was developed for this disclosures type. The following introduces both disclosure types.

10-Ks Form 10-K is an annual report that all publicly traded U.S. firms with above $10M in assets and more than 500 shareholders have to file. The report gives a comprehensive summary of a company's activities throughout the preceding year. A 10-K can contain up to 15 sections, with Sections 1 & 2 "Business and Property Description," Section 7 "Management's Discussion and Analysis (MD&A)," and Section 8 "Financial Statements" most commonly included (Dyer et al., 2017). Figure 3 presents a 10-K excerpt showing typical characteristics of these documents, such as linguistic uncertainty. This 10-K belongs to a relatively young and small company with "limited experience," thus explaining its business description's large share of uncertain wording. Past literature shows that 10-Ks are important informative disclosures for shareholders (Lehavy et al., 2011; Li, 2008) and can explain the uncertainty of the information environment (Li and Zhao, 2015; Loughran and McDonald, 2014). Importantly, the state-of-the-art lexical resource for financial NLP has been developed based on and targeted to form 10-K. This lexicon was developed by Loughran and McDonald (2011) and, given its importance for the domain in general and this thesis in particular, will be described in greater detail in the subsequent Chapter 3.

EARNINGS CALLS Earnings calls are quarterly public teleconferences and webcasts that typically occur shortly after disclosing the

EMMANUEL ROSNER (ANALYST): My first question is about your in-house cell manufacturing efforts. So in addition to building up capacity, some of the goals you highlighted was to cut the pricing or the cost by about 50%, boost the range by about 50% over a number of years. So wanted to know if your initial efforts are trending in that direction? What is sort of like the timeline to achieve these goals? And maybe related to this, how are you thinking about the time line for the cheaper Tesla, the entry model, eventually?

ELON MUSK (CEO): I think we feel very confident about achieving those targets, let's say, over a three-year time frame. I don't know it grew—it's not like year one. So three, maybe four years, give ourselves a little room. But for three or four years, I'd say.

ZACH KIRKHORN (CFO): Yeah. We put together the trajectory in the Battery Day, and we're on that trajectory still. I think that's probably the best reference for the cost trajectory that we are on.

ELON MUSK (CEO): Yeah. We're aspiring to do better than Battery Day, but we are confident of at least for doing what we presented at Battery Day.

Figure 4: Excerpt of Tesla's Q4 2020 earnings call Q&A.

EPS figure in the quarterly earnings report (10-Q). They are a tool for the company management to inform the public, most prominently investors and banking analysts, about its company's financial performance in the closing business quarter. Typically, one or more C-level executives—e.g., the CEO and the Chief Financial Officer (CFO)—along with investor relations representatives represent the company management. Apart from those and banking analysts, an operator participates, who takes care of technical requirements such as opening and ending the call or moderating the Q&A.

Earnings calls start with a scripted presentation by the management, which usually resembles the accompanying 10-Q closely and is thus relatively formal with little opportunity for the executives to speak freely. Afterward, a Q&A follows, in which the participating banking analysts can obtain more profound information about the company's economic position or prospects. Due to this structure, earnings calls are characterized by relatively spontaneous interaction (Larcker and Zakolyukina, 2012, p.499). Therefore, the disclosed information tends to be more novel than the one presented in a 10-K, which should already largely be priced into the underlying asset at the time of publication.[5] Another advantage is that earnings calls allow for a mapping of utterances to individual speakers. This is not possible for 10-Ks, which are written by a collective of company representatives. Therefore, earnings calls allow for an analysis of personal style or personality (cf. Chapter 8). Figure 4 contains a brief excerpt of Tesla's Q4 2020 earnings call displaying the linguistic features of

5 This assumption is supported by the fact that, on average, only about 30 investors download a 10-K immediately after its publication from the SEC's database Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) (Loughran and McDonald, 2017).

this disclosure type, such as speaker turns, colloquial language, or fragmented speech.

These characteristics make earnings calls an interesting disclosure type for detecting linguistic uncertainty such as hedging, modality, or vagueness. Past literature explored related phenomena such as indirectness (Crawford Camiciottoli, 2009), persuasion (Crawford Camiciottoli, 2011, 2018), or deception (Larcker and Zakolyukina, 2012) in earnings calls. That is also why (except for Chapter 7), we will focus on earnings calls in this thesis.

This concludes our discussion of the theoretical background of this thesis. In the following chapter, we will localize our work in the context of related literature.

3

RELATED WORK

* With the emergence of NLP methods in finance, the discipline gradually broadened its focus from the numerical or factual content of financial disclosures to formal choices in linguistic style that modulate such facts, for example, concerning linguistic tone (Loughran and McDonald, 2011; Tetlock, 2007) or uncertainty (Doshi et al., 2021; Dzieliński et al., 2021). This reflects in a growing number of literature surveys on financial NLP (Khadjeh Nassirtoussi et al., 2014; Loughran and McDonald, 2016, 2020; Xing et al., 2018).[1]

The following provides an overview of papers related to the three main tasks of this thesis: §3.1 covers the automatic detection of linguistic uncertainty (T_1) in finance and other domains; §3.2, which addresses the causality of uncertainty and risk (T_2), summarizes literature about decision-making under risk and uncertainty constraints; §3.3, finally, presents literature about text-based risk regression from 10-Ks and earnings calls (T_3).

3.1 UNCERTAINTY DETECTION (T_1)

This section discusses related work on automatic uncertainty detection. We start with approaches to uncertainty detection by financial researchers, which primarily focus on count- and dictionary-based methods.[2] In this regard, the discussion of the state-of-the-art resource, the Loughran and McDonald (2011) dictionary, is of particular interest for our work. Then, we move to supervised classifiers suggested by researchers of other domains, which motivated us to explore the performance of such an approach for our problem.

3.1.1 *Uncertainty Detection in Finance*

SIMPLE WORD COUNTS Two works have experimented with a concise set of manually selected words related to economic uncertainty.

* This chapter expands on the "Related Work" sections of Theil, Daube, and Stuckenschmidt (2022), Theil and Stuckenschmidt (2020), Theil, Štajner, and Stuckenschmidt (2020), Theil, Broscheit, and Stuckenschmidt (2019) and Theil, Štajner, Stuckenschmidt, and Ponzetto (2017).

1 Other literature surveys on the topic include: Das (2014), El-Haj et al. (2019), Fisher et al. (2010, 2016), Kearney and Liu (2014), Kumar and Ravi (2016), Li (2011), and Man et al. (2019).

2 If not otherwise mentioned, all of these works use cross-sectional event study regressions, i.e., multiple linear regression on lagged outcome variables with controls for time- and firm- or industry-fixed effects.

Li (2006) analyze risk sentiment in 34K 10-K filings spanning 1994–2004. They measure risk sentiment by counting the tokens "risk" and "uncertainty" (together with lexical variants), log1p-transforming the value, and taking the annual difference on a company-level. Using linear regression, they find that risk sentiment has a highly significant and negative effect on post-filing earnings and stock returns.

Baker et al. (2016) develop an economic policy uncertainty (EPU) index based on keyword matches in articles by the ten leading US-American newspapers (e.g., *Wall Street Journal* or *New York Times*) in the period 1985–2009. They count the monthly number of articles containing the terms "uncertain[ty]" and "economic" or "economy" together with at least one of the following terms: ["Congress," "deficit," "Federal Reserve," "legislation," "regulation," "White House"].[3] Baker et al. show that the such obtained count-based index is positively related to stock return volatility on a firm-level and negatively related to investment and employment in policy-related industries such as defense or finance.

THE LOUGHRAN MCDONALD DICTIONARY The *de facto* standard for financial sentiment analysis in general and uncertainty detection in particular was established with a comprehensive dictionary by Loughran and McDonald (2011).[4] This dictionary spans the categories POSITIVE, NEGATIVE, UNCERTAIN, LITIGIOUS, STRONG MODAL, and WEAK MODAL. The authors created those lists manually by inspecting the most frequent terms in a sample of 50K documents 10-K published between 1994 and 2008. They focus on terms appearing in at least 5% of the documents and aim to "create a relatively exhaustive list of words that makes avoidance [by managers] much more challenging." (Loughran and McDonald, 2011, p. 44). The underlying assumption is that if a specific term (e.g. "decline") correlates with decreased returns, this should incentivize the management to avoid it in their communication; avoidance becomes harder with a more exhaustive dictionary.

Out of these six lists, the most relevant for this thesis is the UNCERTAIN one. This list aims to capture "the general notion of imprecision rather than exclusively focusing on risk" (Loughran and McDonald, 2011, p. 45). Containing a total of 297 terms, it thus covers both the aleatory and the epistemic facet of uncertainty inspected in this thesis: The aleatory uncertainty or risk words include, e.g., "anomaly," "fluctuate," "volatility," and "risky." The epistemic or linguistic uncertainty side is represented by words such as "almost," "may," "possible," and "suggest." Using their own-developed sentiment dictionary, Loughran and McDonald (2011) find that the cumulative term frequency–inverse document frequency (tf–idf) of UNCERTAIN words

3 The method also accounts for lexical variants such as "uncertainties" or "Fed."

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

in 10-Ks is positively and highly significantly related to future stock return volatility. To allow comparisons to the state-of-the-art, this dictionary is investigated as a feature for uncertainty detection and risk regression in Chapters 4–7.

As a subset of the of the Loughran and McDonald (2011) UNCERTAIN dictionary, WEAK MODAL consists of 27 terms focusing on the epistemic (i.e., vagueness) component of uncertainty. Example terms include "may," "might," "depending," or "possibly." The STRONG MODAL list, on the other hand, covers the opposite linguistic aspect (i.e., certainty), which is measured with 20 tokens such as "always," "definitely," "must," or "will." We will investigate these lists in greater detail within our work on modality in financial disclosures (Chapter 5).

RECENT WORKS ON UNCERTAINTY IN FINANCE In the following, we discuss the most recent financial papers on uncertainty detection. Although we do not consider them in our work on the matter published prior (Theil et al., 2017; Theil and Stuckenschmidt, 2020), their discussion seems prudent to present a complete picture of the recent developments in the scientific community.

Doshi et al. (2021) apply the Loughran and McDonald (2011) UNCERTAIN dictionary to a sample of 25K forms 10-K and 10-Q. They find that an uncertain tone positively impacts the price spread of credit default swaps and asset volatility. This effect is the strongest for firms with high leverage and shorter maturities.

Dzieliński et al. (2021) analyze the linguistic clarity (measured by the absence of UNCERTAIN words) of CEO communication in a sample of 105K earnings calls and 6K CEOs. They find that clarity is independent of the company's business uncertainty in terms of fundamental variables and thus reflects a CEO's personal communication style. Another finding is that stock prices and earnings forecasts react stronger to clear CEOs; the authors interpret this as a dislike of investors and analysts for uncertain language. Lastly, they show that companies with clear CEOs tend to receive more positive earnings forecasts and larger valuations.

Related, Barth et al. (2021) explore a set of 40K answered questions in 2.1K earnings calls from 2002–2019. Using a financial jargon dictionary[5] as a proxy for precise language, they observe that NEGATIVE questions according to the Loughran and McDonald (2011) dictionary are usually answered less precisely. Empirically, imprecise answers seem to be interpreted negatively by investors as they are met with decreased Cumulative Abnormal Returns (CARs) and increased implied volatility.

5 https://people.duke.edu/~charvey/classes/wpg/bfglosm.htm

3.1.2 *Uncertainty Detection in Other Domains*

Researchers in domains outside of finance have experimented with supervised classifiers for linguistic uncertainty detection to complement purely dictionary-based approaches. This inspired us to explore the applicability of supervised models in the financial domain (cf. Part II) while incorporating the established Loughran and McDonald (2011) as a feature. We now describe these approaches in greater detail.

BIOMEDICAL DOMAIN Light et al. (2004) create six gold standard datasets with a total of 3.4K annotated sentences sampled from biomedical paper abstracts. They annotate sentences with ternary labels (*high speculative*, *low speculative*, and *definite*) and represent them as unweighted count vectors of stemmed terms. In a 10-fold Cross-Validation (CV) setting, a substring matching approach with 14 manually selected speculation cues (e.g., "suggest," "at least," and "putative") slightly outperforms a majority class baseline and a Support Vector Machine (SVM) classifier with Bag-of-Words (BoW) vectors.

Medlock and Briscoe (2007) create a gold standard test set of 1.5K sentences sampled from genomics papers and annotate them with binary labels (*speculative* and *non-speculative*). They train a weakly supervised classifier on 300K unlabeled sentences. This classifier automatically selects the keywords maximizing the probability $P(\text{spec}|x)$ of a sentence belonging to the *speculative* class. The classifier outperforms both a linear SVM and the approach by Light et al. (2004).

Exploring the applicability of a maximum entropy classifier to Medlock and Briscoe's dataset, Szarvas (2008) show that including bi- and trigrams into the BoW representation further increases the performance in terms of F_1 score. As an additional contribution, they also experiment with a feature selection method just retaining such terms with a frequency $\geq 5{\times}10^{-5}$ and a conditional probability of $P(\text{spec}|x) > 0.94$ for the *speculative* class.

Zerva (2019) detect uncertainty in mentions of molecule interactions within biomedical papers. Their definition of linguistic uncertainty is similar to epistemic modality (Portner, 2009), as they measure it as an author's degree of confidence in the truth of a statement. Their best approach is rule-based and uses dependency n-grams enriched with the corresponding dependency type (e.g., preposition, object, or participle construction) as features. Confirming findings by Rubin (2007), they show that a binary distinction into certain and uncertain texts is preferable to a multi-level measurement, owing to the increased subjectivity of the latter.

Therefore, in our works on uncertainty detection (Part II), we also consider a binary classification of uncertainty. However, we acknowl-

edge that a multi-class or continuous representation would be an interesting (if challenging) application for future work.

ENCYCLOPEDIC DOMAIN Uncertainty detection has also been investigated in the context of Wikipedia articles. Although these articles are formally and contentually different from financial disclosures, the works studying them feature publicly available lists of uncertainty cues, which motivated us to explore their applicability to our task (Chapter 4).

Ganter and Strube (2009) introduce the first classifier for encyclopedic uncertainty detection. As training data, they sample 169K *weasel*-tagged sentences with 457 *weasel* *n*-grams (e.g., "I think," or "it has been suggested") from Wikipedia dumps in 2006–2008. Four annotators co-annotated two sets of 500 and 246 sentences with a binary uncertainty score as validation and test data. The proposed models are: (1) a term weighing method considering both the relative frequency and their average distance of a *weasel* word; (2) Part-of-Speech (PoS) tag–based syntactic patterns and finite-state automata automatically extracted from the test set. While both methods perform comparable on the validation set, the syntactic method outperforms the corpus statistics method by a large margin on the test set. Ganter and Strube (2009) hypothesize that this is due to the syntactic method detecting previously unidentified weasel tags in the training data.

Subsequently, the CoNLL-2010 shared task (Farkas et al., 2010) led to the resurgence of uncertainty detection in the biomedical and encyclopedic domains. 23 submitted classifiers were benchmarked on the BIOSCOPE corpus (Vincze et al., 2008) and a set of 4.5K Wikipedia paragraphs with *weasel* tags. Therein, an extensive array of features was explored: dictionaries, orthographic token information, lemmas/stems, PoS tags, syntactic chunk information, dependency parsing, and relative token position within the document. Evaluated algorithms include: Sequence Labeling (SL), token classification, and BoW approaches. The best-performing biomedical system uses an SL approach, while the best classifier for the encyclopedic domain is an SVM with dictionary features.

The findings of this task motivated us to also explore the performance of an SVM with dictionaries (including the CoNNL-2010 *weasel* list) on our dataset and to benchmark the generalizability of our classifier on the CoNLL-2010 test set (Chapter 4).

MONETARY POLICY DOMAIN Štajner et al. (2017) explore speculation detection in the monetary policy domain using a dataset of Federal Open Market Committee (FOMC) DEBATES and policy DECISIONS transcripts. Treating the problem as a binary sentence classification task, they use an SVM with the following features: BoW vectors, the uncertainty triggers of the CoNLL-2010 training set (Farkas

et al., 2010), the UNCERTAIN lexicon developed by Loughran and McDonald (2011), and an own-developed list of 85 speculation triggers (e.g., "a number of," "fairly," and "suppose"). Furthermore, they experiment with a Convolutional Neural Networks (CNNs) using BoW vectors. These two classifiers are benchmarked on three test sets: the CoNLL-2010 WIKIPEDIA set, 130 annotated sentences (44 *speculative*) from the DEBATES transcripts, and 139 annotated sentences (71 *speculative*) from the policy DECISIONS closing the FOMC meetings. The best performing approaches are as follows: for WIKIPEDIA, the CNN yielded the largest F_1 score with 0.50, compared to 0.45 with the SVM. For DEBATES, the SVM with BoW and debate triggers yielded $F_1 = 0.65$. On DECISIONS, finally, the CNN achieved $F_1 = 0.72$ (vs. 0.58 with the SVM). For the feature-engineered approach, especially the domain-specific dictionary led to performance gains.

As this work can be considered the most closely related to our efforts on uncertainty detection, we also explore the performance of its suggested feature sets in our classifier (Chapter 4).

3.2 UNCERTAINTY AND RISK (T_2)

This section presents related literature to our second task, causal modeling of uncertainty and risk (T_2). Although no work has specifically investigated the causal effect of linguistic uncertainty on perceptions of risk, there are behavioral economic studies analyzing decision-making under numerical uncertainty constraints. We now discuss the existing body of evidence in more detail and discuss relations to our work on the matter (Chapter 6), where applicable.

CHOICES UNDER UNCERTAINTY AND RISK A key phenomenon regarding decision-making under numerical uncertainty is the *Ellsberg paradox* of choice (Ellsberg, 1961). This paradox was observed in a thought experiment, in which participants had to envision the following setup:

> Suppose there are two urns, from one of which a ball will be drawn. Each urn contains 100 balls. Urn 1 contains 50 black and 50 red balls; urn 2 contains an unknown distribution of red and black balls. Four bets are possible: red_1, black_1, red_2, and black_2, where red_1 means to draw from urn 1 and to receive \$100 if the outcome is red and \$0 otherwise.[6]

Then, participants were asked to indicate their preferences regarding the following four choices: (1) red_1 vs. black_1, (2) red_2 vs. black_2, (3) red_1 vs. red_2, and (4) black_1 vs. black_2.

6 Experimental setup paraphrased from Ellsberg (1961).

Empirically, participants were usually indifferent in cases (1) and (2). This result is consistent with traditional *expected utility theory*[7] (von Neumann and Morgenstern, 1953), as the expected value for both bets in (1) is $50 and unknowable for the bets in (2). For cases (3) and (4), however, red_1 and black_1 (i.e., the safe bets) were preferred by a majority of participants. This result is inconsistent with expected utility theory because if red_1 is preferred over red_2, a subject should expect less than half of the balls in urn 2 to be red. As this implies that more than half of the balls in urn 2 are black, the subject should also prefer black_2. Hence, Ellsberg (1961) concludes that expected utility theory only applies to choice problems under risk and fails to account for choice problems under uncertainty. Rather, individuals usually prefer certain over uncertain bets—in some cases, even if the certain bet yields lower utility.

Building on the efforts by Ellsberg (1961), a major theoretical contribution was brought forth in the development of *prospect theory* (Kahneman and Tversky, 1979). Different from expected utility theory, it asserts that the utility across losses and gains does not follow a linear but rather an asymmetric S-shaped function. This implies that for some individuals, the negative utility from losing $100 can only be outweighed by the utility of earning $200. This phenomenon is called *loss aversion* (Kahneman and Tversky, 1979). In a follow-up refinement of prospect theory, Tversky and Kahneman (1986) call the irrational preference for bets under certainty the *certainty effect*.

A large body of subsequent research has found empirical support for the principle of loss aversion and the certainty effect (Diecidue and van de Ven, 2008; Holzmeister et al., 2020; Klos et al., 2005; Zeisberger, 2021, *inter alia*). Taken together with the evidence presented in §1, we assume that the findings regarding numerical uncertainty perceptions should port to perceptions of linguistic uncertainty. Hence, we expect financial disclosures with uncertain language to have a detrimental causal effect on investments in the underlying asset (Chapter 6).

INDIVIDUAL DRIVERS OF RISK PERCEPTION Apart from these fundamental principles, which individual characteristics affect perceptions of risk? Findings by Heath and Tversky (1991) suggest that investors with high self-perceived domain knowledge and competence may systematically underestimate risk. Related, Diacon (2004) and Sachse et al. (2012) show that experts generally have a lower risk perception than laypeople. In our laboratory study (Chapter 6), we

7 Expected utility theory (EUT) is an economic decision-making framework for investments under uncertainty and risk. It assumes individuals to maximize their expected utility, which is defined as an expected loss or gain multiplied by its outcome probability (von Neumann and Morgenstern, 1953). Crucially, EUT assumes a linear utility function such that a loss of $100 can be outweighed by a gain of $100.

therefore expect investors with high self-assessed financial literacy to have a decreased risk perception.

Moreover, studies show that various personality traits influence risk perception. Analyzing a sample of financial adviser clients ($n = 364$), Nguyen et al. (2019) show that risk perception is lower in male, high-income, and young individuals. Furthermore, they find that risk perception is impacted by individual risk tolerance. Surveying 342 business students, Oehler and Wedlich (2018) find that risk perception is also influenced by Big 5 personality[8] (Goldberg, 1990): High extraversion combined with low conscientiousness and low neuroticism (i.e., high emotional stability) are associated with decreased risk perception.

These findings motivated us to not only explore linguistic uncertainty and risk perception in our laboratory study, but also their interaction with financial literacy, gender, income, age, risk tolerance, and Big 5 personality (Chapter 6).

3.3 RISK REGRESSION (T_3)

Finally, we move our attention to studies related to our last task: risk regression (T_3) from 10-Ks and earnings calls. A large number of papers explore predictions of risk (gauged with volatility) from these financial disclosures. The following provides an overview of the central findings and developments in the literature.

3.3.1 *10-Ks*

Due to their wide-spread availability on the SEC's database EDGAR, 10-Ks are a popular source for text-based risk regression. Most past works consider BoW vectors and the Loughran and McDonald (2011) dictionary or variants thereof as features.

An early study often used as a reference point by subsequent work was conducted by Kogan et al. (2009). They collect a corpus of 60K 10-Ks spanning years 1996–2006 to predict stock return volatility in the post-filing year with a linear Support Vector Regression (SVR). Evaluating performance in terms of Mean Squared Error (MSE), they report a significant performance increase of a model incorporating BoW vectors over a baseline consisting only of the volatility in the preceding year. The best-performing feature representation is achieved with log1p-transformed bigram tf–idf vectors.

Using the same task specification, dataset, and regression model, Wang et al. (2013) find that a BoW representation based on the entire vocabulary of the reports performs comparably to one where only

8 The Big 5 personality traits are an established psychometric model (McCrae and Costa, 1989; McCrae et al., 2010) evaluating personality on the five continuous scales *openness*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism*.

the terms of the Loughran and McDonald (2011) dictionary are retained. They show that this result still holds when using a ranking SVM algorithm (Joachims, 2006) and evaluating it in terms of the rank correlation coefficients Spearman's ρ and Kendall's τ.

In a follow-up work, Tsai and Wang (2014) find that, for the same task and data, the predictive performance can be further improved by expanding the Loughran and McDonald (2011) dictionary using word embedding models. They expand the dictionary by training a `word2vec` model with Continuous Bag-of-Words (CBOW) architecture (Mikolov et al., 2013) on the 10-K corpus and retrieving the twenty most cosine similar terms to each dictionary term. Their best-performing expansion consists of term–PoS tuples and has 13K entries.

Re-investigating the task, Tsai et al. (2016) show that the same dictionary expansion predicts not only return volatility but also post-event volatility, abnormal trading volume, and CAR to varying degrees. Post-event volatility is measured with the Fama–French 3-factor model, which estimates volatility in excess of market risk, firm size, and firm valuation (Fama and French, 1993); abnormal trading volume is defined as the excess traded stock in days $[t_0, t_3]$ compared to the past quarter $[t_{-65}, t_{-6}]$; CAR, finally, is the company's stock return in excess of the overall market return in days $[t_0, t_3]$. The results of this experiment are decisive for all dependent variables except CAR.

Rekabsaz et al. (2017) outperform all of the prior approaches by including additional financial features and contrasting different term weighting and feature fusion methods. Their dictionary expansion method is similar, yet focuses on the POSITIVE, NEGATIVE, and UNCERTAIN lists and a set of 8.5K 10-Ks from 2006–2015. Their best model is a stacked SVR learned on two separate support vector regressors, all with Gaussian kernel: one is trained on BoW vectors weighted with Okapi Best Matching 25 (BM25) and the other on a vector of current volatility, Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model predictions (Bollerslev, 1986), and eleven NASDAQ sector[9] dummies. Apart from the previously explored MSE, they also benchmark the coefficient of determination R^2 and achieve a value of 0.53.[10]

These works motivated us to explore a similar automatic expansion of the Loughran and McDonald (2011) UNCERTAIN dictionary on our dataset of 10-Ks. As an unsupervised expansion expectedly contains irrelevant terms negatively affecting the results, we were also interested in experimenting with a semi-supervised term filtering method (Chapter 7). Furthermore, we consider above-mentioned works (Ko-

9 `https://www.nasdaq.com/market-activity/stocks/screener`

10 On a cautionary note, R^2 only is an adequate measure for evaluating non-linear regression if the distribution of prediction errors is approximately normal (Kvalseth, 1985; Magee, 1990; Spiess and Neumeyer, 2010).

gan et al., 2009; Rekabsaz et al., 2017; Tsai and Wang, 2014) as baselines for our DL-based risk regressor (Chapter 9).

3.3.2 *Earnings Calls*

In recent years, a growing number of NLP papers have leveraged the content of earnings calls for predictive models of volatility. In this regard, a central publication for this thesis is the one by Wang and Hua (2014), as it was the only existing one at the time when we investigated this task (Theil et al., 2019). Nonetheless, subsequent studies have experimented with interesting applications such as multimodal models exploiting speech, which may serve as methodological inspirations for future work.

Using a dataset of 11K earnings call transcripts, Wang and Hua (2014) propose a semiparametric Gaussian copula model for volatility prediction in days $[t_0, t_5]$. As features, they consider uni- and bigrams, PoS tags, Named Entity (NE) tags, and frame-level semantic annotations. Evaluating model performance in terms of Pearson's r, Spearman's ρ, and Kendall's τ, they report a significant performance increase of the Gaussian copula over a linear regression, a linear SVM, and a Gaussian SVM.

Wang and Hua's study motivated us to explore the following in our 2019 work on the same task: (1) comparing a model based on Feed-Forward Neural Networks (FNNs) to their approach; (2) re-assessing the task at a large scale by assembling a new dataset of 90K call transcripts; and (3) exploring a model jointly learning from financial data and contextualized language representations (Chapter 9). We now discuss more recent investigations of risk prediction from earnings calls, which were published during or after our work concerning this task.

Keith and Stent (2019) gather 12K earnings call transcripts and find that pragmatic and lexico-semantic features are moderately predictive of analysts' price forecast targets following the call dates. Investigated pragmatic features include an n-gram dictionary of 118 hedging terms (e.g., "basically," "kind of," "more or less") by Prokofieva and Hirschberg (2014) and the Loughran and McDonald (2011) dictionary; lexico-semantic features include BoW and `doc2vec` (Le and Mikolov, 2014) vectors. For a regression on analyst price targets, the best performing model in terms of MSE and R^2 is a ridge regressor learned on `doc2vec` vectors. For a ternary classification of price targets, a logistic regression with BoW vectors works best in terms of Accuracy and F_1 score.

Ye et al. (2020) collect 6.5K earnings calls between 2015 and 2018. On this set, a multi-round Q&A attention network improves over the sparse approach by Rekabsaz et al. (2017) and PRoFET, a model based on Bidirectional Long Short-Term Memories (BiLSTMs) with at-

tention (Theil et al., 2019).[11] They conclude that the volatility prediction model strongly benefits from a contextualized text representation as it can capitalize on the Q&A turn structure and the dependence of Q&A answers on preceding questions.

MULTIMODAL MODELS Qin and Yang (2019) are the first to explore multimodal volatility regression from earnings call transcripts and speech recordings. Their dataset,[12] which consists of 579 transcripts and recordings from the year 2017, constitutes an often-used testbed by subsequent work. The proposed approach, a multimodal deep regression model (MDRM), is based on contextual BiLSTMs and learns jointly from 300-dimensional GloVe (Pennington et al., 2014) embeddings and a 26-dimensional vector of sparse audio features extracted with PRAAT (Boersma and van Heuven, 2001). Predicting volatility between t_0 and $t_n \in \{3, 5, 15, 30\}$, MDRM outperforms sparse baselines and a state-of-the-art multimodal model (Poria et al., 2017) significantly in terms of MSE.

Yang et al. (2020) re-examine the dataset and task suggested by Qin and Yang (2019). As architecture, they propose a hierarchical Transformer-based multi-task learning model (HTML). For the multi-task learner, volatility prediction in the four windows suggested by Qin and Yang (2019) is used as the main task, and logarithmic stock return prediction from t_0 to t_1 as the auxiliary task. Similar to the previous approach, audio features are represented with PRAAT (Boersma and van Heuven, 2001), but textual representations are based on Whole Word Masked BERT (Devlin et al., 2019). Yang et al. (2020) show that HTML outperforms MDRM (Qin and Yang, 2019) and a set of econometric, sparse, and DL baselines.

Most recently, the data and task proposed by Qin and Yang (2019) were re-assessed by Sawhney et al. (2020). While main and auxiliary tasks are the same as Yang et al.'s, the used architecture is based on graph convolution networks. Other differences include a text representation with FINBERT embeddings (Araci, 2019) and adding past volatility as a vector of n-day average volatilities with $n \in \{-30, -2\}$. For the multimodal prediction, the proposed method outperforms a past volatility baseline, a BiLSTM (Poria et al., 2017), MDRM (Qin and Yang, 2019), and HTML (Yang et al., 2020) for all windows in terms of MSE and R^2.

11 Note that Ye et al.'s replication focuses on PRoFET's text representation. However, the full model features both a textual and a financial learner, the latter of which contributes the largest share of overall predictive power (Chapter 9, Figure 17a).

12 https://github.com/GeminiLn/EarningsCall_Dataset

Part II

UNCERTAINTY DETECTION

This part introduces methods for automatic **uncertainty detection (T_1)** in finance. In particular, it explores the questions: How can we detect linguistic uncertainty automatically in financial disclosures? What are its economic and linguistic determinants? Chapter 4 introduces a binary sentence classifier to detect uncertainty based on lexico-syntactic features. In Chapter 5, we present a classifier to predict linguistic uncertainty in a financial Q&A setting based on the lexico-semantic content of the preceding question and financial features of the discussed company.

4

LINGUISTIC UNCERTAINTY DETECTION IN FINANCIAL DISCLOSURES

* The past chapters have discussed the important role of linguistic uncertainty in the agency dilemma and its assumed influences on financial risk via analyst and investor perceptions. As a first step, therefore, we were interested in developing systems to automatically detect linguistic uncertainty in financial disclosures (T_1). So far, such systems have been developed in the biomedical (Light et al., 2004; Medlock and Briscoe, 2007; Szarvas, 2008), the encyclopedic (Farkas et al., 2010; Ganter and Strube, 2009), and the monetary policy (Štajner et al., 2017) domains, yet lack in finance. Hence, we were motivated to close this research gap and to investigate which features and algorithms used in related work could be transferred to this domain. In particular, we wanted to know which other features apart from the state-of-the-art approach in finance, the Loughran and McDonald (2011) dictionary, could work for this task.

4.1 INTRODUCTION

The automatic detection of linguistically uncertain statements can benefit NLP tasks such as deception detection (Bachenko et al., 2008; Fitzpatrick et al., 2015; Larcker and Zakolyukina, 2012), information extraction (Medlock and Briscoe, 2007; Szarvas, 2008), and summarization (Riloff et al., 2003). Furthermore, automatic uncertainty detection can enable new analyses in social sciences where the quantification of uncertainty or risk plays a substantial role. Disciplines like business and economics would profit from an automatically extractable measure of uncertainty that does not depend on manual analysis. As of now, automatic uncertainty detection has been limited to detecting hedges (as opposed to our broader concept of uncertainty) in biomedical scientific texts and Wikipedia articles.

This chapter addresses this issue and explores the automatic classification of linguistic uncertainty in a financial context. As we plan to explain or predict market reactions in future work, we establish a definition of uncertainty that extends the concept of *hedging*, established by Lakoff (1973) and Hyland (1998). *Hedging* is defined as "any linguistic means used to indicate either (a) a lack of complete com-

* This chapter is based on "Automatic Detection of Uncertain Statements in the Financial Domain" (Theil, Štajner, Stuckenschmidt, and Ponzetto, 2017), presented in April 2017 at the *18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)* in Budapest.

mitment to the truth value of an accompanying proposition, or (b) a desire not to express that commitment categorically" (Hyland, 1998, p. 1). As we aim to capture not only the intention of managers but also the reception of their utterances by shareholders, we expand this definition. Thus, we consider statements as uncertain if they meet the following criteria:

- They reveal ignorance of the speaker
- Their truth value cannot be determined (e.g., statements about the future)
- They refer to uncertain factors in the real world (e.g., statements about market volatility)

4.1.1 *Contributions*

This work is the first to explore the automatic classification of linguistic uncertainty in the financial domain. In doing so, we gather and annotate a new dataset of executive remarks with binary uncertainty labels. Broadening the definition of "linguistic hedging," we introduce a new concept of uncertainty fitting the domain-specific needs. In contrast to previous work, our definition encompasses how uncertain statements can impact other social agents and thus enables predictions of market reactions.

As we aim to benchmark the performance of different feature sets and to explore their domain-dependence, we pose the following two research questions:[1]

- **RQ1**: Which linguistic features work best for uncertainty detection in financial disclosures?
- **RQ2**: How well do the features of our financial uncertainty detector port to the encyclopedic domain?

4.2 DATA

In this work, we analyze earnings calls, a spoken kind of financial disclosure consisting of a presentation and a Q&A.[2] Since the presentation usually follows the accompanying press release closely, it is highly formalized and provides little opportunity for the executives to speak freely. Hence, our analyses focus on the second part of the call, the Q&A. Moreover, as we were interested in gaining insight into the company's financial uncertainty itself, we focus on the answers uttered by executives instead of the analyst questions.

1 Different from the published version of this chapter, we collapsed the original four research questions into two for clarity. This is a purely formal choice with no influence on our subsequent analyses or findings.

2 See §2.2.3 for an in-depth discussion of earnings calls.

Since we analyze free speech instead of written, formalized text, we expect our problem to be more challenging to solve than, e.g., classifying biomedical or encyclopedic sentences. Consider, for example, the following statement:

Example 4.1 *"And increasingly look as you are sort of describe us [sic] as well, we look to focus where we can really make a difference [...]"*

Despite containing a hedge, this sentence was annotated as *certain*, according to our methodology. The first part, "as you [...] sort of describe us," is a colloquial formulation in which the hedge "sort of" is used as a filler with no speculative meaning. In contrast, consider the following example:

Example 4.2 *"Now, what we don't know is what's going to happen at the end of the third quarter."*

While this sentence does not contain any hedges such as adverbials of degree or of possibility, it indicates a lack of knowledge of the speaker, which is why we annotated it as *uncertain*. These examples may serve to illustrate the complexity of automatic uncertainty detection in spoken financial disclosures.

As a basis for the dataset, we used the Standard and Poor's 500 (S&P 500),[3] which is one of the most essential equity indices. In the next step, we obtained all earnings call transcripts of the S&P 500 companies available on the financial database SeekingAlpha. This yielded a dataset of 7,725 transcripts of 217 different companies hailing from a wide array of industries such as finance, manufacturing, or information technology.

From this set of transcripts, we randomly sampled 1,800 sentences and annotated them with *certain* and *uncertain* labels. We excluded habitual utterances such as greetings from the sampling process, as they are irrelevant for uncertainty detection. Out of the 1,800 sentences, 100 were randomly sampled and independently annotated by a second annotator of financial background. The Inter-Annotator Agreement (IAA) measured as Cohen's κ (Cohen, 1960) was 0.81, which we deemed sufficiently high compared to related work (Ganter and Strube, 2009; Štajner et al., 2017) and established magnitude guidelines (Fleiss et al., 2003; Landis and Koch, 1977). Hence, the remainder of the annotation was completed by the first annotator of linguistic background. Afterwards, the set was split in two: We used 800 sentences (683 *certain*, 117 *uncertain*) to develop a set of lexico-syntactic uncertainty rules (see §4.3.2) and the remaining 1,000 (829 *certain*, 171 *uncertain*) as gold standard in the classification experiments.

3 http://us.spindices.com/indices/equity/sp-500

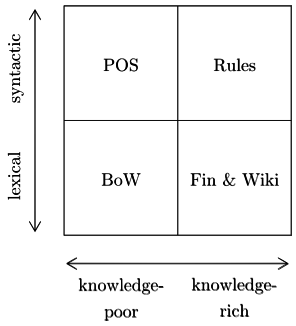


Figure 5: 2x2 matrix containing all feature sets used by our binary sentence classifier. Feature sets are categorized according to the dimensions knowledge-intensity (x) and lexicality vs. syntacticity (y).

4.3 METHODOLOGY

We addressed the problem of automatic uncertainty detection as a binary sentence classification task. In addition to BoW vectors, PoS tags, and the uncertainty cues proposed in related work (Farkas et al., 2010), we explored the following novel features: Due to its domain-specificity for finance, we used the uncertainty dictionary by Loughran and McDonald (2011). In addition, we applied a set of lexico-syntactic rules created by us. All features can be classified along lexical vs. syntactic and knowledge-poor vs. knowledge-rich dimensions, yielding a feature set matrix as depicted in Figure 5.

We lemmatized the BoW vectors with the Natural Language Toolkit (NLTK) 3.2.1's `WordNetLemmatizer` (Bird et al., 2009) and normalized them via tf–idf weighting. Additionally, we extracted PoS tags with NLTK's standard PoS tagger. §4.3.1 and §4.3.2 further elaborate on the features used in our approach.

4.3.1 *Uncertainty Dictionaries*

Within the experiments, we used the following uncertainty dictionaries:

- **Fin**: The UNCERTAIN dictionary developed by Loughran and McDonald (2011).[4] This dictionary was manually extracted from a sample of 10-Ks and contains 297 unigrams indicating uncertainty in the financial domain (e.g., "fluctuation," "recalculation"). After lemmatization, the list totaled 192 items.

- **Wiki**: 1,984 uncertainty triggers of arbitrary length (e.g., "a matter in dispute," "some prehistoric cultures") were extracted from the CoNLL-2010 shared task's (Farkas et al., 2010) Wikipedia

4 `http://www3.nd.edu/~mcdonald/Word_Lists.html`

Table 2: Categorization of the lexico-syntactic rules ($n = 95$).[6]

Category	n	Example
Expectation	29	"I expect our maintenance capital [...] to probably be"
Assumption	25	"I think it's pretty mature"
Probability	12	"perhaps by the end of this year"
Ignorance	10	"we really don't know what [...] it's going to sell"
Subjunction	9	"it might be a few hundred thousand dollars"
Risk	6	"the volatility of where we are"
Unspecificity	4	"somewhere in the 40% range"

training set.[5] After lemmatization, the list totaled 1,868 unique items.

4.3.2 *Rules*

We developed a set of 95 lexico-syntactic rules according to which a sentence can be classified as *uncertain* based on 800 randomly sampled earnings call sentences. These rules are defined by syntactic (PoS tags, phrase chunks) and lexical features (lemmas, dictionaries). The dictionaries define more granular word classes such as *adverbs of degree* (e.g., "kind of," "quite"), *adverbs of probability* (e.g., "potentially," "probably"), *fuzzy quantifiers* (e.g., "about half of," "close to 100"), and *verbs of expectation* (e.g., "anticipate," "expect"). Table 2 categorizes all rules according to seven categories of uncertainty defined by us (e.g., "assumption," "ignorance," or "risk"). According to our methodology, *Example 2* ("we don't know [...] what's going to happen") discussed in §4.2 belongs to the "ignorance" category.

In addition, we applied the rules to 30 random samples of 1,000 Wikipedia test sentences (Farkas et al., 2010) to test their applicability for a general domain as opposed to our domain-specific dataset. To ensure comparability, we sampled the sentences in a stratified manner, thus ensuring the original class distribution of our dataset (*certain*: 82.9%, *uncertain*: 17.1%). The results of this experiment are shown in Table 3.

Empirically, our rules match substantially fewer sentences on the Wikipedia test set ($\overline{x} = 9.70$) than on our financial domain dataset (54 matches). This is an expected result given the formal and contentual difference between both domains. However, this provided us with a first indication that the rule-based approach would probably underperform in terms of recall on the Wikipedia set.

5 http://rgai.inf.u-szeged.hu/conll2010st/download.html

6 Different from the published version of this chapter, "uninformedness" was renamed to "ignorance" and "volatility" to "risk."

Table 3: Descriptive statistics of rule matches on 30 random samples of the Wikipedia test set. $\hat{x}$ is the mode and g_1 is the Fisher–Pearson coefficient of skewness.

$x_{\min}$	$x_{\max}$	$\bar{x}$	$\tilde{x}$	$\hat{x}$	s_x	g_1
3.00	15.00	9.70	9.50	8.00	2.76	0.04

4.3.3 *Experiments*

We transformed each sentence into a vector of feature occurrences. Afterwards, we applied the following Machine Learning (ML) algorithms in the Waikato Environment for Knowledge Analysis (WEKA) Experimenter (Hall et al., 2009) using a 10-fold CV with 10 repetitions: Logistic Regression (le Cessie and van Houwelingen, 1992), Naïve Bayes (John and Langley, 1995), SVM (Platt, 1999) with a first-order polynomial kernel, k-Nearest Neighbors (Aha et al., 1991), Repeated Incremental Pruning to Produce Error Reduction (RIPPER) (Cohen, 1995), C4.5 decision tree (Quinlan, 1993), and Random Forest (Breiman, 2001).[7] We evaluated the performance for all feature sets used in the subsequent experiments and compared the weighted average F_1 scores. Since the SVM achieved the best results in all cases, we used this algorithm for the subsequent experiments.

Addressing our research questions (see §4.1.1), we benchmarked the performance of several feature set combinations across the matrix presented in Figure 5 (RQ1). Furthermore, we were interested in evaluating the performance of our financial domain classifier on the encyclopedic domain (RQ2). To this end, we used the 30 random samples of the Wikipedia test set as shown in Table 3 and calculated the means of the respective performance measures.

As features, we used pure BoW vectors as they are a strong feature set in the encyclopedic domain (Farkas et al., 2010) and additionally experimented with PoS-enriched BoW vectors ("POSBoW"). Apart from benchmarking all different features (RQ1, see 4.4.1), we were interested in exploring how well these features port to the previously researched encyclopedic domain and vice versa (RQ2, see 4.4.2).

7 This work has been published before Transformer-based models such as BERT (Devlin et al., 2019) emerged and hence does not consider them. Furthermore, the dataset size of 1K sentences and 23K tokens would be insufficient for the fine-tuning of such models, as they generally require thousands of training instances for stable results (Devlin et al., 2019; Dodge et al., 2020). However, further experimenting with zero- or few-shot learning approaches would be an interesting project for future research.

4.4 RESULTS AND DISCUSSION

In this section, we present our experimental results. We evaluated the classification performance in terms of precision P, recall R, and F_1 and used corrected paired t-tests[8] ($\alpha = 0.05$) to check for significant differences across feature sets.

4.4.1 *RQ1: Financial Domain*

On our financial gold standard (see Table 4), the rules significantly outperform all other individual features in terms of precision for the *uncertain* class ($P = 0.77$). As expected, this comes at the cost of a low recall (0.13). BoW reaches a significantly higher recall (0.37) than all features apart from POSBoW (0.35). In terms of F_1, POSBoW significantly outperforms all other features except the rules (0.41 vs. 0.40).

Looking at feature categories, syntactic features perform slightly better than lexical ones, with insignificant improvements across all performance measures. Furthermore, knowledge-rich features have a slightly larger precision than the knowledge-poor POSBoW (0.58 vs. 0.53). This comes at the cost of a significantly lower recall (0.24 vs. 0.35) and an insignificantly lower F_1 score (0.32 vs. 0.41).[9] Combining all lexical features with the rules yields the best F_1 for the *uncertain* class (0.47). However, added value of the rules seems negligible with an insignificant increase in terms of P ($\Delta_P = +0.02$) and F_1 ($\Delta_{F_1} = +0.01$).

COMPARISON TO THE LM UNCERTAINTY DICTIONARY In summary, the combination of PoS-augmented BoW tags with the two uncertainty dictionaries (Farkas et al., 2010; Loughran and McDonald, 2011) seems to provide a strong benchmark on our new financial gold standard. Compared to the state-of-the-art in financial uncertainty detection (i.e., using the Loughran and McDonald (2011) dictionary alone), this classifier yields a significantly increased recall (0.40 vs. 0.14), which translates into a twice as high F_1 (0.46 vs. 0.21). These results show that the manual creation of exhaustive in-domain dictionaries is a non-trivial task. Critically, the Loughran and McDonald (2011) dictionary was developed for disclosure type 10-K (and not for earnings calls which we use here), which should additionally dampen the performance of this feature set.

8 We use the corrected resampled t-test proposed by Nadeau and Bengio (2003) that accounts for the violated independence assumption. This correction allows to get a meaningful (if conservative) estimate of significance when comparing the performance of ML algorithms (*ibid.*).

9 Although seemingly large, this difference is indeed insignificant due to the small dataset size and the tendency of the corrected resampled t-test to overestimate variance (Nadeau and Bengio, 2003).

Table 4: Classification results on our financial domain dataset.

Features	Uncertain			Certain			Average		
	P	R	F_1	P	R	F_1	P	R	F_1
BoW	0.46	**0.37**	0.40	0.87	0.91	0.89	0.80	0.82	0.81
POSBoW	0.53	0.35	**0.41**	**0.88**	0.93	0.90	0.82	0.83	**0.82**
Fin	0.53	0.14	0.21	0.85	0.97	**0.91**	0.80	0.83	0.79
Wiki	0.48	0.17	0.23	0.84	0.97	**0.91**	0.78	0.83	0.79
Fin+Wiki	0.53	0.26	0.34	0.87	0.95	**0.91**	0.81	0.83	0.81
Rules	**0.77**	0.13	0.21	0.84	**1.00**	**0.91**	**0.83**	**0.85**	0.79
Lexicality vs. Syntacticity									
BoW+Fin+Wiki	0.53	**0.39**	**0.44**	**0.88**	0.92	0.90	0.82	0.83	0.82
POSBoW+Rules	**0.56**	0.37	0.43	0.87	**0.94**	**0.91**	0.82	0.84	0.83
Knowledge-Intensity									
POSBoW	0.53	**0.35**	**0.41**	**0.88**	0.93	0.90	0.82	0.83	0.82
Fin+Wiki+Rules	**0.58**	0.24	0.32	0.86	**0.96**	**0.91**	0.81	0.84	0.81
Rules									
POSBoW+Fin+Wiki	0.57	**0.40**	0.46	**0.88**	0.93	0.91	0.83	0.84	0.83
POSBoW+Fin+Wiki+Rules	0.59	**0.40**	**0.47**	**0.88**	0.94	**0.92**	0.83	0.85	0.84
Majority Class	0.00	0.00	0.00	0.83	1.00	0.90	0.50	0.50	0.50

Table 5: Classification results on the Wikipedia set (Farkas et al., 2010).

Features	Uncertain			Certain			Average		
	P	R	F_1	P	R	F_1	P	R	F_1
BoW	0.59	0.34	0.42	0.87	0.95	0.91	0.82	0.85	0.83
POSBoW	0.63	0.31	0.41	0.87	0.96	0.91	0.83	0.85	0.82
Fin	0.41	0.05	0.09	0.83	0.99	0.90	0.76	0.83	0.76
Wiki	**0.66**	0.40	**0.49**	**0.89**	0.96	**0.92**	**0.85**	**0.86**	**0.85**
Fin+Wiki	**0.66**	**0.41**	**0.49**	**0.89**	0.95	**0.92**	**0.85**	**0.86**	**0.85**
Rules	0.13	0.01	0.02	0.83	**1.00**	0.91	0.71	0.83	0.76
Lexicality vs. Syntacticity									
BoW+Fin+Wiki	**0.63**	**0.39**	**0.47**	**0.88**	0.95	**0.92**	0.84	0.85	0.84
POSBoW+Rules	**0.63**	0.31	0.41	0.87	**0.96**	0.91	0.83	0.85	0.82
Knowledge-Intensity									
POSBoW	0.63	0.31	0.41	0.87	**0.96**	0.91	0.83	0.85	0.82
Fin+Wiki+Rules	**0.65**	**0.41**	**0.49**	**0.89**	0.95	**0.92**	0.85	0.86	0.85
Rules									
POSBoW+Fin+Wiki	**0.66**	0.37	**0.47**	**0.88**	0.96	**0.92**	0.84	0.86	0.84
POSBoW+Fin+Wiki+Rules	**0.66**	**0.38**	**0.47**	**0.88**	0.96	**0.92**	0.84	0.86	0.84
Majority Class	0.00	0.00	0.00	0.83	1.00	0.90	0.50	0.50	0.50

4.4.2 *RQ2: Encyclopedic Domain*

Moving to the encyclopedic domain (see Table 5), the Wiki dictionary outperforms all other features. Given that it was designed specifically for this domain, this is expected. Similarly, the financial domain rules are the weakest feature set by far. As shown in Table 3, they match only comparably few Wiki sentences, which reflects in an overall poor performance.

Looking at the *uncertain* class, lexical features (BoW+Fin+Wiki) perform noticeably better than syntactic ones (POSBoW+Rules). This holds especially for recall (0.39 vs. 0.31) and F_1 score (0.47 vs. 0.41) and is likely attributable to the high performance of the isolated Wiki dictionary. Furthermore, given the highly formalized sentence structure of encyclopediae, rule-based and other syntactic features likely have little applicability. Instead, lexical choices seem to reflect degrees of uncertainty better in this case. Since the opposite case holds for our financial gold standard (i.e., the sentence structure is relatively free and spontaneous), the results align with our expectations.

Regarding knowledge intensity, the knowledge-rich features yield a slightly increased precision (0.65 vs. 0.63) and a distinctively higher recall (0.41 vs. 0.31) and F_1 score (0.49 vs. 0.41) than the PoS-enriched BoW vectors. Again, this is likely due to the strong isolated performance of the Wiki dictionary. Different from the in-domain application of the Loughran and McDonald (2011) dictionary or our manually created rules, the Wiki dictionary yields a relatively high recall (0.40 vs. 0.13). This is likely owing to its size (1,984 features), which is orders of magnitude larger than the financial dictionary (297 features) and our rules (95 features). Lastly, as previously shown in §4.3.2, the rules are not applicable to the Wikipedia test set. This is also why they have no noticeable performance impact when added to the set POSBoW+Fin+Wiki.

4.5 CONCLUSION

Given the impact of linguistic uncertainty on analyst and market reactions (see §2.2), uncertainty detection is an important yet underexplored task in finance. In this chapter, we propose the first classifier to tackle this problem. In doing so, we present a new financial domain gold standard created from earnings call transcripts. Furthermore, we introduce a set of manually created lexico-syntactic rules matching uncertain language. Comparing various classification algorithms, we find that an SVM yields competitive results. In terms of features, a combination of BoW, PoS, a general-domain dictionary (Farkas et al., 2010), and the domain-specific Loughran and McDonald (2011) dictionary provided a strong benchmark on our dataset. In the next chapter, we will explore linguistic uncertainty prediction leveraging the Q&A

structure of earnings calls. Moreover, we will experiment with including joint semantic and financial features into the classification.

5

IDENTIFYING DRIVERS OF LINGUISTIC UNCERTAINTY IN FINANCIAL DISCLOSURES

* Preliminary evidence shows that linguistic uncertainty in earnings calls can be classified from lexico-syntactic features (see Chapter 4). However, we lack a complete picture of CEOs' motivation to use uncertain language. For example, to which degree do uncertain answers depend on the semantic content of the preceding question? What influence do secondary criteria like the uncertainty or negativity of a question have? Or rather, is the tendency to answer questions uncertainly a reflection of underlying economic or financial uncertainty?

To answer these questions, we retrieve a set of Q&A pairs from earnings calls and predict the uncertainty of answers depending on linguistic representations of their preceding question and fundamental financial data. To that end, we use the best-performing lexical features from the last work and augment them with semantic representations. Furthermore, given the relationship of disclosure language to market movements, we include a comprehensive set of financial features to control for, e.g., past volatility or overall market risk.

5.1 INTRODUCTION

In this chapter, we predict modality, a specific kind of linguistic uncertainty, based on linguistic and financial features in a financial Q&A setting. The phenomenon of modality is closely related to politeness (Danescu-Niculescu-Mizil et al., 2013) and hedging (Hyland, 1998; Lakoff, 1973). Three kinds of modality exist: *dynamic*, *priority*, and *epistemic modality* (Portner, 2009, p. 47). Here, we focus on epistemic modality, which expresses a speaker's confidence in the truth of their proposition (*ibid.*): high epistemic modality (expressed with markers such as "certainly," "must") indicates high confidence and low modality ("probably," "might") indicates low confidence.

We focus on modality, as it only covers the vagueness component of uncertainty, which differentiates it from the last chapter, which also aimed to account for verifiability of statements (e.g., speculations about the future) or real-world uncertainty (e.g., statements about

* This chapter is based on "Predicting Modality in Financial Dialogue" (Theil and Stuckenschmidt, 2020), presented virtually in December 2020 at the *1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation (FNP)* collocated with the *28th International Conference on Computational Linguistics (COLING)* in Barcelona.

market risk). Hence, modeling modality should arguably be a better defined and more easily automatable task.

Another motivation stems from past literature, where Loughran and McDonald (2016) specifically suggested to explore modality in earnings calls. They hypothesize that a large share of weak modal words in CEO utterances might worsen stock or operating performance (*ibid.*, p. 1224). Confirming this hypothesis, Dzieliński et al. (2021) found that that executive modality is explanatory of stock price as well as analyst's earnings forecasts and firm valuations.

However, little is understood about the motivation to use weak (i.e., linguistically uncertain) or strong (certain) modal language in financial dialogue. Hence, we were interested in modeling its linguistic and financial determinants in the Q&A part of earnings calls. To the best of our knowledge, this is the first NLP study to address this task. 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 in the financial domain (Loughran and McDonald, 2011).

Similar to the last chapter, we analyze transcripts of quarterly earnings calls.[1] The spontaneous structure and open dialogue form of their Q&A is particularly suitable for a modality classification task. This is substantiated by past research exploring the Q&A for analyzing the related phenomena of indirectness (Crawford Camiciottoli, 2009), persuasion (Crawford Camiciottoli, 2011, 2018), and deception (Larcker and Zakolyukina, 2012). As task, we extract question–answer pairs from the Q&A to predict the modality of an answer depending on (1) the content of the preceding question and (2) a comprehensive set of financial features.

5.1.1 *Contributions*

We provide the following contributions to the community:

- We publish a new silver standard dataset of 5K question–answer pairs for modality prediction.
- We introduce the first modality classifier learning from both lexico-symantic and financial features.
- We provide interpretable results by visualizing the importance and effect of the used features.

1 See §2.2.3 for an in-depth discussion of earnings calls.

5.2 DATA

We obtain 20K earnings call transcripts from SeekingAlpha[2] and sample all Q&A pairs from them. Numbers are identified with `spaCy`'s (Honnibal et al., 2020) Named Entity Recognition (NER) and replaced with uniform placeholder tokens. We remove Q&A pairs with inaudible parts, audio gaps, 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 largest 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 largest share of weak modal tokens and assign them a *weak modal* label.

  Example 5.1 *"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 largest share of these tokens and assign them a *strong modal* label.

  Example 5.2 *"It will. That's right, it will."*

This yields a balanced dataset of 5K (2.5K *weak* and 2.5K *strong modal*) instances; Table 6 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.[3]

5.3 METHODOLOGY

In this section, we motivate our features sets (§5.3.1) and introduce a binary classifier for modality prediction in financial dialogue (§5.3.2).

5.3.1 *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.

2 `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.

3 See Appendix A.

Table 6: Descriptive statistics of our silver standard dataset for modality prediction in financial Q&A.

Semantic unit	n
types	7.7K
tokens	232.1K
sentences	15.1K
utterances	5.0K

SURFACE FEATURES In the SURFACE feature set, we explore the following:

- **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. §5.2 contains examples of strong and weak model tokens according to the LM lexicon.
- **Uncertainty** is again measured by the respective LM lexicon which contains 297 tokens referring to linguistic imprecision or risk, e.g., "hypothesis" and "volatility."

LEXICAL (SEMANTIC) FEATURES In the LEXICAL category, we compare term frequency (tf) and tf–idf vectors, which performed strongly in our previous uncertainty detection experiments (Chapter 4). 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. §5.2). We evaluate dimensions $d_{w2v} \in \{100, 200, 300\}$ with both the CBoW and the Skip-Gram (SG) architecture. Finally, we represent all questions as embedding centroids. Our results indicate that tf–idf vectors with $d_{BoW} = 300$ are optimal for the given task.

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 PP to $n \in \{5, 10, ..., 45, 50\}$ in

a five-fold CV setup on our training set. Our results indicate that an optimal l and PP are obtained for $n = 5$.[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 of 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 Chicago Board Options Exchange Volatility Index (VIX),[5] 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 an important 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,[6] which distinguishes between e.g., "energy" or "healthcare."

5.3.2 *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.[7] In sum, we consider: (Gaussian) Naïve Bayes,

4 $l = -145218.44$ and $PP = 1782.15$.

5 http://www.cboe.com/vix

6 http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

7 As we were interested to obtain interpretable weights, we chose not to explore Transformer-based models like BERT (Devlin et al., 2019) at the time of publication.

Logistic Regression, SVM (with Radial Basis Function (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`.

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.

EVALUATION We evaluate the performance of our classifiers with precision, recall, and F_1 score metrics. Furthermore, to quantify relative feature importance in case of the early fusion approaches, we use Shapley Additive Explanations (SHAP), 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.

5.4 RESULTS AND DISCUSSION

5.4.1 *Feature Performance*

Table 7 shows the results of our classification task in terms of precision (P), recall (R), and F_1 score for both the *strong* and the *weak modal* class and 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

Nevertheless, upon re-examination during writing this thesis, fine-tuning BERT and RoBERTa with a learning rate of 3×10^{-5} for three epochs led to no decisive performance increases over the LEXICAL baseline. This is likely attributable to the restricted dataset size of 230K tokens (compare, e.g., to the 2.5M tokens in Chapter 8). However, similar to Chapter 4, future work could explore zero- or few-shot learners in more depth.

Table 7: Class-wise and average classification results on our silver standard Q&A dataset ($n = 5$K) with balanced *weak modal* and *strong modal* labels. All results are obtained with XGBoost and ALL$_{late}$ uses an additional Gaussian Naïve Bayes meta-classifier.

Features	Weak Modal			Strong Modal			Average		
	P	R	F_1	P	R	F_1	P	R	F_1
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

measures. The SURFACE, LEXICAL, SEMANTIC, and FINANCIAL feature sets are defined as outlined in §5.3.1 and the fused features are represented by ALL with separate subscripts for the *early* and the *late* fusion approach.

All feature sets (except 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. $R = 0.86$) and in terms of recall on the *strong modal* class ($R = 0.89$ vs. $R = 0.85$), 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. Related work (Loughran and McDonald, 2016) and our research (Chapter 7) show that, compared to financial features, textual information has a relatively small impact on financial risk; 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.

5.4.2 *Feature Importance*

One advantage of the early fusion approach is its interpretability: since all features are represented in the same vector, we can quantitatively obtain a notion of relative feature importance. To do so, we

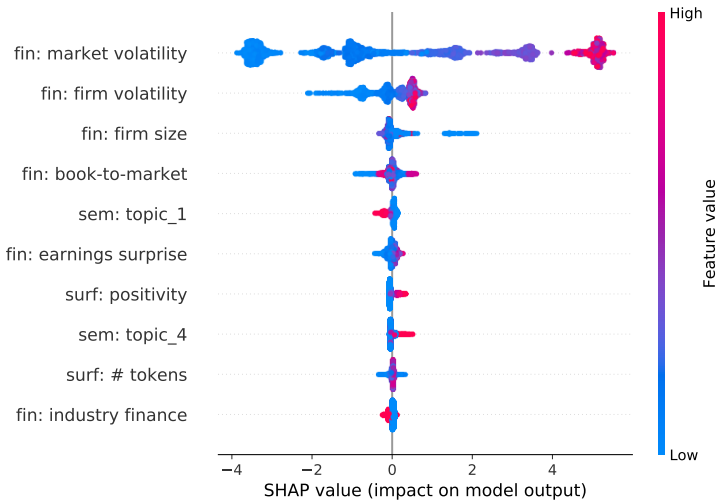


Figure 6: Violin plot of SHAP values for the top-10 features in the binary classification with early fusion.

calculate the SHAP values (cf. §5.3.1) for all features and present the results in Figure 6. 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.

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.

5.5 CONCLUSION

Understanding the drivers of linguistic uncertainty in the financial domain is an important up-stream task for risk regression and inference (see Chapter 2.2). In this chapter, we explore the prediction of modality (a linguistic concept denoting the (un)certainty of utterances) in a financial Q&A setting. We present a new silver standard dataset and introduce a binary classifier exploiting the multimodality of the setting. In our experiments, we perform a systematic comparison 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 BoW, the Loughran and McDonald (2011) dictionary, or word embeddings) of the preceding question.

This chapter concludes our efforts to explore detection and prediction methods for linguistic uncertainty in earnings calls. The results indicate that the financial situation of the company and the economy as a whole are intertwined with the language choices of CEO. However, little is understood about the impact of uncertain disclosure language on analysts and the market. Part III will introduce an experimental approach to address this problem.

Part III

CAUSALITY OF UNCERTAINTY AND RISK

Addressing the **causal modeling of uncertainty and risk (T_2)**, Chapter 6 aims to answer the questions: How can we quantify the influence of linguistic uncertainty on risk perception and investment behavior? Which personal characteristics of investors play a role in this setting? We conducted a laboratory experiment to explore how investors react in the face of linguistic uncertainty. This experiment is the first to establish a causal link between vague disclosure information, investor risk perception, and investment behavior.

6

LINGUISTIC UNCERTAINTY AND RISK PERCEPTION IN FINANCIAL DISCLOSURES

6.1 INTRODUCTION

* Adding to the existing evidence concerning the interrelationship of financial risk and linguistic uncertainty (see Part I), Chapter 5 has shown that economic and fundamental financial variables impact the linguistic choices of CEOs. Nevertheless, up until now, no experimental study has analyzed the direct influence of linguistic uncertainty on the financial market in form of investor behavior. In this chapter, therefore, we present the first causal laboratory study to approach this problem. A laboratory experiment is a natural choice for detecting causal patterns as it provides "the most convincing method of creating the counterfactual because it directly constructs a control group via randomization," which permits to assume causality (Floyd and List, 2016, pp. 442–443). Such claims about causality are "particularly strong [...], because manipulated independent variables preclude reverse causality and elicited dependent variables allow tests of theoretical constructs" which could not be detected differently (Bloomfield et al., 2016, p. 383).

In our laboratory study, we exposed the participants to a randomized set of earnings call[1] snippets of either uncertain or certain linguistic style. Participants then forecasted the underlying companies' future performance and invested a fictional sum of money. Finally, they answered an extensive questionnaire covering socio-demographics, Big 5 personality, financial literacy,[2] and risk tolerance.[3] We used causal mediation analysis as the main analytical tool. Conceptually, we explored whether risk perception[4] mediates a causal effect of linguistic uncertainty on investment sum. In an exploratory analysis, we searched for further investor characteristics with a significant influence on their risk perception. Overall, we find causal evidence that linguistic uncertainty in disclosure language has a significant posi-

* This chapter is based on the working paper "Linguistic Uncertainty and Risk Perception in Financial Disclosures" (Theil, Daube, and Stuckenschmidt, 2022).

1 See §2.2.3 for an in-depth discussion of earnings calls.

2 *Financial literacy* is one's "ability to process economic information and make informed decisions about financial planning, wealth accumulation, debt, and pensions" (Lusardi and Mitchell, 2014, p. 6).

3 *Risk tolerance* is "the maximum amount of uncertainty someone is willing to accept when making a financial decision" (Grable, 2000, p. 625).

4 *Risk perception* is "an individual's assessment of how risky a situation is in terms of probabilistic estimates of the degree of situational uncertainty, how controllable that certainty is, and confidence in those estimates" (Sitkin and Weingart, 1995, p. 1575).

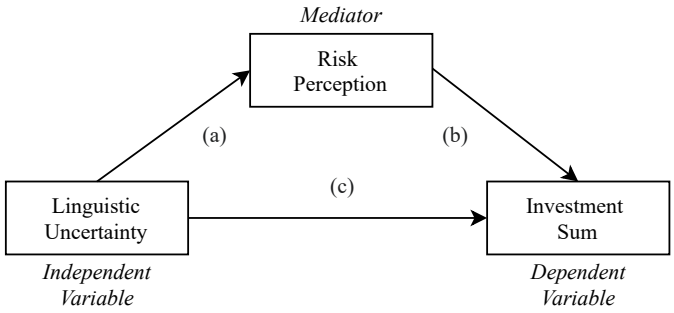


Figure 7: Hypothesized causal diagram for our subsequent causal mediation analysis. Linguistic Uncertainty (*IV*) is assumed to affect the Investment Sum (*DV*) via the *Mediator* Risk Perception.

tive impact on risk perception and a significant negative impact on the investment sum. Empirically, risk perception amplifies the effect of uncertainty on investments. Individual investor characteristics that affect risk perception include gender, age, income, extraversion, and neuroticism. Our results are robust to subject- and text-specific effects.

6.2 RESEARCH QUESTIONS AND HYPOTHESES

Based on the related literature finding correlations between linguistic uncertainty and risk (Barth et al., 2021; Doshi et al., 2021; Loughran and McDonald, 2011, 2013) and between risk perception and investment behavior (Byrne, 2005; Nguyen et al., 2019), we assumed that risk perception mediates the effect of linguistic uncertainty on the amount invested (cf. Figure 7). In particular, we searched for answers to the following main research question:

Research Question 1 (RQ1) *How does linguistic uncertainty in financial disclosures influence individual investment decision-making?*

To that end, we tested hypotheses concerning the causal effect of the linguistic uncertainty in financial disclosures on the risk perception and the investment decision. Moreover, we assessed the mediating role of risk perception between linguistic uncertainty in financial disclosures and the investment decision. As previous research suggests that various individual characteristics are associated with uncertainty and risk perception, we posed the following secondary and exploratory research question:

Research Question 2 (RQ2) *How do individual characteristics of an investor influence the risk perception regarding an investment?*

Addressing RQ2, we developed hypotheses based on past findings that risk perception is associated with socio-demographics such as

age, gender, and income (Nguyen et al., 2019); the Big Five personality traits (Oehler and Wedlich, 2018); financial literacy (Wang et al., 2011); and risk tolerance (Nguyen et al., 2019). The following provides more details on the specific hypotheses we developed to answer our two research questions.

6.2.1 *Linguistic Uncertainty, Risk Perception, and Investment Decision (RQ1)*

Related work hints that linguistic uncertainty increases financial risk proxies: Dzieliński et al. (2021) find that vague earnings call communication increases analyst uncertainty, as reflected in the revision frequency of analyst forecasts. Barth et al. (2021) find that earnings call vagueness increases volatility and decreases abnormal returns. Moreover, studies show that investors tend to be risk averse: For example, Byrne (2005) asserts negative relationships between risk propensity and risk perception and between risk perception and investment allocation. Nguyen et al. (2019) find that risk perception is negatively correlated with the investment sum and hence conclude that investors invest more into perceived low-risk stocks. Based on the findings of the presented studies, we posited the following hypotheses:

Hypothesis 1 (H1) *Linguistic uncertainty in financial disclosures causes investors to invest smaller investment sums. This effect is mediated by risk perception.*

Hypothesis 1a (H1a) *Linguistic uncertainty has a positive effect on risk perception.*

Hypothesis 1b (H1b) *Risk perception has a negative effect on investment sum.*

Hypothesis 1c (H1c) *Linguistic uncertainty has a negative effect on investment sum.*

Multiple studies have shown that uncertainty in financial disclosures may increase uncertainty across analysts and investors (Barth et al., 2021; Doshi et al., 2021; Loughran and McDonald, 2011, 2013). Therefore, we hypothesized that an increased linguistic uncertainty should reflect in a larger variation of investment sums across individuals:

Hypothesis 2 (H2) *Linguistic uncertainty causes a higher variation of the investment sum between individuals.*

Hypotheses H1 (with its sub-hypotheses H1a, H1b, and H1c) and H2 were identified to address RQ1, the main research question of this chapter. Below, we discuss the hypotheses developed to address the exploratory RQ2.

6.2.2 *Investor Characteristics and Risk Perception (RQ2)*

SOCIO-DEMOGRAPHICS A study by Powell and Ansic (1997) suggests that men are more risk-seeking than women. Hartog et al. (2002) find that women and low-income or -wealth individuals are more risk-averse. Nicholson et al. (2005) show that age has a negative effect on risk propensity and that men have a higher propensity than women. Nguyen et al. (2019) confirm the effects of gender, age, and income on the related construct risk tolerance. Although the measures explored by the past literature are related but distinct to risk perception, we expect the observed effects to be transferable. In summary, we hypothesize the following:

Hypothesis 3 (H3) *Men have a lower risk perception than women.*

Hypothesis 4 (H4) *Age has a positive effect on risk perception.*

Hypothesis 5 (H5) *Income has a negative effect on risk perception.*

BIG FIVE Past literature has examined the effect of personality traits on risk perception. Nicholson et al. (2005) find that extraversion and openness have a positive, while neuroticism, agreeableness, and conscientiousness have a negative relationship with risk propensity in financial settings. More closely related to this work, Oehler and Wedlich (2018) find that extraverted individuals are less risk-averse, while the opposite holds for neurotic ones. Furthermore, they find that conscientious individuals are more risk-averse and perceive investments to be riskier. Based on the findings of past literature, we posited the following hypotheses:

Hypothesis 6a (H6a) *Conscientiousness has a positive effect on risk perception.*

Hypothesis 6b (H6b) *Extraversion has a negative effect on risk perception.*

Hypothesis 6c (H6c) *Neuroticism has a positive effect on risk perception.*

FINANCIAL LITERACY Diacon (2004) finds that experts perceive products to be less risky than lay investors. Furthermore, Wang et al. (2011) find that investors perceive easier-to-understand and familiar products as less risky. Sachse et al. (2012) attest a negative relationship between financial literacy and risk perception. Based on these results, we posited the following hypothesis:

Hypothesis 7 (H7) *Financial literacy has a negative effect on risk perception.*

RISK TOLERANCE Following Nguyen et al. (2019), who found a negative relationship between risk tolerance and risk perception, we hypothesized the following:

Hypothesis 8 (H8) *Risk tolerance has a negative effect on risk perception.*

6.3 EXPERIMENTAL DESIGN

The following describes our experimental design in more detail. §6.3.1 summarizes the procedure and items; §6.3.2 describes the data and its preparation; §6.3.3 shows the used materials; §6.3.4, finally, explains the analytical strategy to spot causal influences of uncertainty on risk perception and investment behavior.

We used a mixed design allowing for both within- and between-subject analyses. After the participants had familiarized themselves with a brief fictional scenario, we showed them four randomly selected earnings calls excerpts. These excerpts consisted of two *certain* and two *uncertain* ones. All excerpts were presented simultaneously with randomized position to avoid order effects. To control for possible text-specific effects, subsequent analyses include text dummies. The two main questions of the study were: (1) how much money subjects would invest based on each of the four snippets and (2) how high they perceived the risk of an investment into the company's stock. Motivated by the literature presented in §6.2, we used several items to assess the participants' personality traits, risk tolerance, financial literacy, and socio-demographics. The questionnaire, including all materials and items, is contained in Appendix B.1, and the assessed measures are presented in more detail below.

6.3.1 *Measures*

INVESTMENT SUM Based on the earnings call excerpts, participants were asked to indicate how much out of €10,000 they would invest into each of the four companies. The investment decision was subject to the following three criteria: It was neither necessary to invest the entire amount nor to invest into all companies; furthermore, the total amount invested into all companies could not exceed €10,000.

RISK PERCEPTION Based on the findings of Nguyen et al. (2019), four items measuring potential loss, potential gain, and stock price volatility were used.

FINANCIAL LITERACY Two subjective questions addressing experience and confidence concerning investing and a one-item financial literacy assessment by Gibson et al. (2013) were used. In addition, three objective questions by Lusardi and Mitchell (2011) were used.

RISK TOLERANCE Risk tolerance was assessed using four items selected by Guillemette et al. (2012) and two items selected by Grable and Lytton (1999).

BIG FIVE PERSONALITY TRAITS Personality was measured using the "Big Five" personality traits (Goldberg, 1990) as operationalized by the brief BFI-10 questionnaire (Rammstedt and John, 2007) with two items for each trait.

SOCIO-DEMOGRAPHICS The following socio-demographics were assessed: gender, age, marital status, housing situation, annual income, academic degree, work status, and work sector.

CONTROL QUESTIONS To avoid priming, participants were asked if they already had participated in the pilot study. Furthermore, participants were asked if they were able to understand all of the instructions, questions, or texts presented in the English language without issues. Additionally, the participants were asked if they had answered all questions thoroughly. If "No" was indicated for any of those questions, the responses were discarded.

6.3.2 *Participants and Data Preparation*

A total of 121 subjects participated in the laboratory study. We recruited participants online via SURVEYCIRCLE[5] and online platforms such as LINKEDIN research groups. Due to the removal of incomplete surveys or those that answered any of the control questions (cf. §6.3.1) with "No," the initial sample was reduced to $n = 81$.

We transformed income and education into ordinal variables. Missing values of income were replaced with the median income of the corresponding academic degree cohort. Gender was binarized since no respondent indicated "diverse." For work status, we created three groups: (1) "Working full-time" and"self-employed;" (2) "working part-time" and "working student;" and (3) "retired," "student," and "unemployed." Items of the multi-item constructs were averaged. Consistent with the findings of Nguyen et al. (2019), we assume that both a belief in increasing and decreasing returns leads to increased risk perception, i.e., we do not invert the scale of the first risk perception item.

6.3.3 *Materials: Earnings Call Excerpts*

Using a sample of 90K earnings call transcripts between 2002 and 2017, we selected four transcripts as a basis for creating the earnings call excerpts presented to the study participants. We focused

5 www.surveycircle.com

on the presentation section of the transcripts since the questions-and-answers (Q&A) contain frequent speaker turns and are formally too heterogeneous. In addition, we limited the initial sample to companies from the automotive industry to control for industry-specific risks. Next, we manually identified the management's discussion of future developments and the outlook for the upcoming year. To avoid confounding effects of sentiment polarity, we only selected passages containing neither words from the Loughran and McDonald (2011) POSITIVE nor from the NEGATIVE dictionary.[6]

MANUAL MODIFICATION We manually set all percentage numbers in the text to "5%." In addition, for each of the four earnings call excerpts, we created one linguistically uncertain and one certain version. For the uncertain version, words from the UNCERTAINTY and WEAK MODAL dictionaries were manually induced. The UNCERTAINTY dictionary contains a total of 297 words measuring either linguistic uncertainty (e.g., "roughly," "vaguely," and "assume") or economic risk ("cautious," "risky," and "volatile"). We focused on the first of the two categories. Furthermore, we used the WEAK MODAL dictionary, which contains 27 words (e.g., "almost," "could," and "may"). Modality is the degree of commitment to a statement, where "weak modality" describes possibilities and "strong modality" describes necessities (Palmer, 2001). All such created uncertain excerpts contain nine words from the two dictionaries. Example 6.3.1 shows one of the uncertain excerpts with the induced UNCERTAIN terms highlighted in red.

Example 6.3.1 *After a promising last year, I would like to address our current outlook for the next year. Despite the macroeconomic data coming out of North America, we still believe that U.S. market sales increase. In particular, we possibly increase our net sales by 5%. The European market seems to be more fragmented, which makes our outlook for the year vague. On a consolidated basis, it appears that our costs decrease by 5%. The automotive industry continues to be dynamic and to depend on world politics. In other words, the development of the industry depends somewhat on future regulations. We are adjusting to this environment and assume to see similar general business trends in the second half of the year that we saw in the first.*

For the certain version, words from the STRONG MODAL word list, consisting of 19 words, were integrated. The list contains words like "definitely," "clearly," and "will," which express certainty. The certain version of an excerpt always contains six words from the STRONG MODAL word list. Example 6.3.2 shows the certain version that corresponds to the uncertain excerpt from Example 6.3.1. Induced CERTAIN terms are highlighted in green.

6 https://sraf.nd.edu/textual-analysis/resources/#LM%20Sentiment%20Word%20Lists

Example 6.3.2 *After a promising last year, I would like to address our current outlook for the next year. Despite the macroeconomic data coming out of North America, we will increase our U.S. market sales. In particular, we definitely increase our net sales by at least 5%. The European market will undoubtedly grow at the same rate. On a consolidated basis, we definitely decrease our costs by 5%. The automotive industry continues to be dynamic. Moreover, the development of the industry goes along with future regulations. We are adjusting to this environment and clearly see the same general business trends in the second half of the year that we saw in the first. Overall, we expect a promising next year for all markets.*

Thus, we created eight earnings call excerpts (4 certain, 4 uncertain) of similar length (120–127 words). The full excerpts are contained in Appendix B.1. To ensure the validity of the linguistic uncertainty construct, we additionally conducted a pilot study[7] with a sample representative of the main study. We exposed the subjects to four randomized snippets and asked them to indicate the perceived level of linguistic uncertainty. We found that the linguistic uncertainty perception of subjects differs highly significantly between the certain and the uncertain snippets, which ensures their validity for our laboratory study.

6.3.4 *Causal Mediation Analysis*

Following the hypothesized causal diagram (cf. Figure 7), we perform a causal mediation analysis with parametric bootstrapping and 1,000 simulation runs. The three regression formulae are as follows:

$$\begin{aligned}\text{RiskPerc}_i = a_0 &+ a_1\text{LincUnc}_i + a_2\text{Gender}_i + a_3\text{Age}_i \\ &+ a_4\text{Income}_i + a_5\text{Open}_i + a_6\text{Cons}_i + a_7\text{Extra}_i \\ &+ a_8\text{Agree}_i + a_9\text{Neuro}_i + a_{10}\text{FinLit} \\ &+ a_{11}\text{RiskTol} + \delta_i + \epsilon_i,\end{aligned} \tag{a}$$

$$\begin{aligned}\text{Inv}_i = b_0 &+ b_1\text{LincUnc}_i + b_2\text{Gender}_i + b_3\text{Age}_i \\ &+ b_4\text{Income}_i + b_5\text{Open}_i + b_6\text{Cons}_i + b_7\text{Extra}_i \\ &+ b_8\text{Agree}_i + b_9\text{Neuro}_i + b_{10}\text{FinLit} \\ &+ b_{11}\text{RiskTol} + b_{12}\text{RiskPerc} + \delta_i + \epsilon_i,\end{aligned} \tag{b}$$

$$\begin{aligned}\text{Inv}_i = c_0 &+ c_1\text{LincUnc}_i + c_2\text{Gender}_i + c_3\text{Age}_i \\ &+ c_4\text{Income}_i + c_5\text{Open}_i + c_6\text{Cons}_i + c_7\text{Extra}_i \\ &+ c_8\text{Agree}_i + c_9\text{Neuro}_i + c_{10}\text{FinLit} \\ &+ c_{11}\text{RiskTol} + \delta_i + \epsilon_i.\end{aligned} \tag{c}$$

7 The specifications and results of this pilot study are contained in Appendix B.2.

Index i represents subjects; δ_i stands for work status, academic education, and text identifier dummies; ϵ_i is the regression error. Regressions weights a, b, and c and equation names match the causal paths in Figure 7. For ordinal variables (here: academic education and income), `R` per default includes linear (`.L`), quadratic (`.Q`), and cubic (`.C`) polynomials. These allow us to estimate higher-order effects on the dependent variables.

For robust results, we assume heterogeneity in treatment effects and cluster standard errors by subject and text identifier. As an additional robustness check, we analyze a model with subject-fixed effects and omitted controls (as controlling for subject identity factors those out).[8] To test H2, we use Levene's test to check for the inequality of variances for investment sums between the *certain* and the *uncertain* group, respectively.

6.4 RESULTS

6.4.1 *Descriptive Statistics*

OVERVIEW Table 8 shows descriptive statistics for all variables. The typical participant was male, about 29 years old, working full-time, held a Bachelor's degree, and earned less than €30K per year. On average, participants invested €7.6K out of the possible 10K. The average risk perception is approximately neutral, which is an expected result given that each participant was presented with the same amount of *certain* and *uncertain* earnings call excerpts.

INVESTMENT SUM Figure 8 provides an overview of the invested money per excerpt and per group. Collectively, the average invested sum for *certain* excerpts is always higher than for *uncertain* ones (cf. Figure 8a). For excerpts 1, 2, and 4, the standard deviations of the certain version exceeds the one of the uncertain version. In excerpt 3, we observe the opposite but smaller effect. In summary, we find first indications that H2 (which assumes an increased dispersion of investment sums for uncertain disclosures) might not hold.

Figure 8b summarizes the overall investment distribution per treatment group (*certain* and *uncertain*). The average invested sum across all certain excerpts is €2540 (s_x = €1950), about twice the sum for uncertain excerpts ($\bar{x}$ = €1250, s_x = €1500). Overall, this indicates a clear preference of subjects for the linguistically certain disclosures.

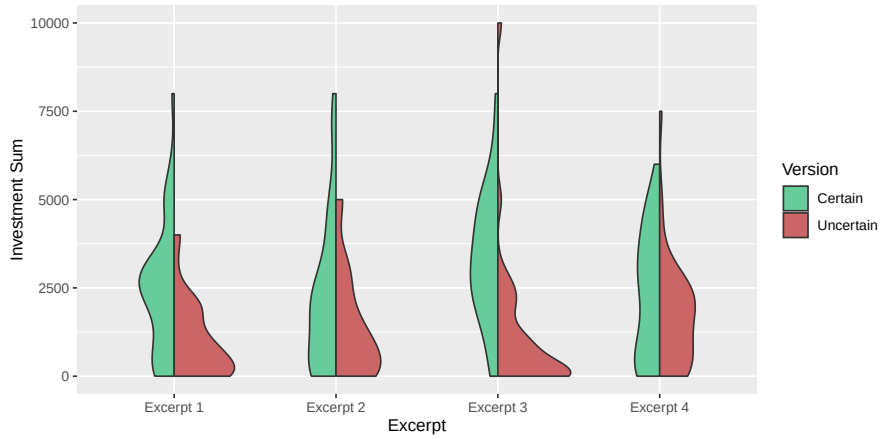
8 Initially, we also considered a mixed-effects model with random effects on a participant-level. However, this model yielded a singular fit while the random effects explained a variance of close to zero. Hence, we decided to use the parsimonious model presented herein.

Table 8: Descriptive statistics for all variables appearing in the laboratory study ($n = 81$). For continuous variables, "center" is the arithmetic mean and for categorical ones the mode. Multi-item constructs are averaged and normalized between 1 and 5.

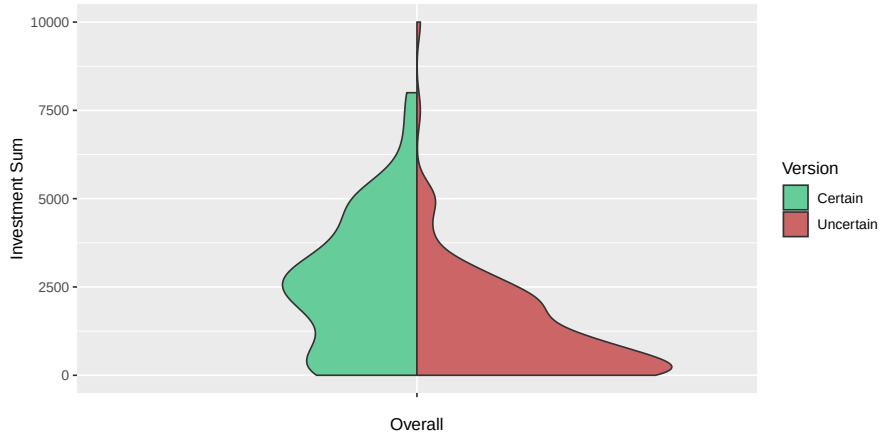
Variable	Center	s_x
Investment Sum	7.6K	2.6K
Risk Perception	3.25	0.42
Gender	male	—
Age	28.88	7.33
Income	<30K	—
Work Status	Full-time	—
Education	Bachelor	—
Openness	3.17	0.95
Conscientiousness	3.81	0.81
Extraversion	3.30	0.86
Agreeableness	3.28	0.75
Neuroticism	2.67	1.00
Financial Literacy	3.27	1.04
Risk Tolerance	3.09	1.11

ITEM RELIABILITY To evaluate the reliability of the multi-item constructs, we inspect their standardized Cronbach's α. Two-item constructs are evaluated in terms of Spearman–Brown's ρ, as this coefficient is more reliable for such cases (Eisinga et al., 2013). Results indicated a good internal consistency for risk perception ($\alpha = 0.81$), financial literacy ($\alpha = 0.83$), and risk tolerance ($\alpha = 0.76$). For the Big 5, extraversion and neuroticism scored acceptable ($\rho = 0.67$, $\rho = 0.69$), but openness, conscientiousness, and agreeableness scored low ($\rho = 0.43$, $\rho = 0.45$, and $\rho = 0.35$). The low score for these items is most probably due to the briefness of the used BF-10 questionnaire (Rammstedt and John, 2007). Hence, future studies could experiment with an increased number of items for Big 5.

CORRELATIONS Figure 9 summarizes correlations across all variables. As can be seen, moderate correlations ($0.4 \leq r \leq 0.6$) exist between gender and the Big 5 ($r_{\text{cons}} = -0.4$, $r_{\text{neuro}} = -0.52$): In our sample, male subjects tend to be score lower on conscientiousness and neuroticism. Furthermore, age and income are positively associated ($r = 0.44$), which is an expected result. More surprisingly, risk tolerance is negatively correlated with neuroticism and openness ($r_{\text{neuro}} = -0.45$, $r_{\text{open}} = -0.52$): risk-tolerant subjects tend to be more emotionally stable, but less open to experience. Finally, financial literacy and risk tolerance are positively associated ($r = 0.49$), aligned



(a) Distributions of investment sums per excerpt and version (certain and uncertain). Excerpt texts can be found in Appendix B.1.



(b) Distribution of investment sums per treatment group (certain and uncertain).

Figure 8: Distributions of investment sums across subjects ($n = 81$).

with the findings of related work (Sachse et al., 2012; Wang et al., 2011).

6.4.2 *RQ1: Linguistic Uncertainty and Investment Decisions (H1–2)*

The results of the mediation analysis are summarized in Table 9 and Figure 10. Empirically, the influence of linguistic uncertainty on investment sum is partially mediated by risk perception with an indirect effect of -16%. Increased linguistic uncertainty is highly significantly associated with a larger risk perception ($a_1 = 0.54$). Aligned with our expectations, both linguistic uncertainty (the independent variable) and risk perception (the mediator) have a highly significant and negative association with investment sum ($b_{12} = -0.29$, $c_1 = -0.69$). Promisingly, we find that the R^2 value of the indirect

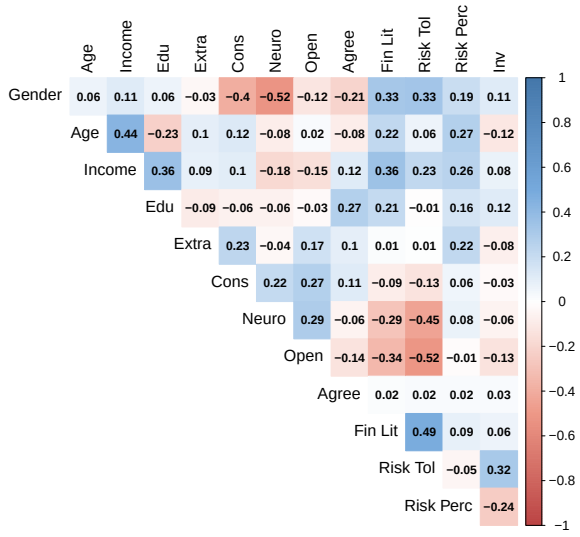


Figure 9: Correlation matrix across all variables considered in our laboratory study. Correlation is measured in terms of Pearson's r.

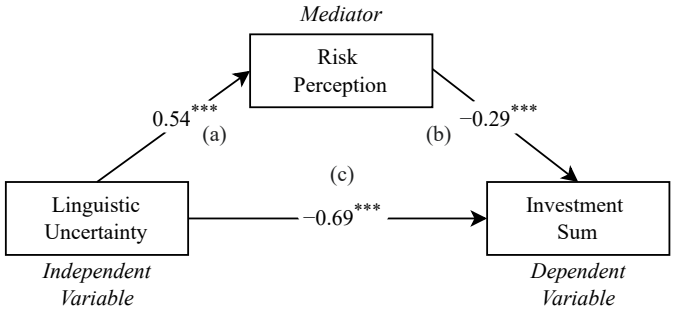


Figure 10: Resulting causal diagram with coefficients according to the causal mediation analysis. $^{*}p \leq 0.05, ^{**}p \leq 0.01, ^{***}p \leq 0.001$.

model (b) including the mediator risk perception increases substantially over the unmediated model (c): Including this causal path explains a variance of 23.94% as opposed to 17.47%.

Regarding RQ1, we find evidence that linguistic uncertainty causes individuals to invest less into a company, mediated by risk perception. However, we can not confirm that linguistic uncertainty in earnings calls is related to increased variation of investment sums. The results of Levene's test yield an F value of 16.23 and $p \leq 0.001$, indicating a highly significant difference across group standard deviations. Contrary to H2, however, we find that the standard deviation of investment sums is larger for the certain than for the uncertain group. This can also be seen in the violin plots of the investment sums in Figures 8a and 8b.

Table 9: Causal mediation results with z-standardized coefficients and t-statistics in parentheses. Standard errors are clustered by subject and text identifier. Regressions include intercepts and fixed effects for text identifier, education, and work status. "L" = linear and "Q" = quadratic fit for the ordinal income. Letters a–c match paths in Figures 7 and 10; "Index" matches coefficients in formulae a–c.

	Risk Perc (a)	Invest Sum (b)	Invest Sum (c)	Index
Ling Unc	0.54^{***}	-0.53^{***}	-0.69^{***}	1
	(0.10)	(0.11)	(0.11)	
Gender	0.45^{*}	0.10	-0.03	2
	(0.18)	(0.18)	(0.18)	
Age	0.26^{**}	0.07	-0.01	3
	(0.08)	(0.08)	(0.09)	
Income.L	0.04	-0.22	-0.23	4
	(0.24)	(0.23)	(0.24)	
Income.Q	0.42^{*}	0.11	-0.02	4
	(0.19)	(0.19)	(0.19)	
Open	-0.10	-0.04	-0.01	5
	(0.07)	(0.07)	(0.08)	
Cons	0.04	-0.01	-0.02	6
	(0.07)	(0.06)	(0.07)	
Extra	0.12^{*}	-0.00	-0.04	7
	(0.06)	(0.06)	(0.06)	
Agree	0.01	0.02	0.01	8
	(0.06)	(0.06)	(0.06)	
Neuro	0.19^{**}	0.08	0.03	9
	(0.07)	(0.07)	(0.07)	
Fin Lit	0.02	-0.06	-0.07	10
	(0.07)	(0.07)	(0.07)	
Risk Tol	-0.10	$0.14^{\dagger}$	0.17^{*}	11
	(0.08)	(0.08)	(0.08)	
Risk Perc		-0.29^{***}		12
		(0.06)		
n	324	324	324	
R^2	21.88%	23.94%	17.47%	

$^{\dagger}p \leq 0.1, ^{*}p \leq 0.05, ^{**}p \leq 0.01, ^{***}p \leq 0.001$

6.4.3 *RQ2: Investor Characteristics and Risk Perception (H3–H8)*

The results of the exploratory RQ2, i.e., which investor characteristics influence risk perception, provide a more mixed picture: We find evidence that the postulated effect of gender (H3) points in the oppo-

site direction than expected ($a_2 = 0.45$): Overall, men have a larger risk perception than women. As this effect opposes the findings of other studies (Hartog et al., 2002; Nguyen et al., 2019; Nicholson et al., 2005; Powell and Ansic, 1997), we additionally considered interaction models with age and personality. We found that gender significantly interacts with age ($\beta = -0.02$, $p \leq 0.05$), extraversion ($\beta = -0.13$, $p \leq 0.05$), and conscientiousness ($\beta = 0.29$, $p \leq 0.01$). The interaction plots are shown in Figure 11. For women, risk perception increases stronger with age than for men. The same applies to female extraverted investors; risk perception of male extraverted investors stays relatively constant, however. In line with past literature, larger conscientiousness has a positive effect on risk perception for men. However, the opposite holds for women. A limiting factor is the overrepresentation of male investors in our analyzed sample.

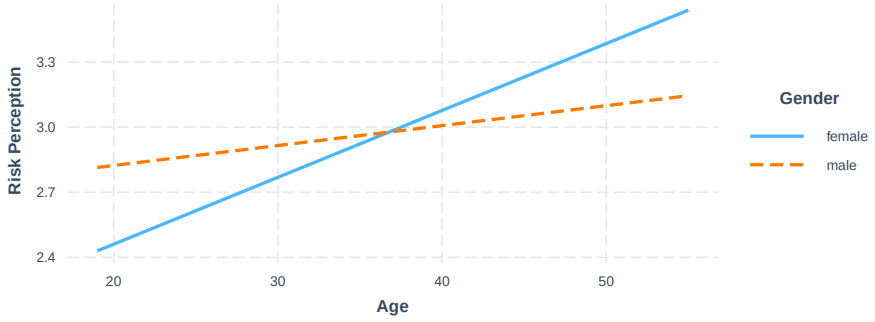
Age behaves as expected (H4), with older people having a larger risk perception ($a_3 = 0.26$). Partly contrary to H5, income has no linear and negative but a quadratic and positive relationship with risk perception ($a_4 = 0.42$). This indicates that an increase in income only has a decreasing effect on risk perception for low-income individuals. Although neither conscientiousness nor extraversion behave as expected (H6a and H6b), we find evidence that neuroticism has a positive and very significant effect on risk perception ($a_9 = 0.19$). Neither the coefficient of financial literacy nor risk tolerance (H7 and H8) is significant. The results are now discussed and interpreted in more detail and in conjunction with the existing body of research.

6.4.4 *Robustness*

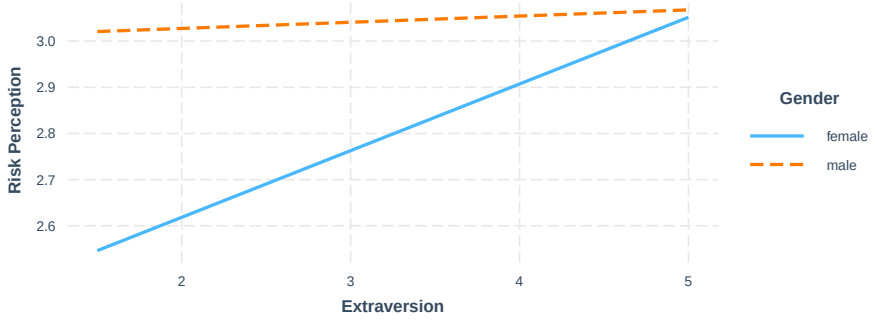
Table 10 contains the result of a causal mediation analysis with subject-fixed effects. Empirically, the causal effect of linguistic uncertainty on investment sum and its mediation via risk perception persists. Therefore, linguistic uncertainty seems to increase risk perception and decrease the amount invested independently of inter-subjective preferences. Overall, we observe increased levels of R^2, which is an expected result: Controlling for subject identifier introduces a factor variable with 81 levels for a dataset with 324 instances. A high explained variance with such a skewed sample-to-term ratio is likely the result of model bias or overfitting.

6.5 DISCUSSION

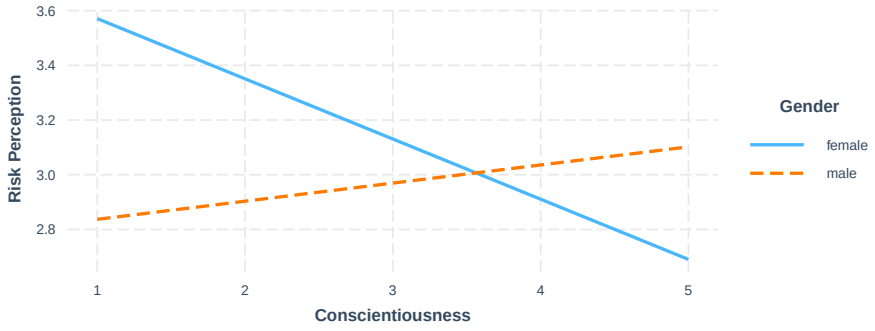
An overview of hypotheses and whether we found support for them is presented in Tables 11 and 12.



(a) Interaction plot for risk perception dependent on gender * age.



(b) Interaction plot for risk perception dependent on gender * extraversion.



(c) Interaction plot for risk perception dependent on gender * conscientiousness.

Figure 11: Interaction plots for risk perception dependent on gender and age, extraversion, and conscientiousness. Regressions adhere to Equation a with the exception of the added interaction term specified in captions 11a–11c.

Table 10: Causal mediation results for a model with subject-fixed effects and omitted controls. Coefficients are z-standardized and t-statistics are presented in parentheses. Standard errors are clustered by subject and text identifier. Regressions include intercepts and fixed effects for text identifier. Letters a–c match paths in Figures 7 and 10.

	Risk Perc (a)	Invest Sum (b)	Invest Sum (c)
Ling Unc	0.54^{***}	-0.51^{***}	-0.69^{***}
	(0.10)	(0.11)	(0.11)
Risk Perc		-0.33^{***}	
		(0.07)	
n	324	324	324
R^2	41.65%	30.63%	24.12%

$^{\dagger} \leq 0.1, ^{*}\ p \leq 0.05, ^{**}\ p \leq 0.01, ^{***}\ p \leq 0.001$

6.5.1 *RQ1: Linguistic Uncertainty and Investment Decisions (H1–H2)*

In line with H1, the results show that the influence of linguistic uncertainty on investment sum is mediated by risk perception. The mediating role of risk perception is a potential explanation that people invest less in a company when the management uses vague communication in earnings calls, as they are at a higher risk of losing the invested money. These findings highlight the important role of managerial language to communicate results and outlooks in earnings calls.

In line with hypothesis H1a, the results suggest that *ceteris paribus*, the frequent use of words reflecting vagueness or other types of linguistic uncertainty in an earnings call, causes an increased risk perception. The analysis indicates that the linguistic uncertainty in earnings calls is by far the most influential factor in predicting risk perception of an investment, compared to other factors like age, annual income, risk tolerance, or financial literacy. Thus, the data contributes to a clearer understanding of the impact of vague communication on risk perception, as this variable was measured directly instead of using stock market figures as a proxy. Furthermore, a causal effect of the linguistic uncertainty in earnings calls on investor risk perception and the amount invested was shown for the first time.

Considering H1b, the analysis supports the theory that a larger risk perception causes investors to invest less money in a company's stock. Thus, the results are following the findings of Byrne (2005) and Nguyen et al. (2019), who find a negative relationship between the perceived risk of a stock and the amount of invested money. We also find support for H1c, which postulates that linguistic uncertainty

Table 11: Hypothesis testing results concerning RQ1: "How does linguistic uncertainty in financial disclosures influence individual investment decision-making?"

Hypothesis		Support
H1	Linguistic uncertainty in financial disclosures causes individuals to allocate smaller investment sums to a company. This effect is mediated by risk perception.	Yes
H1a	Linguistic uncertainty has a negative effect on risk perception.	Yes
H1b	Risk perception has a negative effect on investment sum.	Yes
H1c	Linguistic uncertainty has a negative effect on investment sum.	Yes
H2	Linguistic uncertainty has a positive effect on the variation of investment sums across individuals.	No

Table 12: Hypothesis testing results concerning RQ2: "How do individual characteristics of an investor influence the risk perception regarding an investment?"

Hypothesis		Support
H3	Men have a lower risk perception than women.	No
H4	Age has a positive effect on risk perception.	Yes
H5	Income has a negative effect on risk perception.	Partly
H6a	Conscientiousness has a positive effect on risk perception.	No
H6b	Extraversion has a negative effect on risk perception.	No
H6c	Neuroticism has a positive effect on risk perception.	Yes
H7	Financial literacy has a negative effect on risk perception.	No
H8	Risk tolerance has a negative effect on risk perception.	No

causes smaller investment sums directly. A negative relationship between perceived risk and expected return seems incongruent with traditional financial theory (Sharpe, 1964). However, this result is in line with recent work on the *affect heuristic* (Kempf et al., 2014; Shefrin, 2001; Weber et al., 2013). Hypothetically, linguistic uncertainty could lead to negative affect due to a perceived lack of confidence or competence. Furthermore, vague language is a frequently used linguistic device for obfuscation and deception (Burgoon et al., 2016; Guo et al., 2017). Investors could anticipate this and hence avoid companies with vaguer disclosure language.

Contrary to hypothesis H2, the results show that linguistic uncertainty causes a lower variation in the amount invested instead of an expected higher variation. These results do not match those observed in prior studies demonstrating that language uncertainty in disclosures is positively associated with investor uncertainty as proxied by volatility (Barth et al., 2021; Doshi et al., 2021; Loughran and McDo-

nald, 2011, 2013). Statistically, this previously unexpected result can be explained by the fact that uncertain disclosures lead subjects to believe in declining returns and hence predominantly attract low investment sums. Therefore, the variance across subjects is low with the majority investing less than 10% of the total amount (cf. Figure 8b). Conceptually, there are further open points that might help explaining this result:

1. The results of past work were based on stock prices or forecasts by professional analysts. In contrast, the study results herein are not based on a specific target group and mostly based on unprofessional investors and laypeople. However, work by Holzmeister et al. (2020) suggests that on over-proportional effect of *loss aversion* on risk seems to be robust across different levels of financial literacy.

2. The psychological mechanism between linguistic uncertainty and investor behavior is still not fully understood. An important facet of personality might be trust of subjects and the perceived honesty or humility displayed by company representatives (cf. the HEXACO model of personality (Ashton et al., 2004)). Company representatives using a mostly certain language may be deemed overconfident, hence reducing the invested amount. Conversely, using mostly uncertain language might be perceived as honest or trustworthy, thus increasing the invested amount. Therefore, the investment sums for linguistically uncertain earnings calls may exhibit less variation than those for certain calls.

While previous research has focused on the relationship between linguistic uncertainty and stock price movements or analyst forecasts, these results demonstrate how the linguistic uncertainty in earnings calls influences individual investment decisions. Other driving factors beyond the influence of the disclosing company are found at the individual level with traits such as personality or risk tolerance. The following will shed light on them.

6.5.2 *RQ2: Investor Characteristics and Risk Perception (H3–H8)*

While the coefficient of gender is highly significant, the association behaves in the opposite direction as hypothesized in H3, i.e., men have a larger risk perception than women. As shown in the interaction analysis (Figure 11), gender seems to moderate the effect of age on risk perception. For men, the positive effect of age on risk perception is relatively stronger than for women. Critically, the analyzed sample is imbalanced regarding gender, with the majority of participants being male (61 vs. 20). Hence, larger sample sizes with a more balanced gender ratio might yield different results. Furthermore, we find that the relationship between age and risk perception is highly significant

and positive, aligned with H4. The size of this effect is approximately half as large as that of linguistic uncertainty.

Regarding income, we find that the relationship to risk perception is quadratic, positive, and highly significant. This indicates an upward-facing parabolic relationship, where increases in income at the low end are followed by strong decreases in risk perception. In contrast, the opposite holds for increases in income at the high end. Thus, while a negative effect of income on risk perception can be confirmed for low incomes, the opposite holds for high incomes.

Regarding the Big 5, conscientiousness is positively associated with risk perception, but insignificant. Similar to age, we find evidence that the effect of conscientiousness is significantly moderated by gender: The risk perception of male investors scoring high on conscientiousness increases stronger than for female investors. Although we find that extraversion has a significant association with risk perception, the direction of this relationship is opposite than expected: Extraverted individuals tend to have an increased risk perception compared to introverts. Therefore, we have to reject H6a and H6b. For neuroticism, we find a highly significant and positive relationship with the mediator; hence, we accept H6c. Finally, although the signs of the coefficients behave as expected (negative), contrary to the findings of (Diacon, 2004; Nguyen et al., 2019) and (Gibson et al., 2013; Nguyen et al., 2019), we find no evidence for a significant association between financial literacy or risk tolerance and risk perception. The result regarding financial literacy is however in line with Holzmeister et al. (2020), who find that drivers of risk perception are relatively consistent between experts and laypeople.

6.6 CONCLUSION

Past literature has addressed the tasks of linguistic uncertainty detection or risk regression (based on uncertainty) in finance. However, up until now no laboratory study has been conducted to establish a causal link between both of these tasks. Introducing the first approach of this kind, this chapter provides evidence that vague or otherwise uncertain communication in earnings calls causes increased risk perception and decreased investment sums. In an experimental study, we presented subjects with randomized excerpts of earnings calls. Based on the Loughran and McDonald (2011) dictionary, we manipulated these excerpts to either contain linguistically certain or uncertain words. After having read the excerpts, participants indicated their risk perception and invested fictional money into the respective companies. Finally, they answered items covering various socio-demographic, psychometric, and financial traits.

The survey findings suggest that linguistic uncertainty in financial disclosures causes investors to invest less money into the stock of

the disclosing company. This effect is mediated by risk perception. We found no evidence that linguistic uncertainty is associated with a larger variation of investment sums between subjects.

The exploratory analysis yielded that both age and neuroticism positively affect risk perception, while none of the other hypotheses could be confirmed. We find partial evidence that income has a negative effect on risk perception since the two variables are related in an upward parabolic shape, indicating that increasing income decreases risk perception for low-income individuals. In summary, this study is the first to analyze the causal effect of linguistic uncertainty in conjunction with individual investor characteristics on risk perception in earnings calls. Part IV will move to the prediction of market and analyst reactions to linguistic uncertainty and other representations of disclosure content.

Part IV

RISK REGRESSION

This part contains works addressing the task of **risk regression (T_3)**. What is the influence of linguistic uncertainty on financial risk measures? What other linguistic phenomena are explanatory or predictive of risk? Chapter 7 introduces an econometric approach to explain stock return volatility and analyst-based uncertainty measures based on linguistic uncertainty. Chapter 8 presents a Transformer-based regressor predicting the Myers–Briggs Type Indicator (MBTI) personality of CEOs and using the such predicted personalities to explain volatility. Chapter 9, finally, discusses an assumption-free DL approach to predict volatility based on earnings call transcripts.

7

RISK REGRESSION FROM LINGUISTIC UNCERTAINTY

* In Part II, methods to detect and predict the use of uncertain language in financial disclosures were proposed; in Part III, it was shown that uncertain language influences investment behavior, mediated by risk perception. But do these effects persist in large-scale observational studies? Can we find influences of uncertain language on financial uncertainty measures such as risk with regression analyses? Moreover, how suitable is the dictionary developed by Loughran and McDonald (2011) for such a task?

7.1 INTRODUCTION

In this chapter, we analyze relationships between financial disclosure uncertainty and market reactions. For this purpose, we assemble a corpus of 10-Ks. Moreover, we propose an automatic expansion and filtering method of the Loughran and McDonald (2011) UNCERTAIN dictionary based on multi-task learning. To assess the validity of the findings, we measure the adapted dictionary's explanatory power of financial risk and analyst uncertainty. Overall, we show that the adapted dictionary outperforms the original dictionary and two expansions suggested in our past works from 2018 and 2020. Empirically, linguistic uncertainty seems to increase financial risk and analyst uncertainty.

7.1.1 *Disclosure Type: 10-Ks*

In this work, we focus on disclosure type 10-K, which are annual reports required to be filed by most U.S. companies.[1] 10-Ks usually consist of up to 15 distinct sections; while some studies have argued to focus on specific sections they deem most informative to stakeholders (typically, Section 1a "Risk Factors" or the MD&A), Loughran and McDonald (2014) use the whole 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. Figure 12 depicts the average linguistic uncertainty measured by the

* This chapter expands on the quantitative part of "Explaining Financial Uncertainty through Specialized Word Embeddings" (Theil, Štajner, and Stuckenschmidt, 2020), published in Volume 1, Issue 1 of the *ACM/IMS Transactions on Data Science*. Additions include: Augmenting the data to cover years 1994–2020 (as opposed to 1994–2015), and introducing a dictionary filtering method based on multi-task learning.

1 See §2.2.3 for an in-depth discussion of 10-Ks.

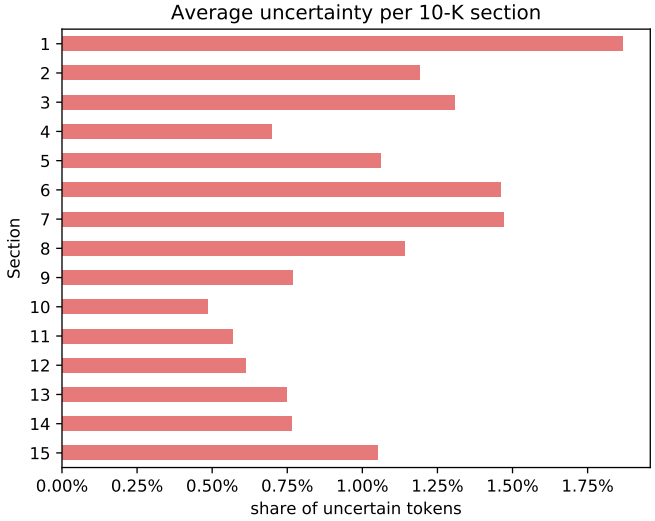


Figure 12: Average share of linguistically *uncertain* tokens (Loughran and McDonald, 2011) per 10-K section. The sample consists of all parsable 10-Ks published between 1994 and 2020 ($n = 217$K). Only sections containing at least 250 words are considered.

Loughran and McDonald (2011) dictionary per 10-K section. As can be seen, other sections than 1a and 7, for example, 3 and 6, can also exhibit considerable uncertainty. Hence, we were further motivated to analyze all documents in their entirety.

7.1.2 *Motivation*

Related work (Rekabsaz et al., 2017; Tsai and Wang, 2014; Tsai et al., 2016) has explored automatic expansions of the Loughran and McDonald (2011) dictionary based on word embeddings. Those approaches consisted of training a word embedding model on a corpus of 10-Ks and adding the top-20 most cosine similar words to each original dictionary term. As the number of related candidates k in related domains is commonly set to 10 (Glavaš and Štajner, 2015; Paetzold and Specia, 2016), we were motivated to explore expansions with $1 \leq k \leq 20$.

Furthermore, we were interested in inducing industry-specific knowledge into the training process of the embedding models. This is motivated by past work suggesting that the specificity of `word2vec` training corpora is more important than their size (Dusserre and Padró, 2017). Conceptually, the financial domain could be understood as a meta-domain comprising various industry-specific sub-domains (e.g., health, mining, or agriculture). Lastly, as an add-on to the experiments presented in 2020, we were motivated to explore automatic term-filtering methods based on multi-task learning. As the size of

the automatic dictionary expansions grows as a function of the number of added candidates and the number of industries considered, the signal-to-noise ratio diminishes. Thus, we propose a method automatically retaining just those terms that are explanatory of the considered financial uncertainty measures. The proposed method is computationally cheap and leverages inter-correlations between the three target variables.

In contrast to prior work, we do not train a predictive model of volatility but instead develop an explanatory one, where the height of the coefficients gauges the estimated effect strength of linguistic uncertainty on financial uncertainty measures. Following the financial methodology of Loughran and McDonald (2014), we use the financial data as an external validation for the assumed correlation. Apart from volatility, we aim to explain two analyst-based measures and include several control variables introduced by the authors mentioned above.

7.1.3 *Contributions*

This work is the first to propose specialized word embedding models which account for industry-specific jargon. Furthermore, a method to retain only relevant terms for explaining external uncertainty measures is introduced. We evaluate the validity and effectiveness of the expansions by providing (1) a brief qualitative analysis of the suggested terms; (2) cross-sectional regression analyses as external validation measuring how well the expanded uncertainty dictionary explains drifts of market and analyst uncertainty. In summary, we contribute to the scientific community by:

- Developing the first word embedding models accounting for industry-specific jargon;
- Proposing an automatic feature selection method based on multi-task learning for retaining the most relevant dictionary candidates;
- Statistically explaining both drifts in stock return volatility and analyst uncertainty with the automatically expanded dictionary.

7.2 DATASET CONSTRUCTION

7.2.1 *Document Parsing*

We download all 10-Ks filed between 1994 and 2020 from the SEC's public filing database EDGAR.[2] For each document, we remove exhibits, graphics, and HTML tags and, following Loughran and McDo-

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

nald (2011), only consider 10-K sections containing at least 250 words. We remove numbers and lowercase all tokens except for proper nouns, which are identified through PoS tagging. Considering only 10-Ks with at least one complete section and dropping duplicates, this leaves us with a dataset of 218K unique documents used to train the word embedding models.

7.2.2 *Data Screens*

VOLATILITY REGRESSION DATA We then perform a set of data screens suggested by Loughran and McDonald (2011, 2014). In particular, we require a match with the financial database Center for Research in Security Prices (CRSP)[3], the stock to be ordinary common equity, a stock price of greater than $3, a positive Book-to-Market (BTM) ratio, as well as at least 60 days of stock return data available for trading day windows t_{-252} to t_{-6} before, one day for t_0 to t_1 around, and at least ten days of data for t_6 to t_{28} after the filing date. This reduces the original dataset to 85K instances available for the volatility regression analyses.

ANALYST REGRESSION DATA For the analyst forecast error regressions, we additionally require at least one analyst forecast of EPS to be present between the filing date and the EPS announcement date on the financial database Institutional Brokers Estimate System (IBES).[4] This dataset has a size of 35K documents. For the analyst dispersion regressions, we require at least two of such forecasts to be present ($n = 23$K).

7.3 METHODOLOGY

We apply the following pipeline to address our task: First, we train word embedding models and expand an existing dictionary of uncertainty triggers (§7.3.1). Afterward, we perform a set of event study regressions as an external validation (§7.3.2).

7.3.1 *Automatic Dictionary Expansion*

We aim to take the established Loughran and McDonald (2011) UNCERTAIN dictionary, which has been shown to possess explanatory power of financial risk (Barth et al., 2021; Dzieliński et al., 2021; Loughran and McDonald, 2011) and to expand it automatically for improved risk regressions. Furthermore, we are interested in testing the applicability of both the original dictionary and our expansions for

3 http://www.crsp.com

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

two other measures of information uncertainty: analyst forecast error and analyst dispersion. These two measures describe how far off EPS forecasts were on average and how dispersed such forecasts were. Larger errors and dispersions indicate increased uncertainty.

As baselines, we use the plain dictionary and expansions developed in our prior work: These include an agnostic expansion (Theil et al., 2018), similar to that of Tsai and Wang (2014) and an industry-specific expansion developed in the published version of this chapter (Theil et al., 2020).

CANDIDATE GENERATION Inspired by Tsai and Wang (2014) and similar to Theil et al. (2018) and Theil et al. (2020), we start by training a `word2vec` embedding model with standard parameters on the full dataset of 218K text documents.[5] Past work proposes to select the top-20 most cosine similar candidates to each original dictionary term. However, those works considered no automatic filtering of such candidates and either took them as-is or conducted manual filtering. Thus, to construct a relatively exhaustive list, we selected a top-k threshold of one order of magnitude larger, i.e., 200. After selecting the 200 most cosine similar terms for each original dictionary entry in the embedding model, we removed stopwords from Porter's dictionary (Porter, 1980) and such words that do not appear in a dictionary of the English language (with 500K terms). This yields an expanded dictionary of 11K terms (compared to the original dictionary with ca. 300 terms).

CANDIDATE SELECTION We were motivated to explore automatic feature selection methods to only retain such terms that are truly explanatory of risk and analyst uncertainty. We frame this task as a regression problem with the tf–idf weighted set of candidate terms (n = 11K) as an input and the three dependent variables *volatility*, *error*, and *dispersion* as output.

We explored regressors meeting two criteria (1) multi-task applicability; and (2) usefulness for feature selection. As we want to explain three output variables, learners able to leverage inter-correlations (i.e., multi-task learners) are particularly suitable for the problem. To achieve results with interpretable coefficients (e.g., *ceteris paribus, X decreases y by p%*), linear regression based on Ordinary Least Squares (OLS) with control variables is most commonly used in finance. There exist two regularized extensions of OLS with out-of-the-box applicability for multi-task learning: the Least Absolute Shrinkage and Selection Operator (LASSO) by Tibshirani (1996) and elastic net (Zou and Hastie, 2005). Our choice fell on the latter, as it entails the LASSO re-

5 In the works from 2018 and 2020, the embedding model was trained only on the residual of files for which no financial data was available (n = 125K and n = 126K, respectively).

gressor while including the advantages of a Ridge regressor (Hoerl and Kennard, 1970). The following provides a formulaic explanation.

The elastic net regressor combines the L_1 penalty of the LASSO regression with the L_2 penalty of the ridge regression via a hyperparameter $0 < \alpha \leq 1$, which specifies the weighing of both. The other tunable hyperparameter is λ, which specifies the shrinkage of the coefficients. The regressor minimizes the following optimization target:

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}} \frac{1}{2n} ||\beta X - Y||_F^2 + \lambda\alpha||\beta||_{2,1} + \frac{\lambda(1-\alpha)}{2}||\beta||_F^2, \qquad (2)$$

where β is the regression coefficient; n is the number of samples (i.e., $23{\times}10^3$); Y is a two-dimensional array of the shape n and the number of tasks, i.e., ($23{\times}10^3$, 3); index F is the Frobenius norm $||\beta||_F = \sqrt{\sum_{ij} \beta_{ij}^2}$; and $||\beta||_{2,1}$ is the product of L_1 and L_2, with $||\beta||_{2,1} = \sum_i \sqrt{\sum_j \beta_{ij}^2}$, i.e., the sum of the norm of each row.[6]

In our experiments, we use the `sklearn 0.24.2` implementation of `MultiTaskElasticNet`. To find a good set of hyperparameters, we run a randomized grid search with a 5-fold temporal CV setup on the subset of data for which all three dependent variables are available ($n = 23$K). We explore $\alpha \in \{0, 1{\times}10^{-1}, 2{\times}10^{-1}, \cdots, 1\}$ and $\lambda \in \{1{\times}10^{-4}, 1{\times}10^{-3}, \cdots, 1{\times}10^{2}\}$ and select the model yielding the largest coefficient of determination R^2.[7] We find $\alpha = 3{\times}10^{-1}$ and $\lambda = 1{\times}10^{-1}$ to be a good set of hyperparameters for the given task. Finally, we run the optimized regressor once on the full dataset and then select all candidate terms for which the coefficients are larger than zero.

COMPARISON TO PRIOR WORK In the work published in 2018, an expansion similar to the ones proposed by Tsai and Wang (2014) and Tsai et al. (2016) and Rekabsaz et al. (2017) performed best for volatility regressions. Here, the 20 most cosine similar terms to each of the ca. 300 terms appearing in the Loughran and McDonald (2011) UNCERTAIN dictionary were added to the list. This yields an expanded dictionary[8] of 4K terms, which we call "agnostic" as it induces no industry-specific jargon.

In addition, we evaluate the industry-specific expansions published in 2020. In this work, such industry-specific expansions performed

6 `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.MultiTaskElasticNet.html`

7 We chose this measure as we aimed to obtain a model maximizing the explained variation of the dependent variables. Furthermore, we were interested in overall explanatory power by considering all covariates (linguistic uncertainty in conjunction with the financial features) for more robust results.

8 All dictionary expansions and data used in this work can be found online (see Appendix A).

better in volatility and analyst forecast regressions than the plain dictionary and the agnostic expansion. As training data, a corpus of 10-Ks in the years 1994–2015 was divided into industry-specific subcorpora based on the established Fama and French (1997) scheme. This scheme comes in different levels of granularity, distinguishing between $\{5, 10, 12, 17, 30, 38, 48, 49\}$ industries (from now on: FF_5 to FF_{49}). Furthermore, for each of the eight levels of granularity, expansions with $1 \leq k \leq 20$ were proposed, i.e., $8 \cdot 20 = 160$ dictionary expansions are evaluated. These dictionaries contain between 14K and 49K terms, depending on the level of industry granularity. The $FF_{49,\,5}$ dictionary, e.g., spans 49 sub-dictionaries (one per industry) and $k = 5$ added candidates per seed term. Thus, for the seed term "anomalous," it contains the candidate "denial-of-service," in the *Hardware* industry, "overheating" in *Lab Equipment* and "unethical" for *Utilities*.

7.3.2 *Event Study Regressions*

For each dictionary and document, we calculate the cumulative tf–idf of uncertain terms to gauge linguistic uncertainty. Specifically, the tf–idf value of a term t in a document d is defined as

$$\text{tfidf}_{t,d} = \text{tf}_{t,d} \cdot \text{idf}_t, \tag{3}$$

where $\text{tf}(t, d)$ is the frequency of t in d normalized by the total number of terms in d and

$$\text{idf}(t) = \log \frac{n}{\text{df}(t)} + 1, \tag{4}$$

where n is the total number of documents and $\text{df}(t)$ the number of documents in which t appears.

We calculate the cumulative tf–idf score for a respective dictionary and document d by summing up the tf–idf scores of the dictionary's terms $w_1, w_2, ..., w_n$ occurring in d. For each dictionary–document combination, this procedure yields a continuous measure of linguistic uncertainty (in the following: *uncertainty*). The documents are the 85K 10-Ks (§7.2).[9]

We evaluate all dictionaries described above (cf. §7.3.1): The automatically filtered expansion retaining only the terms with an above-zero coefficient in a multi-task elastic net regression (cf. §7.3.1); the

9 In line with related work (Rekabsaz et al., 2017; Tsai and Wang, 2014; Tsai et al., 2016), the dataset for candidate selection ($n = 23$K) is a subset of the data used for the regression setup ($n = 85$K). This is a valid methodological choice as the aim of this study is a regression in an explanatory and not a predictive sense (see Shmueli (2010) for the conceptual differences). An acknowledged limitation is a decreased out-of-sample applicability of the induced dictionary. As the proposed model is computationally inexpensive, however, we suggest future work to re-train it on a case-by-case basis instead of applying our learned dictionary out-of-domain without automatic adaptation.

plain uncertainty dictionary (Loughran and McDonald, 2011); the industry-agnostic expansion proposed in 2018; and the 160 industry-specific expansions along the eight industry schemes (FF_5 to FF_{49}) and $k = [1, 20]$ proposed in 2020 (cf. §7.3.1).

To compare which dictionary provides the best assessment of *uncertainty*, for each dictionary and document, we follow Loughran and McDonald (2014) by performing regressions of the financial variables volatility, analyst forecast error (*error*), and analyst dispersion (*dispersion*). Note that different from them, our main explanatory measure is linguistic uncertainty and not readability. The calculation of these three financial variables is outlined below under "Volatility" and "Analyst-Based Measures." All regressions adhere to the following formulae (Equations 5–7):

$$\text{Volatility}_i = \alpha_i + \beta_i \cdot \text{Uncertainty}_i + \delta_i \tag{5}$$

$$\text{Error}_i = \alpha_i + \beta_i \cdot \text{Uncertainty}_i + \delta_i \tag{6}$$

$$\text{Dispersion}_i = \alpha_i + \beta_i \cdot \text{Uncertainty}_i + \delta_i \tag{7}$$

In these equations, α_i is the estimated regression intercept, β_i the estimated slope coefficient for the independent variable *uncertainty*, and δ_i is a vector of control variables which we obtained as outlined under "Control Variables" below. 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.

VOLATILITY We consider the filing date of a 10-K as the event date after which we measure stock return fluctuation (i.e., volatility) attributable to the 10-K disclosure. We follow Loughran and McDonald (2014) and calculate subsequent volatility as the RMSE of a post-filing market model (Sharpe, 1963) using trading days t_6 to t_{28} (approximately a month) relative to the 10-K filing date. To control for historical volatility, we additionally estimate a pre-filing market model using trading days t_{-252} to t_{-6} (approximately a year).

A market model is estimated by regressing a respective company's returns (r_i) on the overall market return (r_m) in said windows. Return data is obtained from the financial database CRSP. As a proxy for r_m, we use the CRSP value-weighted index. The market model regressions adhere to the following formula:

$$r_i = \alpha_i + \beta_i \cdot r_m \tag{8}$$

These regressions yield intercepts α_i and slope coefficients β_i. We use these two variables to estimate expected returns in the given window. We further calculate the volatility, our primary independent variable, as the 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 (Bonsall IV et al., 2017; Loughran and McDonald, 2014, *inter alia*) to obtain a measure of *idiosyncratic* i.e., unsystematic *risk*. Using market model return estimates (called *expected returns*) and quantifying the differences towards 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.

ANALYST-BASED MEASURES In addition to this market-based measure, using data from IBES, we deploy two standard measures of information uncertainty based on analyst forecasts: *analyst forecast error* and *analyst dispersion*. These measures focus on the critical figure *earnings per share*, which indicates the proportion of company profit allocated to each outstanding share. Due to the lower data availability, the sample size gets reduced to 35K for the analyst forecast error and to 23K for the analyst dispersion regressions.

We follow the definitions of analyst forecast error and analyst dispersion by (Loughran and McDonald, 2014): We calculate analyst forecast error as the absolute value of the SUE, defined as

$$\frac{\text{actual earnings} - \text{average expected earnings}}{\text{stock price}}. \tag{9}$$

The *actual earnings* are the earnings per share as published in the earnings announcement. We obtain *average expected earnings* by taking the mean of all earnings forecasts issued by banking analysts between the 10-K filing date and the earnings announcement date. We acquire both figures from the IBES unadjusted data files. We consider only the forecast closest to the filing date for analysts with more than one forecast reported between the 10-K filing and the earnings announcement. Finally, the variable is winsorized at the 1^{st} and at 99^{th} percentile. We calculate analyst dispersion as the standard deviation of analyst forecasts in the forecast error estimate divided by stock price. We retain only firms with at least two analyst forecasts and winsorize the result.

CONTROL VARIABLES Beyond these three independent variables, following (Loughran and McDonald, 2014), we use the following set of control variables within our regressions:

- intercepts α and the RMSE from market model regressions with trading days t_{-252} to t_{-6} as indicators of *historical performance* and *historical volatility* (see §7.3.2 for details);
- *CAR* as the absolute value of the buy-and-hold return in trading days t_0 to t_1 minus the buy-and-hold return of the market index.
- log-transformed *firm size* calculated as current stock price multiplied by the number of outstanding shares;

- log-transformed BTM ratio, calculated as the book value of equity according to Compustat[10] divided by the market value of equity according to CRSP (firms with a negative book value are removed and the variable is winsorized at the 1st and at the 99th percentile);
- a *NASDAQ dummy* set to one if the firm is listed on the NASDAQ and 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 (1997) 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. All variables are z-standardized by subtracting the mean and dividing by the standard deviation.

7.4 RESULTS AND DISCUSSION

The discussion of our results is two-fold: §7.4.1 presents a brief qualitative analysis exploring the suitability of the dictionary expansions to capture uncertainty. §7.4.2 provides regression analyses assessing how well the models are suited to explain financial uncertainty in terms of volatility, analyst forecast error, and analyst dispersion.

7.4.1 *Qualitative Analysis*

We start by showcasing the candidate terms to which the multi-task elastic net regressor assigned coefficients above zero (cf. §7.3.1). This yields 175 terms, two orders of magnitude below the original set of candidates ($n = 11$K). The top-10 terms according to their average coefficient are depicted in Table 13. The total list of terms can be found in Appendix C, Table 32.

As evident in Table 13, the selected set of candidates captures aleatory ("coronavirus," "burglaries") or epistemic components of uncertainty ("nonhomogenous," "postulates"). Nevertheless, it also contains noise in the form of likely overfit candidates such as "amputated" or "hyperglycemia." Furthermore, the inclusion of the term "coronavirus," which did not appear in our work from 2020, reveals the time-dependence of the model. Hence, we suggest training such a model on a case-by-case basis instead of applying our learned dictionary (cf. Appendix C) out-of-domain without adaptation.

An advantage over the previously proposed model in 2020 is the greatly reduced number of candidates and computational complexity:

10 http://www.crsp.com/products/research-products/crspcompustat-merged-database

Table 13: Top-10 terms with a positive coefficient according to the multi-task elastic net regressor. β stands for regression coefficient and the subscripts represent the three primary independent variables: post-filing RMSE, SUE, and analyst forecast dispersion (DIS). Terms are sorted in descending order according to their mean coefficient ($\bar{x}_\beta$).

Term	β_{RMSE}	β_{SUE}	β_{DIS}	$\bar{x}_\beta$
territoriality	0.001	0.347	0.132	0.160
disreputable	0.015	0.230	0.193	0.146
amputated	0.017	0.105	0.114	0.079
warrants	0.052	0.033	0.031	0.039
hyperglycemia	0.005	0.051	0.049	0.035
coronavirus	0.058	0.023	0.015	0.032
nonhomogeneous	0.001	0.077	0.013	0.030
raise	0.027	0.030	0.032	0.029
postulates	0.002	0.047	0.037	0.028
burglaries	0.012	0.049	0.019	0.027

Instead of evaluating $8 \cdot 20 = 160$ different parameter combinations or training $5 + 10 + 12 + 17 + 30 + 38 + 48 + 49 = 209$ different word embedding models, only 40 combinations of hyperparameters had to be evaluated, and one single word embedding model had to be trained. In the following, we will move to the performance gains obtained with such a model.

7.4.2 *Quantitative Analysis: Event Study Regressions*

How valid are our expansions in a real-world application? In this section, we analyze the beneficial effect of our automatic dictionary expansions in event study regressions of the financial uncertainty measures volatility, analyst forecast error, and analyst dispersion. The results of these regressions are summarized in Tables 15a–15c.

VOLATILITY REGRESSIONS Consistent with prior research, linguistic uncertainty and volatility are positively related (see Table 15a). The best industry-specific expansion is achieved with FF_{12} and $k = 8$ (coefficient $\beta = 0.02$, significant at the 1%-level). However, the value is equal to the one of the plain dictionary. The expansion with automatic filtering based on multi-task learning yields the largest coefficient ($\beta = 0.08$, significant at the 1%-level). This value is four times higher than both the plain dictionary and the industry-specific expansion without candidate filtering.

Table 14: Regression results. "LM" is the UNCERTAINTY dictionary, "Agnostic" an industry-agnostic expansion, "$FF_{i,j}$" an industry-specific expansion (with i industries and j candidates), and "ENet" the automatically filtered expansion. Regressions include NASDAQ, industry, and year dummies. $^{\dagger}p \leq 0.1$,* $p \leq 0.05$,** $p \leq 0.01$,*** $p \leq 0.001$.

(a) Regression results for the dependent variable volatility, measured as RMSE of a post-filing market model using trading days t_6 to t_{28} (n = 85K). Largest effect size of uncertainty depicted in bold.

	Post RMSE			
	LM	Agnostic	$FF_{12,8}$	ENet
Uncertainty	0.02^{*}	0.03^{**}	0.02^{***}	$\mathbf{0.08}^{***}$
Alpha	-0.07^{**}	-0.07^{**}	-0.08^{**}	-0.07^{***}
Prior RMSE	0.43^{***}	0.43^{***}	0.46^{***}	0.42^{***}
CAR(0,1)	0.10^{***}	0.10^{***}	0.11^{***}	0.10^{***}
Size	-0.18^{***}	-0.18^{***}	-0.16^{***}	-0.17^{***}
BTM	-0.05^{**}	-0.05^{**}	-0.07^{**}	-0.05^{**}
Adjusted R^2	0.47	0.47	0.47	0.48

(b) Regression results for the dependent variable analyst forecast error, measured as SUE (n = 35K). Largest effect size of uncertainty depicted in bold.

	SUE			
	LM	Agnostic	$FF_{49,1}$	ENet
Uncertainty	-0.03^{*}	-0.00	-0.03^{*}	$\mathbf{0.06}^{*}$
Alpha	-0.11^{***}	-0.11^{***}	-0.11^{***}	-0.11^{***}
Prior RMSE	0.27^{***}	0.27^{***}	0.27^{***}	0.26^{***}
CAR(0,1)	0.05^{***}	0.05^{***}	0.05^{***}	0.05^{***}
Size	-0.22^{***}	-0.22^{***}	-0.22^{***}	-0.21^{***}
BTM	0.13^{***}	0.13^{***}	0.13^{***}	0.13^{***}
# Analysts	-0.03^{*}	-0.03^{*}	-0.03^{*}	-0.03^{*}
Adjusted R^2	0.18	0.18	0.18	0.18

(c) Regression results for the dependent variable analyst dispersion, measured as the standard deviation of analyst forecasts (n = 23K). Largest effect size of uncertainty depicted in bold.

	DIS			
	LM	Agnostic	$FF_{48,4}$	ENet
Uncertainty	0.00	0.01	$0.05^{\dagger}$	$\mathbf{0.09}^{**}$
Alpha	-0.09^{***}	-0.09^{***}	-0.09^{***}	-0.09^{***}
Prior RMSE	0.23^{***}	0.23^{***}	0.23^{***}	0.21^{***}
CAR(0,1)	0.04^{***}	0.04^{***}	0.04^{***}	0.04^{***}
Size	-0.16^{***}	-0.16^{***}	-0.15^{***}	-0.14^{***}
BTM	0.11^{***}	0.11^{***}	0.11^{***}	0.11^{***}
# Analysts	0.02	0.01	0.02	0.02
Adjusted R^2	0.14	0.14	0.14	0.14

Furthermore, the control variables behave similarly for all volatility regressions: Firms with high pre-filing performance and market value are subject to less post-filing volatility. Firms with a low BTM ratio, with higher pre-filing volatility and larger unexpected returns around the filing date experience higher volatility.

ANALYST MEASURE REGRESSIONS Using the augmented dataset also covering years 2016–2020, our expansions developed in 2020 perform weaker than before with decreased significance thresholds. For the analyst forecast error regressions (cf. Table 15b), they yield a negative relationship between uncertainty and analyst forecast error ($\beta = -0.03$, significant at the 5%-level); this result is achieved with the FF_{49} scheme and $k = 1$. Consistent with theory and prior empirical results, the automatically filtered expansion yields a positive relationship ($\beta = 0.06$, significant at the 5%-level); that is, 10-Ks with more linguistic uncertainty tend to be followed by more erroneous analyst forecasts.

For the analyst dispersion regressions (cf. see Table 15c), the plain dictionary and the agnostic expansion yield insignificant coefficients of 0 and 0.01, respectively. The best industry-specific expansion (FF_{48}, $k = 4$) features a significant coefficient at the 10%-level of 0.05, i.e., uncertainty and dispersed analyst forecasts seem to be positively related. Again, the automatically filtered expansion leads to the most decisive results with a coefficient of 0.09 that is significant at the 1%-level.

For both sets of regressions, the control variables follow similar patterns again: firms with higher analyst uncertainty tend to be smaller and subject to lower performance, more volatility before the filing, and higher BTM ratio. The only differing control variable is the number of analysts, which is negatively related to forecast error and positively to dispersion.

DISCUSSION We now discuss 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 best-performing expansion, the automatically filtered one) is related to an increase of 6% to 9% of financial uncertainty's standard deviation. While these coefficients might appear small, they are well in line with recent research: For example, Bonsall IV et al. (2017) 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" (Loughran and McDonald, 2016, p. 1202). In summary, all regressions of financial uncertainty on linguistic uncertainty benefit substantially

from an automatically expanded dictionary. This beneficial effect is most profound for the volatility and analyst dispersion regressions.

7.5 CONCLUSION

The Loughran and McDonald (2011) dictionary is considered to be the state-of-the-art for financial uncertainty detection in finance. As it is hard to manually create an exhaustive dictionary (and given the time-dependence of terms deemed to be uncertain, cf. §7.4.1) past financial domain literature proposed methods to expand this dictionary automatically. Such expansions can improve risk regressions; however, no work has considered an expansion balancing the inherent bias–variance trade-off of such a task.

This chapter introduces a hybrid approach addressing this trade-off for the first time. In particular, we select an exhaustive set of candidates and propose an automatic filtering method. This method is based on a multi-task learner jointly maximizing the explanatory power of regressors of risk and two measures of analyst uncertainty. We have shown that such a method improves over the previously developed unfiltered methods.

In the following chapters, we will expand the focus of the risk regression task (T_3) by exploring other variables apart from linguistic uncertainty. Specifically, the next chapter introduces a risk regression model based on CEO personality.

8

RISK REGRESSION FROM CEO PERSONALITY

* Earnings calls, the main financial disclosure explored in this thesis, are characterized by high spontaneity and authenticity. This is because in these calls, company managers talk directly to analysts and investors which differentiates them from other disclosures like 10-Ks that are written by a large group of executives and investor relations specialists. So far, however, we have not capitalized on this advantage: Modeling personal characteristics of individual speakers and using them to explain financial risk might prove as a fruitful avenue of research. This chapter introduces predictive model of CEO personality and explores which components of personality have the most explanatory power of risk with event study regressions.

8.1 INTRODUCTION

How much influence does the personality of a CEO have on their company's performance? The personal news and antics of famous CEOs like Elon Musk, Jeff Bezos, or Bill Gates make headlines, and their personalities sometimes generate a cult-like following. But what measurable effect do they really have? The *upper echelons theory* (Hambrick and Mason, 1984) suggests that the personalities of CEOs also reflect in the organizational outcomes of their companies. However, presumably due to the lack of labeled data, no supervised models exist to detect CEOs' personalities from text and infer their effect on the financial performance of companies. In this paper, we close this research gap by presenting the first Transformer-based model to predict the impact of CEOs' MBTI personality on financial risk.

Ideally, personality is assessed with self-reported questionnaires. However, it is technically infeasible to request executives such as Elon Musk to fill out targeted pen and paper questionnaires. We were therefore motivated to explore crowd-sourced data. This approach is supported by past research showing that observer reports are an inexpensive and valid alternative to self-reports (Vazire, 2006), as they usually agree with them (Kim et al., 2019), and are particularly suitable for the assessment of top management personality (Connolly et al., 2007).

The dominant personality model is the Big 5, which presents personality on a continuum along the dimensions *openness*, *conscientious-*

* This chapter is based on "Top-Down Influence? Predicting CEO Personality and Risk Impact from Speech Transcripts" (Theil, Hovy, and Stuckenschmidt, 2023), accepted at the 17^{th} *International Conference on Web and Social Media (ICWSM)* in Budapest.

ness, *extraversion*, *agreeableness*, and *neuroticism* (McCrae and John, 1992). The available data source we use lacks Big 5 ratings, so as proxy, we explore the MBTI (Briggs-Myers and Myers, 1995), which has been shown to correlate along the main dimensions with the Big 5 (Furnham, 1996; Furnham et al., 2003; McCrae and Costa, 1989). This model represents personality via the categories *extraversion–introversion*, *sensing–intuition*, *thinking–feeling*, and *judging–perceiving*. Addressing methodological criticism of the MBTI (McCrae and Costa, 1989), we

- explore an alternative MBTI representation as a vector of continuous values (§8.3.1);
- find a high internal and external validity of this measure (§8.3.1);
- show that it can be predicted from text (§8.3.3);
- and demonstrate that it is predictive of financial risk (§8.4.3).

Overall, our findings lend empirical support to the *upper echelons theory* of management.

8.2 BACKGROUND AND RELATED WORK

Various personality measures exist in the literature. This section describes the personality model we explore (MBTI), the de-facto standard model (Big 5), and approaches to predict both representations of personality from text.

8.2.1 *MBTI*

The MBTI is named after Katherine Cook Briggs and Isabel Briggs Myers. They developed it based on the work of the analytical psychologist Carl Jung (Briggs-Myers and Myers, 1995). The MBTI classifies personalities binarily along the following axes:

- *extraversion* vs. *introversion* (E–I): describing an out- or inward-oriented social attention;
- *sensing* vs. *intuition* (S–N): information processing based on perceivable/known facts or conceptualization and imagination;
- *thinking* vs. *feeling* (T–F): decision-making based on logic and rationality or emotions and empathy;
- *judging* vs. *perceiving* (J–P): quick judgement and organized action or observation and improvisation on-the-go.

Combined, the four labels form one of 16 personality types (e.g., "ENTJ"). The MBTI is widely used in human resources management and by laypeople as a tool for self-exploration.

Psychological literature, however, has called assumptions of the MBTI into question. For example, McCrae and Costa (1989) find no evidence that personality can be binarized or distinguished into 16 different types. In addition, they find moderate to strong correlations between MTBI and Big 5 (McCrae et al., 2010), which is described in greater detail below (§8.2.2). We re-assess these correlations in our dataset and explore a continuous representation of the MBTI in line with the Big 5.

MBTI PREDICTION FROM TEXT In a literature study on text-based personality detection and a subsequent annotation study, Štajner and Yenikent (2020, 2021) conclude that predicting the MBTI from textual data is a difficult task. They hypothesize that this is due to the theoretical and qualitative origin of the index, which distinguishes it from the empirical and quantitative Big 5. In particular, the dimensions *sensing* vs. *intuition* (S–N) and *judging* vs. *perceiving* (J–P) depend on behavioral rather than linguistic signals (Štajner and Yenikent, 2020, p. 6291).

In a field survey of project managers, Cohen et al. (2013) show that managers are significantly more often of the *intuitive* (N) and *thinking* (T) type than the general population. We observe a similar pattern in our dataset (§8.3.1, Figure 14). Classifying the MBTI of Twitter users based on count-based features, gender, and tweet *n*-grams, Plank and Hovy (2015) outperform a majority class baseline for the E–I and the T–F dimensions. Gjurković and Šnajder (2018) predict the self-reported MBTI of Redditors with SVM and Multilayer Perceptron (MLP) models based on linguistic and activity-level features. Their model outperforms a majority class baseline across all dimensions with the best results for E–I, followed by S–N, J–P, and T–F.

We compare the best-performing approaches identified by prior MBTI prediction studies (*n*-grams and Linguistic Inquiry and Word Counts (LIWC) dictionaries with SVMs and MLPs) to Transformer architectures. Furthermore, we consider a different domain (spoken financial disclosures) and perform a regression instead of a classification.

8.2.2 *Big 5*

The Big 5 are the established psychometric model. Here, personality is represented as a continuum along the five axes *openness*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism* (McCrae and John, 1992).

BIG 5 PREDICTION FROM TEXT As part of the *myPersonality* project, Kosinski et al. (2015) find that liked Facebook pages predict Big 5, IQ, and other personal characteristics to varying degrees. Mairesse et al.

(2007) create a text-based Big 5 prediction tool based on student essays and speech recordings.

Benischke et al. (2019) show that CEOs' Big 5 personalities moderate the relationship between CEO compensation and risk-taking. Hrazdil et al. (2020) use IBM WATSON PERSONALITY INSIGHT to predict the Big 5 of C-level executives in earnings calls and find that an executive's personality is associated with their risk tolerance and company audit fees. Harrison et al. (2020) find that CEO Big 5 are related to perceived firm risk and shareholder value. Another finding is that CEO *conscientiousness* moderates the effect of financial risk on returns positively, while the opposite holds for *extroversion* and *neuroticism*.

Different to these approaches, we focus on the MBTI rather than the Big 5. We create the first supervised model to predict CEOs' MBTI personality from text by collecting a new dataset of crowd-annotated MBTI profiles. This sets us apart from prior work using unsupervised approaches trained on out-of-domain corpora.

8.3 PERSONALITY PREDICTION

Using transcribed speech data as an input, we predict the MBTI personality of CEOs via text regression. The following sheds light on the dataset collection and validation, methodology, and results.

8.3.1 *Dataset Curation*

For this task, we collect data from two sources: (1) text data and (2) crowd-sourced personality data.

TEXT DATA We obtain 88K earnings call[1] transcripts spanning years 2002–2020 from REFINITIV EIKON.[2] Earnings calls are quarterly teleconferences consisting of a scripted presentation and a spontaneous Q&A session, in which CEOs such as Elon Musk answer open questions of banking analysts. Due to the improvised nature of these answers, earnings calls are particularly suitable for detecting personal style (Malhotra et al., 2018). Figure 13 shows an excerpt of Tesla's Q1 earnings call in 2020.

Given the dialogue nature of the calls, we need to map utterances to individual CEOs as we are not interested in the personality of the analysts. We identify CEO names with regular expressions and minimal preprocessing (e.g., stripping middle name initials or titles). Next, we require a match with the executive database COMPUSTAT EXECUCOMP for age and gender data (§8.4.2),[3] reducing our initial sample to 22K

1 See §2.2.3 for an in-depth discussion of earnings calls.

2 https://eikon.thomsonreuters.com/index.html

3 https://wrds-www.wharton.upenn.edu

ELON MUSK (CEO): Thank you. So Q1 ended up being a strong quarter despite many challenges in the final few weeks. This is the first time we have achieved positive GAAP net income in a seasonally weak first quarter. Even with all the challenges, we achieved a 20% automotive gross margin, excluding regulatory credits, while ramping 2 major products. What we've learned from this is that—we've obviously learned a lot here.

Figure 13: Excerpt of Tesla's Q1 2020 earnings call.

Table 16: CEO examples for each MBTI dimension from our dataset.

MBTI	CEO Examples
Extraversion	Steve Jobs (Apple), Lisa Su (AMD), Mary Barra (General Motors)
Introversion	Rupert Murdoch (Fox), Mark Zuckerberg (Facebook), Sheldon Adelson (Las Vegas Sands)
Sensing	Jack Dorsey (Twitter), John Schnatter (Papa John's), Marcus Lemonis (Camping World)
Intuition	Marissa Mayer (Yahoo), Bob Iger (Disney), Evan Spiegel (Snap)
Thinking	Elon Musk (Tesla), Tim Cook (Apple), Steve Ballmer (Microsoft)
Feeling	Sundar Pichai (Google), Howard Schultz (Starbucks), Naveen Jain (Infospace)
Judging	Jeff Bezos (Amazon), Larry Ellison (Oracle), Martha Stewart (Martha Stewart Living)
Perceiving	Larry Page (Alphabet), Martin Shkreli (Retrophin), Donald Trump (Trump Entertainment)

calls and 1.7K CEOs. For these, we retrieve all of their utterances in the presentation and the Q&A session of the calls.

PERSONALITY DATA We obtain MBTI personality labels for the CEOs from PERSONALITY DATABASE,[4] which provides crowd-sourced personality profiles for celebrities, managers, and other noteworthy people. While each profile features vote results for the four dimensions of the MBTI, a minority also contains results for the Big 5. We find that 32 CEOs (e.g., Elon Musk and Steve Jobs) from our earnings call sample have at least three MBTI votes available. The minimum, maximum, and mean votes per CEO are 3, 1.8K, and 140, respectively. These CEOs participate in a total of 736 earnings calls. Table 17 gives the descriptive statistics of the merged text–personality data, and Table 16 contains example CEOs from our dataset across the MBTI.

Instead of representing each personality as one of 16 types, we represent each personality profile as a vector of 4 continuous variables ranging from 0 to 1, based on the crowd-sourced votes. We normalize the votes for the right-hand side of a scale s by the total votes:

$$\text{personality}_s = \frac{\text{votes}_{1,s}}{\text{votes}_{0,s} + \text{votes}_{1,s}}. \tag{10}$$

For example, for the E–I scale, we divide the votes for introversion (I) by the total votes for E and I. The resulting number is thus the likelihood of the CEO being intro- or extroverted. This representation

4 https://www.personality-database.com/

Table 17: Statistics of the CEO–call data considered for the personality prediction. Sums (Σ_x), averages ($\bar{x}$), minima ($\min_x$), and maxima ($\max_x$) are computed across all earnings calls ($n = 736$).

Unit	Σ_x	$\bar{x}$	$\min_x$	$\max_x$
utterances	13,183	17.91	2	124
sentences	111,781	151.88	2	563
tokens	2,526,473	3,432.71	22	9,968

Table 18: IAA per MBTI dimension in terms of percentage agreement (p_a), Krippendorff's α, Brennan–Prediger coefficient (κ_{bp}), and Gwet's γ.

MBTI	p_a	α	κ_{bp}	γ
E–I	87.45	0.40	0.75	0.76
S–N	80.20	0.43	0.60	0.62
T–F	83.33	0.14	0.67	0.71
J–P	90.62	0.17	0.81	0.88

is similar to the Big 5 model (excluding the *neuroticism* dimension) and allows for a more granular representation of personality than the usual operationalization of the MBTI. Figure 14 shows the distributions of the such obtained continuous labels. Most CEOs in our sample are rather *extroverted*, *intuitive*, *thinking*, and *judging* (Figure 14), which corresponds to the ENTJ "Decisive Strategist" MBTI type.[5]

INTERNAL VALIDATION To assess the validity of the crowd-sourced votes, we analyze the IAA between the MBTI raters of the 32 CEOs (Table 18). While p_a is high with values ranging between ca. 80 and 90%, Krippendorff's α (Krippendorff, 2013) yields only slight to moderate values between 0.14 and 0.43. Quarfoot and Levine (2016) call this phenomenon the "frequency distribution paradox," where highly skewed label distributions combined with high percentage agreements can lead to low values of α. As measures robust to this undesirable property, they suggest the Brennan–Prediger coefficient κ_{bp} (Brennan and Prediger, 1981) and Gwet's γ (Gwet, 2008), which in our case yield a high IAA between 0.60 to 0.88.

EXTERNAL VALIDATION To get a notion of external validity, we construct a correlation matrix between the crowd-based MBTI and Big 5 votes of *all* 2.2K profiles with more than three votes available on PERSONALITY DATABASE (Figure 15). According to McCrae and Costa (1989) and subsequent work (Furnham, 1996; Furnham et al., 2003),

5 https://eu.themyersbriggs.com/en/tools/MBTI/MBTI-personality-Types/ENTJ

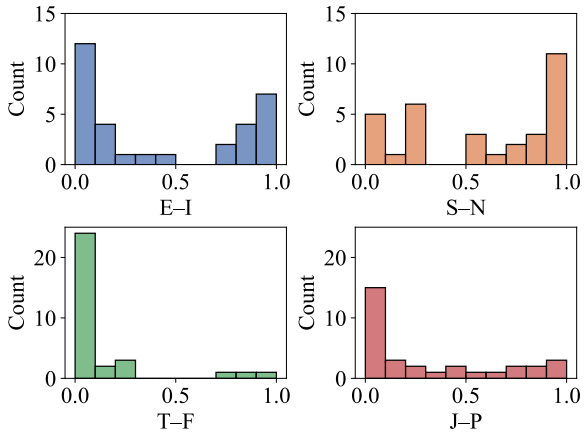


Figure 14: Label distributions for all CEOs considered in the personality prediction ($n = 32$) across the MBTI dimensions *extraversion–introversion* (E–I), *sensing–intuition* (S–N), *thinking–feeling* (T–F), and *judging–perceiving* (J–P).

strong correlations should exist between MBTI *introversion* and Big 5 *extraversion* ($r = -0.74$) as well as between MBTI *intuition* and Big 5 *openness* ($r = 0.72$). Furthermore, moderate correlations should exist between MBTI *feeling* and Big 5 *agreeableness* ($r = 0.44$) and between MBTI *perceiving* and Big 5 *conscientiousness* ($r = -0.49$). Our results confirm the findings of McCrae and Costa (1989) with similar correlations in the first two rows and stronger correlations in the third and fourth rows. This is most likely due to our increased sample size ($n = 2.2$K vs. $n = 267$).

8.3.2 *Methodology*

For each of the 32 CEOs appearing in 736 CEO–call instances, we compare sparse approaches suggested by past literature to Transformer architectures for a regression of MBTI personality.[6]

DATA SPLIT We apply an 80:10:10 split to our data to obtain separate training ($n = 568$), validation ($n = 84$), and test sets ($n = 84$). To avoid overfitting, we use `sklearn`'s `GroupShuffleSplit` with the CEO names as group splitting criterion, i.e., we split the data such that no CEO present in the training data appears in the validation or test data.

6 Appendix A contains a link to our implementation along with the earnings call identifiers. Using those, our corpus can be re-assembled from REFINITIV EIKON, SEEKING ALPHA, or alternative sources.

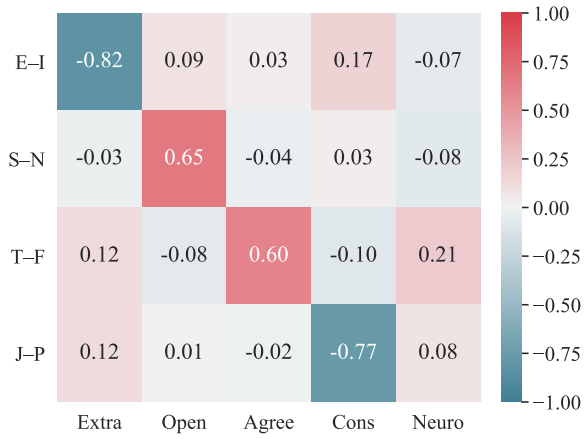


Figure 15: Correlation of MBTI (y-axis) and Big 5 (x-axis) scales for all profiles on the PERSONALITY DATABASE with at least three votes ($n = 2.2$K).

NORMALIZATION Given the highly skewed distributions, after the train–validation–test split, we apply a Box-Cox transformation (Box and Cox, 1964) to y with the following formula:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0, \\ \ln(y) & \text{for } \lambda = 0. \end{cases} \tag{11}$$

We obtain λ via maximum-likelihood estimation. The resulting transformation makes the four label distributions more Gaussian-like by stabilizing variance.

TRANSFORMERS We explore cased-vocabulary $\text{BERT}_{\text{base}}$ (12-layer, 768-hidden, 12-heads, 109M parameters) (Devlin et al., 2019) and $\text{RoBERTa}_{\text{base}}$ (12-layer, 768-hidden, 12-heads, 125M parameters) (Liu et al., 2019) models with a linear regression head. The models are trained with a maximum sequence length of 512 and a sliding window approach. We determine the training batch size and learning rate by running a Bayesian optimization over the grid of batch sizes $b \in \{32, 64, 128, 256\}$ and learning rates $l \in [0, 5.00 \times 10^{-5}]$.[7] We train a model for up to 10 epochs and early stopping with a patience of one epoch. For each of the four MBTI dimensions, we evaluate 40 combinations of hyperparameters and select the model with minimal loss on the validation set. Different to the MSE loss, which is implemented per default in the 🤗 Transformers (Wolf et al., 2020) regressors, we minimize the L1 or alternatively called Mean Absolute Error (MAE) loss, which is less sensitive to outliers.

7 Final hyperparameter choices and results on our validation set can be found in Appendix D.

SPARSE METHODS We also explore the sparse representations suggested by Plank and Hovy (2015) and Gjurković and Šnajder (2018). These include tf–idf vectors with n-grams of length $n \in \{1, 2, 3\}$ and dictionary features across all dimensions of LIWC 2015 (Pennebaker et al., 2015) fed into SVM and three-layer MLP regressors. We compare all possible feature–algorithm combinations with respect to their average MAE on the validation set and select the combination with the lowest error (SVM with trigram tf–idf).

EVALUATION The final model performance is evaluated by inspecting the correlation and error between test set ground truth and prediction. As measures, we explore the linear correlation coefficient (i.e., Pearson's r) and the rank correlation coefficients Spearman's ρ and Kendall's τ. Instead of linear relationships, the latter two measure monotonic relationships and are more robust to outliers. In addition, we consider the error measure MAE, which is the minimized loss function of the Transformers. In case of a tie, we give precedence to τ, as this measure is least sensitive to outliers and particularly suited for small sample sizes.

8.3.3 *Results and Discussion*

The results of the personality prediction task are depicted in Table 19. An SVM performs competitive, especially for the dimensions E–I ($\tau = 0.44$) and S–N ($\tau = 0.20$). While the SVM outperforms BERT for all dimensions except for J–P, RoBERTa achieves the best results in most cases.

The largest correlations across all models are achieved for the *extraversion–introversion* (E–I) scale with strong linear and rank correlations for the RoBERTa regressor ($r = 0.70$, $\rho = 0.66$). This result is not surprising, as distinguishing between *extra-* and *introverted* CEOs based on linguistic style should be comparably easy. This is followed by the *sensing–intuition* (S–N) scale with moderate to strong correlations ($r = 0.45$, $\rho = 0.53$) and the *judging–perceiving* (J–P) scale with weak to moderate correlations ($r = 0.40$, $\rho = 0.36$).

The worst results are obtained for the *thinking–feeling* (T–F) scale, with the SVM and RoBERTa obtaining correlations of around zero and BERT even obtaining weak to moderate negative correlations. There are several possible explanations for this: Conceptually, it could be the case that this dimension simply can not be captured by analyzing linguistic data. Furthermore, the predictive power could be low due to the comparably small sample size. Lastly, we hypothesize that the skewness of the label distribution, which was the highest across all MBTI dimensions for the T–F scale (Figure 14), has contributed to the weak performance. This warrants further research exploring whether

Table 19: Correlation results of the personality regression task. CEO personality is predicted across the MBTI dimensions *extraversion–introversion* (E–I), *sensing–intuition* (S–I), *thinking–feeling* (T–F), and *judging–perceiving* (J–P). SVM is trained on trigram tf–idf vectors, $\text{BERT}_{\text{base}}$, and $\text{RoBERTa}_{\text{base}}$ on text. Best results in bold.

MBTI	Model	r	ρ	τ	MAE
E–I	SVM	0.57	0.58	0.44	0.38
	BERT	0.39	0.35	0.22	0.59
	RoBERTa	**0.70**	**0.66**	**0.52**	**0.34**
S–N	SVM	0.32	0.36	0.20	0.30
	BERT	0.08	0.23	0.16	0.46
	RoBERTa	**0.45**	**0.53**	**0.38**	**0.28**
T–F	SVM	**0.03**	−0.12	−0.08	**0.37**
	BERT	−0.47	−0.41	−0.27	0.41
	RoBERTa	0.01	**−0.10**	**−0.07**	0.39
J–P	SVM	−0.05	0.04	0.02	**0.35**
	BERT	0.39	**0.38**	**0.25**	0.52
	RoBERTa	**0.40**	0.36	0.21	0.36

our findings hold for larger datasets with less skewed label distributions.

Štajner and Yenikent (2020) hypothesize that the S–N and J–P dimensions should theoretically make for the worst candidates in a text-based personality prediction task since they capture behavioral rather than linguistic dimensions of personality. Although our regressors perform worse on these dimensions than for the *extraversion–introversion* dimension, they still achieve moderate to strong correlations, showing that even the more latent dimensions of personality can be predicted from text.

8.4 RISK REGRESSION

According to *upper echelons theory* (Hambrick and Mason, 1984), strategic choices and performance measures of organizations can be predicted by characteristics of their top management. As a use case for our personality prediction task, we explore whether we can find empirical support for this theory. We hypothesize that having a different personality to most CEOs (i.e., ENTJ, see Figure 14 and Cohen et al. (2013)) should translate into increased financial risk.

8.4.1 *Dataset Curation*

As a basis for the risk regression task, we take the sample of 22K earnings calls and merge it with data obtained from the databases CRSP, IBES, and COMPUSTAT EXECUCOMP, which we access via WRDS.[8] As main label, we use risk in the business week following each call, which was proposed by earlier risk regression studies based on earnings calls Theil et al., 2019; Wang and Hua, 2014. To evaluate whether effects on risk persist in larger time frames, we follow related work Qin and Yang, 2019; Yang et al., 2020 and additionally consider risk in the period up to a business month after the call date. In particular, we measure risk as the stock return volatility from business day 0 to business day $\tau \in \{5, 10, 15, 20\}$, where 20 days correspond to a business month. We use the sample standard deviation of logarithmic stock returns for more robust measures.

As features, we incorporate a comprehensive set of risk proxies suggested by Price et al. (2012) and Theil et al. (2019).[9] Furthermore, we include sentiment gauged via the Loughran and McDonald (2011) dictionary as well as CEO age and gender to control for possible confounding effects (e.g., being introverted could have a different effect for male than for female CEOs). Definitions of all used controls are given in Table 20.

8.4.2 *Methodology*

We use the best-performing personality prediction model (RoBERTa) to infer the personality of the 1.7K unlabelled CEOs present in the 22K calls. Together with the financial covariates (see above), the predicted CEO MBTI is then used to explain short-term stock return volatility following the calls with multiple linear regression.[10] Volatility is the most common financial risk measure, and its prediction is an essential task for firm valuation and financial decision-making. Importantly, "risk" is a purely descriptive concept in finance, as it measures the fluctuation of stock returns.

8.4.3 *Results and Discussion*

The results of this risk regression task are shown in Table 21. We find that the first three MBTI dimensions are significantly associated with risk in the business week following the call. This significance is high ($p \leq 0.001$) for E–I and T–F. The direction of this association

8 `https://wrds-www.wharton.upenn.edu`

9 We initially also considered including a market volatility index (VIX), but decided against it as its low explanatory power and high variation inflation factor (VIF) indicated redundancy of this variable (Johnston et al., 2018).

10 Appendix A contains a link to our dataset and implementation.

Table 20: Controls used in the risk regression task. BTM is calculated following (Fama and French, 2001) and firms with a negative value are removed. Size, BTM, and volume are log1p-transformed.

Feature	Definition
Positivity	sentiment according to the LM Positive dictionary by Loughran and McDonald (2011)
Negativity	LM Negative sentiment
Uncertainty	LM Uncertain sentiment gauging economic and linguistic uncertainty
Age	CEO age on the call date
Gender	CEO gender
Past Vola	Standard deviation of logarithmic returns in the business quarter before the call
Size	Market value of the firm, i.e., the number of outstanding shares times stock price one day before the call
Volume	Stock trading volume on the call date
Leverage	Total liabilities divided by assets
Spread	Difference between the stock's bid and ask price on the call date
BTM	Book-to-Market = book value of the firm divided by market value
SUE	Mean absolute deviation of analysts' earnings-per-share forecasts from the actual value in the preceding quarter
ROA	Return on Assets, i.e., net income divided by assets
Industry	Fama–French 12 industry dummies
Time	Year–quarter dummies

behaves as expected: a CEO communicating in an *introverted* and *feeling* manner is associated with increased risk ($\beta_i = 0.03$, $\beta_f = 0.10$, while an *intuitive* communication is associated with decreased risk ($\beta_s = -0.02$). Less surprisingly, positive sentiment tends to decreases risk highly significantly, while the opposite (albeit to a lesser extent) holds for negative sentiment. Notably, all results are robust to age- and gender-fixed effects.

Although seemingly small, the size of the personality effect (i.e., the coefficient height) is in line with that observed by related work (Harrison et al., 2020). It is expectable that fundamentals such as past risk or firm size have a stronger impact on future risk than, e.g., CEO extraversion. Remarkably, T–F has the third-largest impact ($\beta_f = 0.10$) out of all considered features. Though only weakly correlated with the ground truth (Table 19), the results suggest that the predictions for this scale contain strong economic signal for risk regression.

For the larger windows of volatility, i.e., in two, three, and four business weeks following the call date, we observe reduced effect sizes and significance levels of personality (cf. Table 21). This suggests that the personality effect mainly impacts short-term risk. Nonetheless, E–

I and T–F continue to have a significant and positive impact on risk. Furthermore, the J–P factor reaches a positive and significant impact on risk for Vola_{15} and Vola_{20}.

In sum, these results provide new empirical evidence to support the *upper echelons theory*. We show that situational aspects of CEO personality, predicted with our MBTI regressor, also reflect firm performance measured by stock return volatility, the most common financial risk measure.

8.5 ETHICAL CONSIDERATIONS

In the following, we discuss possible biases and environmental considerations.

SOCIAL DESIRABILITY BIAS Past literature has shown that some Big 5 personalities are more socially desirable than others, which paves the way to discrimination: Overall, it is socially desirable to score low on *neuroticism* (an omitted scale in the MBTI) and high on *conscientiousness* and *agreeableness*. To a lesser extent, it is socially desirable to score high on *extraversion* and *openness* Ones et al., 1996, Table 2. For the MBTI, in contrast, there exist no "bad" personality traits. As shown in §8.3.1, however, the Big 5 and the MBTI correlate. Therefore, the points raised about social desirability, albeit to a lesser extent, should apply here, too.

SAMPLE BIASES Critically, our gold standard consists of just 32 CEOs of large American (mostly tech) companies. While these companies (Alphabet, Facebook, Apple, etc.) constitute a large share of the American market, this renders the personality prediction model less applicable to non-American, small, or non-tech companies. Only four (i.e., 12.5%) of the 32 CEOs are female. While this gender ratio is twice as high as that of the S&P 500 (Catalyst, 2021), this highlights that the findings of this study might generalize poorly to non-male CEOs. In addition, as shown in §8.3.1, Figure 14, CEOs as a social cohort share a distinct distribution of personality traits, which is why we argue that the MBTI regressors should only be applied with caution, if at all, to non-executive samples.

ENERGY CONSUMPTION Training neural models can have substantial financial and environmental costs (Strubell et al., 2019), which motivates us to discuss the computational efficiency of the Transformers. Using an NVIDIA Tesla P100 GPU, we run a hyperparameter optimization over 40 configurations per MBTI dimension for both BERT and RoBERTa. The average power consumption is 200W and the optimization takes ca. 16 hours, i.e., 3.2 kilowatt hours (kWh) with an

electricity cost of 40 cents per model.[11] Labeling the 22K earnings call instances with no available ground truth takes ca. 4.5 hours and 140W, i.e., 0.63 kWH of GPU time and 8 cents, respectively. Training time of the SVM with trigram tf–idf is negligible (ca. 2 minutes on a quad-core processor with 8GB RAM). Whether the performance increases of the Transformers over a sparse method justify the added computational costs should be considered carefully on a case-by-case basis.

8.6 CONCLUSION

This chapter adds to two unresolved problems in the literature: The first problem is raised by the *upper echelons theory*, which suggests that characteristics of top managers reflect in the outcomes of their company. The second problem is the text-based prediction of MBTI personality, something that past literature has contested. Here, we introduce the first Transformer-based MBTI prediction model of CEO personality. Furthermore, for the first time, we show that CEOs' MBTI personality has explanatory power of financial risk. For the personality prediction task, we observe moderate to strong correlations with the ground truth for three out of four dimensions. Empirically, *extroverted*, *intuitive*, *thinking*, and *judging* CEOs seem to incur less future financial risk. Notably, the *thinking–feeling* factor of CEO personality has a stronger effect on future risk than all other financial features except for past risk and firm size.

Although focusing on different independent variables, the risk regression approaches of this chapter and the previous one have two commonalities: (1) they are based on feature-engineering and (2) they perform regression for explanatory instead of predictive purposes. The next and last chapter of Part IV, therefore, explores an assumption-free prediction model of financial risk from financial disclosures.

11 Calculations assume the average U.S. electricity rate of 12.55 cents per 15 November 2021: `https://www.electricchoice.com/electricity-prices-by-state`

Table 21: Risk regression results with z-standardized coefficients and t-statistics in parentheses. The sample includes 22K earnings calls with 1.7K CEOs in years 2002–2020. Vola_τ is log return volatility in trading days 0 to τ. Features are the MBTI *extraversion–introversion* (E–I), *sensing–intuition* (S–N), *thinking–feeling* (T–F), and *judging–perceiving* (J–P); other features are defined in §8.4.1.

Feature	Vola_5	Vola_{10}	Vola_{15}	Vola_{20}
E–I	0.03^{***}	0.02^{***}	0.01^{*}	$0.01^{\dagger}$
	(5.01)	(3.20)	(2.53)	(1.80)
S–N	-0.02^{**}	$-0.01^{\dagger}$	-0.01	-0.00
	(−2.69)	(−1.65)	(−1.28)	(−0.22)
T–F	0.10^{***}	0.08^{***}	0.07^{***}	0.05^{***}
	(13.67)	(11.60)	(9.97)	(8.99)
J–P	-0.00	0.01	0.01^{*}	0.01^{*}
	(−0.22)	(0.88)	(1.96)	(2.49)
Positivity	-0.02^{**}	-0.02^{**}	-0.01^{**}	-0.01^{**}
	(−2.96)	(−3.17)	(−2.58)	(−2.58)
Negativity	0.01^{*}	0.01^{*}	$0.01^{\dagger}$	0.01
	(2.17)	(2.05)	(1.75)	(1.75)
Uncertainty	0.01	0.00	0.00	0.00
	(0.83)	(0.64)	(0.66)	(0.66)
Age	-0.01	-0.00	-0.00	-0.00
	(0.38)	(−1.18)	(−0.70)	(−1.41)
Gender	-0.02	-0.02	-0.00	0.00
	(−0.75)	(−0.95)	(−0.12)	(0.01)
Past Vola	0.43^{***}	0.46^{***}	0.45^{***}	0.44^{***}
	(44.72)	(53.07)	(55.57)	(58.08)
Size	-0.19^{***}	-0.21^{***}	-0.21^{***}	-0.21^{***}
	(−19.83)	(−23.62)	(−24.50)	(−26.46)
Volume	0.05^{***}	0.08^{***}	0.09^{***}	0.10^{***}
	(5.36)	(10.63)	(12.10)	(14.12)
Leverage	-0.05^{***}	-0.03^{***}	-0.02^{**}	$-0.01^{\dagger}$
	(−6.88)	(−4.13)	(−2.64)	(−1.83)
Spread	0.03^{***}	0.05^{***}	0.04^{***}	0.04^{***}
	(4.10)	(8.53)	(8.52)	(8.50)
BTM	-0.02^{*}	-0.01	0.01	0.01^{*}
	(−2.92)	(−1.45)	(1.27)	(2.34)
SUE	-0.00	-0.00	-0.01	-0.01
	(−0.73)	(−0.66)	(−1.15)	(−1.49)
ROA	0.00	-0.00	-0.01	-0.02^{***}
	(0.46)	(−0.89)	(−2.47)	(−3.34)
Adj. R^2	34.10%	46.40%	52.80%	59.20%

$^{\dagger}p \leq 0.1$,* $p \leq 0.05$, $^{**}p \leq 0.01$, $^{***}p \leq 0.001$

9

ASSUMPTION-FREE RISK REGRESSION FROM TEXT

* While the previous chapters described explanatory studies of risk based on engineered features (uncertainty and CEO personality), we were also interested in using assumption-free DL architectures in a truly predictive setting. Such architectures have the power of modeling highly non-linear relationships while being based on latent variables rather than pre-defined features. Often, this may come at the cost of interpretability. Hence, we were also interested in exploring visualizations opening up the black box of DL-based NLP models.

9.1 INTRODUCTION

In this chapter, we combine established knowledge from the financial domain with recent advancements in NLP to create PRoFET, a neural model jointly exploiting financial and textual data for financial risk prediction. We collect a comprehensive set of historical financial data and enrich it with natural language information revealed in recurring events: earnings calls; in these calls, the performance of publicly traded companies is summarized and prognosticated by their management. We then train a joint model to predict short-term risk following these calls.

We present a new dataset of 90K calls and thus re-assess the task of text-based risk regression at a large scale. Moreover, we propose a model jointly learning from both semantic text representations and a comprehensive set of financial features. Given the advancements of neural networks and their capabilities in automatic feature learning (Baroni et al., 2014), we were motivated to apply such methods instead of a traditional, feature-engineered approach. Lastly, since performance gains by neural networks have "typically come at the cost of our understanding of the system" (Linzen et al., 2018, p. iii), we were interested in obtaining humanly interpretable results by visually explaining volatility fluctuations in a sample use case.

Introducing PRoFET, we present the following contributions to the academic community:

MODEL Our neural architecture, which jointly learns from semantic text representations and a comprehensive set of financial features,

* This chapter is based on "PRoFET: Predicting the Risk of Firms from Event Transcripts" (Theil, Broscheit, and Stuckenschmidt, 2019), presented in August 2019 at the *28th International Joint Conference on Artificial Intelligence (IJCAI)* in Macao.

significantly outperforms the previous state-of-the-art and other baselines. In an ablation study, we further show that the joint model significantly outperforms models using either of both feature types alone and inspect the performance impact of different document sections.

DATA We present a new dataset of 90K earnings call transcripts and address the task of text-based risk prediction at a large scale.

INTERPRETABILITY The performance increases provided by neural models often come at the cost of interpretability. We address this issue by visualizing the predictive power of contextualized tokens with a heatmap. This demonstrates a use case of PRoFET as a tool for investment decision support.

9.2 DATA

We collect 90K earnings call[1] transcripts from the database Thomson Reuters Eikon.[2] The data covers ca. 4.3K distinct companies and spans the years 2002–2017. The approximate numbers of tokens and types are 675M and 200K, respectively. We divide all transcripts into the Presentation and the Q&A parts; Table 22 describes this dataset's surface features. The average transcript contains 400 sentences and 7.7K tokens.

We retrieve all utterances except technical remarks (e.g., closing the call) by the teleconference Operator and tokenize the documents with `spaCy` (Honnibal et al., 2020). We identify dates, points of time, percentages, monetary values, measurements (as of weight or distance), and cardinal numbers with `spaCy`'s (Honnibal et al., 2020) NER and replace them with uniform placeholder tokens, e.g., "{PERCENTAGE}" or "{CARDINAL}". Since the transcribed text data is the intellectual property of Thomson Reuters, we can not share it publicly in its raw form. However, our word embedding models and the financial data (as defined in §9.3) are available online.[3] Our dataset can be reassembled using the contained identifiers (stock tickers, PERMNOs, and Eikon IDs).

To prevent look-ahead bias, we use a temporal 80/10/10 percentage split to divide the 90K instances into separate training, validation, and test sets. The training data spans from January 2002 to August 2015, validation from August 2015 to November 2016, and test from November 2016 to December 2017.

1 See §2.2.3 for an in-depth discussion of earnings calls.
2 `https://eikon.thomsonreuters.com/index.html`
3 See Appendix A.

Table 22: Surface features for our dataset of 90K documents.

Part	n of sentences	n of tokens
Presentation	12.5M	276.3M
Q&A	22.6M	398.9M
Total	35.1M	675.2M
Per document	0.4K	7.7K

9.3 METHODOLOGY

9.3.1 *Label: Volatility*

Given a firm's transcript and a set of financial features, we perform the prediction of a continuous label in the week following the firm's earnings call. As label, we consider a firm's financial risk, represented as stock return volatility.[4] Volatility is defined as follows: Let $r_t = \frac{p_t}{p_{t-1}} - 1$ be the return of a stock with price p_t on day t. Then the volatility between days t and $t + \tau$ is the sample standard deviation of stock returns in this period:

$$v_{[t,t+\tau]} = \sqrt{\frac{1}{\tau - 1} \sum_{i=0}^{\tau} (r_{t+i} - \bar{r})^2} \tag{12}$$

Here, $\bar{r}$ is the sample mean of r_t over the period. We use the volatility $v_{[1,5]}$ in the business week after the call as the label.

9.3.2 *Features*

Our model jointly learns from various textual and financial features defined as outlined below.

TEXTUAL FEATURES We segment the transcripts into three sections: presentation, questions, and answers. We represent each of these sections as a vector $\mathbf{t}$, i.e., $\mathbf{t}_P$, $\mathbf{t}_Q$, $\mathbf{t}_A$: tokens $w_1, w_2, ..., w_n$ of the transcript sections are represented with embeddings $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, ..., \mathbf{w}^{(n)}$, and two distinct model variants compose those tokens into a representation $\mathbf{t}$ (see Section 9.3.3). Word embeddings with dimensions $d \in \{100, 200\}$ are trained with fastText (Bojanowski et al., 2017) on our dataset of 90K transcripts, 675M tokens, and 200K types.

FINANCIAL FEATURES We retrieve a comprehensive set of financial features for each call. If not stated otherwise, we obtain all data

4 See §2.2.1 for an in-depth discussion of volatility in the context of the agency dilemma.

from the databases CRSP and CRSP/Computstat Merged, which we access via the Wharton Research Data Services (WRDS) platform:[5]

- *Past volatility* should expectedly be a strong predictor of future volatility (Kogan et al., 2009), which is why we add the volatility $v_{[-64,-1]}$ (see Eq. 12) in the business quarter before the call as a feature;
- *Market volatility* as aggregated by the VIX,[6] is a predictor of volatility (Blair et al., 2001). We retrieve the VIX value at the day before the call to factor in market moves affecting all companies;
- *Size* is represented by the total market value of equity (or: "market capitalization"), defined as the number of outstanding shares times stock price. We include the firm size on the day before the call as a feature since it is a well-known driver of risk (Fama and French, 1992);
- *BTM* is the ratio of firm value according to its balance sheet ("book value") over market value (see above) and measures the current degree of over- or undervaluation. This ratio is a well-proven risk factor (Fama and French, 1992), which is why we incorporate it in our model;
- *Earnings surprise* is the difference between the actual and the expected earnings per share (i.e., the profit allocated per individual stock) and obtained from the database IBES. Empirically, high surprises tend to be followed by high volatility (Price et al., 2012), which is why include it as a feature;
- *Industry*-specific characteristics are an important risk driver (Fama and French, 1997). To account for them, we categorize each firm according to the Fama–French 12-industry scheme,[7] which distinguishes between twelve industries (e.g., "energy" or "healthcare").

9.3.3 *Proposed Model:* PROFET

PROFET is a neural model incorporating word embeddings, Long Short-Term Memories (LSTMs) (Hochreiter and Schmidhuber, 1997), and an attention-based text representation. Our implementation (as elaborated below) can be found online.[8]

5 https://wrds-web.wharton.upenn.edu/wrds
6 http://www.cboe.com/vix
7 http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
8 See Appendix A.

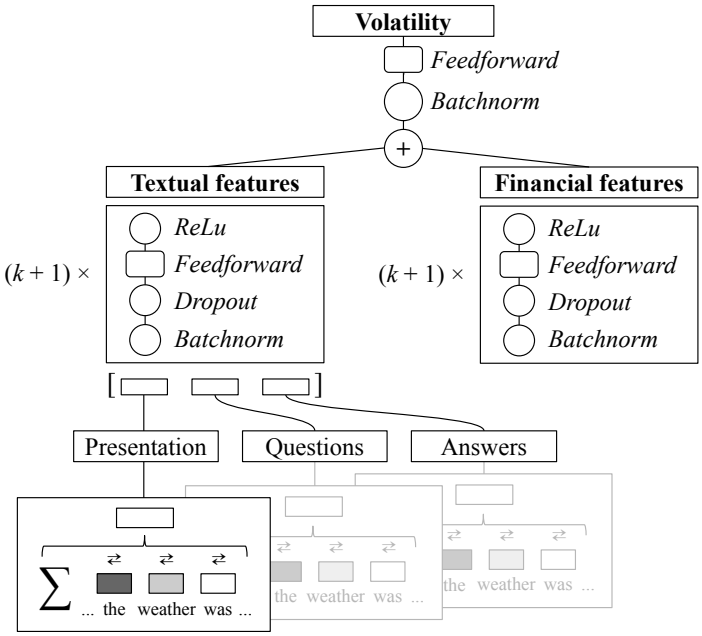


Figure 16: PRoFET's architecture. Representations of Presentation, Questions, and Answers are weighted averages of the token embeddings, contextualized by a BiLSTM and then weighted by an attention score. Text representations are fed into the left FNN, the financial features into the right FNN. Both FNNs are executed $k+1$ times, with k being a hyperparameter. In the last layer, both feature sets are fused for the final prediction.

ARCHITECTURE Figure 16 provides a sketch of PRoFET's architecture. A representation $\mathbf{t}$ is computed for each section: Each token $\mathbf{w}^{(i)}$ is transformed into a contextualized representation $\mathbf{c}^{(i)}$ with a BiLSTM by concatenating the left-to-right and the right-to-left LSTM's hidden state vector of $\mathbf{w}^{(i)}$, i.e.:

$$\mathbf{c}^{(i)} = [\overrightarrow{\text{BiLSTM}}(\mathbf{w}^{(1)}, ..., \mathbf{w}^{(i)}), \overleftarrow{\text{BiLSTM}}(\mathbf{w}^{(n)}, ..., \mathbf{w}^{(i)})] \,. \tag{13}$$

An attention score $s^{(i)}$ is computed for each of these contextualized representations with a learned attention vector $\mathbf{a}$ (Bahdanau et al., 2015), where

$$s^{(i)} = \frac{\exp(\mathbf{a}^T \mathbf{c}^{(i)})}{\sum_j \exp(\mathbf{a}^T \mathbf{c}^{(j)})} \,. \tag{14}$$

A separate $\mathbf{a}$ is learned for presentations, questions, and answers (i.e., $\mathbf{a}_P$, $\mathbf{a}_Q$, and $\mathbf{a}_A$) and the BiLSTM weights are shared among these sections. Finally, each section is represented as the weighted sum

$$\mathbf{t} := \sum_{i=1}^{n} s^{(i)} \mathbf{c}^{(i)} \,. \tag{15}$$

The such obtained text representations $\mathbf{t}_P$, $\mathbf{t}_Q$, and $\mathbf{t}_A$ are concatenated and fed into a FNN with k hidden layers (with k being a hyperparameter). Each of its layers uses dropout, batch normalization (Ioffe and Szegedy, 2015), and a Rectified Linear Unit (ReLU) activation function (Nair and Hinton, 2010). Thus, a single distributed text representation $\mathbf{t}_{\text{dist}}$ is created. A separate FNN with the same architecture calculates a distributed representation $\mathbf{f}_{\text{dist}}$ from the financial data. These representations are summed up, yielding a single vector $(\mathbf{t}_{\text{dist}} + \mathbf{f}_{\text{dist}})$ which is fed into a final hidden layer with batch normalization. The output of this layer is a prediction of the continuous label v, i.e., volatility in the week after the call.

OPTIMIZATION A range of hyperparameters influences the performance of neural architectures. To choose a set of hyperparameters for our FNN, we explore: the number of hidden layers $k \in \{1, 2, 3\}$, hidden layer sizes $n \in \{128, 256, 512, 1024, 2048\}$, and whether to use batch normalization for layers $l_{\text{in}} = 0$, $1 \leq l_{\text{hid}} < k$ and $l_{\text{out}} = k$. For the BiLSTM, we consider: the number of hidden layers $k \in \{1, 2, 3\}$, hidden layer sizes $n \in \{50, 100\}$, learning rate $\lambda \in \{10^{-1}, 10^{-2}\}$, dropout $\delta \in \{0.0, 0.1, \ldots, 0.5\}$, weight decay $\omega \in \{10^{-4}, 10^{-5}, 10^{-6}\}$, embedding size $d \in \{100, 200\}$, and whether the embeddings are adjusted.

To find a suitable configuration of hyperparameters, we perform a Bayesian optimization minimizing MSE on the validation set.[9] We start the search with ten random samples from the hyperparameter grid and then alternate between (1) choosing the next unseen set yielding the lowest loss minus one standard deviation; or (2) sampling a new configuration from the grid. In total, we evaluate 60 hyperparameter configurations. We train a model for up to 20 epochs with Adagrad (Duchi, 2011) and a batch size of 112. We determine the best model with early stopping and use its hyperparameter configuration for subsequent training.

9.3.4 *Baselines*

AVERAGE POOLING As a simple neural benchmark, we train an average pooling model which obtains the text representations $\mathbf{t}$ by averaging all contextualized token representations $\mathbf{c}^{(i)}$:

$$\mathbf{t} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{c}^{(i)} \tag{16}$$

GARCH As one of the most popular econometric models for volatility prediction, GARCH (Bollerslev, 1986) performs well in various settings (Hansen and Lunde, 2005). We train such a model as a baseline

9 The Bayesian optimization is implemented with the `GaussianProcessRegressor` of `sklearn 0.20.1` with RBF kernel and 20 restarts.

for each call. Our joint models should exceed its performance to provide real value. We use all available historical return data up to the call date to predict volatility in the week following the call.

SPARSE METHODS From the related domain of risk prediction from 10-Ks, we replicate the sparse methods by Kogan et al. (2009), Tsai and Wang (2014), and Rekabsaz et al. (2017). They all use different variants of BoW vectors and past volatility[10] as features in a prediction based on the SVR model. Our findings confirm that among these approaches, the most recent one by Rekabsaz et al. (2017) performs best in our domain as well.

This approach consists of training a word embedding model to expand a financial sentiment dictionary (Loughran and McDonald, 2011) with similar terms; this expanded dictionary is used to filter and retain the matching terms in BoW vectors weighted with BM25. Vector sparsity is reduced with Principal Component Analysis (PCA) and separate SVR models with RBF kernel are learned on both the financial and the textual data. The results of these models are fused ("stacked") in a final prediction with an additional SVR. To compare PROFET's performance to the previous state-of-the-art, we report the results of this method on ten folds of our test set.

9.3.5 *Evaluation Metrics*

To evaluate the predictive performance, we analyze the following metrics: the linear correlation coefficient Pearson's r, the non-linear rank correlation coefficients Spearman's ρ and Kendall's τ used in the previous literature (Wang and Hua, 2014), and the MSE. Optimizing the models on our validation set, we noticed consistently higher values for the rank correlation coefficients over r. This indicates a monotonic but non-linear relationship between the predicted values $\hat{y}$ and the actual values y. An inability to capture non-linear relationships and a proneness to outliers are well-known undesirable properties of r (Anscombe, 1973). To obtain more robust correlation estimates in such settings, a log-transformation can be applied to $\hat{y}$ and y. Hence, we report $r_{\log}$, which is the linear correlation measured on the log-transformed data.

9.4 RESULTS AND DISCUSSION

We start by demonstrating that a neural model performs competitively to the previous state-of-the-art, even when using previously proposed data and features (§9.4.1). We continue by benchmarking the performance of different models on our new dataset (§9.4.2), pro-

10 To provide a fair comparison to our approach, we additionally use the comprehensive set of financial features proposed by us (see §9.3.2) in all replication experiments.

Table 23: Performance of an FNN compared to different regression models on the dataset of Wang and Hua (2014) in terms of Pearsons $r_{\log}$, Spearman's ρ, Kendall's τ, and MSE multiplied by 100.

Model	$r_{\log}$	ρ	τ	MSE
Copula	—	0.407	0.302	—
Ridge	0.326	0.356	0.245	0.976
Huber	0.346	0.382	0.262	0.926
FNN	**0.395**	**0.446**	**0.309**	**0.829**

ceed with an ablation study (§9.4.3), and conclude with a showcase of the visualized attention mechanism of PRoFET (§9.4.4).

9.4.1 *Comparison to Previous Work*

We compare the best-performing previously researched model, a Gaussian Copula regression, with a set of regression models selected by us: a Ridge regression, a Huber regression, and a simple FNN. The goal of this comparison is not to present a model which outperforms the previous state-of-the-art but is to show that an FNN poses a competitive alternative to the model proposed by Wang and Hua (2014), which is not publicly available. To stay comparable, we use the dataset published by them.[11] This dataset contains 11K instances with 500 language features and volatility in the week following the call as a label. We explored several neural network architectures with different hyperparameters using a randomized search with 3-fold cross-validation on the full dataset.[12] Table 23 provides an overview of the performance across all regression models. As can be seen, the neural network performs competitively to the Gaussian Copula, especially in terms of Spearman's ρ.

9.4.2 *Model Benchmark*

Table 24 summarizes PRoFET's performance in terms of the evaluation metrics described in §9.3.5. All values are averages of ten runs on our test set. If applicable, we perform (paired) t-tests with $\alpha \in \{0.05, 0.01, 0.001\}$ to test for significant performance increases over the baselines described in §9.3.4: the econometric GARCH; the best-performing sparse method (Rekabsaz et al., 2017); and the average pooling model.

11 https://www.cs.ucsb.edu/~william/data/earningscalls.zip

12 The best performance was achieved with three hidden layers (with 500, 250, and 150 neurons), a logistic activation function, and an L2 penalty parameter of 10^{-3}.

Table 24: Performance of PRoFET compared to the baseline GARCH, the best-performing sparse method (Rekabsaz et al., 2017), and the average pooling model on our test set in terms of Pearson's $r_{\log}$, Spearman's ρ, Kendall's τ, and MSE.

Model	$r_{\log}$		ρ		τ		MSE	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
GARCH	0.437	—	0.531	—	0.368	—	7.236	—
Rekabsaz et al. (2017)	0.560	0.024	0.596	0.020	0.422	0.017	0.504	0.084
Average Pooling	0.571	0.017	0.598	0.018	0.419	0.015	0.870	0.209
PRoFET	**0.622**	0.013	**0.641**	0.013	**0.454**	0.011	**0.485**	0.086

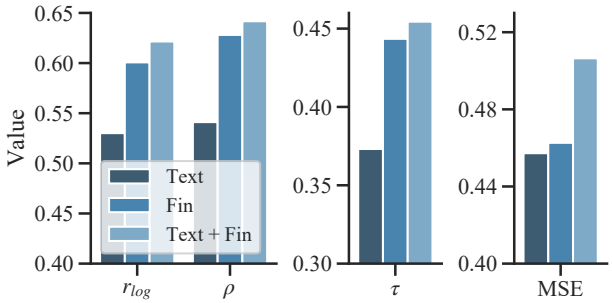
The approach by Rekabsaz et al. (2017) outperforms the econometric model GARCH, which indicates that: (1) it should pose a competitive reference for our neural models, and (2) even sparse methods without a representation of semantic context can lead to considerable performance increases over purely financial models.

The average pooling model and the previous state-of-the-art reach similar performance with insignificant differences across all metrics apart from MSE. For this metric, the average pooling model falls behind by a highly significant margin (0.870 vs. 0.504, $p \leq 0.001$), albeit with a comparably high standard deviation (0.209 vs. 0.084). In summary, these results indicate that a simple averaging of the word embeddings does not appropriately reflect the complexity of the problem.

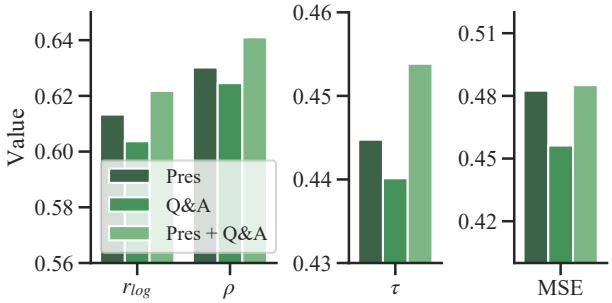
Our proposed model PRoFET exceeds both the average pooling model as well as the previous state-of-the-art across all evaluation metrics. This improvement is highly significant ($p \leq 0.001$) in terms of the linear correlation, and very significant ($p \leq 0.01$) in terms of the rank correlation coefficients. Moreover, PRoFET's performance also exhibits the largest robustness in terms of standard deviation out of all approaches which we consider. In conclusion, our findings suggest that for the given task, fine-grained modeling of semantic context—in our case, with a separate attention mechanism weighting the contextualized token representations—leads to profound performance increases over both traditional econometric as well as state-of-the-art sparse NLP models.

9.4.3 *Ablation Study*

We perform a systematic ablation study to answer the questions: How do textual and financial features influence the prediction? What are the influence of the scripted presentation and the spontaneous Q&A?



(a) Ablation of PRoFET trained on textual, financial, and textual + financial features.



(b) Ablation of PRoFET trained on the presentation, the Q&A, and the presentation + Q&A sections.

Figure 17: Ablation analyses of PROFET over features and sections. Performance is measured in terms of Pearson's $r_{\log}$, Spearman's ρ, Kendall's τ, and MSE.

FEATURE ABLATION We start by comparing the performance of a purely financial model to both a purely textual model and a joint model. The results of this ablation are depicted in Figure 17a. Using only textual features yields noticeable performance drops in terms of $r_{\log}$, ρ, and τ compared to both the financial features and a joint model; however, using textual features alone yields the lowest MSE out of all models that we consider. Although seemingly small, the performance increase of a joint model over a purely financial model is highly significant ($p \leq 0.001$) in terms of $r_{\log}$ and very significant ($p \leq 0.01$) in terms of ρ and τ. In summary, this experiment exemplifies that for the given task, the performance of textual features can only be assessed meaningfully in conjunction with financial features.

SECTION ABLATION We proceed by comparing the influence of different sections on the predictive power. It could be expected that the presentation and the Q&A as structurally different sections also differ concerning their informativeness to the market. Our results (see Figure 17b) show that using only the presentation yields better results than using only the Q&A. While the joint model still performs best in

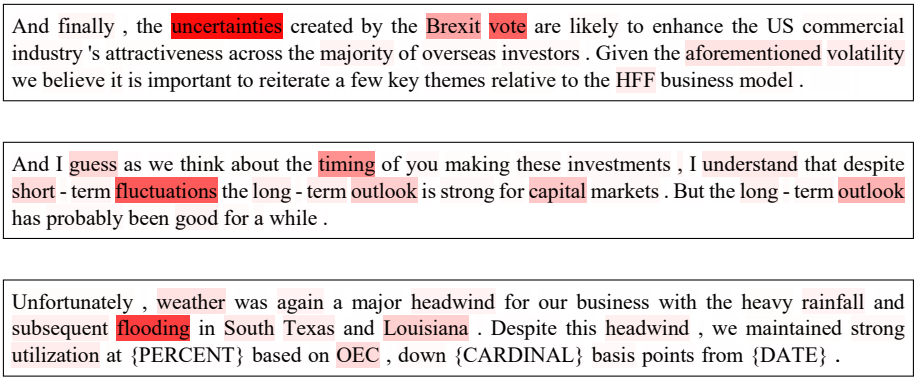


Figure 18: Exemplary text snippets from the validation data with visualized attention per token according to PRoFET. Increasing intensity of red indicates higher attention (i.e., a higher predictive power for risk).

terms of $r_{\log}$, ρ, and τ, it is the model trained on the presentation alone, which achieves the lowest MSE; this difference is insignificant, however. In sum, these results show that the transcripts have to be analyzed in their entirety to achieve the best performance.

9.4.4 *Attention Visualization*

As a concluding use case, we show how the attention mechanism (see §9.3.3) can be visualized on the token-level as a tool for investment decision support. In Figure 18, we present three real-data text snippets to which PRoFET assigned a noticeably above-average attention per token.

As the first snippet indicates, PRoFET allocates high attention to "uncertainties" created by the "Brexit vote". When ordered according to their average attention, the latter collocation appears in the top-10 percentile of tokens, indicating a strong correlation with risk. The second snippet, taken from the Q&A answers given by company executives, is about short-term fluctuations and their implications for investment risk. Notably, the term "outlook" gets assigned slightly different attention levels depending on the verbal context. The last snippet covers severe environmental conditions, namely "heavy rainfall" and "subsequent flooding" with the latter displaying the highest allocated attention.

9.5 CONCLUSION

This work introduced the first neural model exploiting the multimodality of a text-based risk prediction task. PRoFET jointly learns from sequential text representations and a comprehensive set of financial features. We have shown that our proposed model outperforms

the previous state-of-art and other strong baselines. PRoFET's architecture leverages an attention mechanism to model verbal context, leading to significant performance increases over simpler sparse or average pooling models. We concluded by showcasing how this attention mechanism can be visualized on the token-level, thus providing interpretable results and offering a tool for investment decision support.

This chapter finishes Part IV of this thesis, which addresses T_3 by identifying linguistic drivers of financial risk. In the next and final Part, we conclude on the presented work, discuss possible limitations, and directions for future research.

Part V

WRAP-UP

10

CONCLUSION

This thesis explored NLP methods for linguistic uncertainty detection in financial disclosures (T_1), a causal model of uncertainty and risk (T_2), and text-based risk regressors based on linguistic uncertainty, CEO personality, and latent linguistic variables (T_3). Thus, this work is the first to study the entire causal chain of companies issuing uncertain financial disclosures, over their influence on risk perception and investor behavior, down to predictions of market reactions.

The following provides a summary of the findings along tasks T_1–T_3 (§10.1) and a discussion of the implications for researchers and practitioners in financial NLP (§10.2).

10.1 SUMMARY

(T1) Uncertainty detection: *How can we detect linguistic uncertainty automatically in financial disclosures? What are its economic and linguistic determinants?*

We introduced two approaches for automatically classifying linguistic uncertainty in speech transcripts of earnings calls. For a binary sentence classification task (Chapter 4), especially lexical features yielded promising results. As a combination of PoS-enriched BoW vectors with the Loughran and McDonald (2011) UNCERTAIN list yielded high precision but only moderate recall, future work may offer additional contributions by developing a new uncertainty dictionary targeted to spoken financial disclosures.

For a prediction of uncertainty in a Q&A setting (Chapter 5), the most predictive features were financial measures. Given the content of the preceding question and several financial covariates in the prior quarter, primarily market and firm risk seem to impact answer uncertainty. Compared to our expectations, this relationship is inversed, with increased market and firm risk being associated with less uncertain answers.

(T2) Causal modeling of uncertainty and risk: *How can we quantify the influence of linguistic uncertainty on risk perception and investment behavior? Which personal characteristics of investors play a role in this setting?*

For the first time in the literature, we established a causal link between linguistic uncertainty and risk perception as well as investment behavior (Chapter 6). In doing so, we conducted a laboratory study

exposing participants to earnings call excerpts with randomized degrees of linguistic uncertainty. Based on those, participants had to indicate their risk perception and invest fictional money. Furthermore, they answered an extensive survey covering behavioral, psychometric, and socio-demographic items.

Our findings confirm that linguistic uncertainty is associated with increased risk perception and decreased investment sums. This result is in accordance with the *certainty effect* coined by Tversky and Kahneman (1986), asserting that investors generally prefer investments framed in a certain to those framed in an uncertain manner. Furthermore, we show that investor age, extraversion, and neuroticism positively affect risk perception. Lastly, we find a positive quadratic effect of income on risk perception, i.e., income seems to decrease risk perception for lower-income individuals.

(T3) Risk regression: *What is the influence of linguistic uncertainty on financial risk measures? What other linguistic phenomena are explanatory or predictive of risk?*

In Part IV, we introduced risk regression models using textual disclosure data. To that end, in Chapter 7, we propose a risk regressor from linguistic uncertainty in 10-Ks. Specifically, we automatically expand the established Loughran and McDonald (2011) dictionary in an unsupervised manner by selecting the top-k cosine similar terms according to industry-specific word embedding models. Beyond that, we employ a method automatically filtering the suggested candidates in a semi-supervised manner by using a multi-task regressor simultaneously predicting three financial uncertainty measures. This model achieved the most decisive regression results, outperforming both the plain dictionary and an unsupervised expansion. Empirically, we show that the such measured linguistic uncertainty is positively and significantly related to future risk and analyst uncertainty.

Furthermore, we developed a text regression model to predict the MBTI personality of CEOs. Exploring crowd-sourced continuously scaled MBTI ratings as an alternative to self-reported Big 5, we find a high internal and external validity of the measure. Although related work illustrates that text-based MBTI prediction is a difficult task (Štajner and Yenikent, 2020, 2021), we show that a fine-tuned RoBERTa regressor succeeds in predicting three out of the four MBTI dimensions and outperforms both a BERT baseline and sparse methods suggested by prior work. The most promising results were achieved for the *extraversion–introversion* scale, followed by *sensing–intuition* and *judging–perception*. The results for *thinking–feeling* indicated that this dimension is challenging to predict with the given data and method. Finally, controlling for various financial covariates as well as CEO gender and age, we found that the MBTI personality of CEOs is significantly associated with future financial risk. In summary, this work lends empirical support to the *upper echelons theory* of management (Hambrick

and Mason, 1984), asserting that the personality of a company's top management has a top-down effect on financial performance.

The last regression model tackling T_3 was PRoFET, an assumption-free volatility prediction model based on latent features and DL. At the time of its publication, this model was the first neural risk regressor jointly learning from financial features and contextualized language representations of earnings calls. We demonstrate that PRoFET outperforms a traditional econometric approach, an average pooling baseline, and sparse methods suggested by previous literature (Kogan et al., 2009; Tsai and Wang, 2014; Tsai et al., 2016). In an ablation study, we have shown that a joint learner yields the best results for this task, followed by a purely financial and a purely textual learner.

10.2 IMPLICATIONS

We now discuss the implications of this thesis for financial NLP researchers and practitioners.

MARKETS DISLIKE FUZZY TALK Our findings highlight the impact of financial disclosure language in general and linguistic uncertainty in particular on risk, analyst uncertainty, and investment behavior. In our laboratory study, we found a positive causal effect of linguistic uncertainty on risk perception and a detrimental causal effect on investment sums. In the light of these effects, we conclude that companies should handle their investor communication with special care. Specifically, using clear and unobfuscated language is an essential tool for reducing information asymmetry between all stakeholders.

This undesirable effect of linguistic uncertainty in disclosure language might incentivize management to use less fuzzy talk in external communication. Different to related work (Loughran and McDonald, 2011, p. 44), however, we do not opine that the construction of uncertainty dictionaries may prompt managers to avoid specific terms contained therein (e.g., "unclear"). As shown in Chapter 7, the statistical relationship between uncertainty and risk is noisy and complex, which prevents us from applying an abstract correlation to single instances. At best, overall trends across large cross-sectional samples of disclosures can be discovered. However, if the presented research helped in discouraging management from using obfuscated language, this would constitute a desirable side effect.

FINANCIAL FEATURES MATTER A major finding of this thesis is that, although it pays off to include text data for financial prediction models, this is by no means a trivial task and "can only be assessed meaningfully in conjunction with financial features" (Chapter 9, p. 188). For example, concerning modality prediction in dialogue, we found market and firm risk to have a larger impact than linguistic

cues (Chapter 5). Related, our explanatory risk regressions show that linguistic signals have a small, but robust impact on volatility (Chapters 7 and 8). This observation also holds for our assumption-free risk prediction model PRoFET, for which the financial learner contributed the largest share of overall predictive power. Hence, in conclusion, we caution financial NLP researchers to carefully select a relatively exhaustive set of financial data serving as competitive baseline and eliminating confounding effects. Omitting or under-exploring financial features may lead to missassments of predictive power or overfit results.

ETHICAL CONSIDERATIONS Chapter 8 summarized the ethical implications of predicting CEO personality from language. A central issue is that text-based personality prediction models for managers are affected by sample biases: Labeled data predominantly exists for large American tech companies and male CEOs. Thus, systems analyzing the personal style or intentions of CEOs should be treated sensibly, especially given privacy concerns and the imminent danger of dual-use. Special consideration should be placed on the data they are trained on and possible biases concerning gender, culture, or other facets of identity. Otherwise, financial NLP systems in deployment may exacerbate existing societal biases and increase social inequality.

DOWNSTREAM USE A key downstream application of the proposed uncertainty detection methods (Chapters 4, 5, and 7) is deception and fraud detection. Vagueness and related phenomena of linguistic uncertainty are a common feature of deceptive strategies (Bachenko et al., 2008; Fornaciari et al., 2021; Larcker and Zakolyukina, 2012). Nevertheless, as the use of vague language in the presence of lacking "unequivocal truth is not always sufficient to declare falsity," (Egré and Icard, 2018) such systems should be applied sensibly. Particularly, empirical validation by a human-in-the-loop is crucial to avoid prejudgments and false positives.

As another practical finding, research building on the visualization application proposed in Chapter 9 could prove helpful for investment decision support systems. However, the influence of possible confounding factors such as linguistic context, the current financial situation of the monitored company, and unobservable factors such as synchronous behavior of other market participants have to be kept in mind. As the developed models are experimental prototypes, they are currently not suitable to be deployed in financial practice without further research and refinement. The following chapter will discuss the current limitations in more depth and point out possible starting points for methodological refinements to overcome them.

11

LIMITATIONS AND FUTURE WORK

* Taking a bird's view, §11.1 discusses the limitations of this research on a high level. Afterwards, §11.2 discusses open points that could be addressed by future work per chapter.

11.1 LIMITATIONS

11.1.1 *Differences in Data and Measures*

A general limitation of this work is that the discussed approaches use similar but slightly different datasets: Differences concern the used disclosure type (earnings calls vs. 10-Ks) and sampling (firm size and time period). Furthermore, the definitions of the explored linguistic and financial uncertainty measures vary to some extent between the chapters.

DATA Except for Chapter 7, all works focus on earnings calls instead of 10-Ks. This chapter differs, as it aims to expand the Loughran and McDonald (2011) dictionary, which is domain-dependent for this disclosure type. As seen in Chapter 4, the plain 10-K dictionary only has moderate precision and low recall for an application to earnings calls. Nonetheless, this poses the question of whether our observational findings from 10-Ks also hold for earnings calls.

Furthermore, Chapters 4 and 5 focus on the transcripts of companies that are part of the S&P 500 (i.e., the 500 largest US companies). In contrast, chapters 6 and 9 use *all* publicly listed US companies. Hence, the findings of the former works might not be transferable to small- and medium-sized companies.

As this dissertation comprises papers written between 2017 and 2022, the respective chapters cover slightly different time periods of disclosure data. Chapters 4 and 5 use earnings calls up to 2016 and up to 2019, respectively; Chapter 9, in contrast, uses data up to 2017, and Chapter 8 expands this dataset to additionally cover years 2018–2020. The same time period is investigated in Chapter 7. However, as shown in the latter chapter, risk regression models are time-sensitive. Therefore, future work needs to validate whether our findings persist when using a consistent sampling period.

* This chapter expands on the "Limitations" and "Future Work" sections of Theil, Hovy, and Stuckenschmidt (2023), Theil, Daube, and Stuckenschmidt (2022), Theil and Stuckenschmidt (2020), Theil, Štajner, and Stuckenschmidt (2020), and Theil, Štajner, Stuckenschmidt, and Ponzetto (2017).

UNCERTAINTY MEASURES Chapters 4 and 7, who discuss earlier works of this dissertation, explore a relatively broad concept of linguistic uncertainty which encompasses not only imprecision but also expressions of economic and financial uncertainty (reflected in terms such as "volatility" or "anomaly"). In contrast, Chapters 5 and 6 limit the scope to linguistic imprecision. There, we decided to narrow the focus to eliminate possible confounding effects. Hence, it would be interesting to isolate both types of uncertainty by creating separate classifiers and exploring how these types of uncertainty conflict or complement each other.

Moreover, the risk regressors in Part IV differ in their definitions of volatility. In accordance with Loughran and McDonald (2014), Chapter 7 uses post-event volatility in the month after the 10-K filing as dependent variable. This measure quantifies risk in excess of overall market risk. In contrast, following Wang and Hua (2014), Chapters 8 and 9 consider stock return volatility in the week after the earnings call.[2] Therefore, this measure represents risk as the standard deviation of firm-level stock returns, independent from overall market risk. However, as we have experimented with controlling for the market risk index VIX in the latter works, we expect the conceptual difference between post-event and stock return volatility to only have a minor impact on the results. Nevertheless, it may be useful to assess the sensitivity of the models in Chapters 7 and 9 to different measurement periods of risk, similar to the ones in Chapter 8.

Beyond that, including the analyst-based financial uncertainty measures of Chapter 7 as dependent variables may be of interest for future refinements of the works in Chapters 8 and 9. This idea is motivated by the findings of Keith and Stent (2019), who show that earnings call content can be leveraged to predict analyst price forecasts.

11.1.2 *Cultural Differences*

The findings of this thesis are based on data from the United States (US). As such, they might not generalize to other cultures. Different speaker communities from the Anglosphere, such as the British, have a different relationship to linguistic phenomena such as hedging. Other differences concern perceptions of politeness, directness, and the use of mitigation strategies (Flöck and Geluykens, 2018). Intercultural differences might be even more apparent for non-English speaking communities such as Arabic or Chinese. Thus, there is potential for creating multi-lingual datasets of financial disclosures and studying linguistic uncertainty and other language phenomena across them.

2 See §2.2.1 for a detailed definition of the these volatility measures.

11.1.3 *Speaker and Listener Information*

As the findings of Chapter 6 and 8 show, the meaning and interpretation of language depends on the characteristics of the engaged speakers and listeners. This raises the question of how including such information into uncertainty detection or risk regression models affects the results. For example, the same message is likely perceived differently depending on the role (CEO vs. CFO) and reputation of the delivering executive. Furthermore, analyst characteristics in earnings call Q&A could be more thoroughly explored, e.g., by encoding their employer, rank, or identity.

11.1.4 *Latent Forms of Uncertainty*

Linguistic uncertainty is often more latent than surface uncertainty markers (e.g., "maybe" or "probably"). Phenomena such as euphemisms in earnings calls—e.g., "challenge" instead of "difficulty" or "issue" instead of "problem" (Suslava, 2021)—are hard to decipher automatically. Advancements in the area of Masked Language Models (Zhu and Bhat, 2021) could be leveraged for an in-depth study of euphemistic language in disclosures language.

Furthermore, the meaning of non-committal statements such as "OK" or "alright" depends on vocal inflection, which requires modeling speech signals. For earnings calls, audio recordings exist, making it possible to map individual utterances to speech. Thus, exploring this modality leaves open room for future work.

11.2 FUTURE WORK

11.2.1 *Uncertainty Detection (T_1)*

CHAPTER 4 Future research could improve the detection of linguistic uncertainty in the following ways: First, a more sizable section of data could be co-annotated by a larger group of co-annotators. Another opportunity is exploring a more granular representation of uncertainty with more than two classes or on a continuous scale. Labels can then be represented as medians and standard deviations across annotator ratings. Using inter-annotator spread as a label or weighing can also be applied to n-ary classification tasks, as past literature has shown that annotator disagreement contains useful information (Paun et al., 2018; Pavlick and Kwiatkowski, 2019).

Given the broad definition of linguistic uncertainty, which encompasses hedging, statements about the future, and economic or financial uncertainty, a multi-label annotation of uncertainty aspects might be helpful. Regarding the method, developing a domain-adapted version of the Loughran and McDonald (2011) dictionary to better cap-

ture informal, spoken communication as occurring in earnings calls would be interesting. Further value can be added by incorporating n-grams with $n > 1$.

Beyond that, classification performance can be improved with feature selection and parameter optimization methods. Lastly, given our definition of uncertainty and its ties to economic risk factors, incorporating real-world knowledge into the classifier (e.g., with a knowledge base or graph) may be a fruitful avenue of research.

CHAPTER 5 A current limitation of the dataset used to predict linguistic uncertainty in dialogue is that it is a silver standard with automatically inferred labels based on STRONG MODAL or WEAK MODAL terms according to Loughran and McDonald (2011). Hence, the validity of the labels can be increased by conducting an annotation study either leveraging a curated set of experts or wisdom of the crowd (e.g., using Amazon Mechanical Turk).[3] To ensure annotation quality, a measure suited for noisy multi-annotations such as MACE (Hovy et al., 2013) would be of use.

Methodologically, it stands to explore whether the strength of the financial features persists when using a larger context window for the textual representations. For example, for a prediction of CEO answer modality, not only the previous question but several Q&A pairs or even the content of the entire call can be used together with the financial data. Furthermore, it would be helpful to deeper explore contextualized representations such as Transformer-based architectures. However, as our preliminary findings suggest, this would also require increasing the dataset size, e.g., using data augmentation techniques.

11.2.2 *Uncertainty and Risk (T_2)*

CHAPTER 6 In our laboratory study addressing the causality of linguistic uncertainty and risk, the earnings call excerpts only consider a binary classification into linguistically certain and uncertain style. However, similar to §11.2.1, an uncertainty representation on a continuous scale would mirror reality more closely. Such a representation would permit estimating a function of risk perception (y) dependent on linguistic uncertainty (x) across individuals.

Given that the participants were lay investors from Germany, validating the results using domain experts and native English speakers would be of use. Moreover, an interesting application are eye-tracking methods to analyze which linguistic cues get increased visual attention from professional readers.

Lastly, future work could investigate other features apart from linguistic uncertainty. For example, given the same disclosure content, voice pitch or prosody can be manipulated using transfer learners

3 https://www.mturk.com/

trained on speech recordings. It would be insightful to explore the correlations of these features with listener attention, perceived competence, and trustworthiness.

11.2.3 *Risk Regression (T_3)*

CHAPTER 7 Concerning the regressions of linguistic uncertainty on financial uncertainty measures, different word embedding models can be used to suggest new dictionary candidates. For example, the context-based representation of `word2vec` could be refined with topical criteria as implemented in `lda2vec` (Moody, 2016). Furthermore, similar to the suggestion regarding uncertainty classification, it would be useful to expand the dictionaries and embedding models to cover *n*-grams. To that end, `ngram2vec` (Zhao et al., 2017) seems to be a suitable model. Finally, it may be worthwhile to explore methods of lexicon induction (Bravo-Marquez et al., 2021; Hamilton et al., 2016; Pryzant et al., 2018).

CHAPTER 8 In the future, the personality prediction task should be re-assessed as a multi-task learning problem, in which one single regressor is trained to predict all four MBTI dimensions at once. In addition, speech signals of executives (e.g., voice modulation, tonality, and silence) or other modalities (e.g., gestures and facial expressions) may be included in the personality predictions. Accordingly, deep speaker embeddings (Bredin et al., 2020; Li et al., 2017; Wan et al., 2018) with auxiliary targets such as age or gender (Luu et al., 2021) appear promising. Exploring different dimensions of personality such as HEXACO's honesty–humility factor (Ashton et al., 2004) or the dark triad (Paulhus and Williams, 2002) is another compelling prospect.

CHAPTER 9 Currently, the assumption-free risk prediction model performs a point estimate of risk. However, the performance and practical use can be increased by using a time-series prediction model learned over a continuous stream of stock returns. Moreover, the model currently employs two separate regressors trained on textual and financial data, which are assumed to be disparate. Nonetheless, a model resembling the actual causality (i.e., financial situation $\rightarrow$ disclosure language $\rightarrow$ risk) seems worth exploring.

Beyond that, the predictive results might be improved by leveraging correlations between risk estimates at a firm- or industry-level. Finally, live predictions of intra-day returns and volatility are a valuable future contribution. As the starting times of earnings calls are known and audio recordings exist, utterances can be mapped to exact timestamps. This would allow the prediction of return movements following these utterances in real-time.

BIBLIOGRAPHY

Aha, David W., Dennis Kibler, and Marc K. Albert (1991). "Instance-Based Learning Algorithms." In: *Machine Learning* 6, pp. 37–66. DOI: 10.1007/BF00153759.

Amaral, Luís and J. Ottino (2004). "Complex Networks: Augmenting the Framework for the Study of Complex Systems." In: *The European Physical Journal B* 38, pp. 147–162. URL: https://amaral.northwestern.edu/publications/complex-networks-augmenting-the-framework-for-the-study-of-complex-systems/.

de Amicis, Chiara, Sonia Falconieri, and Mesut Tastan (2020). "Sentiment Analysis and Gender Differences in Earnings Conference Calls." In: *Journal of Corporate Finance* 71.C. DOI: 10.1016/j.jcorpfin.2020.101809.

Amihud, Yakov and Lev Baruch (1981). "Risk Reduction as Managerial Motive for Conglomerate Mergers." In: *Bell Journal of Economics* 12.2, pp. 605–617. DOI: 10.2307/3003575.

Andersen, Torben G., Tim Bollerslev, Peter F. Christoffersen, and Francis X. Diebold (2006). "Volatility and Correlation Forecasting." In: *Handbook of Economic Forecasting*. North-Holland. Chap. 15, pp. 778–878. DOI: 10.1016/S1574-0706(05)01015-3.

Anscombe, Francis J. (1973). "Graphs in Statistical Analysis." In: *The American Statistician* 27.1, pp. 17–21. DOI: 10.2307/2682899.

Araci, Dogu Tan (2019). "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models." In: *arXiv e-prints* 1908.10063. URL: https://arxiv.org/abs/1908.10063.

Arthur, W. Brian (2014). "Complexity Economics: A Different Framework for Economic Thought." In: *Complexity and the Economy*. Oxford University Press, pp. 1–29. DOI: 10.2469/dig.v43.n4.70.

Ashton, Michael C., Kibeom Lee, Marco Perugini, Piotr Szarota, Reinout E. de Vries, Lisa di Blas, Kathleen Boies, and Boele de Raad (2004). "A Six-Factor Structure of Personality-Descriptive Adjectives: Solutions From Psycholexical Studies in Seven Languages." In: *Journal of Personality and Social Psychology* 86.2, pp. 356–366. DOI: 10.1037/0022-3514.86.2.356.

Bachenko, Joan, Eileen Fitzpatrick, and Michael Schonwetter (2008). "Verification and Implementation of Language-Based Deception Indicators in Civil and Criminal Narratives." In: *Proceedings of COLING*. Manchester, pp. 41–48. DOI: 10.3115/1599081.1599087.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning

to Align and Translate." In: *Proceedings of ICLR*. URL: https://arxiv.org/abs/1409.0473.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis (2016). "Measuring Economic Policy Uncertainty." In: *The Quarterly Journal of Economics* 131.4, pp. 1593–1636. DOI: 10.1093/qje/qjw024.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski (2014). "Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors." In: *Proceedings of ACL*, pp. 238–247. DOI: 10.3115/v1/P14-1023.

Barrett, Jonathan, Eric G. Cavalcanti, Raymond Lal, and Owen J. E. Maroney (2014). "No ψ-Epistemic Model Can Fully Explain the Indistinguishability of Quantum States." In: *Physical Review Letters* 112.25. DOI: 10.1103/PhysRevLett.112.250403.

Barth, Andreas, Sasan Mansouri, Fabian Woebbeking, and Severin Zörgiebel (2021). "How to Talk Down Your Stock Performance." In: *SSRN*. DOI: 10.2139/ssrn.3336671.

Beck, James L. (2009). "A Probability Logic Framework for Treating Model Uncertainty for Prior and Posterior Robust Predictive System Analyses." In: *Workshop on Statistical Methods for Dynamic System Models*. Vancouver, CA. URL: http://people.stat.sfu.ca/~dac5/workshop09/James_Beck.html.

Benischke, Mirko H., Geoffrey P. Martin, and Lotte Glaser (2019). "CEO Equity Risk Bearing and Strategic Risk Taking: The Moderating Effect of CEO Personality." In: *Strategic Management Journal* 40.1, pp. 153–177. DOI: https://doi.org/10.1002/smj.2974.

Bird, Steven, Edward Loper, and Ewan Klein (2009). *Natural Language Processing with Python*. O'Reilly Media. URL: https://www.nltk.org/book/.

Blair, Bevan J., Ser-Huang Poon, and Stephen J. Taylor (2001). "Forecasting S&P 100 Volatility: The Incremental Information Content of Implied Volatilities and High-Frequency Index Returns." In: *Journal of Econometrics* 105.1, pp. 5–26. DOI: 10.1016/S0304-4076(01)00068-9.

Bloomfield, Robert, Mark W. Nelson, and Eugene Soltes (2016). "Gathering Data for Archival, Field, Survey, and Experimental Accounting Research." In: *Journal of Accounting Research* 54.2, pp. 341–395. DOI: 10.1111/1475-679X.12104.

Boersma, Paul and Vincent J. van Heuven (2001). "Speak and unSpeak with PRAAT." In: *Glot International* 5.9/10. URL: https://www.fon.hum.uva.nl/paul/papers/speakUnspeakPraat_glot2001.pdf.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching Word Vectors with Subword Information." In: *Transactions of the Association for Computational Linguistics (TACL)* 5, pp. 135–146. DOI: 10.1162/tacl_a_00051.

Bollerslev, Tim (1986). "Generalized Autoregressive Conditional Heteroskedasticity." In: *Journal of Econometrics* 31.3, pp. 307–327. DOI: 10.1016/0304-4076(86)90063-1.

Bonsall IV, Samuel B., Andrew J. Leone, Brian P. Miller, and Kristina Rennekamp (2017). "A Plain English Measure of Financial Reporting Readability." In: *Journal of Accounting and Economics* 63.2-3, pp. 329–357. DOI: 10.1016/j.jacceco.2017.03.002.

Box, George E. P. and David R. Cox (1964). "An Analysis of Transformations." In: *Journal of the American Statistical Association* 26.2, pp. 211–252. DOI: https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Bravo-Marquez, Felipe, Arun Khanchandani, and Bernhard Pfahringer (2021). "Incremental Word Vectors for Time-Evolving Sentiment Lexicon Induction." In: *Cognitive Computation*. DOI: 10.1007/s12559-021-09831-y.

Bredin, Hervé, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill (2020). "pyannote.audio: Neural Building Blocks for Speaker Diarization." In: *Proceedings of ICASSP*. DOI: 10.1109/ICASSP40776.2020.9052974.

Breiman, Leo (2001). "Random Forests." In: *Machine Learning* 45, pp. 5–32. DOI: 10.1023/A:1010933404324.

Brennan, Robert L. and Dale J. Prediger (1981). "Coefficient Kappa: Some Uses, Misuses, and Alternatives." In: *Educational and Psychological Measurement* 41.3, pp. 687–699. URL: https://doi.org/10.1177/001316448104100307.

Briggs-Myers, Isabel and Peter B. Myers (1995). *Gifts Differing: Understanding Personality Type*. Davies-Black.

Burgoon, Judee, William J. Mayew, Justin Scott Giboney, Aaron C. Elkins, Kevin Moffitt, Bradley Dorn, Michael Byrd, and Lee Spitzley (2016). "Which Spoken Language Markers Identify Deception in High-Stakes Settings? Evidence From Earnings Conference Calls." In: *Journal of Language and Social Psychology* 35.2, pp. 123–157. DOI: 10.1177/0261927X15586792.

Byrne, Kathleen (2005). "How Do Consumers Evaluate Risk in Financial Products?" In: *Journal of Financial Services Marketing* 10.1, pp. 21–36. DOI: 10.1057/palgrave.fsm.4770171.

Catalyst (2021). *Women CEOs of the S&P 500*. URL: https://www.catalyst.org/research/women-ceos-of-the-sp-500/.

Cazier, Richard A., Kenneth J. Merkley, and John S. Treu (2019). "When are Firms Sued for Qualitative Disclosures? Implications of the Safe Harbor for Forward-Looking Statements." In: *The Accounting Review* 95.1, pp. 31–55. DOI: 10.2308/accr-52443.

le Cessie, Saskia and Hans C. van Houwelingen (1992). "Ridge Estimators in Logistic Regression." In: *Applied Statistics* 41.1, pp. 191–201. DOI: 10.2307/2347628.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System." In: *Proceedings of ACM SIGKDD*, pp. 785–794. URL: https://dl.acm.org/citation.cfm?id=2939785.

Chowdhary, Kamaljit and Paul Dupuis (2011). "Distinguishing and Integrating Aleatoric and Epistemic Variation in Uncertainty Quantification." In: *arXiv e-prints* 1103.1861v2. URL: https://arxiv.org/abs/1103.1861.

Cohen, Jacob (1960). "A Coefficient of Agreement for Nominal Scales." In: *Educational and Psychological Measurement* 20.1, pp. 37–46. DOI: 10.1177/001316446002000104.

— (1998). *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. New York: Lawrence Erlbaum Associates. DOI: https://doi.org/10.4324/9780203771587.

Cohen, William W. (1995). "Fast Effective Rule Induction." In: *Proceedings of IMCL*, pp. 115–123. DOI: 10.1016/B978-1-55860-377-6.50023-2.

Cohen, Yuval, Hana Ornoy, and Baruch Keren (2013). "MBTI Personality Types of Project Managers and Their Success: A Field Survey." In: *Project Management Journal* 44.3, pp. 78–87. DOI: 10.1002/pmj.21338.

Connolly, James J., Erin J. Kavanagh, and Chockalingam Viswesvaran (2007). "The Convergent Validity Between Self and Observer Ratings of Personality: A Meta-Analytic Review." In: *International Journal of Selection and Assessment* 15.1, pp. 110–117. DOI: https://doi.org/10.1111/j.1468-2389.2007.00371.x.

Crawford Camiciottoli, Belinda (2009). "Just Wondering if You Could Comment on That: Indirect Requests for Information in Corporate Earnings Calls." In: *Text & Talk* 29.6, pp. 661–681. DOI: 10.1515/TEXT.2009.034.

— (2011). "Ethics and Ethos in Financial Reporting: Analyzing Persuasive Language in Earnings Calls." In: *Business Communication Quarterly* 74.3, pp. 298–312. DOI: 10.1177/1080569911413810.

— (2018). "Persuasion in Earnings Calls: A Diachronic Pragmalinguistic Analysis." In: *International Journal of Business Communication* 55.3, pp. 275–292. DOI: 10.1177/2329488417735644.

CSA (2017). *CSA Staff Notice 51-348—Staff's Review of Social Media Used by Reporting Issuers*. Tech. rep. URL: https://www.osc.ca/en/securities-law/instruments-rules-policies/5/51-348/csa-staff-notice-51-348-staffs-review-social-media-used-reporting-issuers.

Danescu-Niculescu-Mizil, Cristian, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts (2013). "A Computational Approach to Politeness with Application to Social Factors." In: *Proceedings of ACL*, pp. 250–259. URL: https://www.aclweb.org/anthology/P13-1025.

Danforth, Chris (2015). "Lorenz's Discovery of Chaos." In: *Mathematics of Planet Earth*. Ed. by Hans G. Kaper and Christiane Rousseau. Philadelphia (PA), USA: Society of Industrial and Applied Mathematics. Chap. 3.2, pp. 39–40.

Das, Sanjiv Ranjan (2014). "Text and Context: Language Analytics in Finance." In: *Foundations and Trends in Finance* 8.3, pp. 145–260. DOI: 10.1561/0500000045.

Davidson, Paul (1996). "Reality and Economic Theory." In: *Journal of Post Keynesian Economics* 18.4, pp. 479–508. DOI: 10.1080/01603477.1996.11490083.

Devlin, Jacob, Ming Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of NAACL*, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423/.

Diacon, Stephen (2004). "Investment Risk Perceptions: Do Consumers and Advisers Agree?" In: *International Journal of Bank Marketing* 22.3, pp. 180–199. DOI: 10.1108/02652320410530304.

Diaconis, Persi, Susan Holmes, and Richard Montgomery (2007). "Dynamical Bias in the Coin Toss." In: *SIAM Review* 49.2, pp. 211–235. URL: https://www.jstor.org/stable/20453950.

Diecidue, Enrico and Jeroen van de Ven (2008). "Aspiration Level, Probability of Success and Failure, and Expected Utility." In: *International Economic Review* 49.2, pp. 683–700. DOI: 10.2307/20486812.

Dodge, Jesse, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith (2020). "Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping." In: *arXiv e-prints* 2002.06305. URL: https://arxiv.org/abs/2002.06305.

Doshi, Hitesh, Saurin Patel, Srikanth Ramani, and Matthew Thomas Sooy (2021). "Uncertain Linguistic Tone and Credit Default Swap Spreads." In: *SSRN*. DOI: 10.2139/ssrn.3311776.

Duchi, John (2011). "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization." In: *Journal of Machine Learning Research* 12, pp. 2121–2159. URL: https://dl.acm.org/doi/10.5555/1953048.2021068.

Dusserre, Emmanuelle and Muntsa Padró (2017). "Bigger Does Not Mean Better! We Prefer Specificity." In: *Proceedings of IWCS*. URL: https://aclanthology.org/W17-6908.

Dyer, Travis, Mark Lang, and Lorien Stice-Lawrence (2017). "The Evolution of 10-K Textual Disclosure: Evidence from Latent Dirichlet Allocation." In: *Journal of Accounting and Economics* 64.2, pp. 221–245. DOI: 10.1016/j.jacceco.2017.07.002.

Dzieliński, Michał, Alexander Wagner, and Richard J. Zeckhauser (2021). "CEO Clarity." In: *Swiss Finance Institute Research Paper Series* 17.13. DOI: 10.2139/ssrn.2965108.

Egré, Paul and Benjamin Icard (2018). "Lying and Vagueness." In: ed. by Jörg Meibauer. Oxford, UK: Oxford University Press. DOI: 10.1093/oxfordhb/9780198736578.013.27.

Eisenhardt, Kathleen M. (1989). "Agency Theory: An Assessment and Review." In: *Academy of Management* 14.1, pp. 57–74. DOI: https://doi.org/10.2307/258191.

Eisinga, Rob, Manfred Grotenhuis, and Ben Pelzer (2013). "The Reliability of a Two-Item Scale: Pearson, Cronbach, or Spearman-Brown?" In: *International journal of public health* 58, pp. 637–642. DOI: 10.1007/s00038-012-0416-3.

El-Haj, Mahmoud, Paul Rayson, Martin Walker, Steven Young, and Vasiliki Simaki (2019). "In Search of Meaning: Lessons, Resources and Next Steps for Computational Analysis of Financial Discourse." In: *Journal of Business Finance & Accounting* 46.3–4. DOI: 10.1111/jbfa.12378.

Ellsberg, Daniel (1961). "Risk, Ambiguity, and the Savage Axioms." In: *The Quarterly Journal of Economics* 75.4, pp. 643–669. DOI: 10.2307/1884324.

Fama, Eugene F. and Kenneth R. French (1992). "The Cross Section of Expected Stock Returns." In: *Journal of Finance* 47.2, pp. 427–465. DOI: 10.1111/j.1540-6261.1992.tb04398.x.

— (1993). "Common Risk Factors in the Returns on Stocks and Bonds." In: *Journal of Financial Economics* 33.1, pp. 3–56. DOI: 10.1016/0304-405X(93)90023-5.

— (1997). "Industry Costs of Equity." In: *Journal of Financial Economics* 43.2, pp. 153–193. DOI: 10.1016/S0304-405X(96)00896-3.

— (2001). "Disappearing Dividends: Changing Firm Characteristics or Lower Propensity to Pay?" In: *Journal of Financial Economics* 60.1, pp. 3–43. DOI: 10.1016/S0304-405X(01)00038-1.

Farkas, Richárd, 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*, pp. 1–12. URL: https://www.aclweb.org/anthology/W10-3001.

Fisher, Ingrid E., Margaret R. Garnsey, Sunita Goel, and Kinsun Tam (2010). "The Role of Text Analytics and Information Retrieval in the Accounting Domain." In: *Journal of Emerging Technologies in Accounting* 7.1, pp. 1–24. DOI: 10.2308/jeta.2010.7.1.1.

Fisher, Ingrid E., Margaret R. Garnsey, and Mark E. Hughes (2016). "Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future

Research." In: *Intelligent Systems in Accounting, Finance and Management* 23, pp. 157–214. DOI: 10.1002/isaf.1386.

Fitzpatrick, Eileen, Joan Bachenko, and Tommaso Fornaciari (2015). "Automatic Detection of Verbal Deception." In: *Synthesis Lectures on Human Language Technologies* 8.3, pp. 1–119. DOI: 10.2200/S00656ED1V01Y201507HLT029.

Fleiss, Joseph L., Bruce Levin, and Myunghee Cho Paik (2003). *Statistical Methods for Rates and Proportions*. 3rd ed. John Wiley. DOI: 10.1002/0471445428.

Flöck, Ilka and Ronald Geluykens (2018). "Preference Organization and Cross-Cultural Variation in Request Responses: A Corpus-Based Comparison of British and American English." In: *Corpus Pragmatics* 2.1, pp. 57–82. DOI: 10.1007/s41701-017-0022-y.

Floyd, Eric and John A. List (2016). "Using Field Experiments in Accounting and Finance." In: *Journal of Accounting Research* 54.2, pp. 437–475. DOI: 10.1111/1475-679X.12113.

Fornaciari, Tommaso, Federico Bianchi, Massimo Poesio, and Dirk Hovy (2021). "BERTective: Language Models and Contextual Information for Deception Detection." In: *Proceedings of ACL*, pp. 2699–2708. URL: https://aclanthology.org/2021.eacl-main.232.

Frazier, Katherine Beal, Robert W. Ingram, and B. Mack Tennyson (1984). "A Methodology for the Analysis of Narrative Accounting Disclosures." In: *Journal of Accounting Research* 22.1, pp. 318–331. URL: http://www.jstor.org/stable/2490713.

Furnham, Adrian (1996). "The Big Five Versus the Big Four: The Relationship Between the Myers–Briggs Type Indicator (MBTI) and NEO-PI Five Factor Model of Personality." In: *Personality and Individual Differences* 21.2, pp. 303–307. DOI: 10.1016/0191-8869(96)00033-5.

Furnham, Adrian, Joanna Moutafi, and John Crump (2003). "The Relationship Between the Revised Neo-Personality Inventory and the Myers–Briggs Type Indicator." In: *Social Behavior and Personality* 31.6, pp. 577–584. DOI: https://doi.org/10.2224/sbp.2003.31.6.577.

Ganter, Viola and Michael Strube (2009). "Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features." In: *Proceedings of ACL–IJCNLP*, pp. 173–176. URL: https://www.aclweb.org/anthology/P09-2044.

Gibson, Ryan, David Michayluk, and Gerhard van de Venter (2013). "Financial Risk Tolerance: An Analysis of Unexplored Factors." In: *Financial Services Review* 22.1, pp. 23–50. URL: https://opus.lib.uts.edu.au/bitstream/10453/23532/1/2011008254OK.pdf.

Gjurković, Matej and Jan Šnajder (2018). "Reddit: A Gold Mine for Personality Prediction." In: *Proceedings of the ACL Workshop on*

Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pp. 87–97. DOI: 10.18653/v1/w18-1112.

Glavaš, Goran and Sanja Štajner (2015). "Simplifying Lexical Simplification: Do We Need Simplified Corpora?" In: *Proceedings of ACL–IJCNLP*, pp. 63–68. DOI: 10.3115/v1/P15-2011.

Goldberg, L. R. (1990). "An Alternative "Description of Personality": The Big-Five Factor Structure." In: *Journal of Personality and Social Psychology* 59, pp. 1216–1229. DOI: https://doi.org/10.1037/0022-3514.59.6.1216.

Grable, John E. (2000). "Financial Risk Tolerance and Additional Factors That Affect Risk Taking in Everyday Money Matters." In: *Journal of Business and Psychology* 14.4, pp. 625–630.

Grable, John and Ruth H. Lytton (1999). "Financial Risk Tolerance Revisited: The Development of a Risk Assessment Instrument." In: *Financial Services Review* 8.3, pp. 163–181. DOI: 10.1016/S1057-0810(99)00041-4.

Grable, John and Ruth Lytton (2003). "The Development of a Risk Assessment Instrument: A Follow-Up Study." In: *Financial Services Review* 12, pp. 257–274. URL: https://www.researchgate.net/publication/285841335_The_development_of_a_risk_assessment_instrument_A_follow-up_study.

Guillemette, Michael, Michael Finke, and John Gilliam (2012). "Risk Tolerance Questions to Best Determine Client Portfolio Allocation Preferences." In: *Journal of Financial Planning* 25.5, pp. 36–44. URL: https://ssrn.com/abstract=2088998.

Guo, Wei, Tieying Yu, and Javier Gimeno (2017). "Language and Competition: Communication Vagueness, Interpretation Difficulty, and Market Entry." In: *Academy of Management Journal* 60.6, pp. 2073–2098. DOI: 10.5465/amj.2014.1150.

Gwet, Kilem Li (2008). "Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement." In: *British Journal of Mathematical and Statistical Psychology* 61.1, pp. 29–48. DOI: 10.1348/000711006X126600.

Hacking, Ian (1975). *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge, UK: Cambridge University Press.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). "The WEKA Data Mining Software: An Update." In: *ACM SIGKDD Explorations Newsletter* 11.1, pp. 10–18. DOI: 10.1145/1656274.1656278.

Hambrick, Donald C. and Phyllis A. Mason (1984). "Upper Echelons: The Organization as a Reflection of Its Top Managers." In: *Academy of Management Review* 9.2, pp. 193–206. DOI: 10.5465/amr.1984.4277628.

Hamilton, William L., Kevin Clark, Jure Leskovec, and Dan Jurafsky (2016). "Inducing Domain-Specific Sentiment Lexicons from Un-

labeled Corpora." In: *Proceedings EMNLP*, pp. 595–605. DOI: 10.18653/v1/D16-1057.

Hansen, Peter R. and Asger Lunde (2005). "A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)?" In: *Journal of Applied Econometrics* 20.2, pp. 873–889. DOI: 10.1002/jae.800.

Harrison, Joseph S., Gary R. Thurgood, Steven Boivie, and Michael D. Pfarrer (2020). "Perception Is Reality: How CEOs' Observed Personality Influences Market Perceptions of Firm Risk and Shareholder Returns." In: *Academy of Management Journal* 63.4, pp. 1166–1195. DOI: 10.5465/amj.2018.0626.

Hartog, Joop, Ada Ferrer-i-Carbonell, and Nicole Jonker (2002). "Linking Measured Risk Aversion to Individual Characteristics." In: *Kyklos* 55.1, pp. 3–26. DOI: 10.1111/1467-6435.00175.

Heath, Chip and Amos Tversky (1991). "Preference and Belief: Ambiguity and Competence in Choice under Uncertainty." In: *Journal of Risk and Uncertainty* 4, pp. 5–28. DOI: 10.1007/BF00057884.

Hiller, Jack H. (1971). "Verbal Response Indicators of Conceptual Vagueness." In: *American Educational Research Journal* 8.1, pp. 151–161. DOI: 10.2307/1161744.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory." In: *Neural Computation* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Hoerl, Arthur E. and Robert W. Kennard (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." In: *Technometrics* 12.1, pp. 55–67. DOI: 10.1080/00401706.1970.10488634.

Holzmeister, Felix, Jürgen Huber, Michael Kirchler, Florian Lindner, Utz Weitzel, and Stefan Zeisberger (2020). "What Drives Risk Perception? A Global Survey with Financial Professionals and Lay People." In: *Management Science* 66.9, pp. 3977–4002. DOI: 10.1287/mnsc.2019.3526.

Honnibal, Matthew, Ines Montani, Sofie van Landeghem, and Adriane Boyd (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. DOI: 10.5281/zenodo.1212303.

Hovy, Dirk, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy (2013). "Learning Whom to Trust with MACE." In: *Proceedings of NAACL*, pp. 1120–1130. URL: https://aclanthology.org/N13-1132.pdf.

Hrazdil, Karel, Jiri Novak, Rafael Rogo, Christine Wiedman, and Ray Zhang (2020). "Measuring Executive Personality Using Machine-Learning Algorithms: A New Approach and Audit Fee-Based Validation Tests." In: *Journal of Business Finance and Accounting* 47.3–4, pp. 519–544. DOI: 10.1111/jbfa.12406.

Hristu-Varsakelis, D. and C. Kyrtsou (2008). "Evidence for Nonlinear Asymmetric Causality in US Inflation, Metal, and Stock Returns."

In: *Discrete Dynamics in Nature and Society* 2008. DOI: 10.1155/2008/138547.

Hyland, Ken (1998). *Hedging in Scientific Research Articles*. John Benjamins. DOI: 10.1075/pbns.54.

Ioffe, Sergey and Christian Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." In: *Proceedings of ICML*, pp. 448–456. URL: https://dl.acm.org/doi/10.5555/3045118.3045167.

Jensen, Michael C. and William H. Meckling (1976). "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." In: *Journal of Financial Economics* 3.4, pp. 305–360. DOI: 10.1016/0304-405X(76)90026-X.

Joachims, Thorsten (2006). "Training Linear SVMs in Linear Time." In: *Proceedings of KDD*. Philadelphia (PA), USA: Association for Computing Machinery, pp. 217–226. DOI: 10.1145/1150402.1150429.

John, George H. and Pat Langley (1995). "Estimating Continuous Distributions in Bayesian Classifiers." In: *Proceedings of UAI*, pp. 338–345. URL: https://dl.acm.org/doi/10.5555/2074158.2074196.

Johnston, Ron, Kelvyn Jones, and David Manley (2018). "Confounding and Collinearity in Regression Analysis: A Cautionary Tale and an Alternative Procedure, Illustrated by Studies of British Voting Behaviour." In: *Quality & Quantity* 52, pp. 1957–1976. DOI: 10.1007/s11135-017-0584-6.

Kahneman, Daniel and Amos Tversky (1979). "Prospect Theory: An Analysis of Decision under Risk." In: *Econometrica* 47.2, pp. 253–291. DOI: https://doi.org/10.2307/1914185.

— (1982). "Variants of Uncertainty." In: *Cognition* 11.2, pp. 143–157. DOI: 10.1016/0010-0277(82)90023-3.

Kearney, Colm and Sha Liu (2014). "Textual Sentiment in Finance: A Survey of Methods and Models." In: *International Review of Financial Analysis* 33, pp. 171–185. DOI: 10.1016/j.irfa.2014.02.006.

Keith, Katherine A. and Amanda Stent (2019). "Modeling Financial Analysts' Decision Making via the Pragmatics and Semantics of Earnings Calls." In: *Proceedings of ACL*, pp. 493–503. DOI: 10.18653/v1/P19-1047.

Kempf, Alexander, Christoph Merkle, and Alexandra Niessen-Ruenzi (2014). "Low Risk and High Return: Affective Attitudes and Stock Market Expectations." In: *European Financial Management* 20.5, pp. 995–1031. DOI: 10.1111/eufm.12001.

Khadjeh Nassirtoussi, Arman, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo (2014). "Text Mining for Market Prediction: A Systematic Review." In: *Expert Systems with Applications* 41.16, pp. 7653–7670. DOI: 10.1016/j.eswa.2014.06.009.

Kim, Hyunji, Stefano I. di Domenico, and Brian S. Connelly (2019). "Self–Other Agreement in Personality Reports: A Meta-Analytic

Comparison of Self- and Informant-Report Means." In: *Psychological Science* 30.1, pp. 129–138. DOI: 10.1177/0956797618810000.

Klir, George J. (1987). "Where Do We Stand on Measures of Uncertainty, Ambiguity, Fuzziness, and the Like?" In: *Fuzzy Sets and Systems* 24.2, pp. 141–160. DOI: 10.1016/0165-0114(87)90087-X.

Klos, Alexander, Elke U. Weber, and Martin Weber (2005). "Investment Decisions and Time Horizon: Risk Perception and Risk Behavior in Repeated Gambles." In: *Management Science* 51.12, pp. 1777–1790. DOI: 10.1287/mnsc.1050.0429.

Knight, Frank Hyneman (1921). *Risk, Uncertainty, and Profit*. Houghton Mifflin. URL: https://fraser.stlouisfed.org/files/docs/publications/books/risk/riskuncertaintyprofit.pdf.

Kogan, Shimon, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith (2009). "Predicting Risk from Financial Reports with Regression." In: *Prcoeedings of NAACL*, pp. 272–280. URL: https://dl.acm.org/doi/10.5555/1620754.1620794.

Konstantinova, Natalia, Sheila C. M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada, and Ruslan Mitkov (2012). "A Review Corpus Annotated for Negation, Speculation and Their Scope." In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 3190–3195. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/533_Paper.pdf.

Kosinski, Michal, Sandra C. Matz, Samuel D. Gosling, Vesselin Popov, and David Stillwell (2015). "Facebook as a Research Tool for the Social Sciences: Opportunities, Challenges, Ethical Considerations, and Practical Guidelines." In: *American Psychologist* 70.6, pp. 543–556. DOI: 10.1037/a0039210.

Krippendorff, Klaus (2013). *Content Analysis: An Introduction to Its Methodology*. 3rd. Thousand Oaks (CA), USA: Sage.

Kumar, B. Shravan and Vadlamani Ravi (2016). "A Survey of the Applications of Text Mining in Financial Domain." In: *Knowledge-Based Systems* 114, pp. 128–147. DOI: https://doi.org/10.1016/j.knosys.2016.10.003.

Kvalseth, Tarald O. (1985). "Cautionary Note about R^2." In: *The American Statistician* 39.4, pp. 279–285. DOI: 10.2307/2683704.

Lakoff, George (1973). "Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts." In: *Journal of Philosophical Logic* 2, pp. 458–508. URL: https://www.jstor.org/stable/30226076.

Landis, J. Richard and Gary G. Koch (1977). "The Measurement of Observer Agreement for Categorical Data." In: *Biometrics* 33.1, pp. 159–174. DOI: 10.2307/2529310.

Larcker, David F. and Anastasia A. Zakolyukina (2012). "Detecting Deceptive Discussions in Conference Calls." In: *Journal of Accounting Research* 50.2, pp. 494–540. DOI: 10.1111/j.1475-679X.2012.00450.x.

Le, Quoc and Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents." In: *Proceedings of ICML*, pp. 272–280. DOI: https://dl.acm.org/doi/10.5555/3044805.3045025.

Lehavy, Reuven, Feng Li, and Kenneth Merkley (2011). "The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts." In: *The Accounting Review* 86.3, pp. 1087–1115. DOI: 10.2308/accr.00000043.

Lewis, N. R., L. D. Parker, G. D. Pound, and P. Sutcliffe (1986). "Accounting Report Readability: The Use of Readability Techniques." In: *Accounting and Business Research* 16.63, pp. 199–213. URL: https://doi.org/10.1080/00014788.1986.9729318.

Li, Chao, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu (2017). "Deep Speaker: an End-to-End Neural Speaker Embedding System." In: *arXiv e-prints* 1705.02304. URL: http://arxiv.org/abs/1705.02304.

Li, Feng (2006). "Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports?" In: DOI: 10.2139/ssrn.898181.

— (2008). "Annual Report Readability, Current Earnings, and Earnings Persistence." In: *Journal of Accounting and Economics* 45.2–3, pp. 221–247. DOI: 10.1016/j.jacceco.2008.02.003.

— (2011). "Textual Analysis of Corporate Disclosures: A Survey of the Literature." In: *Journal of Accounting Literature* 29, pp. 143–165. URL: http://www.cuhk.edu.hk/acy2/workshop/20110215FengLI/Paper1.pdf.

Li, Jun and Xiaofei Zhao (2015). "Complexity and Information Content of Financial Disclosures: Evidence from Evolution of Uncertainty Following 10-K Filings." In: *SSRN*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2516622.

Light, Marc, Xin Ying Qiu, and Padmini Srinivasan (2004). "The Language of Bioscience: Facts, Speculations, and Statements in Between." In: *Proceedings of the NAACL Workshop on Linking Biological Literature, Ontologies and Databases (BioLINK)*. Boston, pp. 17–24. URL: https://www.aclweb.org/anthology/W04-3103.

Linus, Wilson (2021). "GPU Prices and Cryptocurrency Returns." In: *SSRN*. DOI: 10.2139/ssrn.3922181.

Linzen, Tal, Grzegorz Chrupała, and Afra Alishahi (2018). "Introduction." In: *Proceedings of the EMNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackBoxNLP)*. URL: https://aclanthology.info/papers/W18-5400/w18-5400.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." In: *arXiv e-prints* 1907.11692. URL: https://arxiv.org/abs/1907.11692.

Loughran, Tim and Bill McDonald (2011). "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." In: *The Journal of Finance* 66.1, pp. 35–65. DOI: 10.1111/j.1540-6261.2010.01625.x.

— (2013). "IPO First-Day Returns, Offer Price Revisions, Volatility, and Form S-1 Language." In: *Journal of Financial Economics* 109.2, pp. 307–9326. DOI: 10.1016/j.jfineco.2013.02.017.

— (2014). "Measuring Readability in Financial Disclosures." In: *The Journal of Finance* 69.4, pp. 1643–1671. DOI: 10.1111/jofi.12162.

— (2016). "Textual Analysis in Accounting and Finance: A Survey." In: *Journal of Accounting Research* 54.4, pp. 1187–1230. DOI: 10.1111/1475-679X.12123.

— (2017). "The Use of EDGAR Filings by Investors." In: *Journal of Behavioral Finance* 18.2, pp. 231–248. DOI: 10.1080/15427560.2017.1308945.

— (2020). "Textual Analysis in Finance." In: *Annual Review of Financial Economics* 12, pp. 357–375. URL: http://dx.doi.org/10.1146/annurev-financial-012820-032249.

Lundberg, Scott M., 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." In: *Nature Machine Intelligence* 2.1, pp. 56–67. DOI: 10.1038/s42256-019-0138-9.

Lundberg, Scott M. and Su-in Lee (2017). "A Unified Approach to Interpreting Model Predictions." In: *Proceedings of NIPS*, pp. 1–10. URL: https://dl.acm.org/doi/10.5555/3295222.3295230.

Lusardi, Annamaria and Olivia S. Mitchell (2011). "Financial Literacy Around the World: An Overview." In: *Journal of Pension Economics and Finance* 10.4, pp. 497–508. DOI: 10.3386/w17107.

— (2014). "The Economic Importance of Financial Literacy: Theory and Evidence." In: *Journal of Economic Literature* 52.1, pp. 5–44.

Lusardi, Annamaria and Olivia Mitchell (2007). "Financial Literacy and Retirement Preparedness: Evidence and Implications for Financial Education." In: *Business Economics* 42, pp. 35–44. DOI: 10.2145/20070104.

Luu, Chau, Peter Bell, and Steve Renals (2021). "Leveraging Speaker Attribute Information Using Multi Task Learning for Speaker Verification and Diarization." In: *arXiv e-prints* 2010.14269. URL: https://arxiv.org/abs/2010.14269.

Magee, Lonnie (1990). "R^2 Measures Based on Wald and Likelihood Ratio Joint Significance Tests." In: *The American Statistician* 44.3, pp. 250–253. DOI: 10.1080/00031305.1990.10475731.

Mairesse, François, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore (2007). "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text." In: *Journal of Arti-*

ficial Intelligence Research 30, pp. 457–500. DOI: https://doi.org/10.1613/jair.2349.

Malhotra, Shavin, Taco H. Reus, PengCheng Zhu, and Erik M. Roelofsen (2018). "The Acquisitive Nature of Extraverted CEOs." In: *Administrative Science Quarterly* 63.2, pp. 370–408. DOI: 10.1177/0001839217712240.

Man, Xiliu, Tong Luo, and Jianwu Lin (2019). "Financial Sentiment Analysis (FSA): A Survey." In: *Proceedings of IEEE ICPS*, pp. 617–622. DOI: 10.1109/ICPHYS.2019.8780312.

Markowitz, Harry (1952). "Portfolio Selection." In: *The Journal of Finance* 7.1, pp. 77–91. DOI: 10.1111/j.1540-6261.1952.tb01525.x.

McCrae, Robert R. and Paul T. Costa (1989). "Reinterpreting the Myers–Briggs Type Indicator from the Perspective of the Five-Factor Model of Personality." In: *Journal of Personality* 57.1, pp. 17–40. DOI: 10.1111/j.1467-6494.1989.tb00759.x.

McCrae, Robert R., Paul T. Costa, and Thomas A. Martin (2010). "The NEO-PI-3: A More Readable Revised NEO Personality Inventory." In: *Journal of Personality Assessment* 84.3, pp. 261–270. DOI: 10.1207/s15327752jpa8403.

McCrae, Robert R. and Oliver P. John (1992). "An Introduction to the Five-Factor Model and Its Applications." In: *Journal of Personality* 60.2, pp. 175–215. DOI: 10.1111/j.1467-6494.1992.tb00970.x.

Medlock, Ben and Ted Briscoe (2007). "Weakly Supervised Learning for Hedge Classification in Scientific Literature." In: *Proceedings of ACL*, pp. 992–999. URL: https://www.aclweb.org/anthology/P07-1125.

Meirowitz, Adam (2005). "Informational Party Primaries and Strategic Ambiguity." In: *Journal of Theoretical Politics* 17.1, pp. 107–136. DOI: 10.1177/0951629805047800.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space." In: *arXiv e-prints* 1301.3781. DOI: 10.1162/153244303322533223.

Moody, Christoph E. (2016). "Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec." In: *arXiv e-prints* 1605.02019. URL: https://arxiv.org/abs/1605.02019.

Nadeau, Claude and Yoshua Bengio (2003). "Inference for the Generalization Error." In: *Machine Learning* 52.3, pp. 239–281. DOI: 10.1023/A:1024068626366.

Nair, Vinod and Geoffrey E. Hinton (2010). "Rectified Linear Units Improve Restricted Boltzmann Machines." In: *Proceedings of ICML*, pp. 807–814. URL: https://dl.acm.org/doi/10.5555/3104322.3104425.

von Neumann, John and Oskar Morgenstern (1953). *Theory of Games and Economic Behavior*. 3rd ed. Princeton (NJ), USA: Princeton University Press. URL: https://archive.org/details/theoryofgameseco00vonn.

Nguyen, Linh, Gerry Gallery, and Cameron Newton (2019). "The Joint Influence of Financial Risk Perception and Risk Tolerance on Individual Investment Decision-Making." In: *Accounting and Finance* 59.S1, pp. 747–771. DOI: 10.1111/acfi.12295.

Nicholson, Nigel, Emma Soane, Mark Fenton-O'Creevy, and Paul Willman (2005). "Personality and Domain-Specific Risk Taking." In: *Journal of Risk Research* 8.2, pp. 157–176. DOI: 10.1080/1366987032000123856.

Nikolić, Hrvoje (2006). "Classical Mechanics Without Determinism." In: *Foundations of Physics Letters* 19.6, pp. 553–566. DOI: 10.1007/s10702-006-1009-2.

Oehler, Andreas and Florian Wedlich (2018). "The Relationship of Extraversion and Neuroticism with Risk Attitude, Risk Perception, and Risk Expectations." In: *Journal of Neuroscience, Psychology and Economics* 11.2, pp. 63–92. DOI: http://dx.doi.org/10.1037/npe0000088.

Ones, Deniz S., Chockalingam Viswesvaran, and Angelika D. Reiss (1996). "Role of Social Desirability in Personality Testing for Personnel Selection: The Red Herring." In: *Journal of Applied Psychology* 81.6, pp. 660–679. DOI: 10.1037/0021-9010.81.6.660.

Paetzold, Gustavo Henrique and Lucia Specia (2016). "Unsupervised Lexical Simplification for Non-native Speakers." In: *Proceedings of AAAI*, pp. 3761–3767. URL: https://dl.acm.org/doi/10.5555/3016387.3016433.

Palmer, F. R. (2001). *Mood and Modality*. 2nd ed. Cambridge, UK: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139167178.

Paulhus, Delroy L. and Kevin M. Williams (2002). "The Dark Triad of Personality: Narcissism, Machiavellianism, and Psychopathy." In: *Journal of Research in Personality* 36.6, pp. 556–563. DOI: 10.1016/C2017-0-01262-4.

Paun, Silviu, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio (2018). "Comparing Bayesian Models of Annotation." In: *Transactions of the Association for Computational Linguistics* 6, pp. 571–585. DOI: 10.1162/tacl_a_00040.

Pavlick, Ellie and Tom Kwiatkowski (2019). "Inherent Disagreements in Human Textual Inferences." In: *Transactions of the Association for Computational Linguistics* 7, pp. 6779–694. DOI: 10.1162/tacl_a_00293.

Pennebaker, James W., Ryan L. Boyd, Kayla Jordan, and Kate Blackburn (2015). *The Development and Psychometric Properties of LIWC2015*. White Paper. University of Texas at Austin. DOI: 10.15781/T29G6Z.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "GloVe: Global Vectors for Word Representation." In: *Proceedings of EMNLP*, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

Petalas, Diamantis Petropoulos, Hein van Schie, and Paul Hendriks Vettehen (2017). "Forecasted Economic Change and the Self-Fulfilling Prophecy in Economic Decision-Making." In: *PLoS ONE* 12.3, pp. 1–18. DOI: 10.1371/journal.pone.0174353.

Plank, Barbara and Dirk Hovy (2015). "Personality Traits on Twitter—or—How to Get 1500 Personality Tests in a Week." In: *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 92–98. DOI: 10.18653/v1/W15-2913.

Platt, John C. (1999). *Fast Training of Support Vector Machines Using Sequential Minimal Optimization*. MIT Press, pp. 185–208. URL: https://dl.acm.org/doi/10.5555/299094.299105.

Poria, Soujanya, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency (2017). "Context-Dependent Sentiment Analysis in User-Generated Videos." In: *Proceedings of ACL*, pp. 873–883. DOI: 10.18653/v1/P17-1081.

Porter, Martin F. (1980). "An Algorithm for Suffix Stripping." In: *Program* 14.3, pp. 130–137.

Portner, Paul (2009). *Modality*. Oxford University Press. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.184.418&rep=rep1&type=pdf.

Powell, Melanie and David Ansic (1997). "Gender Differences in Risk Behaviour in Financial Decision-Making: An Experimental Analysis." In: *Journal of Economic Psychology* 18.6, pp. 605–628. DOI: 10.1016/S0167-4870(97)00026-3.

Price, S. McKay, James S. Doran, David R. Peterson, and Barbara A. Bliss (2012). "Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone." In: *Journal of Banking and Finance* 36.4, pp. 992–1011. DOI: 10.1016/j.jbankfin.2011.10.013.

Prokofieva, Anna and Julia Hirschberg (2014). "Hedging and Speaker Commitment." In: *Proceedings of the LREC International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data (ES³LOD)*, pp. 10–13. URL: http://www.cs.columbia.edu/~prokofieva/Prokofieva_Hirschberg_Paper.pdf.

Pryzant, Reid, Kelly Shen, Dan Jurafsky, and Stefan Wagner (2018). "Deconfounded Lexicon Induction for Interpretable Social Science." In: *Proceedings of ACL*. New Orleans, Louisiana, pp. 1615–1625. DOI: 10.18653/v1/N18-1146.

Qian, Bo and Khaled M. Rasheed (2005). "Hurst Exponent and Financial Market Predictability." In: *Proceedings of the IASTED International Conference on Financial Engineering and Applications*, pp. 203–

209. URL: https://c.mql5.com/forextsd/forum/170/hurst_exponent_and_financial_market_predictability.pdf.

Qin, Yu and Yi Yang (2019). "What You Say and How You Say It Matters: Predicting Financial Risk Using Verbal and Vocal Cues." In: *Proceedings of ACL*, pp. 390–401. DOI: 10.18653/v1/P19-1038.

Quarfoot, David and Richard A. Levine (2016). "How Robust Are Multirater Interrater Reliability Indices to Changes in Frequency Distribution?" In: *The American Statistician* 70.4, pp. 373–384. DOI: https://doi.org/10.1080/00031305.2016.1141708.

Quinlan, J. Ross (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers. URL: https://dl.acm.org/doi/book/10.5555/152181.

Rammstedt, Beatrice and Oliver P. John (2007). "Measuring Personality in One Minute or Less: A 10-Item Short Version of the Big Five Inventory in English and German." In: *Journal of Research in Personality* 41.1, pp. 203–212. DOI: 10.1016/j.jrp.2006.02.001.

Rekabsaz, Navid, 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*, pp. 1712–1721. DOI: 10.18653/v1/P17-1157.

Riloff, Ellen, Janyce Wiebe, and Theresa Wilson (2003). "Learning Subjective Nouns using Extraction Pattern Bootstrapping." In: *Proceedings of CoNLL*, pp. 25–32. URL: https://www.aclweb.org/anthology/W03-0404.

Rogers, Jonathan L. (2008). "Disclosure Quality and Management Trading Incentives." In: *Journal of Accounting Research* 46.5, pp. 1265–1296. DOI: 10.1111/j.1475-679X.2008.00308.x.

van Rooij, Maarten, Annamaria Lusardi, and Rob Alessie (2011). "Financial Literacy and Stock Market Participation." In: *Journal of Financial Economics* 101.2, pp. 449–472. DOI: 10.3386/w13565.

Rubin, Victoria L. (2007). "Stating with Certainty or Stating with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements." In: *Proceedings of NAACL*, pp. 141–144. URL: http://www.aclweb.org/anthology/N/N07/N07-2036.

Sachse, Katharina, Helmut Jungermann, and Julia M. Belting (2012). "Investment Risk – The Perspective of Individual Investors." In: *Journal of Economic Psychology* 33.3, pp. 437–447. DOI: 10.1016/j.joep.2011.12.006.

Sawhney, Ramit, Mihir Goyal, Prakhar Goel, Puneet Mathur, and Rajiv Ratn Shah (2021). "Multimodal Multi-Speaker Merger & Acquisition Financial Modeling: A New Task, Dataset, and Neural Baselines." In: *Proceedings of ACL*, pp. 6751–6762. DOI: 10.18653/v1/2021.acl-long.526.

Sawhney, Ramit, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Ratn Shah (2020). "VolTAGE: Volatility Forecasting via Text Audio Fusion with Graph Convolution Networks for Earnings Calls." In: *Proceedings of EMNLP*, pp. 8001–8013. DOI: 10.18653/v1/2020.emnlp-main.643.

Schumpeter, Joseph A. (1950). *Capitalism, Socialism and Democracy*. 3rd ed. London, UK: Routledge.

Serra-Garcia, Marta, Eric van Damme, and Jan Potters (2011). "Hiding an Inconvenient Truth: Lies and Vagueness." In: *Games and Economic Behavior* 73.1, pp. 244–261. DOI: 10.1016/j.geb.2011.01.007.

Sharpe, William F. (1963). "A Simplified Model for Portfolio Analysis." In: *Management Science* 9.2, pp. 277–293. DOI: 10.1287/mnsc.9.2.277.

— (1964). "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." In: *The Journal of Finance* 19.3, pp. 425–442. DOI: 10.1111/j.1540-6261.1964.tb02865.x.

Shefrin, Hersh (2001). "Do Investors Expect Higher Returns from Safer Stocks than from Riskier Stocks?" In: *Journal of Psychology and Financial Markets* 2.4, pp. 176–181. DOI: 10.1207/S15327760JPFM0204_1.

Shmueli, Galit (2010). "To Explain or to Predict?" In: *Statistical Science* 25.3, pp. 289–310. DOI: 10.1214/10-STS330.

Sitkin, Sim B. and Laurie R. Weingart (1995). "Determinants of Risky Decision-Making Behavior: A Test of the Mediating Role of Risk Perceptions and Propensity." In: *The Academy of Management Journal* 38.6, pp. 1573–1592.

Smithson, Michael (2012). *Ignorance and Uncertainty: Emerging Paradigms*. Springer Science & Business Media. DOI: 10.1007/978-1-4612-3628-3.

Spiess, Andrej-Nikolai and Natalie Neumeyer (2010). "An Evaluation of R^2 as an Inadequate Measure for Nonlinear Models in Pharmacological and Biochemical Research: a Monte Carlo Approach." In: *BMC Pharmacology* 10.6. DOI: 10.1186/1471-2210-10-6.

Štajner, Sanja, Goran Glavaš, Simone Paolo Ponzetto, and Heiner Stuckenschmidt (2017). "Domain Adaptation for Automatic Detection of Speculative Sentences." In: *Proceedings of IEEE ICSC*, pp. 164–. DOI: 10.1109/ICSC.2017.35.

Štajner, Sanja and Seren Yenikent (2020). "A Survey of Automatic Personality Detection from Texts." In: *Proceedings of COLING*, pp. 6284–6295. DOI: 10.18653/v1/2020.coling-main.553.

— (2021). "Why Is MBTI Personality Detection from Texts a Difficult Task?" In: *Proceedings of EACL*, pp. 3580–3589. URL: https://aclanthology.org/2021.eacl-main.312/.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019). "Energy and Policy Considerations for Deep Learning in NLP." In: *Proceedings of ACL*, pp. 3645–3650. DOI: 10.18653/v1/p19-1355.

Suslava, Kate (2021). ""Stiff Business Headwinds and Uncharted Economic Waters": The Use of Euphemisms in Earnings Conference Calls." In: *Management Science (to appear)*. DOI: 10.1287/mnsc.2020.3826.

Szarvas, György (2008). "Hedge Classification in Biomedical Texts with a Weakly Supervised Selection of Keywords." In: *Proceedings of ACL*, pp. 281–289. URL: https://www.aclweb.org/anthology/P08-1033.

Taylor, John Robert (1999). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. Sausalito (CA), USA: University Science Books.

Tetlock, Paul C. (2007). "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." In: *Journal of Finance* 62.3, pp. 1139–1168. DOI: 10.1111/j.1540-6261.2007.01232.x.

Theil, Christoph Kilian, Samuel Broscheit, and Heiner Stuckenschmidt (2019). "PRoFET: Predicting the Risk of Firms from Event Transcripts." In: *Proceedings of IJCAI*, pp. 5211–5217. DOI: 10.24963/ijcai.2019/724.

Theil, Christoph Kilian, Sanja Štajner, and Heiner Stuckenschmidt (2020). "Explaining Financial Uncertainty through Specialized Word Embeddings." In: *ACM/IMS Transactions on Data Science* 1.1, pp. 1–19. DOI: 10.1145/3343039.

Theil, Christoph Kilian, Sanja Štajner, Heiner Stuckenschmidt, and Simone Paolo Ponzetto (2017). "Automatic Detection of Uncertain Statements in the Financial Domain." In: *Proceedings of CICLing*, pp. 642–654. DOI: 10.1007/978-3-319-77116-8_48.

Theil, Kilian, Jens Daube, and Heiner Stuckenschmidt (2022). *Linguistic Uncertainty and Risk Perception in Financial Disclosures*. Working Paper. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4012946.

Theil, Kilian, Dirk Hovy, and Heiner Stuckenschmidt (2023). "Top-Down Influence? Predicting CEO Personality and Risk Impact from Speech Transcripts." In: *Proceedings of ICWSM (forthcoming)*. URL: https://arxiv.org/abs/2201.07670.

Theil, Kilian and Heiner Stuckenschmidt (2020). "Predicting Modality in Financial Dialogue." In: *Proceedings of the COLING Workshop on Financial Narrative Processing (FNP)*, pp. 226–234. URL: https://www.aclweb.org/anthology/2020.fnp-1.35/.

Theil, Kilian, 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)*, pp. 32–37. DOI: 10.18653/v1/W18-3104.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso." In: *Journal of the Royal Statistical Society* 58.1, pp. 267–288. URL: http://www.jstor.org/stable/2346178.

Tsai, Ming-Feng and Chuan-Ju Wang (2014). "Financial Keyword Expansion via Continuous Word Vector Representations." In: *Proceedings of EMNLP*, pp. 1453–1458. DOI: 10.3115/v1/D14-1152.

Tsai, Ming-Feng, Chuan-Ju Wang, and Po-Chuan Chien (2016). "Discovering Finance Keywords via Continuous-Space Language Models." In: *ACM Transactions on Management Information Systems* 7.3, pp. 1–17. DOI: 10.1145/2948072.

Turner, Andrew L. and Eric J. Weigel (1992). "Daily Stock Market Volatility: 1928–1989." In: *Management Science* 38.11, pp. 1586–1609. DOI: 10.1287/mnsc.38.11.1586.

Tversky, Amos and Daniel Kahneman (1986). "Rational Choice and the Framing of Decisions." In: *The Journal of Business* 59.4, pp. 251–278. URL: https://www.jstor.org/stable/2352759.

Vazire, Simine (2006). "Informant Reports: A Cheap, Fast, and Easy Method for Personality Assessment." In: *Journal of Research in Personality* 40.5, pp. 472–481. DOI: https://doi.org/10.1016/j.jrp.2005.03.003.

Vincze, Veronika, György Szarvas, Richárd Farkas, György Móra, and Janos Csirik (2008). "The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and Their Scopes." In: *BMC Bioinformatics* 9, pp. 1–9. DOI: 10.1186/1471-2105-9-S11-S9.

Wan, Li, Quan Wang, Alan Papir, and Ignacio Lopez Moreno (2018). "Generalized End-to-End Loss for Speaker Verification." In: *Proccedings of ICASSP*. URL: https://arxiv.org/abs/1710.10467.

Wang, Chuan-Ju, Ming-Feng Tsai, Tse Liu, and Chin-Ting Chang (2013). "Financial Sentiment Analysis for Risk Prediction." In: *Proceedings of IJCNLP*, pp. 802–808. URL: https://aclanthology.org/I13-1097/.

Wang, Mei, Carmen Keller, and Michael Siegrist (2011). "The Less You Know, the More You Are Afraid of—A Survey on Risk Perceptions of Investment Products." In: *Journal of Behavioral Finance* 12.1, pp. 9–19. DOI: 10.1080/15427560.2011.548760.

Wang, William Yang and Zhenhao Hua (2014). "A Semiparametric Gaussian Copula Regression Model for Predicting Financial Risks from Earnings Calls." In: *Proceedings of ACL*, pp. 1155–1165. DOI: 10.3115/v1/P14-1109.

Weber, Martin, Elke U. Weber, and Alen Nosić (2013). "Who Takes Risks When and Why: Determinants of Changes in Investor Risk Taking." In: *Review of Finance* 17.3, pp. 847–883. DOI: 10.1093/rof/rfs024.

Wolf, Thomas et al. (2020). "Transformers: State-of-the-Art Natural Language Processing." In: *Proceedings of EMNLP*, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Xing, Frank Z., Erik Cambria, and Roy E. Welsch (2018). "Natural Language Based Financial Forecasting: A Survey." In: *Artificial Intelligence Review* 50.1, pp. 49–73. DOI: 10.1007/s10462-017-9588-9.

Yang, Linyi, Tin Lok Ng, Barry Smyth, and Riuhai Dong (2020). "HTML: Hierarchical Transformer-Based Multi-Task Learning for Volatility Prediction." In: *Proceedings of WWW*, pp. 441–451. DOI: 10.1145/3366423.3380128.

Ye, Zhen, Yu Qin, and Wei Xu (2020). "Financial Risk Prediction with Multi-Round Q&A Attention Network." In: *Proceedings of IJCAI*, pp. 4576–4582. DOI: 10.24963/ijcai.2020/631.

Zeisberger, Stefan (2021). "Do People Care About Loss Probabilities?" In: *Journal of Risk and Uncertainty* forthcoming. DOI: 10.2139/ssrn.2169394.

Zerva, Chrysoula (2019). "Automatic Identification of Textual Uncertainty." PhD Thesis. University of Manchester. URL: https://www.research.manchester.ac.uk/portal/files/86864517/FULL_TEXT.PDF.

Zhao, Zhe, Tao Liu, Shen Li, Bofang Li, and Xiaoyong Du (2017). "Ngram2vec: Learning Improved Word Representations from Ngram Co-Occurrence Statistics." In: *Proceedings of EMNLP*, pp. 244–253. DOI: 10.18653/v1/d17-1023.

Zhu, Wanzheng and Suma Bhat (2021). "Euphemistic Phrase Detection by Masked Language Model." In: *arXiv e-prints* 2109.04666. URL: https://arxiv.org/abs/2109.04666.

Zou, Hui and Trevor Hastie (2005). "Regularization and Variable Selection via the Elastic Net." In: *Journal of the Royal Statistical Society* 67, pp. 301–320. DOI: 10.1111/j.1467-9868.2005.00503.x.

A

PUBLISHED RESOURCES

In table 25, we provide an overview of the publicly available implementations and datasets created during this thesis.

Table 25: Data and code published along with this thesis. Column 1 contains the corresponding task, 2 the referencing chapter, 3 a brief description, 4 the resource type, and 5 the link to its online location.

Task	Ch.	Description	Type	Location
T_1	4	gold standard	Data	`http://data.dws.informatik.uni-mannheim.de/theil/theil_2017_uncertainty_detection.zip`
T_1	5	uncertainty classifier, silver standard	Model, Data	`http://data.dws.informatik.uni-mannheim.de/theil/theil_2020_modality.zip`
T_2	6	`R` scripts, survey data	Code, Data	`http://data.dws.informatik.uni-mannheim.de/theil/theil_2022_risk_perception.zip`
T_3	7	Python scripts, text data, dictionaries, financials	Code, Data	`http://data.dws.informatik.uni-mannheim.de/theil/theil_2022_risk_regression.zip`
T_3	8	Python scripts, financials	Code, Data	`http://data.dws.informatik.uni-mannheim.de/theil/theil_2023_ceo_personality.zip`
T_3	9	PRoFET, financials	Model, Data	`https://github.com/samuelbroscheit/neural-profet`

B

APPENDIX TO CHAPTER 6

This is the appendix to Chapter 6 "Linguistic Uncertainty and Risk Perception in Financial Disclosures." The following contains the used text materials (B.1), details on the pilot study (B.2), and details on the main study (B.3.2).

B.1 MATERIALS

B.1.1 *Scenario*

The asset allocation decision was assessed by using the following fictional scenario:

"After working for three years, you have saved €10,000. Your goal is to increase your capital over the next two years. For this purpose, you will be presented with excerpts of earnings calls from four different, but similar companies. The companies are competitors and had similar size, profit margins, and revenue growth in the past. Based on the four excerpts of earnings calls, you should decide whether and how much you want to invest in each company."

B.1.2 *Earnings Call Excerpts*

In the following, all excerpts from earnings calls used in this study are presented in their uncertain and certain version, respectively.

- **Excerpt 1 (uncertain)**: After a promising last year, I would like to address our current outlook for the next year. Despite the macroeconomic data coming out of North America, we still believe that U.S. market sales increase. In particular, we possibly increase our net sales by 5%. The European market seems to be more fragmented, which makes our outlook for the year vague. On a consolidated basis, it appears that our costs decrease by 5%. The automotive industry continues to be dynamic and to depend on world politics. In other words, the development of the industry depends somewhat on future regulations. We are adjusting to this environment and assume to see similar general business trends in the second half of the year that we saw in the first.

- **Excerpt 1 (certain)**: After a promising last year, I would like to address our current outlook for the next year. Despite the macroeconomic data coming out of North America, we will increase our U.S. market sales. In particular, we definitely increase our net sales by at least 5%. The European market will undoubtedly grow at the same rate. On a consolidated basis, we definitely decrease our costs by 5%. The automotive industry continues to be dynamic. Moreover, the development of the industry goes along with future regulations. We are adjusting to this environment and clearly see the same general business trends in the second half of the year that we saw in the first. Overall, we expect a promising next year for all markets.

- **Excerpt 2 (uncertain)**: Due to the somewhat unclear industry outlook, it is hard to predict the development of the next year. However, we may increase our market share by 5% and our net sales by 5%. Moreover, we possibly increase our profit by roughly 5%. Several drivers exist that give us confidence in our outlook, which include the sales of our new cars, which could exceed our expectations and possibly increase over the next years. In general, we recognize the market has become more competitive, but we try to follow our plan and we are taking the necessary actions to help mitigate these headwinds. Finally, concerning our company strategy, we are very much on plan and could increase our annual profit and margin growth.

- **Excerpt 2 (certain)**: Due to the industry outlook, we will sustain our last year performance throughout the next year. Specifically, we expect to increase our market share by 5% and our net sales by 5%. Moreover, we will increase our profit by 5%. Several drivers exist that give us confidence in our outlook, which include the sales of our new cars, which are undoubtedly exceeding our expectations and strongly increase over the next years. In general, we recognize the market has become more competitive, but we definitively follow our plan and we are taking the necessary actions to help mitigate these headwinds. Finally, concerning our company strategy, we are very much on plan and clearly increase our annual profit and margin growth.

- **Excerpt 3 (uncertain)**: In the last year, we introduced new cars at the Hanover Car Show, maybe with the latest technology, including the cars D1, K3, and B4. The new cars possibly affect our aftermarket support and our manufacturing facilities. Moreover, our company appears to expand the market share in the global car market by developing new products and services in the industry, investing in new geographic regions, and occasionally providing returns to our shareholders. Looking at future figures, we might increase our total sales by almost 5%. For the emerging market, we see a possible increase in profit of 5%. In line with the emerging markets, we perhaps increase our numbers by roughly 5% in profit in our main markets, due to lower cost structures.

- **Excerpt 3 (certain)**: In the last year, we undoubtedly introduced new cars with the most advanced technology at the Hanover Car Show, including the cars D1, K3, and B4. The new cars will affect our aftermarket support and manufacturing facilities. Moreover, our company continues to expand the market share in the global car market by developing new products and services in the industry, investing in new geographic regions, and always providing returns to our shareholders. Looking at future figures, we will increase our total sales by 5%. For the emerging market, we clearly see an increase in profit of 5%. In line with the emerging markets, we definitely increase our numbers by 5% in profit in our main markets, due to lower cost structures.

- **Excerpt 4 (uncertain)**: Looking forward, we could see our net sales to increase by roughly 5% due to the appearing economic trends that are generating an increased number in sales. In Europe, we expect a possible increase of nearly 5% in net sales. The UK economy, which is our largest market in the region, may grow by almost 5% in the coming year. On the other side, the U.S. economy may grow due to minor regulating and supervisory mechanisms. In addition, we expect gross margins to increase by 5%, reflecting the benefits of higher production levels. That means that we are heading towards the next year with a possible momentum in the U.S. the car industry. Finally, I want to thank of our employees for their work.

- **Excerpt 4 (certain)**: Looking forward, we definitely expect our net sales to increase by 5% due to the economic trends that are generating an increased number in sales. In Europe, we clearly expect an increase of 5% in net sales. The UK economy, which is our largest market in the region, will grow by 5%

> in the coming year. On the other side, the U.S. economy will strongly grow primarily due to minor regulating and supervisory mechanisms. In addition, we expect gross margins to increase by 5%, reflecting the benefits of higher production levels. That means that we are heading towards the next year with an undisputed momentum in the U.S. the car industry. Finally, I want to thank of our employees for their work.

B.2 PILOT STUDY

The pilot study was conducted with a sample representative of the main study. We expected the excerpts containing words from the UNCERTAINTY and WEAK MODAL dictionaries by Loughran and McDonald, 2011 to be perceived as uncertain, while excerpts containing words from the STRONG MODAL dictionary to be certain.

B.2.1 *Pilot Study Design*

Using the eight earnings call snippets (cf. §B.1.2) as a basis, each participant saw two certain and two uncertain snippets. Snippets were randomly drawn and simultaneously shown in a 2 x 2 layout with randomized positions. Then, participants were asked to rate the linguistic uncertainty of each snippet.

B.2.1.1 *Participants*

Assuming a large Cohen's d of 0.8, an *a priori* power analysis yielded a minimum sample size of 19 for the pilot study. To allow for additional room, 28 participants were recruited. No data was removed due to the completeness and high quality of the data. Subjects were acquaintances of the second author. The average age of the participants was 28.53 years ($s_x = 6.2$), and the sample consisted of 19 men and 9 women. The majority of participants worked full-time (71.42%) with a median annual income of €50,000–€60,000.

B.2.1.2 *Items*

LINGUISTIC UNCERTAINTY "How do you perceive the language in this excerpt?" was measured on a 7-point Likert scale ranging from (1) "extremely uncertain" to (7) "extremely certain." In addition, participants were provided with the following definition: "Uncertainty is generally characterized by ambiguity, vagueness, or probability." The definition is based on the taxonomy of uncertainty developed by Smithson (2012). Only a brief definition was used to avoid biasing participants and to obtain an intuitive judgement.

SOCIO-DEMOGRAPHICS Age was measured as a numerical value between 18 and 99. Gender was assessed via three categories: "male,"

Table 26: Results of the pilot study on uncertainty perception (assessed on a 7-point Likert scale from (1) "extremely uncertain" to (7) "extremely certain") for all uncertain and certain excerpts. $\tilde{x}$ stands for median and R for range.

	Uncertain		Certain	
	$\tilde{x}$	R	$\tilde{x}$	R
Excerpt 1	2.00	1.00	5.00	2.00
Excerpt 2	2.00	2.00	2.00	2.00
Excerpt 3	3.50	2.00	6.00	1.00
Excerpt 4	4.00	2.00	5.00	2.00
Total	3.00	3.00	5.50	2.00

"female," and "diverse." Annual income was assessed with six different levels: less than €30,000; €30,000–€40,000; €40,000–€50,000; €50,000–€60,000; €60,000–€70,000; more than €70,000. Academic degree was measured with an ordinal scale with academic degrees ranging from secondary school to PhD. Work status was assessed by using a categorical variable with eight different levels.

CONTROL QUESTIONS As the study was conducted in Germany but consisted of English-language instructions and questions, the participants were asked if they had any problems in language comprehension during the experiment. Additionally, the participants were asked if they had answered all questions thoroughly. Results of participants who answered the control questions with "No" were excluded.

B.2.2 *Pilot Study Results*

Table 26 contains the median and range of the pilot study responses per text excerpt 1–4 and group (*uncertain* vs. *certain*). Wilcoxon signed-rank tests were conducted to check for significant differences regarding linguistic uncertainty between *certain* and the *uncertain* excerpts. The test indicated a highly significant difference ($p \leq 0.001$) for each excerpt with Cohen's $d \in \{5.12, 4.85, 4.50, 2.73\}$, indicating a "large" effect (Cohen, 1998). In summary, the pilot study indicates that the modified earnings call excerpts used in the main study are valid concerning their linguistic uncertainty.

B.3 MAIN STUDY

B.3.1 *Items*

SOCIO-DEMOGRAPHICS Socio-demographic variables were assessed using the same items as for the pilot study.

RISK PERCEPTION The following items were measured with a 5-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (5):

- "I believe that the price of the company's stock will rise in the future."
- "I believe that the price of the company's stock will decline in the future."
- "I think that an investment in the company's stock is risky."
- "I think the value of the company's stock will fluctuate."

FINANCIAL LITERACY The following two general questions addressing experience and confidence were posed:

- "Do you have experience with investing in shares, ETFs, or funds?"
  - "Yes"
  - "No"
- "How confident are you about your ability to invest?"
  - "Not confident at all"
  - "Confident"
  - "Very confident"

In addition, the one-item financial literacy assessment by Gibson et al. (2013) addressing knowledge was used:

- "How knowledgeable are you about investing?"
  - "I have very little knowledge"
  - "I have reasonable knowledge"
  - "I have good knowledge"
  - "I am an expert"

Lastly, three knowledge questions developed by Lusardi and Mitchell (2011) that have been used in a variety of studies (Lusardi and Mitchell, 2007; Nguyen et al., 2019; van Rooij et al., 2011) were used (correct answers are marked):

- "Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?"
  - "More than $102" (x)

  - "Exactly $102"
  - "Less than $102"

- Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, how much would you be able to buy with the money in this account?
  - "More than today"
  - "Exactly the same"
  - "Less than today" (x)

- Please tell indicate whether this statement is true or false. "Buying a single company's stock usually provides a safer return than a stock mutual fund."
  - "True"
  - "False" (x)

RISK TOLERANCE Four items from the established risk tolerance questionnaire FINAMETRICA,[1], which were selected by Guillemette et al. (2012), were used:

- "When faced with a major financial decision, are you more concerned about the possible losses or the possible gains?"
  - "Always the possible losses"
  - "Usually the possible losses"
  - "Usually the possible gains"
  - "Always the possible gains"

- "Investments can go up and down in value, and experts often say you should be prepared to weather a downturn. By how much could the total value of all your investments go down before you would begin to feel uncomfortable?"
  - "Any fall in value would make me feel uncomfortable"
  - "10%"
  - "20%"
  - "33%"
  - "50%"
  - "More than 50%"

- "What risks have you taken with your financial decisions in the past?"
  - "Very small"
  - "Small"
  - "Medium"
  - "Large"

1 https://www.riskprofiling.com/

  - "Very large"

- " What risks are you currently prepared to take with your financial decisions?"
  - "Very small"
  - "Small"
  - "Medium"
  - "Large"
  - "Very large"

In addition, two items developed by Grable and Lytton (1999) that have been proven to be reliable and valid by subsequent literature (Grable and Lytton, 2003; Guillemette et al., 2012) were used:

- "In addition to whatever you own, you have been given $1,000 to invest. You are now asked to choose between:"
  - "A sure gain of $500"
  - "A 50% chance to gain $1,000 and a 50% chance to gain nothing"

- "In addition to whatever you own, you have been given $2,000. You are now asked to choose between:"
  - "A sure loss of $500"
  - "A 50% chance to lose $1,000 and a 50% chance to lose nothing"

BIG FIVE PERSONALITY TRAITS We used the BF-10 developed by Rammstedt and John (2007). Participants could rate the following items on a 5-point Likert scale, from "strongly disagree" (1) to "strongly agree" (5):

- "I see myself as someone who ..."
  - "... is reserved." (*extraversion* $-$)
  - "... is generally trusting." (*agreeableness* $+$)
  - "... tends to be lazy." (*conscientiousness* $-$)
  - "... is relaxed, handles stress well." (*neuroticism* $-$)
  - "... has few artistic interests." (*openness* $-$)
  - "... is outgoing, sociable." (*extraversion* $+$)
  - "... tends to find fault with other." (*agreeableness* $-$)
  - "... does a thorough job." (*conscientiousness* $+$)
  - "... gets nervous easily." (*neuroticism* $+$)
  - "... has an active imagination." (*openness* $+$)

B.3.2 *Full Results*

RISK PERCEPTION Table 27 provides a detailed overview of the results for the risk perception items. On average, participants were more convinced that the stock price of the certain excerpts would

Table 27: Descriptive statistics for the risk perception for all excerpts divided into certain and uncertain version. The items were measured using a 5-point Likert scale ranging from "strongly disagree" (1) to "strongly agree" (5).

(a) "I believe that the price of the company's stock will rise in the future."

	Uncertain		Certain	
	$\bar{x}$	s_x	$\bar{x}$	s_x
Text 1	3.02	0.92	3.74	0.75
Text 2	2.83	0.84	3.71	1.03
Text 3	2.85	1.01	4.10	0.76
Text 4	3.41	0.84	3.83	0.78
Total	3.03	0.93	3.85	0.85

(b) "I believe that the price of the company's stock will decline in the future."

	Uncertain		Certain	
	$\bar{x}$	s_x	$\bar{x}$	s_x
Text 1	3.00	0.96	2.49	0.85
Text 2	3.20	0.76	2.32	0.76
Text 3	3.13	0.95	2.17	0.70
Text 4	2.71	0.72	2.42	0.84
Total	3.01	0.87	2.35	0.79

(c) "I think that an investment in the company's stock is risky."

	Uncertain		Certain	
	$\bar{x}$	s_x	$\bar{x}$	s_x
Text 1	3.67	0.90	3.15	0.84
Text 2	3.83	0.81	3.10	0.97
Text 3	3.79	0.80	3.05	1.01
Text 4	3.46	0.90	2.83	1.03
Total	3.69	0.86	3.03	0.97

(d) "I think the value of the company's stock will fluctuate."

	Uncertain		Certain	
	$\bar{x}$	s_x	$\bar{x}$	s_x
Text 1	3.74	0.83	3.46	0.88
Text 2	3.85	0.80	3.34	0.91
Text 3	3.72	0.72	3.29	1.02
Text 4	3.61	0.80	3.20	0.88
Total	3.73	0.79	3.32	0.92

rise than the price of the uncertain ones. The opposite was true for the belief about declining stock prices. Participants thought that an investment into the uncertain companies was riskier than for the certain ones.

SOCIO-DEMOGRAPHICS Table 29 presents an overview of the focal socio-demographics as absolute and relative frequencies. All 81 participants were from Germany. The average age was 28.88 years ($s_x = 7.33$) ranging from 19 years to 55 years. The sample consisted of 61 men (75.31%) and 20 women (24.69%). The majority of participants was working full-time or was self-employed (54.32%), the second largest group were students (22.22%), the third largest group were working students (16.05%), and the others were either working part-time (4.94%) or unemployed (2.47%). Considering the high-

Table 29: Counts and percentages of socio-demographic characteristics (n = 81).

Variable	Count	Percentage
Gender		
Male	61	75.31%
Female	20	24.69%
Diverse	0	0.00%
Age		
18–24	16	19.75%
25–30	48	59.26%
31–40	11	13.58%
Over 40	6	7.41%
Income		
Less than €30,000	31	38.27%
€30,000–€40,000	7	8.64%
€40,000–€50,000	6	7.41%
€50,000–€60,000	2	2.47%
€60,000–€70,000	14	17.28%
More than €70,000	16	19.75%
Unspecified	5	6.18%

est academic qualification, it was observed that most of the participants were academics (35.80% Bachelor degree, 32.10% Master degree, and 7.41% PhD). The other participants either had Abitur (12.35%), a secondary school certificate (6.17%), or another academic qualification (6.17%). The annual income of the most participants was below €30,000 (38.27%), followed by above €70,000 (19.75%).

BIG FIVE PERSONALITY TRAITS After inverting the negative items, each factor was assessed as the average of both items. Descriptive statistics can be found in Table 30.

FINANCIAL LITERACY The results of the financial literacy questions are presented in Table 31. More than half of the participants (n = 49, 60.49%) already had experience with investing in shares, ETFs, or funds. However, no participant considered themselves as an expert investor. 35 participants (43.21%) had reasonable knowledge about investing, 25 participants (30.86%) had very little knowledge, and 21 participants (25.93%) had good knowledge. A little more than half of the participants were confident about their ability to invest (n = 42, 51.85%), whereas the remaining participants were not confident at all (n = 39, 48.15%). None of the participants was “very confident" about their ability to invest. Regarding the knowledge test,

Table 30: Descriptive statistics of the Big Five personality traits after inverting the negative items ($n = 81$). Negative items are denoted by $-$ and positive items by $+$. Items were measured with a 5-point Likert scale from "strongly disagree" (1) to "strongly agree" (5).

Big 5	I see myself as someone who ...	$\bar{x}$	s_x	$\bar{x}$	s_x
O	... has few artistic interests. $(-)$	2.83	1.26	3.17	0.95
	... has an active imagination. $(+)$	3.52	1.11		
C	... tends to be lazy. $(-)$	3.65	1.16	3.81	0.81
	... does a thorough job. $(+)$	3.98	0.84		
E	... is reserved. $(-)$	2.96	1.04	3.30	0.86
	... is outgoing, sociable. $(+)$	3.64	0.95		
A	... tends to find fault with others. $(-)$	3.28	1.00	3.28	0.75
	... is generally trusting. $(+)$	3.28	0.91		
N	... is relaxed, handles stress well. $(-)$	2.60	1.14	2.67	1.00
	... gets nervous easily. $(+)$	2.74	1.15		

63 participants (77.78%) answered all three questions correctly, 16 participants answered two questions correctly (19.75%). One participant answered one question correctly, and one participant did not answer any question correctly.

RISK TOLERANCE When faced with a major financial decision, most of the participants were either usually concerned about possible losses ($n = 37$, 45.68%) or usually concerned about possible gains ($n = 29$, 35.80%). Six participants (7.41%) were always concerned about possible gains and nine (11.11%) were always concerned about the possible losses. When it comes to a downturn of an investment, 14.81% of the participants would feel uncomfortable with any fall in value, 13.85% would feel uncomfortable with a decrease of 10% (33.33% with 20%, 24.69% with 33%, 9.88% with 50%, and 3.70% with more than 50%). In respect to the risk that participants have taken with their financial decisions in the past, 45.68% haven taken a small or very small degree of risk, 29.63% a large or very large degree of risk, and 24.69% a medium degree of risk. Considering the risk they were currently prepared to take with their financial decision, 45.67% are prepared to take a small or very small degree of risk, 18.52% a large or very large degree of risk, and 35.80% for a medium degree of risk. When choosing between a certain gain and an uncertain gain, 74.07% would choose a certain gain. In contrast, only 46.91% would choose a certain loss instead of an uncertain loss.

Table 31: Resulting absolute and relative frequencies of the results divided into self-assessed and objective financial literacy.

Question	Count	Percentage
Experience		
No	32	39.51%
Yes	49	60.49%
Knowledge		
Very little	25	30.86%
Reasonable	35	43.21%
Good	21	25.93%
Expert	0	0.00%
Confidence		
Not confident	39	48.15%
Confident	42	51.85%
Very Confident	0	0.00%
Interest rate		
Correct	80	98.77%
Wrong	1	1.23%
Inflation		
Correct	77	95.05%
Wrong	4	4.94%
Risk of products		
Correct	65	80.25%
Wrong	16	19.75%

C

APPENDIX TO CHAPTER 7

This is the appendix to Chapter 7 "Risk Regression from Linguistic Uncertainty." Table 32 contains the uncertain candidate terms selected by the multi-task learner.

Table 32: Candidate terms with an above-zero coefficient as reported by the elastic net regressor. Regressions were fit using 23K 10-Ks spanning years 1994–2020 with the cumulative tf–idf of each uncertain candidate term (n = 11K) as an input and the three dependent variables volatility, analyst error, and dispersion as output. Terms are sorted in descending order according to their average coefficient across the three independent variables.

territoriality, disreputable, amputated, warrants, hyperglycemia, coronavirus, nonhomogeneous, raise, postulates, burglaries, million, establishes, able, deterioration, vie, personalize, achievability, setback, mastheads, understatement, lysins, widening, unprecedented, valuation, currently, worsening, enmeshed, deteriorating, downgraded, preferred, setbacks, reduced, result, date, substantially, subfreezing, efforts, ratoon, betting, persist, complete, terminate, obtain, closing, tightening, chariots, dissatisfy, generate, hijacking, adynamic, codification, reduce, rescinded, upthrown, expedients, suspend, experienced, foreseeable, sufficient, dramatic, foreclosures, weakness, acoustically, bushels, topline, downturn, miscalculations, install, turnaround, fallen, candidates, untoward, yellowhammer, headlining, altitudes, oedema, bears, catalyze, decreased, osseous, announcements, reclamation, expands, originate, reducing, measuring, retender, noncompensatory, prorating, presage, rems, undetected, liquated, correctible, declining, harvestable, chromosomes, roofs, dramatically, substantial, gallons, receivable, contempt, scheduled, lack, designate, speculators, predictions, dilutive, channels, frequent, future, delays, incompleteness, kilobits, deficiency, presences, worst, projection, defines, crisis, repriced, realizability, depletable, deploying, significantly, contingency, decrease, converted, fonds, constrained, futility, elevated, moieties, misleading, geological, motived, ungulate, discourages, icons, challenging, rebuttals, questions, preselected, scripts, accretes, rescission, capitalize, availability, milestone, turmoil, endorse, shortfall, redetermining, rescind, pitfalls, crosscut, retain, deflectors, nonproductive, targeted, unreliable, dizziness, unsophisticated, unobtrusive, flew, steps, unemployable, contango, cannons, curtail, voyages, pursue, assurances, entrant

D

APPENDIX TO CHAPTER 8

This is the appendix to Chapter 8 "Risk Regression from CEO Personality."

D.1 FINAL HYPERPARAMETERS

Using a Bayesian hyperparameter optimization as specified in §8.3.2, we obtain the configurations specified in Table 34a and 34b with minimal loss on the validation set.

Table 33: Final hyperparameter configurations found by the Bayesian optimization searching over 40 configurations per MBTI dimension.

(a) Hyperparameters for BERT.

MBTI	Batch Size	Learning Rate
E–I	128	4.8×10^{-5}
S–N	32	4.9×10^{-5}
T–F	32	1.0×10^{-6}
J–P	256	8.6×10^{-6}

(b) Hyperparameters for RoBERTa.

MBTI	Batch Size	Learning Rate
E–I	256	4.3×10^{-5}
S–N	32	4.6×10^{-5}
T–F	128	9.4×10^{-8}
J–P	128	4.7×10^{-5}

D.2 RESULTS ON THE VALIDATION SET

Table 35: Validation results of the personality prediction task. CEO personality is predicted from earnings call transcripts across the four MBTI dimensions: *extraversion–introversion* (E–I), *sensing–intuition* (S–I), *thinking–feeling* (T–F), and *judging–perception* (J–P). The models are an SVM trained on trigram tf–idf vectors, $\text{BERT}_{\text{base}}$, and $\text{RoBERTa}_{\text{base}}$.

MBTI	Model	r	ρ	τ	MAE
E–I	SVM	0.70	0.69	0.55	0.38
	BERT	0.46	0.42	0.28	0.62
	RoBERTa	0.72	0.60	0.48	0.35
S–N	SVM	0.34	0.48	0.30	0.28
	BERT	0.20	0.35	0.24	0.53
	RoBERTa	0.43	0.61	0.43	0.27
T–F	SVM	0.13	−0.05	−0.03	0.33
	BERT	−0.43	−0.32	−0.22	0.38
	RoBERTa	0.11	−0.07	−0.03	0.36
J–P	SVM	−0.05	0.05	0.03	0.35
	BERT	0.32	0.28	0.19	0.53
	RoBERTa	0.25	0.14	0.06	0.40