Band 103

Marina Kölbl

Textual Analysis of Corporate Disclosures: The Case of REITs

International Real Estate Business School
Universität Regensburg

Schriften zu Immobilienökonomie und Immobilienrecht

Marina Kölbl

Textual Analysis of Corporate Disclosures: The Case of REITs

Die Deutsche Bibliothek – CIP Einheitsaufnahme
Marina Kölbl
Textual Analysis of Corporate Disclosures: The Case of REITs
Regensburg: Universitätsbibliothek Regensburg 2021
(Schriften zu Immobilienökonomie und Immobilienrecht; Bd. 103)
Zugl.: Regensburg, Univ. Regensburg, Diss., 2021
978-3-88246-458-0

978-3-88246-458-0

Verlag: Universitätsbibliothek Regensburg, Regensburg 2021
Zugleich: Dissertation zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft, eingereicht an der Fakultät für Wirtschaftswissenschaften der Universität Regensburg
Tag der mündlichen Prüfung: 16. Juli 2021
Berichterstatter: Prof. Dr. Wolfgang Schäfers
Prof. Dr. Gregor Dorfleitner

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 General Motivation

While practitioners and researchers have for decades been carefully considering how quantitative data affects financial markets, we are just now beginning to explore the narratives that contextualize quantitative data in corporate disclosures. This is surprising, given that efficient market theorists like Fama (1970) assume that prices always "fully reflect" available information – whether quantified or described qualitatively. Considering the benefits derived from the analysis of quantitative data, it would be negligent to ignore the ever-increasing amount of qualitative information. This is especially so, since the Securities and Exchange Commission (SEC) is pursuing its mission of ensuring efficient markets by mandating publicly traded firms to disclose significant information in numeric and textual form.[1] The latter is hereby intended to explain and complement quantitative figures. It provides insights that extend beyond financial projections and measures, and reveals, for example, manager expectations or important qualifiers and caveats that are not evident from numerical data (Ferris et al., 2013). Thereby, the qualitative, textual form is superior to numeric data in that it gives managers more flexibility to express information. They decide how precisely and with which words a certain aspect shall be described. By contrast, the presentation and calculation of numeric data are often determined by accounting standards like GAAP (Davis and Tama-Sweet, 2012).

Yet, discovering and quantifying information from unstructured text is a nontrivial task. For a long time, the only way to process such data was to simply read the flood of disclosures provided by companies. Manual handling of the large number of textual documents available is, however, extremely time-consuming. In fact, stakeholders have been criticizing qualitative disclosures for being too long and difficult to understand. Even a distinguished expert like Berkshire Hathaway CEO Warren Buffett once lamented: "Too often, I've been unable to decipher just what is being said or, worse yet, had to conclude that nothing was being said” (SEC, 1998). Thereby, not only pointing out difficulties in processing qualitative data, but also challenging the SECs mission by raising the question of whether qualitative information in corporate disclosures is truly informative. This is a challenging issue to solve, given the temporal and cognitive limitations of human beings that preclude large-scale analysis.

[1] For more information on the SEC’s mission, visit: https://www.sec.gov/Article/whatwedo.html.

Only with the rise of computational power and the tremendously increasing online-availability of text over the last few decades, computer-based techniques have emerged as a new way to analyze the vast amount of qualitative information provided by a plethora of firms. Easy access to data and analytical tools[2] has enabled practitioners and researchers alike to process a variety of texts from different disclosure media (e.g., voluntary and mandatory corporate disclosures, news articles, and internet postings). In order to collect and quantify information from the narratives, they increasingly rely on machine-assisted methods, and analyze, for example, readability, sentiment, targeted phrases, or measures of document similarity. As a consequence, researchers have presented promising results in predicting the reactions of financial markets, underlining the tremendous importance of new ways of processing qualitative data. Textual sentiment has been linked, for instance, to abnormal stock returns (e.g., Feldman et al., 2010; Chen et al., 2014; Jegadeesh and Wu, 2013), subsequent stock return volatility (e.g., Loughran and McDonald, 2011 & 2015), and future earnings and liquidity (e.g., Li, 2010). Readability on the other hand, is associated with stock return volatility (Loughran and McDonald, 2014), earnings persistence and earnings surprise (Li, 2008; Loughran and McDonald, 2014), analyst dispersion (Lehavy et al., 2011; Loughran and McDonald, 2014), as well as investors' trading behavior (Miller, 2010). Thus, these studies provide evidence that market participants incorporate more than just quantitative data in their decision making and would benefit from quick and efficient processing of disclosures.

Although textual analysis has received considerable attention in the accounting and finance literature, the new research opportunities provided by computerization and digitization have not been used with the same intensity in the real estate industry. In fact, Real Estate Investment Trusts (REITs) were explicitly excluded from the analysis, because of their different nature and regulatory requirements (e.g., underlying asset, distribution requirement). However, it is precisely these specific regulatory requirements of US REITs that make them an interesting starting point for an in-depth analysis. For example, the high dividend payout requirement of at least 90% of their taxable earnings allows only a limited cash reserve. Thus, REITs are highly dependent on external equity capital and must turn to the capital markets repeatedly, in order to take advantage of growth opportunities. As a result, US REIT managers have an unusually strong incentive to be transparent and maintain investor trust (e.g., Price et al., 2017), which ensures relatively high disclosure

[2] A large number of corporate disclosures are publicly available on the SEC's EDGAR database, and open-source software such as Python and R make the analysis of qualitative data accessible to everyone. In addition, market participants can buy the expertise they seek from leading information providers like S&P Global, Thomson Reuters, Bloomberg, and Dow Jones/Factiva.

quality. In addition, the significance of the US REIT market, as evidenced by its recent market capitalization of $1,249,186.3 million as of December 31, 2020 for 223 listed REITs, suggests that the sector is worth investigating while providing an ideal laboratory to close this gap in the literature.

This dissertation aims to provide first important insights in this regard by exploring whether qualitative data revealed in corporate disclosures of US REITs conveys valuable information to market participants, helping them to make well-informed investment decisions. In particular, the first paper comprising this thesis examines the informativeness of the Management Discussion and Analysis (MD&A) of US REITs. It investigates whether sentiment conveyed through the narrative is associated with future firm performance and whether it generates a market response. The second study exploits an unsupervised machine learning algorithm to identify the risk factors discussed in 10-Ks and evaluates the validity of this approach by analyzing the relation between the extracted risk-factor topics and stock return volatility after the filing date. The third article assesses whether linguistic issues are associated with underpricing, the phenomenon that an Initial Public Offerings' (IPOs') closing price at the end of the first trading-day is higher than its offer price. To do so, it determines the level of uncertain language in IPO prospectuses, as well as their similarity to prospectuses filed up to six months prior to the document in question.

Overall, this thesis is the first to conduct a comprehensive analysis on the informativeness of qualitative information provided in US REIT corporate disclosures. In particular, it investigates multiple disclosure media, uses different tools to process textual data, and analyzes various qualitative dimensions of text to assess whether and how the market reacts to qualitative information revealed by US REITs.

1.2 Research Questions

While all three research articles focus on the informativeness of qualitative data revealed in corporate disclosures of US REITs, they differ in terms of disclosure media, methodologies, textual features, and outcomes investigated. This section provides an overview of the subject matter and the superordinate research questions addressed in the respective papers of this dissertation.

Paper 1 | Is the MD&A of US REITs informative? A Textual Sentiment Study

Paper 1 examines whether language disclosed in the MD&A of US REITs provides signals regarding future firm performance and thus generates a market response. Specifically, the Loughran and McDonald (2011) financial dictionary, and a custom dictionary for the real estate industry created by Ruscheinsky et al. (2018), are employed to determine the inherent sentiment, that is, the level of pessimistic or optimistic language for each filing. The central considerations can be stated as such:

- Does textual sentiment in the MD&A of US REITs reveal managers' expectations regarding future firm performance?
- Does the market efficiently process the information conveyed through textual sentiment in the MD&A of US REITs?
- Are the domain-specific wordlists by Ruscheinsky et al. (2018) more appropriate to capture sentiment in the MD&A of US REITs?
- Does the market react more strongly to negative information in line with the negativity bias shown in large body of psychology literature?
- Do US REITs behave similarly to their industrial firm counterparts?

Paper 2 | Can Risks be Good News? Revealing Risk Perception of Real Estate Investors using Machine Learning

Paper 2 expands the analysis on annual reports and investigates the risk section. This section requires US REITs to specify risk factors that could adversely affect the company's business and operations. To identify the individual risk topics disclosed in 10-Ks, this study uses a machine-assisted topic detecting modeling, specifically the Structural Topic Model (STM), and investigates the validity of this approach to quantifying risk in narrative form. The research questions are as follows:

- Do the STM extracted risk factors help to explain the perceived risk in the stock market?
- How do the identified risk topics impact on investor risk perceptions? Are they risk-reducing, risk-increasing, or boilerplate?
- Is the risk-factor influence consistent with the efficient market hypothesis, mostly in the short-run, whereas fundamentals dominate the risk perception of investors in the long run?
- Can a hybrid model, combining machine-assisted topic modelling and a classic factor, namely the number of words a firm allocates towards a specific risk, more efficiently explain investor risk perceptions?
- Is the STM superior to more common approaches like Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM) in identifying risk factors revealed in risk disclosures of US REITs?
- Are the STM-identified risk factors capable of capturing changes in firms' reporting behavior during or after extreme events such as the global financial crisis (2007-2009)?

Paper 3 | Can Textual Analysis solve the Underpricing Puzzle? A US REIT Study

Paper 3 concentrates on Form S-11, the initial document for registering US REIT IPOs with the SEC. Form S-11 constitutes one of the most important documents during the IPO process which informs investors about the firms' business model, financial situation, potential problems or risks, and other important information. Recognizing that the vast majority of theories attribute underpricing to asymmetric information between the parties involved in the IPO, which can be reduced by both quantitative and qualitative information, this study investigates whether and how qualitative data in Form S-11 helps to explain initial-day returns of US REITs. The central research questions can be summarized as follows:

- Does the qualitative information revealed in Form S-11 help to explain the underpricing of US REIT IPOs?
- Does the language used in Form S-11 impact investor capability to price the issue?
- Do higher levels of uncertain language in Form S-11 make it more difficult for investors to price the issue and thus increase initial-day returns of US REIT IPOs?
- Does the level of uncertain language in Form S-11 serve as a proxy for *ex-ante* uncertainty?

- Does higher similarity to prior filings indicate that the filing contains a relatively high proportion of standardized content, but little useful information?
- Does higher similarity to prior filings suggest that information asymmetries persist, leading to increased underpricing?

1.3 Submissions and Conference Presentations

This section summarizes the authorship, submission process, and conference participation for each of the three papers comprising this dissertation.

Paper 1 | Is the MD&A of US REITs informative? A Textual Sentiment Study

Authors:

Marina Koelbl

Submission:

Journal:	Journal of Property Investment and Finance
Submission Date:	12/08/2019
Current Status:	published in Volume 38 Number 3

Conference participation:

The paper was presented at the Center of Finance Workshop "Artificial Intelligence and Finance" at the University of Regensburg in July 2019, at the 35th Annual Conference of the American Real Estate Society (ARES) in Paradise Valley, US, as well as at the 26th Annual Conference of the European Real Estate Society (ERES) in Cergy-Pontoise Cedex, France.

Paper 2 | Can Risks be Good News? Revealing Risk Perception of Real Estate Investors using Machine Learning

Authors:

Marina Koelbl, Ralf Laschinger, Bertram I. Steininger, and Wolfgang Schaefers

Submission:

Journal:	European Accounting Review
Submission Date:	03/01/2021
Status:	Under Review until 05/18/2021

Journal:	Contemporary Accounting Review
Submission Date:	11/12/2020
Status:	Under Review until 01/30/2021

Conference participation:

An earlier version of this paper was presented at the ARES Virtual Session 2020 and the IRES Symposium for Doctoral Students 2020, where the contribution earned the best presentation award. In 2021, the paper was presented at the AREUEA-ASSA Conference, the Cambridge Real Estate Seminar, the ERES Conference, and will be presented at the AREUEA-SINGAPORE Conference, and the annual conference of the "Verein für Socialpolitik" later this fall.

Paper 3 | Can Textual Analysis solve the Underpricing Puzzle? A US REIT Study

Authors:

Nino Paulus, Marina Koelbl, and Wolfgang Schaefers

Submission:

Journal:	Journal of Property Investment and Finance
Submission Date:	06/04/2021
Current Status:	Under Review

Conference participation:

The paper will be presented at the Center of Finance Workshop "Artificial Intelligence and Finance" at the University of Regensburg in July 2021.

1.4 References

Chen, H., De, P., Hu, Y. and Hwang, B.H. (2014), "Wisdom of crowds: The value of stock opinions transmitted through social media", *Review of Financial Studies*, Vol. 27 No. 5, pp. 1367–1403.

Davis, A.K. and Tama-Sweet, I. (2012), "Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A", *Contemporary Accounting Research*, Vol. 29 No. 3, pp. 804–837.

Fama, E.F. (1970), "Efficient Capital Markets: A Review of Theory and Empirical Work", *Journal of Finance*, Vol. 25 No. 2, pp. 383-417.

Feldman, R., Govindaraj, S., Livnat, J. and Segal, B. (2010), "Management's tone change, post earnings announcement drift and accruals", *Review of Accounting Studies*, Vol. 15 No. 4, pp. 915–953.

Ferris, S.P., Hao, Q. and Liao, M.Y. (2013), "The Effect of Issuer Conservatism on IPO Pricing and Performance", *Review of Finance*, Vol. 17 No. 3, pp. 993–1027.

Jegadeesh, N. and Wu, D. (2013), "Word power: A new approach for content analysis", *Journal of Financial Economics*, Vol. 110 No. 3, pp. 712–729.

Lehavy, R., Li, F. and Merkley, K. (2011), "The effect of annual report readability on analyst following and the properties of their earnings forecasts", *Accounting Review*, Vol. 86 No. 3, pp. 1087–1115.

Li, F. (2008), "Annual report readability, current earnings, and earnings persistence", *Journal of Accounting and Economics*, Vol. 45 No. 2–3, pp. 221–247.

Li, F. (2010), "The information content of forward- looking statements in corporate filings-A naïve bayesian machine learning approach", *Journal of Accounting Research*, Vol. 48 No. 5, pp. 1049–1102.

Loughran, T. and McDonald, B. (2011), "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks", *Journal of Finance*, Vol. 66 No. 1, pp. 35–65.

Loughran, T. and McDonald, B. (2014), "Measuring Readability in Financial Disclosures", *Journal of Finance*, Vol. 69 No. 4, pp. 1643–1671.

Loughran, T. and McDonald, B. (2015), "The Use of Word Lists in Textual Analysis", *Journal of Behavioral Finance*, Vol. 16 No. 1, pp. 1–11.

Miller, B.P. (2010), "The Effects of Reporting Complexity on Small and Large Investor Trading", *Accounting Review*, Vol. 85 No. 6, pp. 2107–2143.

Price, S.M., Seiler, M.J. and Shen, J. (2017), "Do Investors Infer Vocal Cues from CEOs During Quarterly REIT Conference Calls?", *Journal of Real Estate Finance and Economics*, Vol. 54 No. 4, pp. 515-557.

Ruscheinsky, J.R., Lang, M. and Schaefers, W. (2018), "Real estate media sentiment through textual analysis", *Journal of Property Investment and Finance*, Vol. 36 No. 5, pp. 410-428.

SEC. Securities and Exchange Commission (1998), Plain English Handbook, available at: https://www.plainlanguage.gov/resources/guides/plain-english-handbook/

2 Is the MD&A of US REITs informative? A Textual Sentiment Study

2.1 Abstract

Purpose: This study examines whether language disclosed in the Management Discussion and Analysis (MD&A) of US REITs provides signals regarding future firm performance and, thus generates a market response.

Design/methodology/approach: This research conducts textual analysis on a sample of approximately 6,500 MD&As of US REITs filed by the SEC between 2003 and 2018. Specifically, the Loughran and McDonald (2011) financial dictionary, and a custom dictionary for the real estate industry created by Ruscheinsky et al. (2018), are employed to determine the inherent sentiment i.e., the level of pessimistic or optimistic language for each filing. Thereafter, a panel fixed effects regression enables investigating the relationship between sentiment and future firm performance, as well as the markets' reaction.

Findings: The empirical results suggest that higher levels of pessimistic (optimistic) language in the MD&A predict lower (higher) future firm performance. Hereby, the use of a domain-specific real estate dictionary, namely that developed by Ruscheinsky et al. (2018) leads to superior results. Corresponding to the notion that the human psyche is affected more strongly by negative than positive news (Rozin and Royzman, 2001), the market responds solely to pessimistic language in the MD&A.

Practical implications: The results suggest that the market can benefit from textual analysis, as investigating the language in the MD&A reduces information asymmetries between US REIT managers and investors.

Originality/value: This is the first study to analyze exclusively for US REITs, whether language in the MD&A is predictive of future firm performance and whether the market responds to textual sentiment.

Keywords: US REITs, Sentiment, Textual analysis, Dictionary-based approach, MD&A, 10-K, 10-Q

2.2 Introduction

In 1968 the US Securities and Exchange Commission (SEC) introduced reporting requirements that mandate publicly traded firms to include a narrative disclosure, called Management Discussion and Analysis (MD&A), in their annual and quarterly reports. The justification for this requirement is as follows: "The Commission has long recognized the need for a narrative explanation of the financial statements, because a numerical presentation and brief accompanying footnotes alone may be insufficient for an investor to judge the quality of earnings and the likelihood that past performance is indicative of future performance. MD&A is intended to give the investor an opportunity to look at the company through the eyes of management by providing both a short and long-term analysis of the business of the company." (SEC, 1987).

Although the MD&A is clearly meant to inform investors, and specific SEC rules for the MD&A as well as subsequent SEC releases give detailed instructions and interpretive guidance to assist companies in preparing their MD&A disclosures (Huefner, 2007), the informativeness of the MD&A has frequently been criticized. For example, Pava and Epstein (1993) show that although most of the companies they study accurately describe historical events, very few provide useful and accurate forecasts in their MD&As. However, early research examining the market implications of qualitative disclosure relied on human coders making item-by-item subjective assessments of tone (e.g., Bryan, 1997; Barron et al., 1999; and Callahan and Smith, 2004). Recognizing the limitations of manual coding (e.g., small sample sizes, subjectivity), whether the MD&A is truly informative remains an open empirical question.

With the rise of behavioral finance in the last decade, this discussion was resumed and textual analysis has garnered increased attention, with the aim of assessing the informativeness of corporate disclosures. Thereby, researchers most frequently relied on textual tone analysis to examine firms' prospectuses (Feldman et al., 2010; Jegadeesh and Wu, 2013; Li, 2010; Loughran and McDonald, 2011 & 2015). Furthermore, most studies reviewed a random sample of firms, restricted only by the availability of necessary data. However, Callahan and Smith (2004) find evidence that the impact of language varies across industries.

This paper adds a new dimension to the discussion by analyzing MD&As for a sample of US REITs. In contrast to a sample randomly drawn from the capital market, the US REIT market provides a number of beneficial characteristics. First, equity REITs are fairly homogeneous regarding characteristics that usually vary widely across different industries (Hartzell et al., 2008). Second, US REITs are required to pay out a minimum of 90% of taxable earnings to shareholders as dividends. Consequently, in order to take advantage of growth opportunities, US REITs must turn to the capital markets. As such, US REIT managers have an unusually strong incentive to be transparent and maintain investor trust (Danielsen et al., 2009; Doran et al., 2012; Price et al., 2017). Third, the underlying assets of US REITs are real estate, which is an illiquid, slow-moving asset and thus more compatible to analysis over a relatively large time-span (e.g., from one quarter to the subsequent quarter).

These unique characteristics suggest that the MD&As of US REITs are particularly informative. However, the asset class's peculiarities also indicate that results from previous studies cannot automatically be extended to US REITs. Thus, we investigate the information content of the MD&A for a US REIT sample by answering the following questions: Does textual sentiment in the MD&A reveal managers' expectations regarding future firm performance? If so, does the market process the information efficiently?

To extract sentiment from the MD&A, we rely on a dictionary-based approach. Specifically, we employ the Loughran and McDonald (2011) financial dictionary and a custom wordlist for the real estate industry created by Ruscheinsky et al. (2018) to determine the overall sentiment inherent in each filing. Our findings suggest that higher levels of pessimistic (optimistic) language in the MD&A are associated with lower (higher) future firm performance. This holds even after controlling for the information released in other concurrent disclosures that may be predictive of future performance. Hereby, the use of a domain-specific real estate dictionary, namely the dictionary developed by Ruscheinsky et al. (2018) leads to superior results. Moreover, we find a significant market response to pessimistic language in the MD&A at the time of the SEC filing. However, corresponding to the notion that individuals are affected more strongly by negative than positive news, we cannot find a significant impact of optimistic language. Overall, to the best of our knowledge, this is the first study providing evidence that the use of language in the MD&A reveals US REIT managers' expectations regarding future firm performance and that the market responds to this information. We demonstrate that the market can benefit from textual analysis, as investigating the language in the MD&A should decrease information asymmetries between US REIT managers and investors.

The remainder of the paper is organized as follows. Section 2.3 discusses related literature. Section 2.4 introduces the data, that is, sample and variables. Section 2.5 defines the specific sentiment measures and presents empirical methods for the analysis. Finally, Section 2.6 reports the empirical results and Section 2.7 concludes.

2.3 Related Literature and Hypothesis Development

Textual analysis has recently attracted increased attention to address many pivotal questions in behavioral finance. Not least because in today's world a huge amount of information is stored as text instead of numeric data (Nasukawa and Nagano, 2001). Apart from annual and quarterly reports or merely the MD&A, researchers have focused on multiple disclosure outlets such as earnings press releases and earnings conference calls (Henry, 2008; Davis et al., 2012; Huang et al., 2014; Davis and Tama-Sweet, 2012), news articles (Tetlock, 2007, 2008; Engelberg et al., 2012; Ferguson et al., 2015; Garcia, 2013; Heston and Sinha, 2017), and internet postings (Antweiler and Frank, 2004; Das and Chen, 2007; Chen et al., 2014). Thereby, researchers have presented promising results in predicting the reactions of financial markets. Textual sentiment has been linked, for example, to abnormal stock returns (Feldman et al., 2010; Chen et al., 2014; Jegadeesh and Wu, 2013, Heston and Sinha, 2017), trading volume (Tetlock, 2007; Garcia, 2013), subsequent stock return volatility (Loughran and McDonald, 2011 & 2015), and future earnings and liquidity (Li, 2010).

Although textual analysis has received considerable attention in the accounting and finance literature, real estate still lags behind. Only recently, have researchers focused on textual sentiment in the context of the real estate market. For example, Doran et al. (2012) and Price et al. (2017) analyze transcripts and audio files of earnings conference calls for a REIT sample, to show that sentiment impacts initial reaction-window abnormal returns. In creating a custom dictionary for the real estate domain, Ruscheinsky et al. (2018) find media sentiment to lead future REIT market movements. Beracha et al. (2019) apply the newly developed wordlists to news abstracts of the Wall Street Journal and provide evidence that news-based sentiment has predictive power for the direct commercial real estate market in the US. In related literature, Hausler et al. (2018) link news-based sentiment to total returns of the US securitized and direct commercial real estate market. Most recently, Carstens and Freybote (2019) investigate how institutional REIT investors react to an 'abnormally' positive tone disclosed in annual and quarterly reports of US REITs. They find institutional REIT investors to respond positively (negatively) to an abnormally positive tone and behave as net buyers (net sellers) in periods of institutional REIT investor optimism (pessimism).

Interestingly, a substantial body of literature focuses on how investor sentiment affects the real estate market. Thereby, researchers usually rely on surveys (Das et al., 2015; Marcato and Nanda, 2016; Freybote, 2016) or market-based indicators, such as the closed-end fund discount (Lin et al., 2009), buy-sell imbalances (Freybote and Seagraves, 2017) or mortgage fund flows (Clayton et al., 2009; Ling et al., 2014) to measure investor sentiment. The major difference between investor sentiment and textual sentiment is that the former presents the subjective judgments and behavioral characteristics of investors, while the latter could include the former, but also contains the more objective reflection of conditions within firms, institutions and markets (Kearney and Liu, 2014). The connection between textual sentiment and investor sentiment is complex and the extent to which they are casually related has not yet been thoroughly examined or understood (Kearney and Liu, 2014). Nonetheless, prior research has linked investor sentiment, among others, to pricing and market returns in the commercial real estate market (Clayton et al., 2009; Ling et al., 2014), REIT returns (Lin et al., 2009; Das et al., 2015), REIT bond pricing (Freybote, 2016), as well as institutional investor trading behavior (Das et al., 2015). In sum, prior findings on investor sentiment indicate that real estate is sensitive to sentiment.

This paper conducts textual analysis on the MD&A section of annual and quarterly reports - a section that mandates managers to provide information on past performance, current financial positions, and future prospects. Since the structure and content of the MD&A are fixed, managers are "legally obligated to touch upon" subjects they would probably avoid in voluntary disclosure outlets like earnings press releases or conference calls (Buffett, 1998). In other words: analyzing a mandatory disclosure medium, namely the MD&A, guarantees a relatively high level of preparation, while still allowing managers considerable leeway on how to present the company's business, financial conditions, and results of operation. Assuming that executives report their views truthfully (under SEC scrutiny and penalty of litigation), it can be argued that statements released by management have high predictive ability.

Given the abovementioned features of the MD&A, and results of prior (investor sentiment) studies indicating that real estate is prone to sentiment, we expect the narrative to provide information regarding future US REIT performance. Against this background we test the the following hypothesis to assess the informativeness of the MD&A for a US REIT sample:

Hypothesis 1: *The proportion of total pessimistic (optimistic) language disclosed in the MD&A of US REITs is negatively (positively) associated with future firm performance.*

If management shares information about the firm's prospects with investors through disclosures in SEC filings, then market reaction should be associated with the nonfinancial information disclosed by management in the MD&A section. To assess whether the market accurately processes information delivered in the MD&A of US REITs, we test the following hypothesis:

Hypothesis 2: *Pessimistic (optimistic) language disclosed in the MD&A is negatively (positively) associated with market returns around the SEC filing date.*

2.4 Data

To answer our research questions, we use two types of dataset: (1) the text corpus given by corporate annual and quarterly reports, as well as (2) real estate return data, and firm fundamentals.

2.4.1 Management Discussion and Analysis

While management can use a variety of formal and informal methods to communicate with investors, we focus on annual and quarterly reports, more precisely on the MD&A. This section is particularly suitable for the analysis, because the SEC mandates publicly traded firms to express expectations regarding future firm performance in this section. Our initial sample contains all Equity REITs present in the FTSE NAREIT All REITs Index at any point of time between January 2003 and December 2018.[3,4] We start the analysis in 2003, because the SEC issued new guidelines for preparing MD&As, and the Sarbanes-Oxley Act increased disclosure requirements in this year. Furthermore, it is important to note that some firms remain in the index all through the sample period, whereas other enter, exit or both enter and exit. Based on NAREIT information, the number of index constituents increased from 144 in 2013 to 186 in 2018 (see Figure 2.1 in the Appendix). At present, leading players in terms of market capitalization are American Tower Corporation, Simon Property Group, Inc., Crown Castle International Corp., Prologis, Inc., and Public Storage (see Figure 2.2 in the Appendix). To access the corresponding filings electronically, we employ the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) database. Thereafter, the MD&A is extracted from the entire SEC filing by filtering for the item number. For annual reports, we capture the portion of the filing that is assigned to "Item 7.", whereas for quarterly reports one has to seek "Item 2.". This procedure yields a sample of almost 10,000 observations consisting of 268 unique firms.

[3] The Sarbanes-Oxley Act is a United States federal law that set new or expanded requirements for all US public company boards, management and public accounting firms.
[4] Mortgage REITs are excluded from the analysis because of their different characteristics.

2.4.2 REIT Return Data and Firm Fundamentals

For each filing in the sample, we further collect a number of accounting and financial market variables. All accounting and financial market data is derived from Compustat, CRSP, or I/B/E/S.[5][6]

Regarding the dependent variables for our tests of Hypothesis 1 and Hypothesis 2, we require measures of future firm performance and market returns centered on the SEC filing date. To measure future firm performance, we use the real-estate-specific measure funds from operations per share (*FFO/Share*) for the quarter after the company filing date. In contrast to common performance variables like *ROA* or *Tobin's Q*, which carry investment properties at historical cost, this metric takes into account the market value of assets. In accordance with NAREIT guidelines, the *FFO* is calculated as net income excluding gains or losses from sales of properties and adding back real estate depreciation.[7] To measure market response to language disclosed in the MD&A, we calculate cumulative abnormal returns (*CARs*) relative to the CRSP value-weighted market portfolio over the three-day window centered on the filing submission date (-1 day to +1 day).[8] We additionally investigate alternative time periods using *CAR*(-1,5), and *CAR*(-1,10), where the distinct reaction-windows are specified within parentheses.

To account for additional quantifiable information in SEC filings besides language, control variables like *Firmsize*, *REV*, *Leverage*, *SUE*, *MB*, *ROE*, *Liquidity*, *Volatility* are included in each regression. Those variables are likely to be associated with future firm performance and the market response around the filing submission date. We describe all control variables below, and provide specific definitions, including Compustat data items in Table 2.10 in the Appendix.

First, we measure the current quarter's standardized unexpected earnings (*SUE*) as the difference in *I/B/E/S* actual earnings per share and the most recent mean analyst forecast estimates, scaled by the stock price. However, when there are no analyst forecasts for the particular quarter, we follow Feldman et al. (2010) and calculate *SUE* as Compustat income before extraordinary items for the current quarter, minus income for the previous quarter, scaled by the market value of equity. We also measure the market-to-book value (*MB*) as the market value of equity scaled by the book value of equity. *SUE* and *MB* serve as control

[5] Center for Research on Securities Prices
[6] Institutional Brokers' Estimate System
[7] As a non US GAAP measure the FFO is rarely reported and definitions differ. Thus, no FFO data is available on Compustat before 2007 and later, many data points are missing.
[8] We also separately run our analysis using the CRSP equal-weighted index for computing abnormal returns. The untabulated results are robust to the variation.

variables for the information in other concurrent disclosures that may predict future performance. Additionally, we collect the current quarter's return on equity (*ROE*) that is calculated as income before extraordinary items scaled by the common equity. *ROE* is included to capture persistence in performance metrics. Finally, we include various firm-structure variables. The variables *Firmsize* and *REV* control for firm size. The variable *Firmsize* is defined as the logarithm of the firm's total assets, whereas *REV* represents the difference between current quarter sales and sales in the previous quarter. *Leverage* and *Volatility* are proxies for firm risk. *Leverage* is defined as total liabilities scaled by total assets, and *Volatility* is calculated as the standard deviation of daily returns over the most recent quarter. We also control for *Liquidity* using the ratio of traded shares to shares outstanding.

Unfortunately, not all accounting and financial market data is available for each observation in the sample. We eliminate any SEC filing for which we do not have the necessary data, which results in a final sample of roughly 6,500 observations.
Table 2.1 presents descriptive statistics of all variables.

Table 2.1: Descriptive Statistics

	Mean	Median	Maximum	Minimum	Std. Dev.
FFO/Share [$]	0.544	0.432	5.451	- 3.449	0.968
CAR(-1,1) [$]	- 0.001	0.000	0.106	- 0.117	0.033
CAR(-1,5) [$]	- 0.003	- 0.003	0.166	- 0.200	0.055
CAR(-1,10) [$]	- 0.003	- 0.002	0.139	- 0.163	0.045
Firmsize [mm $]	3.360	3.423	4.442	1.332	0.579
REV [mm $]	3.180	1.084	121.768	- 112.235	27.219
Leverage	0.575	0.574	1.111	0.044	0.168
SUE	0.000	0.000	0.220	- 0.266	0.036
MB	2.544	1.968	17.579	- 6.859	2.832
ROE	0.014	0.014	0.285	- 0.263	0.054
Liquidity	0.459	0.388	1.829	0.000	0.325
Volatility [$]	6.132	2.447	50.692	0.071	9.310

2.5 Methodology

2.5.1 Dictionary-based Approach

The pioneering work of Tetlock (2007) is arguably among the first applications of the dictionary-based approach in the financial domain. Examining news articles from *The Wall Street Journal*, Tetlock emphasizes that high values of media pessimism induce downward pressure on market prices, and that unusually high or low pessimism predicts temporarily high market trading volumes. Subsequently, several papers followed his example and relied on dictionary-based approaches to extract qualitative information from text. Thereby, researchers frequently employed general English language dictionaries such as the Diction or the Harvard GI wordlists (Tetlock, 2007; Tetlock et al., 2008; Davis et al., 2012; Davis and Tama-Sweet, 2012). However, Loughran and McDonald (2016) argue that most researchers used only the GI and Diction wordlists, because these lists were the first ones readily available. Henry and Leone (2016) claim that general dictionaries tend to lack predictive power in a financial context, since the language used in the accounting and finance domain tends to be very specific. To overcome this issue and to enable the generation of accurate and efficient sentiment scores, researchers developed dictionaries specific to the finance domain. The first of these is the Henry wordlist (Henry, 2008). However, the 85 negative words included in Henry's dictionary are sparse to gauge sentiment efficiently (Loughran and McDonald, 2016). Subsequently, Loughran and McDonald (2011) created a dictionary specifically for financial text by examining word usage in 10-Ks (thereafter referred to as LM dictionary). In contrast to the Henry list, this dictionary is quite extensive and contains 354 positive and 2,355 negative words.[9] Only recently, Ruscheinsky et al. (2018) adjusted the LM dictionary for their investigation on the impact of media-sentiment in real estate markets. Comparing headlines from *Bloomberg*, *The Financial Times*, *Forbes* and *The Wall Street Journal* to the wordlists of the LM dictionary, the researchers added words that are considered positive or negative in the context of real estate, and removed other words that have a rather different or unclear classification. In total, 197 words were added to the wordlists, whereas 43 words were deleted. Overall, the real estate dictionary contains 408 positive and 2,455 negative words. Indeed, recent studies recognize the researchers' efforts and confirm the superiority of domain-specific wordlists (Rogers et al., 2011; Doran et al., 2012; Henry and Leone, 2016).

In sum, both, the LM dictionary as well as the custom wordlists of Ruscheinsky et al. (2018) are derived from the word use in corporate filings, and thus match the subject of our textual corpus very well. While the first dictionary focusses solely on the language generally used in financial texts, the latter takes into account terminology specific to the real estate industry. For example, words like 'bubble', 'crash', or 'depression' that clearly convey sentiment in the context of real estate are included in the adjusted wordlists of Ruscheinsky et al. (2018), although missing from the original LM wordlists. Therefore, this paper employs both dictionaries: (1) the financial dictionary of Loughran and McDonald (2011) and (2) the real estate-specific dictionary of Ruscheinsky et al. (2018).

[9] Revised version currently available at: https://sraf.nd.edu/textual-analysis/resources/

2.5.2 Sentiment Measures

To extract sentiment from the textual corpus, i.e., the MD&A, we employ the abovementioned dictionaries, to categorize all words in the narrative into either positive or negative. However, in order to facilitate the procedure and to reduce linguistic complexity, firm disclosures are pre-processed. In this respect, punctuation, numbers, symbols, and stopwords are removed. The remaining text is organized in a document-term matrix (DTM) in order to visualize the occurrence of each word in the textual corpus. The rows represent the documents or filings, columns are the words appearing in the document and cells are designed to capture how often each term was used in the document. Finally, applying either dictionary to the matrix allows easy computation of the number of positive and negative words in each filing. This raw count of words is inappropriate as a measure of tone however, because simple frequencies obviously depend on document length. To solve this problem, we use the ratio of negative (positive) words to the total number of words in the MD&A to assess the disclosure tone:

$$NEG = \frac{Negative}{Total} \tag{2.1}$$

$$POS = \frac{Positive}{Total} \tag{2.2}$$

Both *NEG* and *POS* range from 0 to 1, whereby a higher value indicates a greater level of pessimism or optimism, respectively.

Additionally, we follow Henry and Leone (2016) and employ a relative measure, which accounts for positive as well as negative language in the narrative:

$$Tone = \frac{(Positive - Negative)}{(Positive + Negative)} \tag{2.3}$$

By construction, *Tone* is bound between –1 and 1. Hence, the variable allows deciding rapidly, whether a text is relatively positive (if greater than 0) or relatively negative (if less than 0).

All measures of language are calculated from the LM dictionary (NEG_{LM}, POS_{LM}, $Tone_{LM}$) and the dictionary created by Ruscheinsky et al. (2018) (NEG_{RE}, POS_{RE}, $Tone_{RE}$). Table 2.2 displays the descriptive statistics of all six sentiment measures.

Table 2.2: Descriptive Statistics of Sentiment Measures

	Mean	Median	Maximum	Minimum	Std. Dev.
NEG_{LM}	0.017	0.016	0.038	0.006	0.006
POS_{LM}	0.010	0.010	0.019	0.004	0.003
$Tone_{LM}$	- 0.224	- 0.237	0.344	- 0.674	0.208
NEG_{RE}	0.018	0.017	0.040	0.006	0.006
POS_{RE}	0.013	0.013	0.025	0.005	0.004
$Tone_{RE}$	- 0.131	- 0.138	0.472	- 0.660	0.233

2.5.3 Model Specification

To test the ability of our measures of language to predict future firm performance, defined as the *FFO/Share* in the quarter subsequent to the current quarter (Hypothesis 1), we estimate the following fixed-effects regression model for all six measures of tone:

$$FFO/Share_{t+90} = \beta_0 + \beta_1 sentiment\ measure_{it} + \beta_2 controls_{it} + a_i + \lambda_t + u_{it} \qquad (2.4)$$

In this model, i denotes the firm, and t the period. In addition to the measure of language, the regression equation includes a vector of control variables consisting of firm characteristics and previous market variables such as *Firmsize*, *REV*, *Leverage*, *SUE*, *MB*, *ROE*, *Liquidity*, and *Volatility*. All control variables are defined in Table 2.10 in the Appendix. The regression variables a_i and λ_t indicate that we account for unobserved firm and time effects. u_{it} represents the error term. To account for heteroscedasticity and serial correlation, the models are estimated using white robust standard errors.

Hypothesis 2 predicts that language in the MD&A is associated with market returns around the filing submission date. We regress *CAR* (i.e., the cumulative abnormal return relative to the CRSP value-weighted market portfolio) over the distinct reaction-windows (*CAR*(-1,1), *CAR*(-1,5), *CAR*(-1,10)) on our measures of sentiment, so as to gauge the incremental market response to language. Again, the model is estimated employing a fixed-effects regression with robust standard errors:

$$CAR = \beta_0 + \beta_1 sentiment\ measure_{it} + \beta_2 controls_{it} + a_i + \lambda_t + u_{it} \qquad (2.5)$$

Apart from a vector of control variables that includes firm characteristics as well as variables that might have information content for predicting future market returns, we include unobservable firm and time effects.

2.6 Results

2.6.1 Main Regression Results

Future Firm Performance

The results of the panel fixed-effects regression regarding the association of language in the MD&A and future firm performance (Hypothesis 1) are displayed in Table 2.3. Regressions are run for all six measures of language (*NEG_{LM}*, *POS_{LM}*, *$Tone_{LM}$*, *NEG_{RE}*, *POS_{RE}*, *$Tone_{RE}$*).

Table 2.3: Future Firm Performance - FFO/Share

	Loughran and McDonald (2011)			**Ruscheinsky et al. (2018)**		
Intercept	- 0.708*	- 0.960**	- 0.972**	- 0.691*	- 1.089**	- 0.788**
	(0.079)	(0.027)	(0.017)	(0.079)	(0.010)	(0.044)
Firmsize	0.375***	0.422***	0.407***	0.372***	0.397***	0.359***
	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)
REV	0.001	0.001	0.001	0.001	0.001	0.001
	(0.513)	(0.486)	(0.484)	(0.510)	(0.498)	(0.525)
Leverage	0.279*	0.284*	-0.306*	0.280*	0.305*	0.299*
	(0.098)	(0.092)	(0.079)	(0.097)	(0.079)	(0.082)
SUE	- 0.020	0.053	0.042	0.020	0.045	0.038
	(0.953)	(0.879)	(0.905)	(0.954)	(0.897)	(0.913)
MB	- 0.001	0.000	0.000	- 0.001	0.000	0.000
	(0.928)	(0.998)	(1.000)	(0.917)	(0.967)	(0.951)
ROE	0.172	0.236	0.214	0.175	0.222	0.179
	(0.445)	(0.310)	(0.352)	(0.439)	(0.344)	(0.431)
Liquidity	- 0.113	- 0.129*	- 0.125*	- 0.114	- 0.124*	- 0.109
	(0.128)	(0.089)	(0.099)	(0.124)	(0.098)	(0.137)
Volatility	0.006**	0.006**	0.006**	0.006**	0.006**	0.006**
	(0.028)	(0.023)	(0.027)	(0.029)	(0.021)	(0.030)
NEG_{LM}	- 9.530***					
	(0.004)					
POS_{LM}		- 6.360				
		(0.315)				
$Tone_{LM}$			0.096			
			(0.403)			
NEG_{RE}				- 9.432***		
				(0.002)		
POS_{RE}					9.698**	
					(0.025)	
$Tone_{RE}$						0.297***
						(0.001)
N	6,648	6,648	6,648	6,648	6,648	6,648
R^2	0.321	0.319	0.319	0.321	0.320	0.321
Adj. R^2	0.292	0.291	0.291	0.292	0.291	0.293

see next page

Table 2.3: continued

White robust p-values are presented in parentheses. *, **, and *** denote statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively, based on a two-tailed t-test. For complete definitions, see Table 2.10 in the Appendix.

Concerning the sentiment measures derived from the LM dictionary, we find a negative and significant coefficient on the measure of negative language (NEG_{LM}), suggesting that pessimistic language disclosed in the MD&A provides information about future performance. However, we cannot find a significant impact of the remaining two sentiment measures. The coefficients on both the measure of optimistic language (POS_{LM}) as well as on the tone measure ($Tone_{LM}$), are statistically insignificant at common levels.

In contrast, we find statistically significant coefficients on all three measures of language derived from the real-estate-specific dictionary of Ruscheinsky et al. (2018). Thus, the measure of negative language (NEG_{RE}), the measure of positive language (POS_{RE}) as well as the tone measure ($Tone_{RE}$) significantly predict future firm performance, and all coefficients have the expected sign. While pessimistic language in the MD&A negatively affects future firm performance, the opposite is true for optimistic language.[10]

Since the measure of positive language, and the tone measure derived from the LM dictionary, both lack statistical significance, we find the dictionary of Ruscheinsky et al. (2018) to be more appropriate for capturing sentiment in the MD&A of US REITs. This is consistent with recent studies by Rogers et al. (2011), Doran et al. (2012), and Henry and Leone (2016) who provide evidence of the superiority of domain-specific wordlists. Although both the LM wordlists as well as the dictionary developed by Ruscheinsky et al. (2018) are suitable for a financial context, the latter benefits from the researchers' efforts to capture terminology specific to the real estate industry.

Market Response

Next, we examine whether language in the MD&A is related to market returns around the filing submission date (Hypothesis 2). Tables 2.4 to 2.6 present the regression results for the alternative reaction-windows (i.e., *CAR*(-1,1), *CAR*(-1,5), *CAR*(-1,10)).

[10] In unreported results, we find that sentiment affects future firm performance up to one year. This corresponds to a sluggish real estate market with inherently illiquid assets and long transaction periods.

Table 2.4: Market Response CAR(-1,1)

	Loughran and McDonald (2011)			Ruscheinsky et al. (2018)		
Intercept	0.007	- 0.001	0.000	0.007	- 0.002	0.000
	(0.394)	(0.858)	(0.973)	(0.348)	(0.835)	(0.988)
Firmsize	- 0.001	0.000	0.000	- 0.001	0.000	0.000
	(0.682)	(0.898)	(0.943)	(0.661)	(0.849)	(0.904)
REV	0.000	0.000	0.000	0.000	0.000	0.000
	(0.253)	(0.221)	(0.224)	(0.251)	(0.211)	(0.231)
Leverage	0.002	0.002	0.003	0.002	0.002	0.002
	(0.626)	(0.657)	(0.584)	(0.635)	(0.652)	(0.615)
SUE	- 0.015	- 0.014	- 0.014	- 0.015	- 0.014	- 0.014
	(0.357)	(0.388)	(0.373)	(0.361)	(0.387)	(0.378)
MB	0.000	0.000	0.000	0.000	0.000	0.000
	(0.250)	(0.304)	(0.301)	(0.244)	(0.309)	(0.297)
ROE	- 0.005	- 0.003	- 0.004	- 0.005	- 0.003	- 0.003
	(0.472)	(0.700)	(0.600)	(0.469)	(0.702)	(0.613)
Liquidity	- 0.001	- 0.001	- 0.001	- 0.001	- 0.001	- 0.001
	(0.811)	(0.683)	(0.723)	(0.803)	(0.670)	(0.731)
Volatility	0.000	0.000	0.000	0.000	0.000	0.000
	(0.759)	(0.659)	(0.699)	(0.770)	(0.658)	(0.696)
NEG_{LM}	- 0.284***					
	(0.004)					
POS_{LM}		- 0.082				
		(0.715)				
$Tone_{LM}$			0.003			
			(0.333)			
NEG_{RE}				- 0.288***		
				(0.002)		
POS_{RE}					- 0.086	
					(0.637)	
$Tone_{RE}$						0.003
						(0.347)
N	6,586	6,586	6,586	6,586	6,586	6,586
R^2	0.104	0.103	0.103	0.105	0.103	0.103
Adj. R^2	0.066	0.065	0.065	0.066	0.065	0.065

White robust p-values are presented in parentheses. *, **, and *** denote statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively, based on a two-tailed t-test. For complete definitions, see Table 2.10 in the Appendix.

Table 2.5: Market Response CAR(-1,5)

	Loughran and McDonald (2011)			Ruscheinsky et al. (2018)		
Intercept	0.016	0.005	0.008	0.016	0.006	0.008
	(0.215)	(0.720)	(0.531)	(0.208)	(0.659)	(0.523)
Firmsize	- 0.004	- 0.003	- 0.003	- 0.004	- 0.003	- 0.003
	(0.246)	(0.458)	(0.368)	(0.244)	(0.483)	(0.347)
REV	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*
	(0.094)	(0.080)	(0.084)	(0.093)	(0.080)	(0.088)

see next page

Table 2.5: continued

Leverage	0.004	0.004	0.004	0.004	0.004	0.004
	(0.626)	(0.624)	(0.595)	(0.631)	(0.640)	(0.618)
SUE	0.019	0.020	0.019	0.019	0.020	0.020
	(0.425)	(0.411)	(0.414)	(0.421)	(0.407)	(0.410)
MB	0.000	0.000	0.000	0.000	0.000	0.000
	(0.586)	(0.674)	(0.657)	(0.581)	(0.664)	(0.651)
ROE	- 0.011	- 0.008	- 0.009	- 0.011	- 0.008	- 0.009
	(0.282)	(0.408)	(0.361)	(0.283)	(0.417)	(0.370)
Liquidity	- 0.002	- 0.002	- 0.002	- 0.002	- 0.002	- 0.002
	(0.707)	(0.612)	(0.640)	(0.698)	(0.605)	(0.642)
Volatility	0.000	0.000	0.000	0.000	0.000	0.000
	(0.865)	(0.969)	(0.924)	(0.856)	(0.964)	(0.927)
NEG_{LM}	- 0.333**					
	(0.026)					
POS_{LM}		0.063				
		(0.835)				
$Tone_{LM}$			0.004			
			(0.393)			
NEG_{RE}				- 0.326**		
				(0.022)		
POS_{RE}					- 0.059	
					(0.810)	
$Tone_{RE}$						0.003
						(0.430)
N	6,586	6,586	6,586	6,586	6,586	6,586
R^2	0.128	0.128	0.128	0.128	0.128	0.128
Adj. R^2	0.091	0.090	0.090	0.091	0.090	0.090

White robust p-values are presented in parentheses. *, **, and *** denote statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively, based on a two-tailed t-test. For complete definitions, see Table 2.10 in the Appendix.

Table 2.6: Market Response CAR(-1,10)

	Loughran and McDonald (2011)			**Ruscheinsky et al. (2018)**		
Intercept	0.010	0.006	0.008	0.009	0.009	0.005
	(0.484)	(0.656)	(0.546)	(0.515)	(0.506)	(0.741)
Firmsize	- 0.004	- 0.004	- 0.004	- 0.004	- 0.003	- 0.003
	(0.254)	(0.314)	(0.278)	(0.269)	(0.466)	(0.409)
REV	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
	(0.048)	(0.045)	(0.047)	(0.047)	(0.042)	(0.044)
Leverage	0.002	0.002	0.002	0.002	0.002	0.002
	(0.813)	(0.803)	(0.797)	(0.814)	(0.855)	(0.826)
SUE	- 0.018	- 0.018	- 0.018	- 0.018	- 0.017	- 0.018
	(0.670)	(0.673)	(0.672)	(0.672)	(0.684)	(0.677)
MB	0.000	0.000	0.000	0.000	0.000	0.000
	(0.543)	(0.563)	(0.555)	(0.547)	(0.545)	(0.569)
ROE	0.000	0.000	0.000	0.000	0.001	0.001
	(0.992)	(0.974)	(0.992)	(0.997)	(0.944)	(0.929)

see next page

Table 2.6: continued

Liquidity	0.003	0.003	0.003	0.003	0.002	0.002
	(0.633)	(0.656)	(0.643)	(0.641)	(0.682)	(0.681)
Volatility	0.000	0.000	0.000	0.000	0.000	0.000
	(0.829)	(0.861)	(0.835)	(0.836)	(0.843)	(0.895)
NEG_{LM}	- 0.094					
	(0.601)					
POS_{LM}		0.085				
		(0.804)				
$Tone_{LM}$			0.002			
			(0.649)			
NEG_{RE}				- 0.067		
				(0.686)		
POS_{RE}					- 0.374	
					(0.182)	
$Tone_{RE}$						- 0.003
						(0.441)
N	6,586	6,586	6,586	6,586	6,586	6,586
R^2	0.135	0.135	0.135	0.135	0.136	0.135
Adj. R^2	0.098	0.098	0.098	0.098	0.099	0.099

White robust p-values are presented in parentheses. *, **, and *** denote statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively, based on a two-tailed t-test. For complete definitions, see Table 2.10 in the Appendix.

We find a negative and significant coefficient on NEG_{LM} and NEG_{RE} for the three-day and the six-day windows, respectively. This suggests that pessimistic language in the MD&A is associated with lower abnormal returns at the time a firm files its 10-K or 10-Q with the SEC. However, the coefficient becomes insignificant during the *CAR*(-1,10) window, indicating that the impact of language is fully incorporated into the stock price by that time. In contrast to the measures of pessimistic language (NEG_{LM}, NEG_{RE}), we cannot find a significant impact of either optimistic language (POS_{LM}, POS_{RE}) or the tone scores ($Tone_{LM}$, $Tone_{RE}$) on market returns.

Given that results on the *FFO/Share* reveal that managers disclose positive as well as negative expectations regarding future firm performance, the asymmetrical relation confirms the notion of an existing negativity bias. A large body of psychology literature shows that negative information has more impact and is more thoroughly processed in a number of contexts (Baumeister et al., 2001; Rozin and Royzman, 2001).

Overall, our findings indicate that US REITs behave similarly to their industrial firm counterparts in some cases and differently in others. On the one hand, prior research in accounting and finance links language disclosed in the MD&A to future performance (Li, 2010; Davis and Tama-Sweet, 2012). On the other hand, analyzing a random sample of firms, Davis and Tama-Sweet (2012) state that tone in corporate disclosures has no impact on abnormal returns. This suggests that REIT investors more efficiently process information disclosed by firm executives. Accordingly, Hausler et al. (2018) argue that real estate markets are especially prone to sentiment because of their characteristics such as low transparency, information asymmetry, illiquidity, and long transaction periods.

2.6.2 Robustness

Performance Measures

In addition to the real-estate-specific performance measure *FFO/Share*, we test our hypothesis employing the return on assets (*ROA*). This measure is used frequently as an independent performance variable, but seems to be less applicable to the REIT industry, as the variable does not reflect the market value of investment properties. Nonetheless, as shown in Table 2.7, the empirical results are consistent with our findings based on the *FFO/Share*. More precisely, our results confirm that higher levels of pessimistic (optimistic) language in the MD&A predict lower (higher) firm performance in the quarter after the filing submission date. Again, the use of the dictionary from Ruscheinsky et al. (2018) improves the predictive power of the sentiment measures.

Table 2.7: Future Firm Performance - ROA

	Loughran and McDonald (2011)			**Ruscheinsky et al. (2018)**		
Intercept	0.016*	0.009	0.012	0.015	0.008	0.014
	(0.095)	(0.390)	(0.232)	(0.136)	(0.421)	(0.143)
Firmsize	- 0.002	- 0.001	- 0.001	- 0.001	- 0.001	- 0.002
	(0.542)	(0.813)	(0.650)	(0.595)	(0.680)	(0.472)
REV	0.000	0.000	0.000	0.000	0.000	0.000
	(0.644)	(0.536)	(0.573)	(0.614)	(0.602)	(0.673)
Leverage	- 0.003	- 0.003	- 0.002	- 0.003	- 0.003	- 0.003
	(0.429)	(0.504)	(0.557)	(0.447)	(0.511)	(0.492)
SUE	0.007	0.007*	0.007	0.007	0.007*	0.007*
	(0.108)	(0.095)	(0.102)	(0.104)	(0.095)	(0.094)
MB	0.000	0.000	0.000	0.000	0.000	- 0.000
	(0.818)	(0.973)	(0.909)	(0.840)	(0.993)	(0.865)
ROE	0.010***	0.012***	0.011***	0.011***	0.012***	0.011***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Liquidity	- 0.002**	- 0.003**	- 0.002**	- 0.002**	- 0.002**	- 0.002**
	(0.038)	(0.024)	(0.032)	(0.033)	(0.028)	(0.043)
Volatility	0.000	0.000	0.000	0.000	0.000	0.000
	(0.633)	(0.539)	(0.628)	(0.622)	(0.516)	(0.658)
NEG_{LM}	- 0.199***					
	(0.000)					
POS_{LM}		0.047				
		(0.619)				

see next page

Table 2.7: continued

$Tone_{LM}$			0.004***			
			(0.002)			
NEG_{RE}				- 0.152***		
				(0.002)		
POS_{RE}					0.191***	
					(0.002)	
$Tone_{RE}$						0.006***
						(0.000)
N	6,649	6,649	6,649	6,649	6,649	6,649
R^2	0.447	0.442	0.445	0.445	0.444	0.448
Adj. R^2	0.424	0.419	0.422	0.422	0.421	0.424

White robust p-values are presented in parentheses. *, **, and *** denote statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively, based on a two-tailed t-test. For complete definitions, see Table 2.10 in the Appendix.

Financial crisis

We further compare observations that occurred during the financial crisis of 2007/2008 to a sample excluding the crisis. For the subsample representing the financial crisis, neither of the sentiment measures is statistically significant at common levels (Table 2.8). This is not surprising, since all measures of language are derived from the MD&A. The narrative is authored by firm executives and as such depends substantially on their truthfulness and expectations regarding future performance. If executives are uncertain about future performance, measures of language suffer predictive power. Especially after the enormous decrease in values and firm results at the beginning of the financial crisis, executives were extremely insecure about future performance. In contrast, the subsample excluding the crisis supports the predictive power of textual sentiment (Table 2.9). On balance, regression coefficients can be interpreted reliably as the true casual effects of language on performance. However, language in the MD&A is less accurate as a predictor of future performance during recessions.

Table 2.8: Future Firm Performance - Financial Crisis of 07/08

	Loughran and McDonald (2011)			**Ruscheinsky et al. (2018)**		
Intercept	- 5.224*	- 5.686*	- 5.884*	- 5.129*	- 5.765*	- 5.664*
	(0.096)	(0.072)	(0.071)	(0.098)	(0.073)	(0.083)
Firmsize	1.093	1.324	1.264	1.073	1.260	1.200
	(0.257)	(0.185)	(0.210)	(0.262)	(0.217)	(0.231)
REV	0.001	0.001	0.001	0.001	0.001	0.001
	(0.375)	(0.431)	(0.383)	(0.369)	(0.387)	(0.378)
Leverage	3.191***	3.128***	3.113***	3.205***	3.178***	3.158***
	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)
SUE	- 0.252	- 0.206	- 0.113	- 0.277	- 0.181	- 0.155
	(0.774)	(0.809)	(0.894)	(0.752)	(0.836)	(0.856)

see next page

Table 2.8: continued

MB	0.007	0.008	0.008	0.007	0.008	0.008
	(0.768)	(0.758)	(0.762)	(0.770)	(0.763)	(0.766)
ROE	- 0.183	- 0.162	- 0.178	- 0.182	- 0.194	- 0.186
	(0.696)	(0.729)	(0.701)	(0.697)	(0.683)	(0.692)
Liquidity	0.154	0.154	0.140	0.157	0.144	0.143
	(0.453)	(0.450)	(0.494)	(0.442)	(0.478)	(0.487)
Volatility	0.035**	0.035**	0.035**	0.035**	0.035**	0.035**
	(0.014)	(0.016)	(0.013)	(0.014)	(0.014)	(0.015)
NEG_{LM}	- 6.499					
	(0.699)					
POS_{LM}		- 34.722				
		(0.143)				
$Tone_{LM}$			- 0.220			
			(0.644)			
NEG_{RE}				- 8.248		
				(0.592)		
POS_{RE}					- 8.082	
					(0.571)	
$Tone_{RE}$						- 0.063
						(0.875)
N	806	806	806	806	806	806
R^2	0.427	0.430	0.427	0.427	0.427	0.427
Adj. R^2	0.321	0.324	0.321	0.321	0.320	0.320

White robust p-values are presented in parentheses. *, **, and *** denote statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively, based on a two-tailed t-test. For complete definitions, see Table 2.10 in the Appendix.

Table 2.9: Future Firm Performance - without Financial Crisis of 07/08

	Loughran and McDonald (2011)			**Ruscheinsky et al. (2018)**		
Intercept	- 0.570	- 0.813*	- 0.808*	- 0.557	- 0.927*	- 0.623
	(0.209)	(0.093)	(0.079)	(0.209)	(0.051)	(0.161)
Firmsize	0.358***	0.399***	0.385***	0.356***	0.373***	0.337***
	(0.004)	(0.002)	(0.002)	(0.004)	(0.003)	(0.005)
REV	0.000	0.000	0.000	0.000	0.000	0.000
	(0.665)	(0.632)	(0.634)	(0.664)	(0.659)	(0.690)
Leverage	0.181	0.183	0.202	0.181	0.198	0.198
	(0.356)	(0.350)	(0.314)	(0.358)	(0.322)	(0.321)
SUE	- 0.019	0.000	- 0.008	- 0.019	- 0.006	- 0.006
	(0.960)	(1.000)	(0.984)	(0.960)	(0.988)	(0.989)
MB	- 0.003	- 0.003	- 0.003	- 0.003	- 0.002	- 0.003
	(0.759)	(0.809)	(0.807)	(0.752)	(0.831)	(0.770)
ROE	0.190	0.256	0.234	0.194	0.241	0.190
	(0.489)	(0.366)	(0.402)	(0.483)	(0.395)	(0.491)
Liquidity	- 0.183**	- 0.196**	- 0.193**	- 0.184**	- 0.192**	- 0.178**
	(0.019)	(0.014)	(0.016)	(0.018)	(0.015)	(0.021)
Volatility	0.005*	0.005*	0.005*	0.005*	0.005*	0.005*
	(0.090)	(0.074)	(0.083)	(0.093)	(0.065)	(0.091)

see next page

Table 2.9: continued

NEG_{LM}	- 9.122***					
	(0.004)					
POS_{LM}		- 4.688				
		(0.481)				
$Tone_{LM}$			0.096			
			(0.418)			
NEG_{RE}				-8.901***		
				(0.004)		
POS_{RE}					10.406**	
					(0.035)	
$Tone_{RE}$						0.312***
						(0.001)
N	5,842	5,842	5,842	5,842	5,842	5,842
R^2	0.319	0.318	0.318	0.319	0.318	0.320
Adj. R^2	0.287	0.286	0.286	0.287	0.287	0.288

White robust p-values are presented in parentheses. *, **, and *** denote statistical significance at the 10 percent, 5 percent, and 1 percent levels, respectively, based on a two-tailed t-test. For complete definitions, see Table 2.10 in the Appendix.

2.7 Conclusion

The SEC instructs firm executives to signal expectations regarding future firm performance in the MD&A. However, it is uncertain whether the narrative provides enough information to enable investors "to look at the company through the eyes of management" (SEC, 1987). We follow a recent stream of literature in behavioral finance and address this issue by employing techniques from other research areas such as computer science and computational linguistics. Specifically, we extract textual sentiment from the narrative to investigate whether language disclosed in the MD&A predicts future firm performance and generates a market response. In so doing, we limit the analysis to a sample of US REITs, which provides a number of beneficial characteristics compared to a sample drawn randomly from the capital market (e.g., industry jargon, homogenous firm characteristics, truthful managers, and slow-moving assets). We use the LM financial dictionary and a custom dictionary for the real estate industry created by Ruscheinsky et al. (2018), to determine the overall sentiment inherent in each filing.

Confirming the informativeness of the MD&A, we find that higher levels of pessimistic (optimistic) language in the MD&A are indeed associated with lower (higher) future firm performance. This holds even after controlling for the information released in other concurrent disclosures that may predict future performance. In accordance with prior literature from Rogers et al. (2011), Doran et al. (2012), and Henry and Leone (2016), the use of a domain-specific real estate dictionary, namely that developed by Ruscheinsky et al. (2018), leads to superior results. Furthermore, we find a significant market response to pessimistic language in the MD&A at the time of the SEC filing. However, we do not find a significant impact of optimistic language. Overall, the empirical results suggest that managers deliver expectations about future performance, although corresponding to the notion that individuals are affected more strongly by negative news, the market does not process optimistic information efficiently.

Our results suggest that the market can benefit from textual analysis, as investigating the language in the MD&A reduces information asymmetries between US REIT managers and investors. Nevertheless, one needs to be cautious when applying sentiment analysis. Employing the dictionary-based approach, researchers typically collapse a document down to a document-term matrix to generate sentiment measures. As a result, the direct context of a word is ignored, which discards all information that can be gained from word sequences. Thus, combinations of words or phrases that might imply different meanings from the constituent words are disregarded. A more recent stream of literature overcomes this issue by utilizing machine-learning approaches such as N-grams, support vector machines or naïve Bayes to classify the content of documents based on statistical inference (Li, 2010; Hausler et al., 2018). Hence, using a machine-learning approach to conduct the analysis would be worthwhile. Moreover, Davis and Tama-Sweet (2012) provide evidence that firm executives use the various available disclosure media to report strategically, or even use language deliberately to mislead investors (Huang et al., 2014). Therefore, it is advisable for future research to investigate multiple disclosure media (i.e., corporate annual or quarterly reports and the corresponding earnings press releases and conference calls) or more spontaneous indicators like pictures, gestures, and facial expressions to outwit executives.

2.8 References

Antweiler, W. and Frank, M.Z. (2004), "Is all that talk just noise? The information content of Internet stock message boards", *Journal of Finance*, Vol. 59 No. 3, pp. 1259–1294.

Barron, O.E., Kile, C.O. and O'Keefe, T.B. (1999), "MD&A Quality as Measured by the SEC and Analysts' Earnings Forecasts", *Contemporary Accounting Research*, Vol. 16 No. 1, pp. 75–109.

Baumeister, R.F., Bratslavsky, E., Finkenauer, C. and Vohs, K.D. (2001), "Bad Is Stronger Than Good", *Review of General Psychology*, Vol. 5 No. 4, pp. 323–370.

Beracha, E., Lang, M., and Hausler, J. (2019), "On the Relationship between Market Sentiment and Commercial Real Estate Performance—A Textual Analysis Examination", *Journal of Real Estate Research*, Vol. 41 No. 4, pp. 605–637.

Bryan, S.H. (1997), "Incremental information content of required disclosures contained in management discussion and analysis", *Accounting Review*, Vol. 72 No. 2, pp. 285–301.

Buffet, W.E. (1998), in "A Plain English Handbook - How to create clear SEC disclosure documents", SEC. Securities and Exchange Commission, available at: https://www.sec.gov/pdf/handbook.pdf.

Callahan, C.M. and Smith, R.E. (2004), "Firm performance and management's discussion and analysis disclosures: An industry approach", available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=588062.

Carstens, R. and Freybote, J. (2019), "Tone in REIT financial statements and institutional investments", *Journal of Property Research*, Vol. 36 No. 3, pp. 227–244.

Chen, H., De, P., Hu, Y. and Hwang, B.H. (2014), "Wisdom of crowds: The value of stock opinions transmitted through social media", *Review of Financial Studies*, Vol. 27 No. 5, pp. 1367–1403.

Clayton, J., Ling, D.C. and Naranjo, A. (2009), "Commercial real estate valuation: Fundamentals versus investor sentiment", *Journal of Real Estate Finance and Economics*, Vol. 38 No. 1, pp. 5–37.

Danielsen, B.R., Harrison, D.M., Van Ness, R.A. and Warr, R.S. (2009), "REIT auditor fees and financial market transparency", *Real Estate Economics*, Vol. 37 No. 3, pp. 515–557.

Das, P.K., Freybote, J. and Marcato, G. (2015), "An Investigation into Sentiment-Induced Institutional Trading Behavior and Asset Pricing in the REIT Market", *Journal of Real Estate Finance and Economics*, Vol. 51 No. 2, pp. 160–189.

Das, S.R. and Chen, M.Y. (2007), "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web", *Management Science*, Vol. 53 No. 9, pp. 1375–1388.

Davis, A.K., Piger, J.M. and Sedor, L.M. (2012), "Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language", *Contemporary Accounting Research*, Vol. 29 No. 3, pp. 845–868.

Davis, A.K. and Tama-Sweet, I. (2012), "Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A", *Contemporary Accounting Research*, Vol. 29 No. 3, pp. 804–837.

Doran, J.S., Peterson, D.R. and Price, S.M. (2012), "Earnings Conference Call Content and Stock Price: The Case of REITs", *Journal of Real Estate Finance and Economics*, Vol. 45 No. 2, pp. 402–434.

Engelberg, J.E., Reed, A. V and Ringgenberg, M.C. (2012), "How are shorts informed? Short sellers, news, and information processing", *Journal of Financial Economics*, Vol. 105 No. 2, pp. 260–278.

Feldman, R., Govindaraj, S., Livnat, J. and Segal, B. (2010), "Management's tone change, post earnings announcement drift and accruals", *Review of Accounting Studies*, Vol. 15 No. 4, pp. 915–953.

Ferguson, N.J., Philip, D., Lam, H.Y.T. and Guo, J.M. (2015), "Media Content and Stock Returns: The Predictive Power of Press", *Multinational Finance Journal*, Vol. 19 No. 1, pp. 1–31.

Freybote, J. (2016), "Real estate sentiment as information for REIT bond pricing*", Journal of Property Research*, Vol. 33 No. 1, pp. 18–36.

Freybote, J. and Seagraves, P.A. (2017), "Heterogeneous Investor Sentiment and Institutional Real Estate Investments", *Real Estate Economics*, Vol. 45 No. 1, pp. 154–176.

Garcia, D. (2013), "Sentiment during Recessions", *Journal of Finance*, Vol. 68 No. 3, pp. 1267–1300.

Hartzell, J.C., Kallberg, J.G. and Liu, C.H. (2008), "The Role of Corporate Governance in Initial Public Offerings: Evidence from Real Estate Investment Trusts", *Journal of Law and Economics*, Vol. 51 No. 3, pp. 539–562.

Hausler, J., Ruscheinsky, J. and Lang, M. (2018), "News-based sentiment analysis in real estate: a machine learning approach", *Journal of Property Research*, Vol. 35 No. 4, pp. 344–371.

Henry, E. (2008), "Are investors influenced by how earnings press releases are written?", *Journal of Business Communication*, Vol. 45 No. 4, pp. 363–407.

Henry, E. and Leone, J.A. (2016), "Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone", *Accounting Review*, Vol. 91 No. 1, pp. 153–178.

Heston, S.L. and Sinha, N.R. (2017), "News versus Sentiment: Predicting Stock Returns from News Stories", *Financial Analysts Journal*, Vol. 73 No. 3, pp. 67–83.

Huang, X., Teoh, S.H. and Zhang, Y. (2014), "Tone management", *Accounting Review*, Vol. 89 No. 3, pp. 1083–1113.

Huefner, B. (2007), "The SEC's MD&A: Does it Meet the Informational Demands of Investors? - A Conceptual Evaluation", *Schmalenbach Business Review*, Vol. 59 No. 1, pp. 58–84.

Jegadeesh, N. and Wu, D. (2013), "Word power: A new approach for content analysis", *Journal of Financial Economics*, Vol. 110 No. 3, pp. 712–729.

Kearney, C. and Liu, S. (2014), "Textual sentiment in finance: A survey of methods and models", *International Review of Financial Analysis*, Vol. 33, pp. 171-185.

Li, F. (2010), "The information content of forward- looking statements in corporate filings-A naïve bayesian machine learning approach", *Journal of Accounting Research*, Vol. 48 No. 5, pp. 1049–1102.

Lin, C.Y., Rahman, H. and Yung, K. (2009), "Investor sentiment and REIT returns", *Journal of Real Estate Finance and Economics*, Vol. 39 No. 4, pp. 450–471.

Ling, D.C., Naranjo, A. and Scheick, B. (2014), "Investor sentiment, limits to arbitrage and private market returns", *Real Estate Economics*, Vol. 42 No. 3, pp. 531–577.

Loughran, T. and McDonald, B. (2011), "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks", *The Journal of Finance*, Vol. 66 No. 1, pp. 35–65.

Loughran, T. and McDonald, B. (2015), "The Use of Word Lists in Textual Analysis", *Journal of Behavioral Finance*, Vol. 16 No. 1, pp. 1–11.

Loughran, T. and McDonald, B. (2016), "Textual Analysis in Accounting and Finance: A Survey", *Journal of Accounting Research*, Vol. 54 No. 4, pp. 1187–1230.

Marcato, G. and Nanda, A. (2016), "Information content and forecasting ability of sentiment indicators: Case of real estate market", *Journal of Real Estate Research*, Vol. 38 No. 2, pp. 165–203.

Nasukawa, T. and Nagano, T. (2001), "Text analysis and knowledge mining system", *IBM Systems Journal*, Vol. 40 No. 4, pp. 967–984.

Pava, M. and Epstein, M. (1993), "How good is MD&A as an investment tool?", *Journal of Accountancy*, Vol. 175 No. 3, pp. 51–54.

Price, S.M., Seiler, M.J. and Shen, J. (2017), "Do Investors Infer Vocal Cues from CEOs During Quarterly REIT Conference Calls?", *Journal of Real Estate Finance and Economics*, Vol. 54 No. 4, pp.515–557.

Rogers, J.L., Van Buskirk, A. and Zechman, S.L.C. (2011), "Disclosure tone and shareholder litigation", *Accounting Review*, Vol. 86 No. 6, pp. 2155–2183.

Rozin, P. and Royzman, E.B. (2001), "Negativity Bias, Negativity Dominance, and Contagion", *Personality and Social Psychology Review*, Vol. 5 No. 4, pp. 296–320.

Ruscheinsky, J.R., Lang, M. and Schäfers, W. (2018), "Real estate media sentiment through textual analysis", *Journal of Property Investment and Finance*, Vol. 36 No. 5, pp. 410–428.

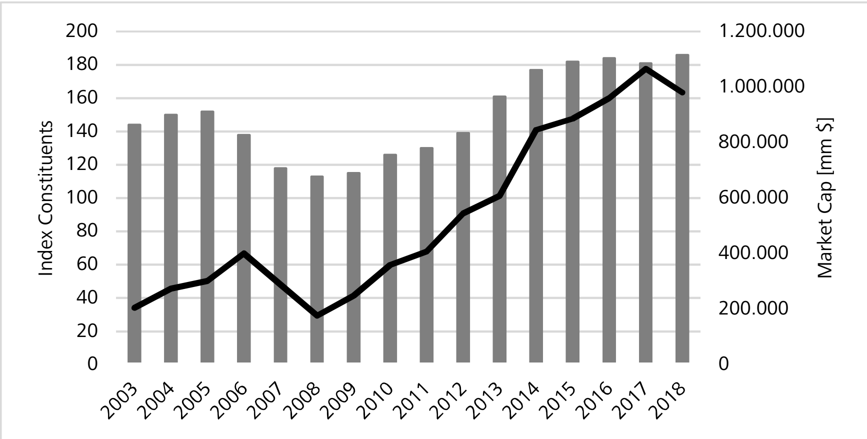
SEC. Securities and Exchange Commission (1987), Securities Act Release No. 6711, available at: https://www.sec.gov/rules/interp/33-6835.htm.

Tetlock, P.C. (2007), "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", *Journal of Finance*, Vol. 62 No. 3, pp. 1139–1168.

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. (2008), "More Than Words: Quantifying Language to Measure Firms' Fundamentals", *Journal of Finance*, Vol. 63 No. 3, pp. 1437–1467.

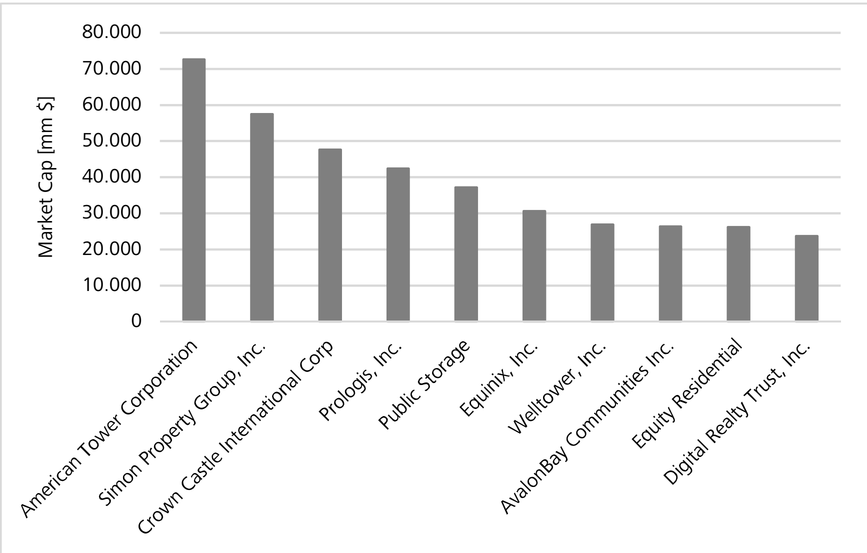
2.9 Appendix

Figure 2.1: NAREIT All Equity Index Constituents and Market Capitalization



This figure shows the number of Equity REITs present in the FTSE NAREIT All REITs Index at any point of time between January 2003 and December 2018 as well as the total market capitalization over years.

Figure 2.2: NAREIT All Equity Index Leading Players



This figure shows the leading players present in the FTSE NAREIT All REITs Index in terms of market capitalization.

Table 2.10: Variable definitions using Compustat data items

Dependent Variables	
FFO/Share	FFO scaled by shares outstanding (NIQ+SPPEY+(DPACREQt- DPACREQt-90))/CSHOQ
AR	Actual return – return on CRSP value-weighted market portfolio
CAR	Sum of AR over the reaction-window
Control Variables	
Firmsize	Logarithm of total assets log(ATQ)
REV	Current quarter sales – sales in the previous quarter (SALEQt-SALEQt-90)
Leverage	Total liabilities/ total assets (LTQ/ATQ)
SUE	Actual earnings – most recent mean analyst forecast from I/B/E/S scaled by the stock price (PRCCQ)
MB	Market to book value of common stock ((PRCCQ*CSHOQ)/CEQQ)
ROE	Income before extraordinary items/common equity as reported (IBQ/CEQQ)
Liquidity	Traded shares/shares outstanding (CSHTRQ/CSHOQ)
Volatility	Standard deviation of daily returns over the most recent quarter (PRCCD)

This table describes the variables used and the corresponding Compustat data items.

3 Can Risks be Good News? Revealing Risk Perception of Real Estate Investors using Machine Learning

3.1 Abstract

The SEC mandates firms to inform investors about their assessment of future contingencies in their 10-Ks. However lengthy and complex disclosures – mostly for dozens of firms in an investor's portfolio – can barely be processed by a human being. To cope with the flood of information, we exploit an unsupervised machine learning algorithm, the Structural Topic Model, to identify the risk factors discussed in 10-Ks. We apply this algorithm to a US REIT sample between 2005 and 2019 to assess whether the probability of appearance of the extracted risk factors helps to explain the perceived risk on the stock market. We find that the majority of risk factors is significantly associated with volatility indicating that our machine-assisted modeling presents a valid approach to quantify risk disclosures in textual form. Furthermore, we investigate in which direction individual topics affect investor risk perception. Even if all kinds of directions exist, uninformative topics with no impact, increasing risk-perception topics, and decreasing risk-perception topics, the latter is clearly predominant. The predominance of the risk-reducing effect indicates that risk disclosures can indeed be considered good news as long as they clarify the implications of already known risk.

Keywords: Risk, Textual Analysis, Machine Learning, Structural Topic Model, 10-K

3.2 Introduction

According to the Securities and Exchange Commission (SEC), investors can find precise risk information for a company of interest in the standardized report form of 10-Ks. The complex (inter-)relationship of risk can hereby be condensed and described quantitatively (e.g., volatility) or qualitatively (e.g. “the corona virus could have a material adverse effect on our business”). However, the qualitative form is the regular form, since companies are required to discuss “the most significant factors that make the company speculative or risky” within “Item 1A - Risk Factors” (SEC, 2005). Although all types of risk – whether quantified or described qualitatively – influence the decisions of managers and investors alike, mandatory risk disclosures in qualitative form (e.g., Item 1A) are less explored than quantitative information. This is mainly driven by how we are trained to make investment decisions – as a quantified trade-off between expected return and risk. Ignoring the information that is stored in textual data instead of numeric data would be negligent since the textual form gives managers more flexibility than numeric data to express or hide information about risk. They decide how precisely, with which words, and to what extent risk factors are described. By contrast, the presentation and calculation of numeric data are often determined by accounting standards like GAAP (Davis and Tama-Sweet, 2012).

The length and complexity make it increasingly difficult and time-consuming for investors to read and interpret disclosures. Our data show that the average length of Item 1A had increased from roughly 41,000 to 93,000 words between 2006 and 2019. The complexity, measured in years of formal education a person needs to understand the text, remains at a constantly high level of more than 20 years on average. Certainly, human investors are not able to handle the large quantity of corporate disclosures provided by dozens or hundreds of firms.

We suggest solving this challenge using an unsupervised machine learning technique, which automatically clusters words around topics found in a large set of documents – the topic model. It assumes that documents are characterized as a collection of topics, and topics as a collection of words. This machine-assisted reading of text corpora enables uncovering latent topics in a relatively short time period in comparison to a human being. The underlying procedure is conceptually similar to factor analysis, where a large and complex dimensionality of variables is reduced to fewer factors (Dyer et al., 2017). However, previous studies indicate that topic modeling has the tendency to find already known factors; for example, the extracted topics across different industries are more likely identifiers for the known industries than common factors across industries.

To overcome this issue, we use advanced topic modeling approaches which allow that the topics correlate within documents (Correlated Topic Model, CTM) and covariate words describing similar (sub-)sectors can be excluded in order to clearly extract the common risk factors across firms (Structural Topic Modeling, STM). The text corpora in Figure 3.1 illustrate why covariate words are important to extract the common risk factors. The identified words defining the topics using the traditional Latent Dirichlet Allocation (LDA) correspond to the already known sectors – corpus A is provided by a firm in the healthcare sector and corpus B by a firm in the residential sector. At the same time, both corpora address the topic "Legal & Litigation Risk" which is identified by STM as the common topic. Thus, STM allows extracting common factors across documents by excluding the already known factors (e.g., healthcare and residential) and their corresponding words.

Figure 3.1: Stylized Illustration of LDA and STM

Corpus A

Our operators and tenants are faced with litigation and may experience rising liability and insurance costs. In some states, advocacy groups have been created to monitor the quality of **care** at **healthcare** facilities and these groups have brought litigation against the operators and tenants of such facilities. Also, in several instances, private litigation by **patients** has succeeded in winning large damage awards for alleged **abuses**. The effect of this litigation and other potential litigation may materially increase the costs incurred by our operators and tenants for monitoring and reporting quality of **care** compliance. In 16 Table of Contents addition, their cost of liability and medical malpractice insurance can be significant and may increase or even not be available at a reasonable cost so long as the present **healthcare** litigation environment continues.

Cost increases could cause our operators to be unable to make their lease or mortgage payments or fail to purchase the appropriate liability and malpractice insurance, potentially decreasing our revenues and increasing our collection and litigation costs. In addition, as a result of our ownership of **healthcare** facilities, we may be named as a defendant in lawsuits allegedly arising from the actions of our operators or tenants, for which claims such operators and tenants have agreed to indemnify, defend and hold us harmless from and against, but which may require unanticipated expenditures on our part.

LDA Topic: Health Care
healthcare, medicaid, correctional, detention, hospitals, hospital, brookdale, seniors, nursing, physicians, patients, payors, medicare, sunrise, inmates, tenants, care, medical, physician, science

STM Topic: Legal & Litigation Risk
plaintiffs, sue, zones, tax-exempt, prejudice, supreme, examine, defendants, federally, defendant, render, oversee, complaint, day, straight-line, exposures, tangible, feature, flood, conform

STM Covariate: Health Care
referral, licensure, patients, false, physician, payors, abuse, healthcare, whistleblower, medicare, medicaid, denial, hospitals, patient, payor, physicians, hipaa, referrals, care, anti-kickback

Corpus B

Potential liability or other expenditures associated with potential environmental contamination may be costly. Various federal, state and local laws subject **apartment** **community** owners or operators to liability for management, and the costs of removal or remediation, of certain potentially hazardous materials that may be present in the land or buildings of an **apartment** **community**. Potentially hazardous materials may include polychlorinated biphenyls, petroleum-based fuels, lead-based paint, or asbestos, among other materials. Such laws often impose liability without regard to fault or whether the owner or operator knew of, or was responsible for, the presence of such materials. The presence of, or the failure to manage or remediate properly, these materials may adversely affect occupancy at such **apartment** **communities** as well as the ability to sell or finance such **apartment** **communities**. In addition, governmental agencies may bring claims for costs associated with investigation and remediation actions, damages to natural resources and for potential fines or penalties in connection with such damage or with respect to the improper management of hazardous materials. Moreover, private plaintiffs may potentially make claims for investigation and remediation costs they incur or personal injury, disease, disability or other infirmities related to the alleged presence of hazardous materials at an **apartment** **community**.

LDA Topic: Residential
communities, apartment, digital, companys, multifamily, realty, housing, freddie, incs, fannie, mac, homes, mae, residents, sale, lps, manufactured, multi-family, excel, partnership

STM Topic: Legal & Litigation Risk
plaintiffs, sue, zones, tax-exempt, prejudice, supreme, examine, defendants, federally, defendant, render, oversee, complaint, day, straight-line, exposures, tangible, feature, flood, conform

STM Covariate: Residential
mae, fannie, residents, homes, mac, freddie, apartment, housing, multifamily, fhaa, household, communities, explore, apartments, home, lawsuits, offers, conservatorship, already, regulating

This figure shows text corpora provided by a firm in the healthcare sector (Corpus A) and a firm in the residential sector (Corpus B). Both corpora address the topic "Legal & Litigation Risk" which is identified by STM as the common topic. Words associated with the topic "Legal & Litigation Risk" are highlighted in yellow. Words associated with the LDA topic "Health Care" are highlighted in blue. Words associated with the LDA topic "Residential" are highlighted in red. Words associated with either the metadata covariate "Health Care" or "Residential" are in bold.

An appealing testing ground for our textual analysis procedure is presented by the REIT industry for multiple reasons. First, while the sector is characterized by relatively homogenous business models and firm characteristics, different investment foci or property types (e.g., healthcare, residential) are salient and distract the LDA from extracting common risk factors (see Figure 3.1). Second, REITs' 10-Ks guarantee a relatively high disclosure quality, given their high dividend payout requirement of at least 90% of their taxable earnings. Consequently, they have a very limited cash reserve and must turn to the capital markets repeatedly to raise funding for new projects. This regulation incentivizes that REITs are transparent, act for the long-term, and sustain investor trust (Danielsen et al., 2009; Doran et al., 2012; Price et al., 2017). Third, the real estate industry is characterized by a high number of institutional investors who are likely to process lengthy and complex disclosures in the form of 10-Ks. In addition, such investors are rarely driven by noise trading or herding behavior which irrationally influence the stock prices. In line with previous research, we expect investors to alter their perception of future firm performance and risk based on the firms' disclosure (e.g., Kravet and Muslu, 2013; Bao and Datta, 2014).

To assess whether our machine-assisted topic detecting modeling presents a valid approach to quantify risk in narrative form, we analyze whether the STM extracted risk factors help to explain the perceived risk on the stock market. Indeed, we find that the majority of risk topics is significantly associated with volatility, confirming the effectiveness of our model. Furthermore, we disentangle how the identified risk topics explain investor risk perception, in addition to traditional firm characteristics. Although an initial intuition suggests that risks are *per se* negative, we primarily expect a risk-reducing effect of risk disclosures, since most risk factors are revealed in a timely manner through press releases or Form 8-K. Instead of disclosing new risk factors, Item 1A primarily provides essential information on risk factors that have already been communicated to investors using more frequent channels thereby reducing uncertainty and risk perception. In line with expectations, most of our identified factors follow the convergence argument indicating a risk-reducing effect. Consequently, it seems like executives' concerns of adverse effects of disclosing "negative" information are baseless and risks described in 10 Ks can indeed be considered good news as long as executives clarify the implications of already known risk.

To the best of our knowledge, this is the first study applying advanced topic modeling (CTM and STM) to the accounting and finance domain. Both models yield superior results than previous textual analysis approaches like keyword counts or the LDA, whereas STM is even better than CTM. We thus add to the literature by proposing STM to quantify the risk described in disclosures at a higher level of granularity than previous methods. By analyzing the topics' generated vocabulary, the nature and scope of the disclosed risk factors can be identified. This allows assessing the stock market reaction to each risk factor. Finally, we close a research gap, since prior studies conducting textual analysis often exclude the banking, insurance, and REIT sectors due to their financial characteristics and legal requirements.

The remainder of the paper is organized as follows. Section 3.3 discusses related literature on mandatory risk disclosures and develops hypotheses. Section 3.4 explains the textual analysis procedures (i.e., LDA, CTM, and STM) and interprets the risk factors. The empirical methods for the analysis are presented in Section 3.5, while Section 3.6 introduces the data used and describes the variables. The empirical results are reported in Section 3.7, and Section 3.8 concludes.

3.3 Previous Literature and Hypotheses Development

3.3.1 Textual Analysis in Accounting and Finance

Fueled by the rise of computational power and the tremendously increasing online-availability of text, a growing body of literature in accounting and finance has focused on computer-based techniques to find and quantify information revealed in qualitative disclosures (e.g., media news, public corporate disclosures, analyst reports, and internet postings). Within the finance research, probably Tetlock (2007) provides the pioneering study by employing automated content analysis to extract sentiment from the *Wall Street Journal's* column "Abreast of the Market". He demonstrates, that media pessimism induces downward pressure on market prices and leads to temporarily high market trading volume. Thereafter, multiple studies analyze how sentiment predicts the reactions of financial markets. For example, Garcia (2013) processes finance news from *The New York Times* and provides evidence that positive words also help to predict stock returns. Tetlock et al. (2008) analyze firm-specific news from the *Dow Jones News Service* and *The Wall Street Journal* and prove that negative words convey negative information about firm earnings above and beyond stock analysts' forecasts and historical accounting data. Antweiler and Frank (2004), Das and Chen (2007), and Chen et al. (2014) investigate the textual sentiment of internet messages. Hereby, Antweiler and Frank (2004) find evidence that the amount of message posting (activity) predicts market volatility and trading volume. Chen et al. (2014) figure out that the fraction of negative words contained in articles published on *Seeking Alpha* negatively correlates with contemporaneous and subsequent stock returns. Das and Chen (2007) make assumptions about the relationship between textual sentiment and investor sentiment when interpreting textual sentiment or tone of internet messages as small investor sentiment. They link market activity to small investor sentiment and message board activity. Regarding the studies addressing corporate disclosures, textual sentiment has been found to be positively related to abnormal stock returns (e.g., Feldman et al., 2010; Chen et al., 2014; Jegadeesh and Wu, 2013), subsequent stock return volatility (e.g., Loughran and McDonald, 2011, 2015), and future earnings and liquidity (e.g., Li, 2010).

Further research investigates the readability of corporate disclosures and provides evidence that lower annual report readability is associated with increased stock return volatility (Loughran and McDonald, 2014), lower earnings persistence and higher earnings surprise (Li, 2008; Loughran and McDonald, 2014), larger analyst dispersion (Lehavy et al., 2011; Loughran and McDonald, 2014), as well as lower trading due to a reduction in small investor trading activity (Miller, 2010). Only recently, Cohen et al. (2020) use sentiment and multiple similarity measures to show that changes to the language and construction of corporate disclosures impact stock prices with a time lag. The authors conclude that investors need time to process complex and lengthy disclosures.

This study contributes to the emerging literature on textual analysis by adopting a new perceptive. Instead of focusing on the tone conveyed through the narrative, the complexity of the language, or document similarity, we examine the individual risk factors disclosed in Item 1A of the annual reports.

3.3.2 Textual Analysis of Risk Disclosures

The literature in accounting and finance has applied various methods to assess a firms' risk disclosure, which we classify in two categories. Within the first and more straightforward category, the entire risk disclosure is observed as a unit and its "size" is considered as a proxy for risk. Within the second and more sophisticated category, the individual risk itself comes to the forefront. The former category comprises studies that count risk keywords (e.g., Li, 2006; Kravet and Muslu, 2013) or rely on the total length of the risk section (e.g., Campbell et al., 2014; Nelson and Pritchard, 2016) to measure firms' risk disclosures. Hereby, increased levels of risk disclosure are linked to higher stock return volatility (Kravet and Muslu, 2013; Campbell et al., 2014), trading volume (Kravet and Muslu, 2013), and lower future earnings and stock returns (Li, 2006). Although these straightforward approaches allow the user to process a large number of textual documents which is beyond human capacity, they obviously lose a lot of information written in the text.

Only recently and within the latter category, researchers have started to focus more on the written content by making use of machine learning approaches to identify and quantify the individual risks. In this context, the unsupervised machine learning approach Latent Dirichlet Allocation (LDA) is most popular for finding the individual risks discussed in firms' filings. The outcomes are manifold: Israelsen (2014), for example, examines the association between the risks disclosed in Item 1A and stock return volatility, as well as betas of the Fama-French Four-Factor model. Employing a variation of the LDA, Bao and Datta (2014) analyze whether and how risk disclosures affect investor risk perceptions. Their findings indicate that some risk factors increase or decrease investor risk perceptions, and thus lead to higher or lower post-filing return volatility, whereas others have no effect at all. Gaulin (2017) uses disclosed risk factors to analyze disclosure habits and suggests that managers time the identification of new risks, as well as the removal of previously identified ones, to match their expectations of adverse outcomes in the future. Recently, Lopez-Lira (2019) demonstrates the importance of risk disclosures by providing a factor model that uses only identified firm risk factors to explain stock returns and performs as least as well as traditional models, without including any information from past prices.

The key benefit of machine learning approaches is that they do not require predefined rules (i.e., *a priori* determined keywords) to identify risk factors. Instead, risk factors or in general topics derive naturally from fitting the statistical model to the textual corpus, based on word co-occurrences in the documents.

3.3.3 Hypothesis Development

Common to all approaches, whether straightforward or sophisticated, is that they attempt to quantify qualitative information in risk disclosures without the need for a human being to read them. However, quantifying risk disclosures in the form of Item 1A is quite challenging given that firms neither reveal the likelihood that a disclosed risk will ultimately affect the company, nor the quantified impact a risk might have on the firm's current and future financial statements. Thus, risk disclosures might inform the reader, for the most part about a vague range, but certainly not the level of future performance (Kravet and Muslu, 2013). Nevertheless, assuming that firm executives truthfully report their views (under SEC scrutiny and penalty of litigation), it can be argued that detailed firm-specific information is provided in Item 1A. In fact, previous research (e.g., Kravet and Muslu, 2013; Bao and Datta, 2014) finds a stock market reaction to risk disclosures confirming its informativeness.

Recognizing that management's discretion entails considerable leeway in deciding which risks to disclose and how much of the filing is allocated to a particular risk-factor topic, we assume that this proportion provides valuable information on how companies assess the probability of occurrence and the extent of the risks. Accordingly, the topic distribution in the filing could serve as a proxy for risk disclosures, allowing investors to quantify the information provided in narrative form.

Hypothesis 1: *The probabilities of the appearance of the extracted topics in textual risk reports present significant explaining factors in empirical models analyzing investor risk perception.*

In line with the intention of mandatory risk disclosures to warn of future adverse outcomes, sophisticated investors process disclosures and alter their expectations of future firm performance accordingly. While intuition might suggest that risks are *per se* negative, Kravet and Muslu (2013) define three opposing arguments regarding how risk disclosures affect the risk perception of investors. The first argument suggests that investor risk perceptions remain unaffected since risk disclosures are vague and boilerplate in nature because managers are likely to disclose all possible risks and uncertainties without considering their impact on businesses just to be on the safe side (null argument). The second argument states that risk disclosures reveal unknown risk factors or risk-increasing facts about known risk factors causing diverging investor opinions and increasing risk perceptions (divergence argument). The third argument assumes that executives use disclosures to resolve firms' known risk factors or give more facts about known risk factors and thus, reduce risk perceptions (convergence argument).

Contradicting the null argument, Loughran and McDonald (2020) find that less than 21% of 10-Ks filed in 2018 forewarn of the adverse effects of a pandemic in general. Given the lessons learned of other disease outbreaks (e.g., SARS or swine flu) over the last decade, it seems like pandemics should have long been identified as a potential global risk and more firms should have addressed them in their filings if they just want to be on the safe side. Furthermore, it seems quite unlikely that previously unknown risk factors and contingencies are revealed in Item 1A (divergence argument), which is published only once a year, given that firms are required to promptly file 8-Ks, whenever "material" events take place.[11] Contingencies that do not require a Form 8-K might be discussed in press releases or even materialize before they are published in Item 1A. Therefore, it is likely that Item 1A primarily provides essential information on risk factors that have already been communicated to investors through more frequent channels, instead of disclosing new risk factors. For example, the current pandemic, which has so far only been presented as a risk

[11] According to the SEC (2012), "the types of information required to be disclosed on Form 8-K are generally considered to be 'material,' … [meaning] there is a substantial likelihood that a reasonable investor would consider the information important in making an investment decision."

factor in very few reports, is now known to investors as a risk anyway. Investors' expectations regarding the adverse effects of the pandemic have already been incorporated into prices when the company finally specifies them in its report. Thus, the detailed information regarding the pandemics' influence on the firm in Item 1A may reduce risk perceptions and converge investors' opinions. We thus formulate our second hypothesis as follows:

Hypothesis 2: *The majority of the risk factors present a risk-reducing effect, supporting the convergence argument.*

3.4 Textual Analysis with Machine Learning

3.4.1 Topic Modeling

The Latent Dirichlet Allocation (LDA) is the most frequently used topic modeling approach in the scientific literature; it is borrowed from genetic science (Pritchard et al., 2000) and transferred to machine learning by Blei et al. (2003), who named it after the Dirichlet distribution. It is a mixture model, thus generating the probabilities of co-occurring topics (subpopulation) within the distribution over all words (population). Put simply, the mixture model aims to break documents down into topics, whereby the words within each topic co-occur most frequently. Thus, applying the LDA to a textual corpus results in two data structures in the output. The former presents the probability of appearance of each topic in each document (θ_d), with documents being indexed by d. The latter lists a set of words and their probabilistic relation with each of the extracted topics (β_k), with topics being indexed by k.

LDA comes with the limitation that the Dirichlet distribution assumes almost uncorrelated topics. However, they are likely correlated since particular topics are likely to occur at the same time. For example, a shopping mall firm pointing out that the loss of the anchor tenant is a potential risk, may also talk about related risks such as the impact of e-commerce on the retail business. These covariances are addressed by Blei and Lafferty (2007) in their Correlated Topic Model (CTM) method. Also, the CTM is a mixture model but replaces the Dirichlet distribution with a logistic normal distribution in order to include the covariance structure among topics. Surprisingly, it is not very often applied even if Blei and Lafferty (2007) show the theoretical and practical importance of a covariance structure by using 16,351 *Science* articles. Employing a smaller collection of articles, they calculate the log likelihoods of the held-out data given a model estimated from the remaining data.

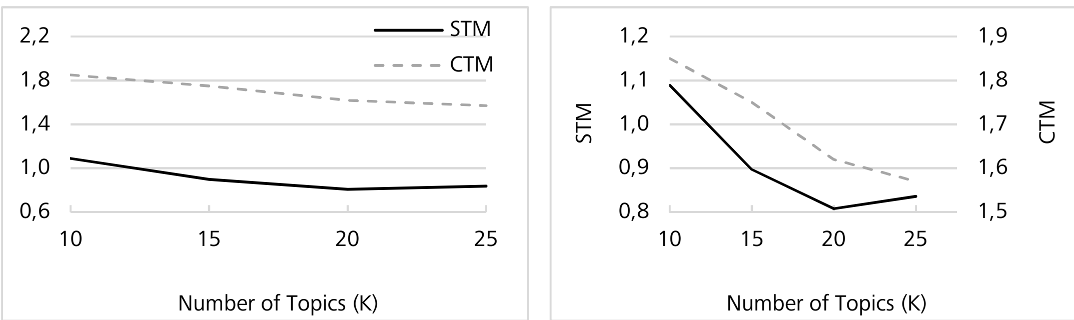
A higher likelihood of the held-out data indicates a better model. They find that CTM is always superior to LDA for altering the number of topics from 5 to 120.[12]

The Structural Topic Modeling (STM) by Roberts et al. (2019) goes even one step further and incorporates metadata of pre-specified covariates, not only covariances, to disentangle the unique topics. Again, it remains a mixture model based on a logistic normal distribution, which is the same as CTM if it ignores the covariates. In order to allow the algorithm to find topics beyond the already known identifiers (see Figure 3.1 and discussion in the Introduction for healthcare vs. residential), we include property types as metadata covariates. More details on the STM algorithm are given in the next sections and Appendix A presents technical details of the STM.

To compare CTM and STM, we conduct an analysis similar to that described above and performed by Blei and Lafferty (2007). We fit a smaller collection of documents to a varying number of topics (between 10 and 25) and calculate the residuals, lower bounds, and log likelihoods of the held-out data. For example, the better a model fits the higher is the probability of the held-out data. All three measures indicate a better fit for STM over CTM for the full range of topic numbers (see Figure 3.2, Panel A-C). Additionally, topic modeling requires an *a priori* determination of the number of topics to be generated. All comparison measures indicate directly or converge to a topic number of 20 as the best number. Consequently, we extract 20 individual risk factors from the risk reports.

Figure 3.2: Comparison of CTM and STM

Panel A: Residuals

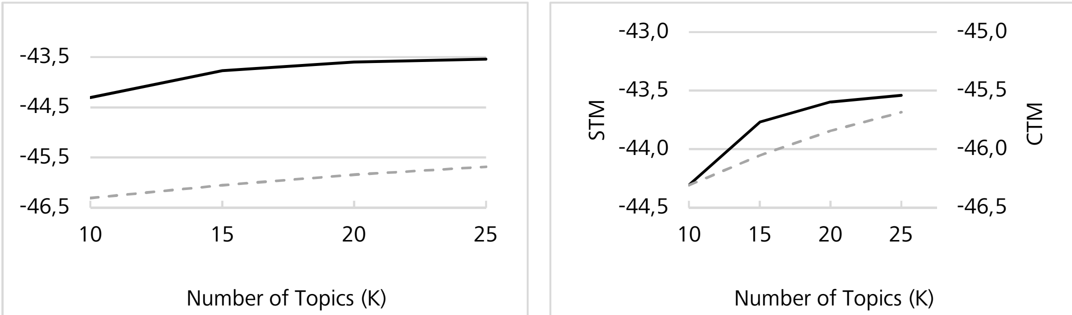


see next page

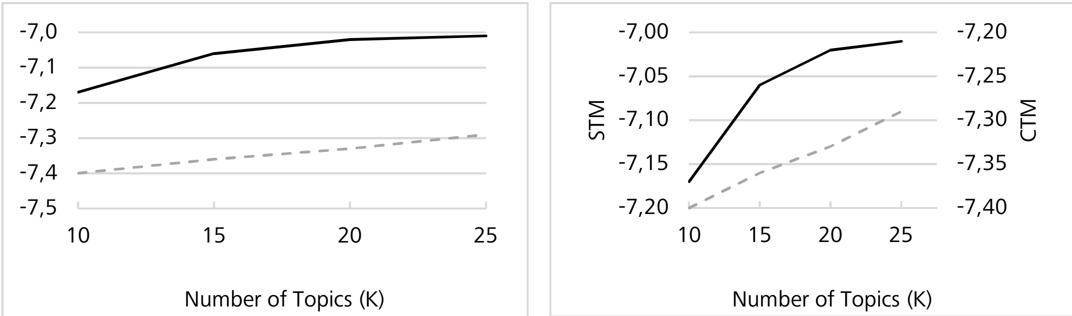
[12] This paper provides only an overview of LDA and CTM); for deeper insights, we refer to the original papers by Blei et al. (2003), Blei and Lafferty (2007).

Figure 3.2: continued

Panel B: Lower Bound (in millions)



Panel C: Held-Out Likelihood



This figure shows the standard criteria for comparing different topic models, namely residuals regarding the text corpora, lower bound and held-out likelihood with a standard of 20 percent. On the left hand side and identical scale, the STM outperforms the CTM on our data set. The right side shows two different scales for each model to clarify the turning points of optimization process for a different number of topics (*K*) within each model.

Based on the superiority of CTM over LDA (see, Blei and Lafferty, 2005), and STM over CTM and LDA (see, Roberts et al. (2014), we focus on the use of STM in or later analysis so that we assume that covariates describe pre-specified subsectors. Nevertheless, we verify our decision by applying LDA and CTM (see Section 3.7.4).

3.4.2 Topic Identifications

To apply the STM to our textual corpus, we use the programming language R (version 4.0.2) and the corresponding package STM, authored by Roberts et al. (2019). As often done in machine-learning-based text analysis methods, several preprocessing steps are necessary before running the topic model. First, we reduce documents to individual words or tokens and clean the data by removing white spaces, numbers, and punctuation. Second, relying on the “stop word” list provided by the STM package, words like ‘and’, ‘or’, and ‘the’ are removed from the corpus, since they lack semantic information, and thus do not help to identify the topics. Third, we eliminate infrequent words (e.g., Argentina, banana, bio-fuel) to avoid their influence. More precisely, we remove words appearing in fewer than 20 disclosures. On the one hand, this threshold rules out words occurring solely in 10-Ks of one particular firm, since we have 14 years of observations (e.g., the firm names). On the other hand, low-frequency words cannot be clearly assigned to an individual topic, and thus introduce noise into the process. Excluding them ensures the robustness of the algorithm, and in addition, increases computational speed (Papilloud and Hinneburg, 2018). Unlike Roberts et al. (2019), we do not stem the words and instead use explicit word inflections for reasons of interpretability. This abandonment is supported by Schofield and Mimno (2016), who find that stemming does not improve topic stability, and possibly even degrades it.

3.4.3 Risk Factor Topics

Although topic-modeling approaches allow classifying textual data without further instruction by the user, the topics created by the algorithm do require interpretation. More specifically, a human being has to assign labels with an assessment of the most plausible content to the algorithm-based topics, which are only equipped with a number and a set of words most frequently associated with each topic. In order to label the risk-factor topics appropriately, we read a random sample of disclosures comprising 2% of the overall sample. Two of us then independently reviewed the word lists comprising the 20 highest associated terms for each risk-factor topic. As recommended by Roberts et al. (2019), we also inspected documents that were considered to be highly associated with a specific topic, and thus, are expected to represent the topic most clearly.

Following this procedure, we entitled, for example, topic #1 “Transaction” including words such as ‘unenforceable’, ‘origination’, ‘repurchases’, and ‘sale-leaseback’. Topic #8 labeled “Capital Products and Market” contains words such as ‘nyse’ (New York Stock Exchange), ‘cdo’ (Collateralized Debt Obligation), ‘servicer’, and ‘electronically’. The frequent appearance of phrases such as ‘plaintiffs’, ‘defendant’, ‘supreme’, and ‘prejudice’ suggests that the corresponding topic #13 is related to “Legal & Litigation risk”. The STM also reveals topics related to the “REIT Status”, “Property Risk”, “IT”, “Tax”, and “Politics”. Table 3.5 in Appendix B presents the 20 highest associated words for each risk factor topic and the corresponding name.

For some topics, however, it is more difficult to find a one-title-fits-all label. For example, topic #10 of contains phrases such as ‘hackers’, ‘terrorists’, ‘libor’, and ‘tcja’ (Tax Cuts and Jobs Act), and thus, the interpretation is somewhat blurry or mixed. In this case, examining disclosures including these keywords can be helpful in finding the missing link among the STM-identified words for a topic, being able to find a generic topic and interpret its meaning. The annual report of Boston Properties, Inc. in 2018 discusses certain ‘risks associated with security breaches through cyber attacks’, ‘terrorist attacks may adversely affect the ability to generate revenues’, and ‘tax changes that could negatively impact financials’ in close proximity to each other. A deeper look into the documents shows that numerous disclosures raise these risks directly one after the other. Given that topic models rely on word co-occurrences and ignore visual clues (e.g., subsection titles, boldface fonts, extra spacing) or logical coherence, the resulting "mixture of topics" is the consequence. At a higher level, however, topic #10 can be subsumed as "Contingencies".

Similarly, polysemy – the capacity for a word to have multiple meanings – makes it harder to label topics, at least for non-professionals. At first glance, the words 'migration' and 'recycling' do not fit with the other words in topic #5 of Item 1A (e.g., 'moodys', 'poors') which intuitively entails the label "Rating". However, the word 'migration' may also be used in the context of 'rating migration' and 'recycling' might refer to 'capital recycling' which may be the reason for a rating upgrade or downgrade.

3.5 Model Specification

When applying the STM to our textual data, we pre-specify risk-factor topics along with their frequency of appearance in each individual 10-K. To assess whether the probability of appearance of the STM extracted risk factors helps to explain the perceived risk on the stock market and how risks affect investor risk perception, we regress these frequencies (*Freq_Topics*) on the firms' stock return volatility (*Vola*) by using the following two-way fixed-effects regression model:

$$Vola_{it} = \beta_0 + \beta_1 Freq\_Topics_{it} + \beta_2 Controls_{it} + a_i + \lambda_t + u_{it} \,, \tag{3.1}$$

where i denotes the firm, and t the year. In addition to the vector of frequencies of the individual risk topics (*Freq_Topics*), the regression equation includes a vector of control variables (*Controls*). The parameters a_i and λ_t incorporate the unobserved firm and time effects and u_{it} is the error term. The two-way fixed effects model is most applicable for our analysis since we aim to investigate the specific differences between individuals in a micro panel dataset covering roughly 14 years (Wooldridge, 2010). To produce consistent, efficient, and unbiased estimates, we examine whether any of the models' assumptions are violated. Employing Variance Inflation Factors (VIF) to check for multicollinearity, we find values greater than 5 for Topic #7, Topic #11, Topic #14, and Topic #18. Thus, these topics are explained by all other topics by at least 80% each, so that we exclude these topics from our later analysis. In doing so, we apply a stricter threshold often applied (greater than 10 or 90% is explained by the other topics), since we prefer to have a parsimonious model with fewer variables, which make it less susceptible to spurious relationships and harder to verify that our topics are significant. The VIFs of the remaining variables are within the range of 1.1 and 4.4.

3.6 Data

To test our hypotheses, we combine multiple datasets: (1) the text corpus given by Item 1A of the annual 10-Ks obtained from EDGAR, (2) investors' risk perception proxied by stock volatility from CRSP, as well as (3) firms' financial fundamentals obtained from Compustat or Thomson Reuters, and information besides the pure risk description disclosed in their filings that could affect investor activity.

3.6.1 Textual Corpus

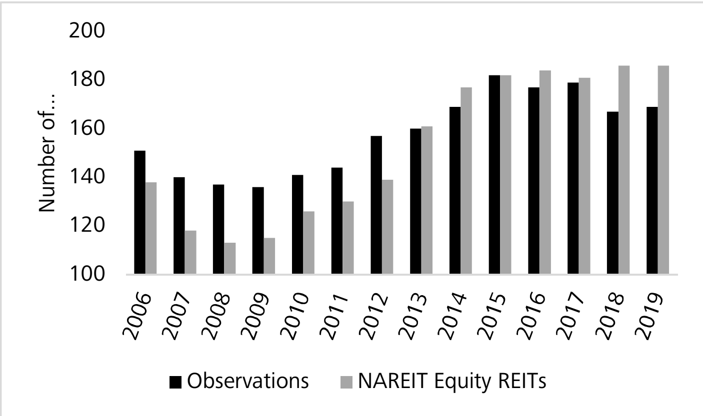
Since this study analyzes mandatory risk disclosures in 10-Ks of US REITs, our sample begins with the earliest date when Item 1A was available (December 1, 2005), and extends through the fiscal year-end 2019.[13] In contrast to other studies focusing on the entire firm-year sample available from the Electronic Data Gathering and Retrieval (EDGAR) database, we limit our examination to a single industry, namely the REIT industry. Our sample consists of all Equity REITs present in the FTSE NAREIT All REITs Index at any point of time during the sample period. Mortgage REITs are excluded from the analysis because they differ in characteristics (e.g., underlying asset, risk structure), exposed risk factors, and are recognized as more difficult to value for external investors (Buttimer et al., 2005). Whereas 25 firms remain in the index throughout the sample period, 221 firms enter, exit, or both enter and exit. The corresponding filings are downloaded from the SEC's EDGAR database. To extract Item 1A, we parse the 10-K forms, employing the edgarWebR package in R, which assigns each line of the report to a particular section of the annual 10-K. Thereafter, the risk factor section can be extracted by searching for the item number (1A). Firm-year observations that lack necessary control variables or stock prices are excluded, resulting in an overall sample of roughly 1,230 observations consisting of 199 unique firms. The limiting variables are the control variables and not the risk factors (see Table 3.1 for more details about *N*).

Figure 3.3 displays the sample composition of the 10-Ks over years; it mostly follows the number of Equity REITs included in the FTSE NAREIT All REITs Index over the same time

[13] Actually, there is a second risk section in the 10-K. Item 7A should list "quantitative and qualitative disclosures about market risk" which are relevant for a company (e.g., interest rate risk or foreign currency exchange risk). However, Item 7A differs from Item 1A in that this section not only names but additionally quantifies the impact of the individual risk factors on future firm performance. Thus, managers usually use numbers to describe how risk factors affect firms' filings in this section. Additionally, with an average length of only 6,680 words, Item 7A is just a tenth of the average length of Item 1A. Given that our method focuses on textual data i.e., the words used to qualitatively describe relevant risks, we exclude Item 7A from the main analyses. This is essential because topic models cannot take numbers into account and shorter documents decrease the robustness of the topic model because it "learns" less from the data (Papilloud and Hinneburg, 2018). However, for reasons of completeness, results for Item 7A are presented in Appendix D.

period. For some years, the observations exceed the number of index constituents, since we include a REIT in our sample if it was a constituent at some point during the period. We thus take into account survivorship bias and address index effects like greater investor and analyst attention to firms listed in an index. The number of observations is also driven by the data availability on CRSP and Compustat.

Figure 3.3: Sample Distribution over Years



This figure shows the number of observations included in the sample and the number of Equity REITs present in the FTSE NAREIT All REITs Index over years.

3.6.2 Metadata Covariates

To apply the STM to the textual corpus, we include the property type of the respective REIT as metadata covariate. We, therefore, assign the property type as classified by CRSP Ziman to each filing. The metadata covariate “Retail” is, for example, accompanied by the words ‘shopping’, ‘goods’, ‘e-commerce’, ‘consumer’, ‘malls’, and ‘anchor’. The words ‘hotels’, ‘leisure’, ‘travelers’, ‘room’, and ‘franchise’ are instead typically associated with the lodging industry. Observations assigned to the category ‘Unknown’, meaning that the firm is not assigned to a type for this year in the Ziman dataset are excluded from the analysis. The group ‘Unclassified’ includes asset classes like Timber, Data Centers, Infrastructure, and Specialty. These STM-derived word sets for each metadata covariate, describe the specifics of each asset class impressively well. Table 3.6 in Appendix B shows the full list of metadata covariates, along with their covariate words.

Contrary, topics identified by LDA highly correspond to the investment types (see Table 3.13 in Appendix C). For example, LDA Topic #1 corresponds to “Health Care”, LDA Topic #4 to “Residential”, and LDA Topic #9 to “Retail” to name a few (see discussion in the Introduction for healthcare vs. residential and Table 3.13 in Appendix C). Moreover, the frequency of appearance for the individual risk topics identified by LDA is closely related to Ziman property types. Specifically, we find that disclosure frequencies are mostly driven by 1-3 property types (see Table 3.14 in Appendix C).

3.6.3 Investors' Risk Perception

Our primary variable of interest is the perceived risk on the stock market. Since risk is typically measured by the stock return volatility, we use volatility after the filing submission date on the dependent side of our model. Volatility is thereby derived from the daily closing prices from CRSP. We assume that investors and analysts need time to process lengthy disclosures and do not immediately update their predictions. However, it is unclear how long it takes until investors read 10-Ks, and new information is incorporated into price changes. Thus, we apply multiple testing periods for firms' stock return volatility after the 10-K filing is published – a 5, 40, and 60 trading-day period. We thereby cover the entire period from the 10-K filing date to the next 10-Q that may report changes in risk factors. The 5 trading-day period gives investors and analysts enough time to read, interpret and react to disclosures while being short enough to minimize the influence of other disruptive events that may also affect volatility. The 60 trading-day period accounts for investors comparing risk factors disclosed in 10-Ks to changes disclosed in 10-Qs.[14]

We define volatility as the standard deviation of daily log returns (over the corresponding testing period) extrapolated to the 5, 40, and 60 trading-day periods after the 10-K filing day.

$$Vola_T = \sqrt{T} * \sqrt{\frac{\sum_{t=1}^{T}(\ln(1+r_t)-\mu_T)^2}{T-1}}, \tag{3.2}$$

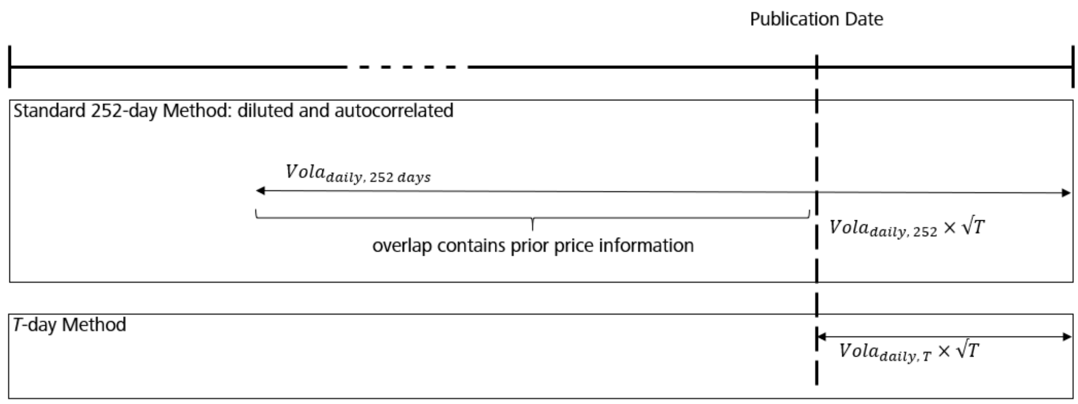
where $T \epsilon \{5, 40, 60\}$.

This procedure enables capturing volatility induced by the information released in the 10-K, but does not, in contrast to the common approach, use the 252 trading-day volatility to calculate current volatility. We see a 252 trading-day window as too diluted since it includes price-sensitive information over the entire prior trading year. Thus, past information that is already known and has been incorporated into prices, would be extrapolated to the testing period. Additionally, the standard deviation over a 252 trading-day window would cause autocorrelation problems after adding lag volatility for the 252 days before the 10-K filing date, since the majority of the time window overlaps, see Figure 3.4, Panel A.

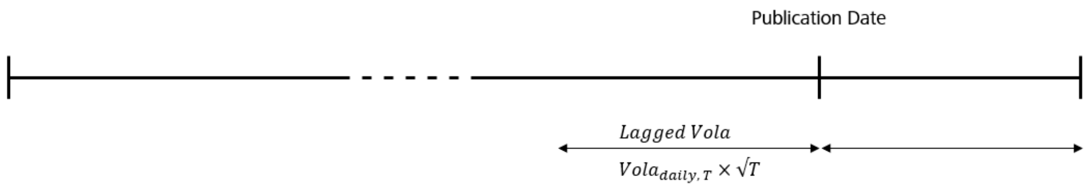
[14] We additionally analyze the 10 and 20 trading-day periods. As expected, the results are in the intermediate ranges.

Figure 3.4: Volatility around Publication Date

Panel A: *T*-day Method



Panel B: Lagged Volatility



This figure contrasts the common approach using the 252 trading-day volatility to calculate current volatility to our *T*-day method (Panel A). Panel B shows the lagged volatility measure.

By contrast, our *T*-day method of volatility testing surveys volatility, starting from the filing publication date until the end of the processing period. To account for the problem of autocorrelation due to volatility clustering around specific dates and other influencing filing events, we include a lagged volatility measure in the model as a control variable. This variable gauges the standard deviation *T* days before the publication date, see Figure 3.4, Panel B.

3.6.4 Control Variables

To control for information beyond the risks revealed in Item 1A that might affect investor activity, a set of control variables is included. Besides firm characteristics, performance, and risk measures, we additionally consider textual 10-K characteristics that previous research has revealed as determinants of return volatility. We describe all control variables below, and provide more specific definitions, including Compustat data items, in Table 3.7 in Appendix B. We cluster the controls into two subsets: accounting-based/market-based and textual.

For the first of the two, we include the REIT-specific performance measure Funds From Operations per share (*FFO/Share*), to incorporate the real-estate-specific income characteristics. Thus, we follow NAREIT’s guidelines by adding amortization and depreciation to the net income and subtracting the net of gains and losses originated by the sale of assets from the net income. Since *FFO/Share* is a performance measure, we expect a negative coefficient sign. The variable *Size*, measured as the natural logarithm of the firm's total assets, is included in each regression. In line with Fama and French (1993), we expect its coefficient to be negative, as small firms are more volatile than large firms. *Leverage* is a common proxy for firm risk, so that we expect the variable to be positively related to volatility. The motivation for the next two factors is purely at the operating level – the annual change in revenue (Δ*REV*) as well as sales growth (*Sales_Growth*). Δ*REV* is defined as current sales or rental income less prior year sales. *Sales_Growth* is calculated as *REV* scaled by total assets in the previous year. We expect a positive influence from both variables.

In the subset of market-based controls, *Beta* proxies the firm risk similar to *Leverage*, so that we expect a positive nexus to volatility. *BTM* is calculated as the book value of equity, scaled by the market capitalization of equity. Our expectations of the coefficient on *Book-to-Market* (*BTM*) are ambiguous. On the one hand, the coefficient could be positive if market participants have little confidence in the future prospects of a firm. On the other hand, the coefficient on *BTM* will be negative if growth opportunities are positively related to firm risk (Fama and French, 1993; Campbell et al., 2014).

Additionally, we include the stock return volatility (*Lag_Vola*) for *T* trading-days before the 10-K filing date, to control for positive correlation in the short-run, and information released in other outlets as the 10-K. We expect a positive relationship between the pre- and post-filing-date volatility. We also include the stock return volatility of the S&P 500 ($Vola^{S\&P}$) for *T* trading-days before the 10-K filing date, as a benchmark for changes in market-level return volatility and expect a positive coefficient. The change of a firms' average daily trading volume from the symmetric period of *T* trading-days before to after the 10-K is filed (Δ*Volume*), serves as a factor of the economic interactions in the financial market. In addition to stock price changes, trading volume conveys important information about the underlying economic forces. We expect that higher changes in the trading volume go in line with higher volatilities. Furthermore, the percentage of institutional ownership (*IO*), defined as the sum of shares held by institutional investors, divided by the shares outstanding, is incorporated as obtained from Thompson Reuters. Institutional investors have higher capacities to process 10-Ks, and thus could react in a timely manner to the disclosed information, causing a positive coefficient on *IO*. Conversely, the coefficient could be negative if the long-term orientation of sophisticated investors is predominant and they behave inertially.

Finally, we include additional controls related to the textual content. Since other research shows that the number of words (in Item 1A) is positively related to stock return volatility after the disclosure date (Campbell et al., 2014), we incorporate the natural logarithm of the total text length of the risk sections (*Text_Length*). To account for higher information-processing costs due to complex language, we additionally incorporate the readability of the risk sections, measured by the Gunning fog index (*FOG*).

3.6.5 Descriptive Statistics

Table 3.1 presents descriptive statistics for all variables – STM's frequencies for the risk factor topics (*Freq_Topic*) of Item 1A and the control variables. STM reveals additionally to the risk-factor topics (β_k), the frequencies for these topics (θ_d), see Section 3.4. These frequencies sum to 1 within each document but not over all documents. We observe rather small topic frequencies for Item 1A by looking at the means; the highest is around 7.6% for Topic #16, the lowest for Topic #14 at 2.2%. An equal distribution over all topics would result in 5% (1/20) for each topic. Focusing on the extreme values (Min and Max), we see that all topics constitute the core of any 10-K filing (lowest Max is 99.8%) or are practically not discussed (highest Min is 0.0004%). The probabilities of all topics are extremely skewed in distribution so that we use a log transformation of these factors in our later regressions, but present here, the percentage figures for easy comprehension. By using the Shapiro and Wilk's test, we can conclude that the logs of the risk factors are normally distributed (Royston, 1982). The correlation coefficients among the logged risk factors are not higher/lower than 0.47/-0.63 (Table 3.8 in Appendix B). Thus, the topics have no direct linear relationship, but as shown in Section 3.5, the VIF for 4 topics (#7, #11, #14, and #18) is high. Thus, these topics are explained substantially by a linear combination of the other topics, so that we exclude them from our later analysis and restrict our model to topics that mostly convey new information.

Table 3.1: Descriptive Statistics

	N	Mean	StDev	Min	Q1	Median	Q3	Max
				Item 1A				
Freq_Topic 1	2,207	5.121	20.447	0.000	0.003	0.007	0.017	99.940
Freq_Topic 2	2,207	5.043	20.626	0.000	0.003	0.007	0.020	99.934
Freq_Topic 3	2,207	2.441	13.409	0.000	0.008	0.018	0.055	99.773
Freq_Topic 4	2,207	3.968	17.793	0.000	0.004	0.012	0.036	99.901
Freq_Topic 5	2,207	3.475	16.227	0.000	0.005	0.014	0.044	99.835
Freq_Topic 6	2,207	4.828	19.686	0.000	0.003	0.009	0.020	99.934
Freq_Topic 7	2,207	3.715	17.584	0.000	0.004	0.010	0.025	99.894
Freq_Topic 8	2,207	4.317	18.118	0.000	0.007	0.014	0.043	99.877
Freq_Topic 9	2,207	4.883	20.521	0.000	0.004	0.008	0.020	99.978
Freq_Topic 10	2,207	4.813	16.571	0.000	0.011	0.024	0.116	99.870

see next page

Table.3.1: continued

Freq_Topic 11	2,207	3.330	15.479	0.000	0.004	0.009	0.025	99.959
Freq_Topic 12	2,207	6.648	23.855	0.000	0.002	0.008	0.024	99.939
Freq_Topic 13	2,207	6.406	22.932	0.000	0.004	0.009	0.028	99.932
Freq_Topic 14	2,207	2.221	13.626	0.000	0.001	0.004	0.012	99.973
Freq_Topic 15	2,207	5.477	21.310	0.000	0.004	0.009	0.022	99.952
Freq_Topic 16	2,207	7.566	25.358	0.000	0.003	0.008	0.019	99.939
Freq_Topic 17	2,207	6.527	23.341	0.000	0.004	0.009	0.023	99.939
Freq_Topic 18	2,207	7.043	23.956	0.000	0.004	0.012	0.036	99.983
Freq_Topic 19	2,207	6.913	23.799	0.000	0.003	0.009	0.025	99.975
Freq_Topic 20	2,207	5.265	21.145	0.000	0.004	0.008	0.020	99.931
Control Variables								
FFO/Share	1,862	170.343	7,264.763	-18.258	0.594	1.386	2.593	313,483.000
Size	2,020	7.759	1.314	-1.931	7.106	7.907	8.558	10.556
Leverage	2,020	0.566	0.181	0.000	0.473	0.560	0.660	1.638
ΔREV	1,876	47.207	204.435	-4,403.782	1.039	21.619	68.020	3,701.640
Sales_Growth	1,862	0.034	0.436	-0.800	0.001	0.011	0.027	16.478
Beta	1,892	0.974	0.495	-0.692	0.622	0.927	1.259	4.661
BTM	1,956	-0.116	3.018	-64.892	-0.049	0.0002	0.001	75.038
IO	1,749	0.760	0.283	0.000	0.637	0.838	0.954	2.383
$Vola^{S\&P}$ (-5, 0 days)	1,543	0.019	0.012	0.002	0.010	0.017	0.025	0.082
$Vola^{S\&P}$ (-40, 0 days)	1,537	0.056	0.030	0.025	0.038	0.047	0.056	0.175
$Vola^{S\&P}$ (-60, 0 days)	1,535	0.068	0.031	0.030	0.052	0.056	0.078	0.193
ΔVolume (0, 5 days)	1,543	0.119	0.893	-4.306	-0.049	0.025	0.183	20.333
ΔVolume (0, 40 days)	1,529	0.052	0.545	-2.601	-0.085	0.001	0.095	7.790
ΔVolume (0, 60 days)	1,519	0.050	0.520	-2.860	-0.082	0.003	0.100	7.646
Text_Length	2,207	68,231	50,034	36	38,302	57,270	87,198	516,463
FOG	2,207	22.460	1.707	5.000	21.665	22.496	23.307	29.698
Dependent Variables								
Vola (0, 5 days)	1,543	0.041	0.047	0.001	0.020	0.032	0.047	1.125
Vola (0, 40 days)	1,537	0.116	0.123	0.030	0.071	0.085	0.110	2.119
Vola (0, 60 days)	1,535	0.142	0.132	0.033	0.088	0.107	0.141	2.130

This table shows the descriptive statistics for the frequencies (in %) for the risk factor topics (*Freq_Topic*) of Item 1A, further control variables, and dependent variables (*Vola*). The definition of all variables is presented in Table 3.7 in Appendix B. *N* is the number of observations, StdDev stands for standard deviation, Q1 is the first and Q3 the third quartile of the distribution, and Min is the minimum and Max the maximum of each variable. *N* is set to the maximal available number of observations for each variable.

The classical fundamentals in the control set show the normal values and are comparable with other REIT studies (e.g., Doran et al., 2012; Price et al., 2017; Koelbl, 2020). The percentage of institutional ownership (*IO*) is on average 76%, with an interquartile range from 64% to 95%. The restriction to shares outstanding in the denominator results in extreme ratios of greater than 1 for a few observations where the institutional investors own more than the outstanding shares. The *Text_Length* counted by words included in Item 1A varies in the interquartile range from 38,302 to 87,198. The extreme values are surprising; the shortest Item 1A has only 36 words, whereas the longest has 516,463 words. The low number of words is driven by small REITs which do not have to publish risk reports according to the SEC requirements; see Example 1-2 in Table 3.9 in Appendix B. In total, we have only 8 reports with fewer than 1600 characters (including stop words) for their reports; see Example 3 in Table 3.9 for a short Item 1A with 374 words. The readability of the text, as measured by the Gunning fog index, is complex. The interquartile range is close with 21.7 to 23.3 and higher than the reading level of a colleague graduate given by 17. What is surprising is the low minimum with 5.0, probably induced by the short reports mentioned above, since the value 10 is only at the level of a high school sophomore (usually aged 15-16).

3.7 Results

3.7.1 Validity of STM to explain Investor Risk Perception

Recognizing that firm executives are free to decide which risks to disclose and how much to discuss a specific risk factor, we hypothesize that the probability of appearance of the STM extracted risk factors helps to explain the perceived risk on the stock market. To test our Hypothesis 1, we regress disclosure frequencies for each risk factor topic on the stock return volatility. We run three model specifications, for which we alternate the dependent variable (*Vola*) according to the time horizon of investor risk perception – 5 trading days (Model 1), 40 trading days (Model 2), and 60 trading days (Model 3) after the respective 10-K filing was published.

After controlling for firm-level characteristics and other textual measures from 10-Ks that have been shown to be associated with volatility in previous studies, we find that the STM extracted risk factors help to explain investor risk perceptions. In line with the efficient market hypothesis, risk factors should mostly be relevant in the short-run, whereas fundamentals dominate the risk perception of investors (after the information of 10-Ks has been incorporated into prices) in the long run. Indeed, 12 of 16 risk topics are significantly associated with volatility in the short-run (5-day window).

If we move to 40 trading days, we observe lower coefficients for the risk factors with the exemption of Topic #1 and #15 which increase. Furthermore, we observe fewer significant coefficients in the long run – 6 factors for 40 days. Regarding the fundamentals, some are never relevant (*FFO/Share*, Δ*REV*, and *Sales_Growth*), others increase their impact over time and mitigate the impact of risk factors. *Leverage* is the only fundamental variable that is significant in the short-run, but insignificant in the long run. This is not surprising since *Beta* already incorporates a large part of the risk. The ratio of institutional owners (*IO*), volatility of the last trading days (*Lag_Vola*), and trading volume (Δ*Volume*) also increase their impact over time. Textual variables (*Text_Length* and *FOG*) are never relevant so that the risk factors convey the information of Item 1A. In the 60-day-window model, 7 risk factors are significant and the coefficients have a comparable magnitude compared to the

40 trading-day period. However, the goodness of fit improves with R^2 increasing to 27% from 18%. This effect is mostly driven by controls (*IO*, *Beta*, and *Lag_Vola*).

Table 3.2: Probability of Appearance – Risk Perception

	Model 1 (0, 5 days)	**Model 2** (0, 40 days)	**Model 3** (0, 60 days)
Freq_Topic 1	-0.006***	-0.015***	-0.014***
Transaction	(0.002)	(0.005)	(0.005)
Freq_Topic 2	0.031***	0.028***	0.030***
Regulation	(0.003)	(0.007)	(0.007)
Freq_Topic 3	-0.011***	-0.004	-0.006
Business Process	(0.002)	(0.004)	(0.004)
Freq_Topic 4	0.039***	0.028***	0.031***
Unsecured Claims and Debts	(0.003)	(0.007)	(0.007)
Freq_Topic 5	0.009***	0.008*	0.008
Rating	(0.002)	(0.005)	(0.005)
Freq_Topic 6	-0.0003	-0.008*	-0.009**
Tax and Capital Contribution	(0.002)	(0.005)	(0.004)
Freq_Topic 8	-0.010***	-0.005	-0.008**
Capital Products and Market	(0.002)	(0.003)	(0.003)
Freq_Topic 9	0.002	-0.004	-0.004
Acquisition	(0.002)	(0.004)	(0.004)
Freq_Topic 10	-0.003***	0.001	0.001
Contingencies	(0.001)	(0.002)	(0.002)
Freq_Topic 12	0.0001	-0.005	-0.004
IT	(0.001)	(0.003)	(0.003)
Freq_Topic 13	-0.017***	-0.008	-0.010**
Legal & Litigation Risk	(0.002)	(0.005)	(0.005)
Freq_Topic 15	-0.013***	0.009**	0.010**
Single Tenant Risk	(0.002)	(0.004)	(0.004)
Freq_Topic 16	-0.007***	-0.003	-0.005
Property	(0.002)	(0.004)	(0.004)
Freq_Topic 17	-0.004**	-0.004	-0.004
Politics	(0.002)	(0.004)	(0.004)
Freq_Topic 19	-0.013***	-0.005	-0.006
Cash-flow	(0.002)	(0.005)	(0.005)
Freq_Topic 20	0.003	0.002	0.003
Property	(0.002)	(0.004)	(0.004)
FFO/Share	0.0005	0.002	0.001
	(0.001)	(0.002)	(0.002)
Size	0.002	0.013*	0.014*
	(0.003)	(0.007)	(0.007)
Leverage	0.025**	0.016	0.006
	(0.013)	(0.028)	(0.027)
ΔREV	0.00000	-0.00001	-0.00002
	(0.00001)	(0.00002)	(0.00002)
Sales_Growth	0.005	-0.004	-0.004
	(0.004)	(0.009)	(0.009)

see next page

Table 3.2: continued

Beta	0.009***	0.024***	0.013*
	(0.003)	(0.008)	(0.008)
BTM	-0.020***	0.056***	0.066***
	(0.002)	(0.006)	(0.006)
IO	-0.018***	-0.043***	-0.036***
	(0.006)	(0.013)	(0.013)
Lag_Vola	0.350***	0.352***	0.542***
	(0.038)	(0.062)	(0.045)
$Vola^{S\&P}$	0.168	0.079	0.316
	(0.123)	(0.231)	(0.212)
Δ*Volume*	0.008***	0.022***	0.022***
	(0.002)	(0.004)	(0.004)
Text_Length	-0.005	0.014	0.011
	(0.004)	(0.010)	(0.009)
FOG	-0.0001	-0.0003	-0.0003
	(0.002)	(0.004)	(0.004)
N	1,228	1,224	1,223
R^2	0.318	0.182	0.272

This table presents the results of fixed-effect models controlling for unobserved firm and time effects for Item 1A. The table reports panel regression results of fixed effects models, which include coefficients and standard errors (in parentheses) of determinants affecting investor's risk perception. The dependent variable (*Vola*) takes a different number of trading days after the 10-K filing date into account – 5 trading days (Model 1), 40 trading days (Model 2), and 60 trading days (Model 3). The definition of all variables is presented in Table 3.7 in Appendix B.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This result is contrary to Cohen et al. (2020) who find that corporate reports of non-REITs impact stock prices with a lag. One possible explanation for the diverging results might be that REITs generally attract more institutional investors, who are able to process official firm filings faster, and thus react more quickly to the disclosed information. By using our findings, we conclude that the topic distribution presents a valid approach to quantify risk disclosures and explain investor risk perception, however, the risk factors' influence is mostly in the short-run.

3.7.2 Risk Disclosures can be Good News

To analyze Hypothesis 2, which predicts a risk-reducing effect for the majority of risk factors, we assign the STM-extracted risk factors to the competing arguments of Kravet and Muslu (2013). Consistent with Bao and Datta (2014), our results provide support for all three arguments, depending on the specific risk types disclosed but primarily reduce the volatility (convergence argument).

In the shortest 5-trading-day model, four risk-factor topics #6, #9, #12, and #20, have an insignificant coefficient, thus supporting the null argument of an uninformative risk factor. Three risk factors, including topics #2, #4, and #5 are positively associated with stock return volatility (divergence argument). Specifically, after controlling for other factors causing stock price movements, a 1% increase in the probability of appearance for topics #2, #4, and #5 will lead to a 3.1, 3.9, and 0.9 basis point increase of the stock return volatility after filing the 10-K, at the 1% significance level.[15] In line with the assumption that executives use 10-Ks to resolve firms' known risk factors or give more facts about known risk factors and thus, reduce risk perceptions, the convergence factors are in the majority (topics #1, #3, #8, #10, #13, #15, #16, #17 and #19). For example, after controlling for the effect of other variables, a 1% increase in disclosure about topics #1, #3, and #8, will lead to a 0.6, 1.1, and 1.0 basis point decrease in volatility after the disclosure date, respectively. Thus, instead of presenting new risk factors, executives use Item 1A extensively to resolve the REITs' known contingencies, thereby reducing investor risk perceptions. The predominance of the convergence arguments suggests that executives' concerns of adverse effects of disclosures are baseless – risk disclosures may even be good news as long as they clarify the implications of already known risk.

In contrast to prior efforts quantifying the information revealed in risk disclosures (i.e., length of Item 1A or the number of risk keywords), an additional advantage of our approach is that it can be interpreted economically, as the STM not only provides frequencies of appearance, but also the corresponding set of words representing the topic. Our results indicate, that risk factors talking about Tax and Capital Contribution, Acquisition, IT, and Property (#6, #9, #12, and #20) have no effect on stock return volatility after the filing submission date. The risk factor topics supporting the divergence argument comprise Regulation, Unsecured Claims and Debts, and Rating (#2, #4, and #5). The convergence factors cover the topics Transaction, Business Process, Capital Products and Market, Contingencies, Legal & Litigation Risk, Single Tenant Risk, Property, Politics, and Cash-flow (#1, #3, #8, #10, #13, #15, #16, #17, and #19). However, STM only provides the set of words representing the risk factor while researchers choose the label. Therefore, labels may not represent topics appropriately. Israelsen (2014) gets to the heart of this dilemma by stating that "it is the words that define the topics, not the title".

3.7.3 Probability of Appearance vs. Absolute Allocation of Words

So far, our analyses focus on the probability of appearance of risk factor topics and ignores the number of words a firm allocates towards a specific risk. For example, even in the extreme case that a firm describes litigation risk with 100% within its 10-word long risk

[15] Given that the STM mostly identifies few topics that represent the majority of the filing, the actual impact of a specific topic on the firms' volatility is higher because the typical increase of the probability of occurrence for a risk factor is higher than our used example of 1%.

disclosure, it seems that this risk is for this firm much less material than for another firm that allocates 20% of its 1000-word long disclosure towards litigation risk. We adapt our target variables by multiplying the probability of appearance for each risk factor (*Freq_Topics*) with the total length of the corresponding disclosure (*Text_Length*). This approach presents a hybrid model using machine learning and widely used wordcount methods. We regress the log transformation of the new target variable (*Abs_Allocation*) on the stock return volatility following the 5, 40, and 60 trading-day windows. The descriptive statistics of *Abs_Allocation* are given in Table 3.3 and the results of the regression model which follows Equation (3.1) are in Table 3.4.

Table 3.3: Descriptive Statistics – Absolute Allocation of Words

	N	Mean	StDev	Min	Q1	Median	Q3	Max
				Item 1A				
Abs_Allocation 1	2,157	4,784.894	20,770.350	0.025	1.368	3.466	9.380	211,302.900
Abs_Allocation 2	2,157	3,180.996	14,577.500	0.001	1.536	3.854	11.476	138,226.600
Abs_Allocation 3	2,157	899.028	6,324.319	0.106	4.675	10.062	28.268	133,751.700
Abs_Allocation 4	2,157	2,200.952	11,153.190	0.003	2.174	6.535	21.225	108,071.900
Abs_Allocation 5	2,157	1,680.289	8,918.072	0.104	3.044	7.509	21.734	142,100.100
Abs_Allocation 6	2,157	4,300.814	20,565.220	0.053	1.514	4.321	11.861	175,507.700
Abs_Allocation 7	2,157	2,074.562	10,261.370	0.001	2.203	5.334	14.812	97,628.020
Abs_Allocation 8	2,157	2,005.718	8,796.460	0.207	4.073	8.368	21.142	87,897.500
Abs_Allocation 9	2,157	4,258.056	23,163.760	0.057	1.985	4.361	9.766	358,091.100
Abs_Allocation 10	2,157	2,517.047	8,149.857	0.156	6.277	12.305	48.238	72,535.240
Abs_Allocation 11	2,157	2,618.542	13,752.160	0.001	1.997	5.100	15.108	186,137.400
Abs_Allocation 12	2,157	3,524.577	14,625.800	0.0001	1.418	4.120	12.151	132,529.400
Abs_Allocation 13	2,157	4,080.354	16,148.920	0.001	1.704	4.595	14.166	173,824.100
Abs_Allocation 14	2,157	2,124.229	14,972.500	0.001	0.593	2.183	6.113	180,428.300
Abs_Allocation 15	2,157	4,613.534	20,843.580	0.023	2.390	5.010	12.168	241,480.400
Abs_Allocation 16	2,157	4,252.121	16,687.200	0.071	1.798	4.441	11.206	159,719.300
Abs_Allocation 17	2,157	4,191.365	16,482.040	0.161	2.496	4.602	12.725	126,125.000
Abs_Allocation 18	2,157	4,892.229	26,515.550	0.001	2.442	7.021	20.794	516,358.900
Abs_Allocation 19	2,157	6,162.992	31,754.840	0.041	1.925	4.782	12.686	410,365.500
Abs_Allocation 20	2,157	3,981.453	17,581.560	0.138	2.208	4.583	10.895	137,661.800

This table shows the descriptive statistics for the frequencies (in %) for the risk factor topics multiplied by the total length of the corresponding disclosure (*Abs_Allocation*). *N* is the number of observations, StDev stands for standard deviation, Q1 is the first and Q3 the third quartile of the distribution, and Min is the minimum and Max the maximum of each variable. *N* is set to the maximal available number of observations for each variable.

Table 3.4: Absolute Allocation of Words – Risk Perception

	Model 1 (0, 5 days)	**Model 2** (0, 40 days)	**Model 3** (0, 60 days)
Abs_Allocation 1	-0.007***	-0.016***	-0.015***
Transaction	(0.002)	(0.005)	(0.005)
Abs_Allocation 2	0.032***	0.027***	0.030***
Regulation	(0.003)	(0.007)	(0.007)
Abs_Allocation 3	-0.011***	-0.005	-0.006
Business Process	(0.002)	(0.004)	(0.004)

see next page

Table 3.4: continued

Abs_Allocation 4	0.038***	0.029***	0.031***
Unsecured Claims and Debts	(0.003)	(0.007)	(0.007)
Abs_Allocation 5	0.009***	0.010**	0.008*
Rating	(0.002)	(0.005)	(0.005)
Abs_Allocation 6	-0.001	-0.009**	-0.009**
Tax and Capital Contribution	(0.002)	(0.004)	(0.004)
Abs_Allocation 8	-0.010***	-0.006	-0.008**
Capital Products and Market	(0.002)	(0.003)	(0.003)
Abs_Allocation 9	0.002	-0.004	-0.004
Acquisition	(0.002)	(0.004)	(0.004)
Abs_Allocation 10	-0.002***	0.001	0.001
Contingencies	(0.001)	(0.002)	(0.002)
Abs_Allocation 12	0.00001	-0.005*	-0.004
IT	(0.001)	(0.003)	(0.003)
Abs_Allocation 13	-0.017***	-0.008*	-0.010**
Legal & Litigation Risk	(0.002)	(0.005)	(0.005)
Abs_Allocation 15	-0.012***	0.009*	0.010**
Single Tenant Risk	(0.002)	(0.004)	(0.004)
Abs_Allocation 16	-0.007***	-0.002	-0.005
Property	(0.002)	(0.004)	(0.004)
Abs_Allocation 17	-0.005***	-0.004	-0.004
Politics	(0.002)	(0.004)	(0.004)
Abs_Allocation 19	-0.012***	-0.005	-0.006
Cash-flow	(0.002)	(0.005)	(0.005)
Abs_Allocation 20	0.003	0.003	0.004
Property	(0.002)	(0.004)	(0.004)
FFO/Share	0.001	0.001	0.001
	(0.001)	(0.002)	(0.002)
Size	0.001	0.012*	0.013*
	(0.003)	(0.007)	(0.007)
Leverage	0.029**	0.018	0.007
	(0.012)	(0.028)	(0.027)
ΔREV	0.00000	-0.00001	-0.00002
	(0.00001)	(0.00002)	(0.00002)
Sales_Growth	0.005	-0.004	-0.006
	(0.004)	(0.009)	(0.009)
Beta	0.009***	0.024***	0.015**
	(0.003)	(0.007)	(0.007)
BTM	-0.020***	0.056***	0.066***
	(0.002)	(0.006)	(0.006)
IO	-0.018***	-0.044***	-0.038***
	(0.006)	(0.013)	(0.013)
Lag_Vola	0.354***	0.328***	0.521***
	(0.037)	(0.058)	(0.043)
$Vola^{S\&P}$	0.866***	1.610***	1.290***
	(0.133)	(0.291)	(0.305)
ΔVolume	0.007***	0.020***	0.020***
	(0.002)	(0.004)	(0.004)

see next page

Table 3.4: continued

Text_Length	-0.005	0.002	-0.001
	(0.005)	(0.010)	(0.010)
FOG	-0.0003	-0.00005	0.00004
	(0.002)	(0.004)	(0.004)
N	1,228	1,224	1,223
R^2	0.345	0.207	0.283

This table presents the results of fixed-effect models controlling for unobserved firm and time effects for Item 1A. The table reports panel regression results of fixed effects models, which include coefficients and standard errors (in parentheses) of determinants affecting investor's risk perception. The dependent variable (*Vola*) takes a different number of trading days after the 10-K filing date into account – 5 trading days (Model 1), 40 trading days (Model 2), and 60 trading days (Model 3). The definition of all variables is presented in Table 3.7 in Appendix B.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Consistent with previous findings 12 of 16 risk topics are significantly associated with volatility in the short-run (5-day window). Again, the risk factor influence varies over the windows. Comparable to the probability model (Section 3.7.1), we observe lower significant coefficients for the risk factors if we move to 40 trading days (8 risk factors instead of 6) or to 60 trading days (8 risk factors instead of 7). In comparison to the probability model (*Freq_Topics*), the absolute allocation of words model explains the variations better; the R^2 is higher for all windows. For example, the model with *Abs_Allocation* explains around 35% of the variation for the 5-day window, whereas *Freq_Topics* explains 32%. The goodness of fit decreases for longer windows – 21% for 40 days and 28% for 60 days – but remains higher than all models using *Freq_Topics*. The classification of risk factors to the competing arguments of Kravet and Muslu (2013) remains unchanged.

Based on the comparable coefficients and the higher explanatory power for the *Abs_Allocation* model, we evaluate this hybrid model as a good alternative to combine machine learning with a classical factor. Thereby, a combination of the number of words and machine-assisted topic modeling helps to explain investor risk perceptions most efficiently. The topics are most important for a short window even after controlling for traditional firms-specific accounting and market control variables.

3.7.4 Alternative of Risk Perception and Alternative Topic Modeling

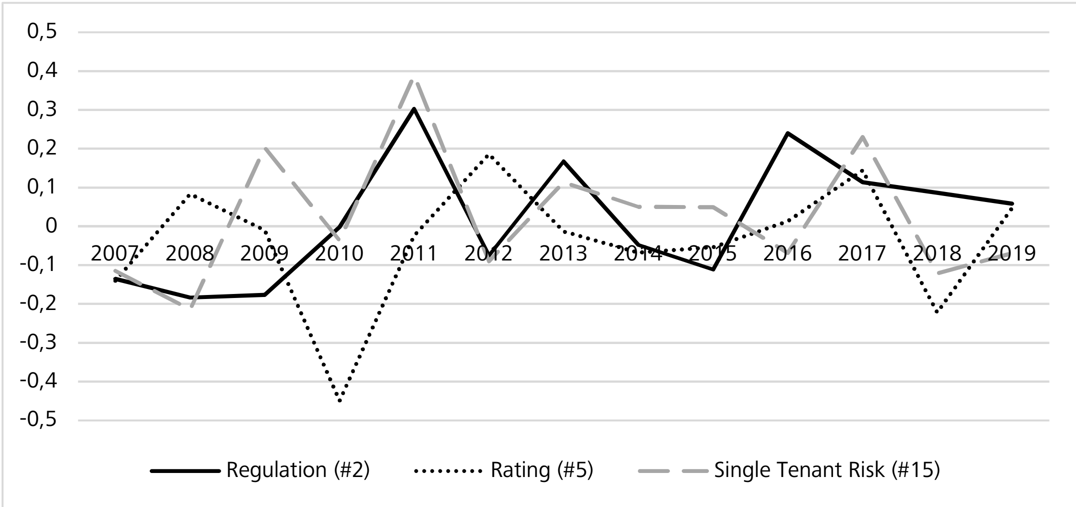
To examine the robustness of our finding that the majority of the risk factors follow the convergence argument, we alter the measure of risk perception and topic modeling approach. For the alternative measure of risk, we follow Kravet and Muslu (2013) and re-run our analysis using the change in the standard deviation of a firms' daily stock returns from the symmetric period of *T* trading-days before to after the 10-K is filed. For example, Kravet and Muslu (2013) calculate the difference between the volatility during the first 60 trading days after the filings and the last 60 trading days before the filings. Higher volatility after the filing goes in line with the divergence argument whereas lower volatility is supported by the convergence argument. Our results are robust to this alternated dependent variable since all coefficients' signs are the same and their magnitudes have a comparable size (see Table 3.10 in Appendix B). Thus, our conclusion that most risk factors follow the convergence argument applies even after using a different measure of risk perception, too.

After presenting an alternative for the dependent side, we change the topic extracting process on the independent side, too. Even if Blei and Lafferty (2007) and Roberts et al. (2014) show that STM and CTM are superior to LDA, we want to stress our results and use all three topic model approaches for our best model (*Abs_Allocation*). Within this robustness check, we additionally run regressions for STM, CTM, and LDA extracted risk factor topics over the 5 trading-day and 60 trading-day periods. Note that the model-specific topics are not directly comparable since their words are different. In the short-run, LDA identifies three risk factors and CTM four risk factors that are significantly associated with investor risk perception; these numbers are lower than the twelve factors for STM. STM also leads in the long run with eight significant risk factors, CTM has no significant factor, and LDA two factors. This relatively low number could also be induced by randomness around the t-value and not from the economic significance of the factors. Additionally, the goodness of fit is highest for STM for both time windows. Thus, we conclude that our empirical findings confirm the theoretical and empirical derived superiority of STM (see Section 3.4) as a topic model approach. The results are presented in Table 3.12 in Appendix B.

3.7.5 Validity of the STM to capture Changes in Reporting Behavior

The lessons of the financial crisis (2007-2009) and the strengthened disclosure requirements of the SEC, changed the reporting behavior of companies. To further assess the validity of our method, we analyze whether the STM identified probabilities of appearance are capable of capturing these changes in 10-Ks. To conduct the analysis, we calculate the yearly growth rate of the probability of appearance for each of the risk factors over all firms. Figure 3.5 illustrates these growth rates for selected topics whose reporting certainly changed during or after the crisis: Regulation (#2), Single Tenant Risk (#15), and Rating (#4).

Figure 3.5: Yearly Growth Rate of the Probability of Appearance



This figure shows yearly growth rates of the probability of appearance for the topics Regulation (#2), Single Tenant Risk (#15), and Rating (#4).

We observe that topic #2 Regulation had decreased before/during the crisis and increased in the aftermath, representing strengthened regulatory requirements after the crisis. Contrary, Single Tenant Risk (#15) peaked in 2009 and 2011 and has increased on average in the aftermath of the global financial crisis. This might be due to strengthened disclosure requirements, or it showcases that risk factors become immanent or even real threats for the company during an economic crisis. Rating (#4) dropped in the year 2010 and has oscillated since then around zero. This trend may reflect the loss of confidence in rating agencies following the events of 2007 and 2008. In summary, probabilities of appearance are time-varying and deviate from their previous level when specific events (e.g., global financial crisis) occur. Thus, disclosure frequencies reflect changes in firms' reporting behavior caused by specific events, confirming the validity of the STM.

3.8 Conclusion

Firms have to inform their shareholders about the expected implications and consequences of adverse events so that the investors are able to monitor the current and future risk factors a firm is facing and integrate them into their decision-making analysis. Specifically, the SEC mandates firms to discuss the most relevant factors that may entail speculative or risky aspects for the firm in their 10-Ks.

Recognizing the temporal and cognitive limitation of human investors to read and react to the massive amount of text, we exploit unsupervised machine-learning approaches (STM, CTM, and LDA), allowing the user to identify and quantify the risk factors discussed in

REITs' 10-Ks. However, since LDA is limited when identifying common risk factors across industries, we focus on the advanced topic modeling approaches (STM and CTM). Hereby, STM has demonstrated superiority over CTM.

To assess whether our machine-assisted topic detecting modeling presents a valid approach to quantify risk in narrative form, we analyze whether the STM extracted risk factors help to explain the perceived risk on the stock market. Indeed, we find that the majority of risk topics is significantly associated with volatility, confirming the effectiveness of our model. Furthermore, we disentangle how the identified risk topics explain investor risk perception, in addition to traditional firm characteristics. Consistent with Bao and Datta (2014), we find evidence supporting all three competing arguments provided by Kravet and Muslu (2013). Specifically, whereas four risk factors support the null argument of uninformative disclosures, three risk factors reveal previously unknown contingencies to investors, thus increasing their risk perceptions (divergence argument), and the majority (9 risk factors) decreases risk perceptions (convergence argument). The predominance of risk-reducing risk factors is in line with expectations since most risk factors are revealed in a timely manner through press releases or Form 8-K. Thus, instead of disclosing new risk factors, Item 1A primarily provides essential information on risk factors that have already been communicated to investors using more frequent channels. Consequently, it seems like executives' concerns of adverse effects of disclosing "negative" information are baseless and risks described in 10 Ks can indeed be considered good news as long as executives clarify the implications of already known risk. Assuming that the number of words a firm allocates towards a specific topic is also important, we test our hypotheses using the absolute allocation of words rather than pure disclosure frequencies. The results are in line with previous analyses. Given that this model explains investor risk perceptions most efficiently, we evaluate the combination of machine learning and a classical factor as beneficial. We further analyze whether the STM identified probabilities of appearance are capable of capturing changes in firms' reporting behavior during or after extreme events such as the global financial crisis (2007-2009). Indeed, the disclosure frequencies change for the topics Regulation (#2), Single Tenant Risk (#15), and Rating (#4). The response of disclosure frequencies to specific events confirms the validity of the STM.

To the best of our knowledge, this is the first study to employ STM in the accounting and finance domain. Clearly, investors would benefit from using machine-learning techniques allowing them to process a huge amount of company information simultaneously, just as the risk items in the 10-Ks. However, STM is not suitable for the entire business sector universe in one model since it is limited in the application of covariates. It is thereby applicable to a variety of industries whose companies operate in the same business and are exposed to various risk factors. The banking sector comprises, for example, commercial banks and investment banks but are exposed to country-specific regulations and accounting standards. Similar to property types in the REIT sector, the sector-specific vocabulary distracts common approaches (LDA and CTM) from discovering underlying topics, whereas the STM accomplishes that.

Further research might focus on other industries (e.g., the banking sector) or investigate other variables that could be affected by risk disclosures, such as bid-ask spread or trading volume. Additionally, it is worth examining whether risk disclosures provide (negative) signals regarding future performance or liquidity.

3.9 References

Antweiler, W. and Frank, M.Z. (2004), "Is all that talk just noise? The information content of Internet stock message boards", *Journal of Finance*, Vol. 59 No. 3, pp. 1259-1294.

Bao, Y. and Datta, A. (2014), "Simultaneously discovering and quantifying risk types from textual risk disclosures", *Management Science*, Vol. 60 No. 6, pp. 1371–1391.

Blei, D.M. (2012), "Probabilistic topic models", *Communications of the ACM*, Vol. 55 No. 4, pp. 77–84.

Blei, D. M. and Lafferty, J. D. (2007), "A correlated topic model of Science", *Annals of Applied Statistics,* Vol. 1 No. 1, pp. 17-35.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022.

Buttimer, R.J., Hyland, D.C. and Sanders, A.B. (2005), "Real Estate REITs, IPO Waves and Long-Run Performance", *Real Estate Economics*, Vol. 33 No. 1, pp. 51–87.

Campbell, J.L., Chen, H., Dhaliwal, D.S., Lu, H. and Steele, L.B. (2014), "The information content of mandatory risk factor disclosures in corporate filings", *Review of Accounting Studies*, Vol. 19, pp. 396–455.

Chen, H., De, P., Hu, Y. and Hwang, B.H. (2014), "Wisdom of crowds: the value of stock opinions transmitted through social media", *Review of Financial Studies*, Vol. 27 No. 5, pp. 1367-1403.

Cohen, L., Malloy, C., and Nguyen, Q. (2020), "Lazy Prices", *Journal of Finance*, Vol. 75 No. 3, pp. 1371-1415.

Danielsen, B.R., Harrison, D.M., Van Ness, R.A. and Warr, R.S. (2009), "REIT auditor fees and financial market transparency", *Real Estate Economics*, Vol. 37 No. 3, pp. 515-557.

Das, S.R. and Chen, M.Y. (2007), "Yahoo! For Amazon: sentiment extraction from small talk on the web", *Management Science*, Vol. 53 No. 9, pp. 1375-1388.

Davis, A.K. and Tama-Sweet, I. (2012), "Managers' use of language across alternative disclosure outlets: earnings press releases versus MD&A", *Contemporary Accounting Research*, Vol. 29 No. 3, pp. 804-837.

Doran, J.S., Peterson, D.R. and Price, S.M. (2012), "Earnings conference call content and stock price: the case of REITs", *Journal of Real Estate Finance and Economics*, Vol. 45 No. 2, pp. 402-434.

Dyer, T., Lang, M. and Stice-Lawrence, L. (2017), "The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation", *Journal of Accounting and Economics*, Vol. 64 No. 2–3, pp. 221–245.

Fama, E.F. and French, K.R. (1993), "Common risk factors in the returns on stocks and bonds", *Journal of Financial Economics*, Vol. 33 No. 1, pp. 3–56.

Feldman, R., Govindaraj, S., Livnat, J. and Segal, B. (2010), "Management's tone change, post earnings announcement drift and accruals", *Review of Accounting Studies*, Vol. 15 No. 4, pp. 915–953.

Garcia, D. (2013), "Sentiment during recessions", *Journal of Finance*, Vol. 68 No. 3, pp. 1267-1300.

Gaulin, M. (2017), "Risk Fact or Fiction: The information content of risk factor disclosures", Working Paper, Rice University.

Israelsen, R.D. (2014), "Tell It Like It Is: Disclosed Risks and Factor Portfolios", available at: https://doi.org/10.2139/ssrn.2504522.

Jegadeesh, N. and Wu, D. (2013), "Word power: A new approach for content analysis", *Journal of Financial Economics*, Vol. 110 No. 3, pp. 712–729.

Koelbl, M. (2020), "Is the MD&A of US REITs informative? A textual sentiment study", *Journal of Property Investment and Finance*, Vol. 38 No. 3, pp. 181–201.

Kravet, T. and Muslu, V. (2013), "Textual risk disclosures and investors' risk perceptions", *Review of Accounting Studies*, Vol. 18 No. 4, pp. 1088–1122.

Kuhn, K.D. (2018), "Using structural topic modeling to identify latent topics and trends in aviation incident reports", *Transportation Research Part C: Emerging Technologies*, Vol. 87, pp. 105–122.

Lehavy, R., Li, F. and Merkley, K. (2011), "The effect of annual report readability on analyst following and the properties of their earnings forecasts", *Accounting Review*, Vol. 86 No. 3, pp. 1087–1115.

Li, F. (2006), "Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports?", available at: https://doi.org/10.2139/ssrn.898181.

Li, F. (2008), "Annual report readability, current earnings, and earnings persistence", *Journal of Accounting and Economics*, Vol. 45 No. 2–3, pp. 221–247.

Li, F. (2010), "The information content of forward- looking statements in corporate filings - A naïve bayesian machine learning approach", *Journal of Accounting Research*, Vol. 48 No. 5, pp. 1049-1102.

Lopez-Lira, A. (2019), "Risk Factors That Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns", available at: http://dx.doi.org/10.2139/ssrn.3313663.

Loughran, T. and McDonald, B. (2011), "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks", *Journal of Finance*, Vol. 66 No. 1, pp. 35-65.

Loughran, T. and McDonald, B. (2014), "Measuring readability in financial disclosures", *Journal of Finance*, Vol. 69 No. 4, pp. 1643–1671.

Loughran, T. and McDonald, B. (2015), "The Use of Word Lists in Textual Analysis", *Journal of Behavioral Finance*, Vol. 16 No. 1, pp. 1–11.

Loughran, T. and McDonald, B. (2020), "Textual Analysis in Finance", available at: http://dx.doi.org/10.2139/ssrn.3470272

Miller, B.P. (2010), "The effects of reporting complexity on small and large investor trading", *Accounting Review*, Vol. 85 No. 6, pp. 2107–2143.

Nelson, K.K. and Pritchard, A.C. (2016), "Carrot or Stick? The Shift from Voluntary to Mandatory Disclosure of Risk Factors", *Journal of Empirical Legal Studies*, Vol. 13 No. 2, pp. 266–297.

Papilloud, C. and Hinneburg, A. (2018), "Qualitative Textanalyse Mit Topic-Modellen. Eine Einführung für Sozialwissenschaftler" Springer, Wiesbaden, Germany.

Price, S.M., Seiler, M.J. and Shen, J. (2017), "Do investors infer Vocal cues from CEOs during quarterly REIT conference calls?", *Journal of Real Estate Finance and Economics*, Vol. 54 No. 4, pp. 515-557.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000), "Inference of Population Structure Using Multilocus Genotype Data", *Genetics*, Vol. 155 No. 2, pp. 945-959.

Roberts M.E., Stewart B.M., Tingley D., Lucas C., Leder-Luis J., Gadarian S.K., Albertson B., Rand D.G. (2014), "Structural Topic Models for Open-Ended Survey Responses", *American Journal of Political Science*, Vol. 58 No.4, pp. 1064–1082.

Roberts, M., Stewart, B. and Tingley, D. (2019), "stm: R Package for Structural Topic Models", *Journal of Statistical Software*, Vol. 91 No. 2, pp. 1-40.

Royston, J.P. (1982), "An Extension of Shapiro and Wilk's W Test for Normality to Large Samples", *Applied Statistics*, Vol. 31, No. 2, pp. 115-124.

Schofield, A. and Mimno, D. (2016), "Comparing Apples to Apple: The Effects of Stemmers on Topic Models", *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 287–300.

Scholes, M. and Williams, J. (1977), "Estimating betas from nonsynchronous data", *Journal of Financial Economics*, Vol. 5 No. 3, pp. 309–327.

SEC (2005), Securities and exchange commission final rule, release no. 33–8591 (FR-75). http://www. sec.gov/rules/final/33-8591.pdf

SEC (2012), Securities and exchange commission, Investor Bulletin: How to Read an 8-K. https://www.sec.gov/files/readan8k.pdf

Tetlock, P.C. (2007), "Giving content to investor sentiment: the role of media in the stock market", *Journal of Finance*, Vol. 62 No. 3, pp. 1139-1168.

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. (2008), "More than words: quantifying language to measure firms' fundamentals", *Journal of Finance*, Vol. 63 No. 3, pp. 1437-1467.

Wooldridge, J.M. (2010), "Econometric Analysis of Cross Section and Panel Data", 2nd ed., MIT Press, Cambridge, MA.

3.10 Appendix A

Technical Details on the STM

The Structural Topic Modeling (STM) by Roberts et al. (2019) incorporates metadata of pre-specified covariates to disentangle the unique topics. The covariates cover for topical prevalence, topical content, or both. The former affects how much a topic is discussed (θ_d), whereas the latter affects which words are used to discuss a particular topic parameter (β_k) (Roberts et al., 2014). In order to allow the algorithm to find topics beyond the already known identifiers (see Figure 3.1 and discussion in the Introduction for healthcare vs. residential), we include property types as metadata covariates. Contrary to the LDA, where the topic proportion θ_d is drawn from a Dirichlet distribution, the STM employs a logistic-normal generalized linear model which is based on document covariates (X_d). Thus, the frequency with which a topic is discussed that is common across all documents in the LDA is now affected by the observed metadata, as indicated by the following equation:

$$\vec{\theta}_d \mid X_d\gamma, \Sigma \sim LogisticNormal(\mu = X_d\gamma, \Sigma)\,, \tag{3.3}$$

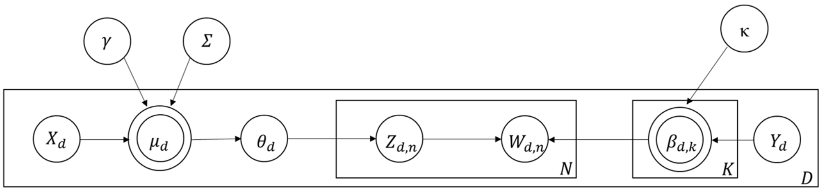
where X_d is a 1-by-p vector, γ is a p-by-$(K-1)$ matrix of coefficients and Σ is $(K-1)$-by-$(K-1)$ covariance matrix.

Whereas LDA assumes that word proportions within each topic (k) are represented by the model parameter β_k, which is identical for all documents (d), STM allows that the words describing a topic vary. Specifically, given a document-level content covariate y_d, the STM forms document-specific distributions of words representing each topic (k) based on the baseline word distribution (m), the topic-specific deviation K_k, the covariate group deviation K_{y_d}, and the interaction between the two $K_{y_d,k}$. The following equation provided by Kuhn (2018), and based on Roberts et al. (2019), summarizes this relationship as follows:

$$\beta_{d,k} \propto exp(m + K_k + K_{y_d} + K_{y_d,k}) \tag{3.4}$$

Figure 3.6 presents the STM in the common plate notation for topic modeling. Hereby, one "plate" exists for each document (D) and its associated topic distribution (θ_d) in the textual corpus. The inner plate, comprising topics ($Z_{d,n}$) and words ($W_{d,n}$), is replicated for each of the N words in the document. Analogously, the plate including the model parameter $\beta_{d,k}$ is replicated for each of the K topics in a textual corpus (Blei, 2012; Kuhn, 2018)

Figure 3.6: Structural Topic Modeling



This figure shows the STM in plate notation following Roberts et al. (2019).

After pre-processing, we estimate the STM, based on a variational Expectation-Maximization algorithm. The maximum number of iterations is set to 100, so that convergence is always reached before this threshold.

3.11 Appendix B

Table 3.5: STM Top Word Lists

Item 1A
Topic 1: Transaction unenforceable, hence, distinct, origination, repurchases, mentioned, artificially, concentrations, spread, sale-leaseback, post, enforceability, action, objective, appreciate, terminating, leads, staff, servicing, imposing
Topic 2: Regulation insufficiency, accumulation, reconfiguration, lessors, precautions, refrain, accommodation, unqualified, batteries, comprise, re-leased, co-members, anything, grants, removing, extinguished, fix, globally, speed, witness
Topic 3: Business Process appointment, probability, contemplate, economical, terrorist-related, started, voluntarily, confirmed, par, unfeasible, caption, execution, discuss, computation, cancelled, dramatically, zero, encumbering, free, please
Topic 4: Unsecured Claims and Debts shares-trust, owing, assign, went, quantities, attached, park, ends, trustees, neglect, commerce, insulate, incumbent, appraised, degrees, adapt, impairs, jersey, correspond, beverage
Topic 5: Rating printing, moodys, migration, recycling, injunction, poors, southeast, complicated, declaring, terminates, obligors, expirations, enforced, interfere, sent, indentures, vulnerabilities, prone, terminology, pendency
Topic 6: Tax and Capital Contribution draft, motivated, earliest, re-characterization, iraq, administering, faults, functions, choosing, affiliation, widening, futures, sensitivity, built, awareness, exercisable, advised, profession, irrevocable, drafts
Topic 7: Financial Risk attestation, recapitalization, dealings, amends, unsatisfactory, fairly, parcel, effectiveness, encumbering, drought, departments, time-consuming, effectuated, reliable, firm, james, taxpayer, endowments, exemptions, document
Topic 8: Capital Products and Market exhibits, non-renewal, shows, nyses, institution, expirations, website, perhaps, correctly, servicer, electronically, nyse, requisite, cdo, outage, earth, advanced, america, pledged, swaps
Topic 9: Acquisition understanding, describe, vendor, discovered, tactics, coordinated, lessen, rated, lps, inherently, works, expects, obama, stores, distributing, emanating, abatement, two-year, clean, co-tenancy
Topic 10: Contingencies correlate, tcja, condominiums, hackers, phasing, functioning, pronouncements, discounted, sent, destruction, launched, libor, encouraging, terrorists, non-business, fires, modifying, confidential, inside, deadlines

see next page

Table 3.5: continued

Topic 11: Capital Markets and Realization of Profit
excludes, changes, unregistered, prospectus, inclined, optimize, unenforceable, loss-generating, participates, eventually, interfere, comprising, list, internalize, registrants, stages, par, twenty, ipo, rata
Topic 12: IT
contingencies, normalized, cyberattacks, restraints, oppose, agents, automated, administered, staffing, inflationary, faces, cybersecurity, concerned, shall, adoption, sheet, indications, ineffective, interpret, recordkeeping
Topic 13: Legal & Litigation Risk
plaintiffs, sue, zones, tax-exempt, prejudice, supreme, examine, defendants, federally, defendant, render, oversee, complaint, day, straight-line, exposures, tangible, feature, flood, conform
Topic 14: REIT Status
nonqualified, revocation, timeframe, mitigated, exploration, referenced, appointment, overpay, follow, jeopardizing, broadly, procedural, committee, reviews, transferees, pronounced, violated, re-electing, capitalizations, owner-operators
Topic 15: Single Tenant Risk
plant, assign, burdensome, involvement, surveyed, non-affiliates, stabilize, greatly, hiring, capacities, owing, cessation, cooperation, side, deficit, reputations, forging, seriously, re-leasing, accomplish
Topic 16: Property
live, internationally, viable, vandalism, trained, corporate-level, cycles, movement, inventories, capitalizing, unionized, served, owner-operators, entry, incorrectly, intervening, union, contractors, equity-related, cercla
Topic 17: Politics
users, interstate, expects, perils, possess, avenue, investigative, reclassified, distributes, richard, bidding, lend, west, nuclear, holds, manages, advertising, systemic, places, philosophy
Topic 18: Tax
drip, itemized, consequence, debt-total, kind, supplemental, passive, tax-free, lease, percent, eligibility, satisfies, minimis, protective, snow, files, buy-sell, bind, commitment, commodity
Topic 19: Cash-flow
establishes, belief, property, trustees, productive, visual, declaration, withdraw, updates, simultaneously, corporate-level, redeemable, capitalized, billed, reviewed, landlords, overruns, noi, secondarily, impairments
Topic 20: Property
page, catastrophe, metro, establishes, notification, reauthorization, nearby, unwillingness, ventilation, distributes, notify, stem, charters, destructive, repositioning, david, insurers, constant, plaza, tcja
This table shows the top 20 words for each of the topics.

Table 3.6: Metadata Covariates

Property Type	# of 10-Ks	Covariate Words
Unknown	0	n.a.
Unclassified	264	generation, equipment, products, pressures, distributing, diversification, appeal, option, letter, planning, finding, uncertain, paying, lesser, oil, larger, capacity, negotiate, satisfying, advantage
Diversified	233	incident, five-year, weaknesses, raised, rating, diluted, accept, vacancies, renewal, valuation, expiring, dealer, tenant, existence, designed, assumptions, terminated, accounting, grade, insolvent
Health Care	215	referral, licensure, patients, false, physician, payors, abuse, healthcare, whistleblower, medicare, medicaid, denial, hospitals, patient, payor, physicians, hipaa, referrals, care, anti-kickback
Industrial/ Office	424	feet, office, square, francisco, evaluation, undisclosed, downgraded, space, units, evict, budgeted, utilities, perceived, enforcing, building, lack, honor, disclosure, geopolitical, settle
Lodging/ Resorts	269	brands, hotels, centralized, leisure, travelers, room, revpar, hotel, rooms, building, franchisors, guests, true, adr, reservation, travel, franchise, alerts, respected, lodging
Residential	277	mae, fannie, residents, homes, mac, freddie, apartment, housing, multifamily, fhaa, household, communities, explore, apartments, home, lawsuits, offers, conservatorship, already, regulating

see next page

Table 3.6: continued

Retail	455	retailers, shopping, retailing, shoppers, goods, retail, e-commerce, consumer, locations, malls, creditworthiness, traffic, vacated, anchor, tanks, stores, premises, convenience, spaces, approvals
Self Storage	78	self-storage, extensively, cyber-attack, penetrate, armed, telephone, destructive, avail, commerce, storage, collecting, shutdowns, changed, disruptive, releases, audits, view, worms, protections, integrating

This table shows the metadata Covariate Words based on 8 of the Ziman Property Types and the number of occurrence within our sample (# of 10-Ks). The STM identifies these covariate words that the algorithm uses to determine the covariate group deviation K_{y_d} and the covariate-topic interactions $K_{y_d,k}$ (see Appendix A).

Table 3.7: Description of Variables

Dependent Variables	
Vola	The standard deviation of daily log returns extrapolated to the *T*-trading-day period after the 10-K filing; $T \in [5, 40, 60]$.
ΔVola	The change in the standard deviation of a firms' daily stock returns from the symmetric period of *T* trading days before to after the 10-K filing.
Control Variables	
FFO/Share	FFO scaled by shares outstanding; (NI+SPPE+($DPACRE_t$–$DPACRE_{t-1}$))/CSHO
Size	Natural logarithm of total assets; log(AT)
Leverage	Ratio of total liabilities to total assets; LT/AT
ΔREV	Change in sales; $SALE_t$–$SALE_{t-1}$
Sales_Growth	Ratio of change in sales to lagged assets; ($SALE_t$–$SALE_{t-1}$)/AT_{t-1}
Beta	This CAPM-based measure of the systematic risk compared to the market is directly obtained from CRSP and calculated using the methods developed by Scholes and Williams (1977).
BTM	Book-to-market ratio of common stock; (TEQ/(AT-LT))+TXDITC-PSTK)/ (CSHPRI*PRCC)
IO	Shares hold by institutional investors from Thomson Reuters divided by the total shares outstanding.
Lag_Vola	The stock return volatility of the last T trading days before the 10-K filing.
$Vola^{S\&P}$	The stock return volatility of the S&P 500 for *T* trading days before the 10-K filing.
ΔVolume	The change of a firms' average daily trading volume from the symmetric period of *T* trading days before to after the 10-K filing.
Text_Length	Total number of words in Item 1A or Item 7A of an annual report (excluding stop words). We use the natural logarithm of the number in our regressions.
FOG	Gunning Fog score for the text in Item 1A or Item 7A of an annual report (excluding stop words); calculated as: *(words per sentence + percent of complex words)*0.4*

This table describes the variables used and the corresponding Compustat data items.

Table 3.8: Correlation of Risk Factor Topics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
(1) *Freq_Topic 1*	1																			
(2) *Freq_Topic 2*	0.233	1																		
(3) *Freq_Topic 3*	0.346	0.195	1																	
(4) *Freq_Topic 4*	0.207	-0.371	0.357	1																
(5) *Freq_Topic 5*	0.354	0.115	0.444	0.351	1															
(6) *Freq_Topic 6*	0.190	0.316	0.267	0.227	0.274	1														
(7) *Freq_Topic 7*	0.364	0.227	0.351	0.178	0.278	0.281	1													
(8) *Freq_Topic 8*	0.253	0.195	0.416	0.306	0.427	0.313	0.090	1												
(9) *Freq_Topic 9*	-0.151	-0.002	0.287	0.324	0.306	0.083	0.244	0.251	1											
(10) *Freq_Topic 10*	0.201	0.120	0.210	0.399	0.472	0.256	0.238	0.248	0.243	1										
(11) *Freq_Topic 11*	0.268	0.205	0.177	0.211	0.038	0.410	0.314	0.293	0.153	0.173	1									
(12) *Freq_Topic 12*	0.032	0.056	0.280	0.281	0.409	-0.120	0.071	0.148	0.211	0.267	-0.629	1								
(13) *Freq_Topic 13*	0.302	0.390	0.177	0.244	0.407	0.408	0.235	0.066	0.084	0.305	0.134	0.128	1							
(14) *Freq_Topic 14*	0.165	0.382	0.366	-0.030	0.149	-0.062	0.159	0.110	0.185	-0.030	-0.001	0.354	-0.284	1						
(15) *Freq_Topic 15*	0.106	0.208	-0.068	0.241	0.253	0.329	0.203	0.234	0.224	0.189	0.231	0.128	0.177	0.303	1					
(16) *Freq_Topic 16*	-0.070	0.115	0.363	0.221	-0.005	0.091	0.247	0.088	0.216	0.073	0.307	0.043	0.112	0.172	0.042	1				
(17) *Freq_Topic 17*	-0.0002	0.155	0.071	0.123	0.041	0.131	0.176	0.217	0.194	0.241	0.276	0.060	0.111	0.113	0.232	0.171	1			
(18) *Freq_Topic 18*	0.161	0.127	0.264	0.298	0.259	0.131	-0.502	0.396	0.131	0.167	0.226	0.078	0.069	0.303	0.136	0.001	0.092	1		
(19) *Freq_Topic 19*	0.055	0.149	0.295	0.235	0.251	0.237	0.213	0.312	0.122	0.235	0.152	0.203	-0.145	0.426	0.228	-0.129	0.020	0.227	1	
(20) *Freq_Topic 20*	0.163	0.062	0.284	0.289	0.214	0.152	0.206	0.320	0.267	0.182	0.265	0.170	0.161	0.112	0.262	0.186	0.272	0.228	0.065	1

This table shows the Bravais-Pearson correlation coefficients of the logged frequencies for the twenty risk factor topics of Item 1A (*Freq_Topics*).

Table 3.9: Short Risk Description

Example 1 (Bluerock Residential Growth REIT, Inc., 2010)
Item 1A. Risk Factors We have omitted a discussion of risk factors because, as a smaller reporting company, we are not required to provide such information. For a discussion of the significant factors that make an investment in our shares risky, see the prospectus that relates to our ongoing Initial Public Offering. (48 words)
Example 2 (Medalist Diversified REIT, Inc., 2019)
ITEM 1A. RISK FACTORS We have omitted a discussion of risk factors because, as a smaller reporting company, we are not required to provide such information. (22 words)
Example 3 (Paragon Real Estate Equity & Investment Trust, 2009)
Item 1A. Risk Factors. This annual report contains historical information, as well as forward-looking statements that involve known and unknown risks and relate to future events, our future financial performance, or our expected future operations and actions. In some cases, you can identify forward-looking statements by terminology such as \"may,\" \"will,\" \"should,\" \"expect,\" \"plan,\" \"anticipate,\" \"believe,\" \"estimate,\" \"future,\" \"intend,\" \"could,\" \"hope,\" \"predict,\" \"target,\" \"potential,\" or \"continue\" or the negative of these terms or other similar expressions. These forward-looking statements are only our predictions based upon current information and involve numerous assumptions, risks and uncertainties. Our actual results or actions may differ materially from these forward-looking statements for many reasons. While it is impossible to identify all of theses factors, the following could cause actual results to differ materially from those estimated by us: \u0095 worsening of national economic conditions, including continuation of lack of liquidity in the capital markets and more stringent lending requirements by financial institutions; \u0095 depressed values for commercial real estate properties and companies; \u0095 changes in local market conditions due to changes in general or local economic conditions and neighborhood characteristics; \u0095 changes in interest rates and in the availability, cost and terms of mortgage funds; \u0095 impact of present or future environmental legislation and compliance with environmental laws; \u0095 ongoing need for capital improvements, particularly in older properties; \u0095 more attractive lease incentives offered by competitors in similar markets; \u0095 increased market demand for newer properties; \u0095 changes in real estate tax rates and other operating expenses; \u0095 decreases in market prices of the shares of publicly traded real estate companies; \u0095 adverse changes in governmental rules and fiscal policies; \u0095 adverse changes in zoning laws; and \u0095 other factors which are beyond our control. 3 Table of Contents In addition, an investment in the Company involves numerous risks that potential investors should consider carefully, including, without limitation: \u0095 we have no operating assets; \u0095 our cash resources are limited; \u0095 we have a history of losses; \u0095 we have not raised funds through a public equity offering; \u0095 our trustees control a significant percentage of our voting shares; \u0095 shareholders could experience possible future dilution through the issuance of additional shares; \u0095 we are dependent on a small number of key senior professionals who are part-time employees; and \u0095 we currently do not plan to distribute dividends to the holders of our shares. (374 words)
This table shows 3 instances of Item 1A for a low number of words since there is no legal requirement for small firms to do that (Example 1 and Example 2) or the risk factors are very short described (Example 3). Stop words are not excluded from these examples.

Table 3.10: Probability of Appearance – Risk Perception measured by the Change in Volatility

	Model 1 (0, 5 days)	**Model 2** (0, 40 days)	**Model 3** (0, 60 days)
Freq_Topic 1	-0.008***	-0.017***	-0.016***
Transaction	(0.003)	(0.005)	(0.005)
Freq_Topic 2	0.034***	0.030***	0.033***
Regulation	(0.004)	(0.007)	(0.007)
Freq_Topic 3	-0.009***	-0.002	-0.005
Business Process	(0.002)	(0.005)	(0.004)
Freq_Topic 4	0.041***	0.030***	0.033***
Unsecured Claims and Debts	(0.004)	(0.008)	(0.007)
Freq_Topic 5	0.008***	0.007	0.006
Rating	(0.003)	(0.005)	(0.005)
Freq_Topic 6	-0.001	-0.009*	-0.008*
Tax and Capital Contribution	(0.002)	(0.005)	(0.005)

see next page

Table 3.10: continued

Freq_Topic 8	-0.012***	-0.008**	-0.011***
Capital Products and Market	(0.002)	(0.004)	(0.004)
Freq_Topic 9	0.002	-0.006	-0.005
Acquisition	(0.002)	(0.005)	(0.005)
Freq_Topic 10	-0.002**	0.003	0.003
Contingencies	(0.001)	(0.002)	(0.002)
Freq_Topic 12	0.00000	-0.006	-0.004
IT	(0.002)	(0.003)	(0.003)
Freq_Topic 13	-0.020***	-0.009*	-0.012**
Legal & Litigation Risk	(0.002)	(0.005)	(0.005)
Freq_Topic 15	-0.012***	0.014***	0.014***
Single Tenant Risk	(0.002)	(0.005)	(0.005)
Freq_Topic 16	-0.008***	-0.004	-0.006
Property	(0.002)	(0.004)	(0.004)
Freq_Topic 17	-0.005***	-0.007	-0.006
Politics	(0.002)	(0.004)	(0.004)
Freq_Topic 19	-0.013***	-0.008	-0.007
Cash-flow	(0.002)	(0.005)	(0.005)
Freq_Topic 20	0.003	0.003	0.004
Property	(0.002)	(0.005)	(0.004)
FFO/Share	0.0004	0.001	0.0002
	(0.001)	(0.002)	(0.002)
Size	0.002	0.015*	0.014**
	(0.004)	(0.008)	(0.007)
Leverage	0.016	-0.006	-0.018
	(0.014)	(0.029)	(0.029)
ΔREV	0.00000	-0.00001	-0.00001
	(0.00001)	(0.00002)	(0.00002)
Sales_Growth	0.006	-0.008	-0.010
	(0.005)	(0.009)	(0.009)
Beta	-0.001	0.001	-0.010
	(0.004)	(0.008)	(0.007)
BTM	-0.018***	0.069***	0.082***
	(0.002)	(0.006)	(0.006)
IO	-0.009	-0.033**	-0.025*
	(0.007)	(0.014)	(0.013)
$Vola^{S\&P}$	-0.208	-0.696***	-0.464**
	(0.138)	(0.230)	(0.208)
ΔVolume	0.011***	0.020***	0.020***
	(0.002)	(0.004)	(0.004)
Text_Length	-0.009*	0.010	0.005
	(0.005)	(0.010)	(0.010)
FOG	-0.001	-0.002	-0.001
	(0.002)	(0.004)	(0.004)
N	1,228	1,224	1,223
R^2	0.230	0.177	0.223

see next page

Table 3.10: continued

This table presents the results of fixed-effect models controlling for unobserved firm and time effects for Item 1A. The table reports panel regression results of fixed effects models, which include coefficients and standard errors (in parentheses) of determinants affecting investor's risk perception. The dependent variable (Δ*Vola*) takes a different number of trading days after the 10-K filing date into account – 5 trading days (Model 1), 40 trading days (Model 2), and 60 trading days (Model 3). The descriptive statistics of Δ*Vola* are given in in Table 3.11 in Appendix B. The definition of all variables is presented in Table 3.7.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.11: Descriptive Statistics – Change in Volatility

	N	Mean	StDev	Min	Q1	Median	Q3	Max
				Dependent Variables				
Δ*Vola* (0, 5 days)	1,543	0.001	0.044	-0.324	-0.012	0.001	0.014	1.114
Δ*Vola* (0, 40 days)	1,529	0.0003	0.102	-2.238	-0.021	-0.003	0.013	2.023
Δ*Vola* (0, 60 days)	1,519	-0.007	0.106	-2.229	-0.029	-0.004	0.015	1.993

This table shows the change in the standard deviation of a firms' daily stock returns from the symmetric period of *T* trading-days before to after the 10-K is filed (Δ*Vola*). *N* is the number of observations, StdDev stands for standard deviation, Q1 is the first and Q3 the third quartile of the distribution, and Min is the minimum and Max the maximum of each variable. *N* is set to the maximal available number of observations for each variable.

Table 3.12: Comparison of STM, CTM, and LDA – Risk Perception

		Model 1 (0, 5 days)			**Model 3** (0, 60 days)	
	STM	**CTM**	**LDA**	**STM**	**CTM**	**LDA**
Abs_Allocation 1	-0.007***	0.001	0.0001	-0.015***	-0.001	0.0005
	(0.002)	(0.004)	(0.001)	(0.005)	(0.008)	(0.001)
Abs_Allocation 2	0.032***	0.007	0.001	0.030***	-0.009	0.001
	(0.003)	(0.005)	(0.001)	(0.007)	(0.011)	(0.001)
Abs_Allocation 3	-0.011***	-0.009	-0.0001	-0.006	0.002	0.0001
	(0.002)	(0.006)	(0.001)	(0.004)	(0.011)	(0.001)
Abs_Allocation 4	0.038***	-0.013*	-0.001	0.031***	-0.007	-0.002
	(0.003)	(0.008)	(0.001)	(0.007)	(0.016)	(0.002)
Abs_Allocation 5	0.009***	0.009**	0.0001	0.008*	0.006	0.0003
	(0.002)	(0.004)	(0.001)	(0.005)	(0.008)	(0.002)
Abs_Allocation 6	-0.001	-0.001	-0.0004	-0.009**	-0.001	-0.00004
	(0.002)	(0.003)	(0.001)	(0.004)	(0.007)	(0.001)
Abs_Allocation 7		-0.011	-0.001		-0.016	-0.001
		(0.008)	(0.001)		(0.017)	(0.002)
Abs_Allocation 8	-0.010***	-0.004	0.001**	-0.008**	0.004	0.0001
	(0.002)	(0.003)	(0.001)	(0.003)	(0.007)	(0.001)
Abs_Allocation 9	0.002	-0.006	0.00003	-0.004	-0.011	-0.001
	(0.002)	(0.005)	(0.0005)	(0.004)	(0.011)	(0.001)
Abs_Allocation 10	-0.002***	0.002	0.001	0.001	-0.001	0.003
	(0.001)	(0.004)	(0.001)	(0.002)	(0.007)	(0.002)
Abs_Allocation 11		0.002	0.001		0.004	0.001
		(0.003)	(0.001)		(0.005)	(0.003)
Abs_Allocation 12	0.00001	-0.006	0.0003	-0.004	-0.011	0.001
	(0.001)	(0.005)	(0.001)	(0.003)	(0.009)	(0.001)

see next page

Table 3.12: continued

Abs_Allocation 13	-0.017***	0.016***	-0.0004	-0.010**	0.015	-0.001
	(0.002)	(0.005)	(0.001)	(0.005)	(0.011)	(0.002)
Abs_Allocation 14		-0.004	0.0001		0.014	0.003*
		(0.011)	(0.001)		(0.022)	(0.001)
Abs_Allocation 15	-0.012***	0.010***	-0.002***	0.010**	0.008	-0.005***
	(0.002)	(0.003)	(0.001)	(0.004)	(0.005)	(0.001)
Abs_Allocation 16	-0.007***	0.005	-0.0004	-0.005	-0.008	0.0005
	(0.002)	(0.005)	(0.001)	(0.004)	(0.011)	(0.003)
Abs_Allocation 17	-0.005***	-0.006	-0.001	-0.004	-0.012	-0.002
	(0.002)	(0.005)	(0.001)	(0.004)	(0.011)	(0.002)
Abs_Allocation 18		0.0001	-0.00004		-0.003	-0.0001
		(0.004)	(0.001)		(0.009)	(0.001)
Abs_Allocation 19	-0.012***	-0.002	0.001	-0.006	-0.002	0.001
	(0.002)	(0.005)	(0.001)	(0.005)	(0.010)	(0.001)
Abs_Allocation 20	0.003	0.020	-0.001*	0.004	0.080	-0.001
	(0.002)	(0.026)	(0.001)	(0.004)	(0.054)	(0.001)
FFO/Share	0.001	0.0001	0.0001	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
Size	0.001	0.0004	-0.0002	0.013*	0.004	0.004
	(0.003)	(0.003)	(0.003)	(0.007)	(0.007)	(0.007)
Leverage	0.029**	0.012	0.018	0.007	-0.025	-0.004
	(0.012)	(0.013)	(0.013)	(0.027)	(0.028)	(0.027)
ΔREV	0.00000	0.00000	0.00000	-0.00002	-0.00002	-0.00002
	(0.00001)	(0.00001)	(0.00001)	(0.00002)	(0.00002)	(0.00002)
Sales_Growth	0.005	0.004	0.002	-0.006	-0.008	-0.007
	(0.004)	(0.004)	(0.004)	(0.009)	(0.009)	(0.009)
Beta	0.009***	0.011***	0.010***	0.015**	0.023***	0.020***
	(0.003)	(0.004)	(0.004)	(0.007)	(0.008)	(0.007)
BTM	-0.020***	-0.020***	-0.019***	0.066***	0.065***	0.065***
	(0.002)	(0.002)	(0.002)	(0.006)	(0.006)	(0.006)
IO	-0.018***	-0.016**	-0.019***	-0.038***	-0.034***	-0.038***
	(0.006)	(0.006)	(0.006)	(0.013)	(0.013)	(0.013)
Lag_Vola	0.354***	0.310***	0.328***	0.521***	0.501***	0.516***
	(0.037)	(0.040)	(0.040)	(0.043)	(0.044)	(0.043)
$Vola^{S\&P}$	0.866***	0.885***	0.893***	1.290***	1.328***	1.278***
	(0.133)	(0.145)	(0.145)	(0.305)	(0.310)	(0.310)
ΔVolume	0.007***	0.007***	0.006***	0.020***	0.023***	0.020***
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.004)
Text_Length	-0.005	-0.011	-0.0003	-0.001	-0.034	0.008
	(0.005)	(0.019)	(0.005)	(0.010)	(0.039)	(0.010)
FOG	-0.0003	-0.0004	0.0003	0.00004	0.001	0.002
	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.004)
N	1,228	1,228	1,228	1,223	1,223	1,223
R^2	0.345	0.234	0.229	0.283	0.268	0.274

see next page

Table 3.12: continued

This table presents the results of fixed-effect models controlling for unobserved firm and time effects for Item 1A. The table reports panel regression results of fixed effects models, which include coefficients and standard errors (in parentheses) of determinants affecting investor's risk perception. The dependent variable (*Vola*) takes a different number of trading days after the 10-K filing date into account – 5 trading days (Model 1) and 60 trading days (Model 3). The variable *Abs_Allocation* is derived using three different machine assisted approaches (i.e., STM, CTM, and LDA). Each approach applies a 20 topic full model to identify and quantify the risks disclosed in Item 1A. The risk topics identified by STM, CTM, and LDA are not identical. The definition of all variables is presented in Table 3.7 in Appendix B.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.12 Appendix C

LDA Topics and Metadata Covariates

We apply the standard LDA and identify the top words for 20 topics analogously to the STM method for Item 1A. Table 3.13 in Appendix C presents the results of this clustering. As assumed given by the optimization criterion of the LDA, the topics are close to investment foci, such as Topic #1 corresponds to "Health Care", Topic #4 to "Residential", and Topic #9 to "Retail" to name a few. LDA identifies the foci as the most substantial distinction within the textual corpus and allocates them as latent topics.

We further regress the investment foci (i.e., Ziman property types) on each of the 20 topics, in order to analyze whether the frequency of appearance for the individual risk factors is associated with property types (see Table 3.14 in Appendix C). We find, for example, that 5 out of 7 Ziman property types are statistically significantly associated with Topic #8 "Infrastructure". A positive coefficient sign suggests that a REIT assigned to the respective property type (e.g., "Unclassified") is likely to allocate a larger proportion of its risk disclosure to Topic #8. On the contrary, the negative relationship indicates that Topic #8 is less likely to occur in filings of REITs which are classified as "Residential", "Health Care", or "Self Storage". The relationship between property type and the probability of appearance for a risk-factor topic shows that we need to consider document-specific metadata (i.e., property types) when using a machine to identify the risk factors discussed by a REIT.

Table 3.13: LDA Top Word List

Item 1A
Topic 1: Health Care
healthcare, medicaid, correctional, detention, hospitals, hospital, brookdale, seniors, nursing, physicians, patients, payors, medicare, sunrise, inmates, tenants, care, medical, physician, science
Topic 2: Taxable REIT Subsidiary
spin, manager, bennett, comments, master, trss, trs, separation, stockholders, reits, treated, tenant, charter, arc, emerging, restaurant, tcja, gain, agreement, withholding

see next page

Table 3.13: continued

Topic 3: Reporting Duties/Auditing reporting, caption, report, discussion, see, analysis, information, management's, expressions, filer, composed, incorporated, rule, relates, underway, sponsoring, jpmorgan, auditors, oxley, sarbanes
Topic 4: Residential staff, single-family, hoa, hoas, homes, homeownership, cdo, loans, mortgage, foreclosure, non-performing, servicers, homeowners, residents, rental, securitizations, borrower, borrowers, stockholders, home
Topic 5: Market and Politics smaller, rules, effecting, collected, disclosure, vendor, weakness, oversight, defined, interim, restate, see, electing, regulation, misstatement, trump, relates, attestation, detected, commission
Topic 6: Investment Universe advisor, cole, stockholders, wells, ira, erisa, co-ownership, tenant-common, sponsored, estate-related, mezzanine, bridge, manager, sponsor, nav, sale-leaseback, internalization, builders, advisory, tenants
Topic 7: Property and Hurricane companies, omitted, professionals, managed, information, rita, controls, investing, commodity, ranks, katrina, adequacy, continuance, client, capitalizations, segment, pursue, pose, calculation, disagree
Topic 8: Infrastructure wireless, towers, disclose, tower, antenna, sprint, billboards, t-mobile, nols, radio, advertising, verizon, att, fcc, communications, nextel, roaming, lighting, broadcast, theatres
Topic 9: Retail host, incs, penn, mall, centers, shopping, separation, entirety, anchor, stores, sears, gaming, outlet, cam, anchors, retailers, malls, retail, lps, shareholders
Topic 10: Cyber Criminality systems, security, information, technology, confidential, cyber, computer, networks, identifiable, breaches, data, arisk, unauthorized, cyber-attacks, reputation, electronic, store, hackers, shutdowns, software
Topic 11: Stock Market/Partnerships stockholders, directors, stockholder, risky, partnership, military, privatization, million, preferred, units, warrants, agreement, andrew, messrs, llc, quoted, approximately, vice, executive, combination
Topic 12: Lodging/Resorts rmr, included, tas, aic, portnoy, sonesta, stars, trustees, star, adam, gov, irc, travel, hotels, barry, hotel, shareholders, marriott, snh, living
Topic 13: Infrastructure adviser, depositary, arc, gas, grand, terminal, corridor, infrastructure, decommissioning, sale-leaseback, percent, convertible, commodities, production, investees, privately-held, stockholders, notes, commodity, preferred
Topic 14: Lodging/Resorts hotels, hotel, permitted, lodging, travel, room, rooms, franchisors, shareholders, marriott, trustees, franchisor, franchise, revpar, reservations, hilton, leisure, intermediaries, guests, lessees
Topic 15: Company/Real Estate requested, partnership, stockholders, tenants, space, mgcl, honolulu, directors, units, charter, rental, tenant, stockholder, self-storage, market, partner, asking, leases, airborne, co-venturers
Topic 16: Timber timber, timberlands, timberland, forest, centers, wood, harvest, species, logs, harvesting, student, connectivity, fiber, logging, data, universities, endangered, hbu, campus, colocation
Topic 17: Residential communities, apartment, digital, companys, multifamily, realty, housing, freddie, incs, fannie, mac, homes, mae, residents, sale, lps, manufactured, multi-family, excel, partnership
Topic 18: REIT Specifics vornado, trustees, shareholders, alexanders, shareholder, gladstone, roth, transitional, declaration, toys, trust, tenants, mandelbaum, wight, maryland, interstate, space, partnership, zell, realty
Topic 19: Retail anchor, shopping, tenants, space, retail, shareholders, centers, self-storage, retailers, tenant, stores, leases, redevelopment, predictions, bankruptcy, rental, retailing, re-lease, development, venture
Topic 20: Property Risk and Terrorism page, securityholders, science, tenants, space, industrial, ofac, manhattan, asbestos, avenue, ifrs, co-investment, tria, indoor, unconsolidated, earthquake, ventures, nbcr, unsecured, partnership
This table shows the top 20 words for each of the topics.

Table 3.14: Regressions for LDA and Property Focus

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Intercept	0.038***	0.0641***	0.0282***	0.0062	0.004	0.0759***	0.0584***	0.0673***	0.0714***	0.0556***
	(0.0141)	(0.0144)	(0.0103)	(0.011)	(0.004)	(0.0132)	(0.0116)	(0.014)	(0.0134)	(0.0134)
Health Care	0.0205	0.0428*	-0.0041	-0.0029	-0.0023	-0.0197	0.0071	-0.0482**	-0.0122	-0.0098
	(0.0203)	(0.0224)	(0.0155)	(0.016)	(0.0056)	(-0.0205)	(0.017)	(-0.0193)	(-0.0203)	(0.0201)
Industrial/Office	0.0315*	0.0021	0.0119	0.031**	0.0232***	-0.0378**	-0.0264*	-0.0307*	-0.0311*	0.0326*
	(0.0176)	(0.0184)	(0.0129)	(0.0138)	(0.0066)	(-0.0161)	(0.0142)	(-0.0167)	(-0.0165)	(0.0184)
Lodging/Resorts	0.0237	0.0225	-0.0167	0.0699***	-0.0036	-0.0392**	-0.0385**	-0.0174	-0.0247	0.0065
	(0.0198)	(0.0212)	(0.0136)	(0.0165)	(0.0054)	(-0.0189)	(0.0158)	(-0.0186)	(-0.018)	(0.0203)
Residential	-0.0003	-0.0357*	0.0087	0.0398**	0.0103	-0.0241	-0.0075	-0.0549***	-0.0212	0.0065
	(0.0191)	(0.0193)	(0.0141)	(0.0156)	(0.0065)	(-0.0173)	(0.0162)	(-0.0181)	(-0.0186)	(0.0185)
Retail	-0.0103	-0.0136	0.0039	0.0647***	-0.003	-0.0456***	-0.025*	-0.0245	-0.0434**	-0.007
	(0.0174)	(0.018)	(0.0122)	(0.0139)	(0.0048)	(-0.0162)	(0.0147)	(-0.0163)	(-0.0169)	(0.017)
Self Storage	0.0171	-0.0611**	-0.0266	-0.002	-0.0021	0.0834***	-0.0442*	-0.0567**	0.0796**	0.0099
	(0.0275)	(0.0275)	(0.0191)	(0.0221)	(0.008)	(0.0315)	(0.0234)	(-0.026)	(0.0331)	(0.0325)
Unclassified	0.047**	-0.0283	0.0308**	-0.0022	0.0005	-0.049***	-0.033**	0.0591***	-0.0262	-0.007
	(0.0192)	(0.0196)	(0.0149)	(0.015)	(0.0054)	(-0.0179)	(0.0166)	(0.0205)	(-0.0178)	(0.0188)

	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
Intercept	0.0332***	0.0378**	0.0417***	0.0273***	0.023*	0.0896***	0.0633***	0.0566***	0.0671***	0.0912***
	(0.0101)	(0.0152)	(0.015)	(0.0093)	(0.0137)	(0.0166)	(0.0147)	(0.0152)	(0.0156)	(0.0156)
Health Care	-0.0118	-0.0175	0.0496**	-0.0006	0.0459**	-0.016	-0.0055	0.0065	0.004	-0.0254
	(0.0148)	(0.0219)	(0.0229)	(0.0133)	(0.0197)	(0.0242)	(0.0218)	(0.0227)	(0.0221)	(0.0215)
Industrial/Office	0.0092	-0.0043	0.0419**	-0.0111	0.0285	-0.05**	-0.0283	0.0478**	0.0085	-0.0481**
	(0.0129)	(0.0191)	(0.0193)	(0.0116)	(0.0178)	(0.0204)	(0.0183)	(0.0198)	(0.0194)	(0.0189)
Lodging/Resorts	0.0039	0.0554***	0.0095	-0.0243*	0.0121	-0.0067	0.0161	-0.0079	-0.0293	-0.0111
	(0.0145)	(0.0208)	(0.0205)	(0.0125)	(0.0189)	(0.022)	(0.0211)	(0.0219)	(0.0206)	(0.0225)
Residential	0.0175	0.0324	0.015	-0.0024	0.004	0.0107	0.0237	0.0006	-0.0234	0.0004
	(0.016)	(0.0212)	(0.0208)	(0.0139)	(0.0186)	(0.0227)	(0.0207)	(0.0209)	(0.0207)	(0.0207)
Retail	-0.0245**	0.0801***	0.0583***	0.0208*	0.0507***	-0.0581***	-0.0111	0.0449**	0.0018	-0.0584***
	(0.0121)	(0.0192)	(0.0198)	(0.0121)	(0.0174)	(0.0198)	(0.0181)	(0.0193)	(0.0198)	(0.0185)
Self Storage	-0.0218	0.0809**	-0.0391	-0.0248	0.0299	-0.0762**	0.0938***	-0.0495*	0.0949***	-0.0854***
	(0.0207)	(0.0319)	(0.0298)	(0.018)	(0.0271)	(0.0312)	(0.0333)	(0.0298)	(0.0327)	(0.0281)
Unclassified	0.009	0.0155	0.0086	0.0029	0.0506***	0.0404*	-0.0052	-0.0325	-0.0221	-0.058***
	(0.014)	(0.021)	(0.0207)	(0.0133)	(0.0187)	(0.0229)	(0.021)	(0.0206)	(0.0213)	(0.0205)

This table shows the relationship between metadata (investment foci) and topics. The topic proportions are the dependent variables of a regression which shows the conditional expectation of topic prevalence given document characteristics, so that the estimation uncertainty is incorporated in the dependent variable. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.13 Appendix D

Risk Perception for Item 7A

The second risk section included in the 10-K is represented by Item 7A. This section should list "quantitative and qualitative disclosures about market risk" which are relevant for a company (e.g., interest rate risk or foreign currency exchange risk). We conduct our analyses additionally for this section describing more long-term risk.

In the first step, we apply the STM to Item 7A and label the topics. Since Item 7A is shorter, we set the number of topics to be identified by the STM to 5. Following the SECs' requirement (Item 305 of Regulation S-K (§ 229.305)) to inform the public on market risk, the risk topics describe more long-term risks like "Politics & Regions" or "(Re-)financing (see Table 3.15 in Appendix D). The descriptive statistics of *Abs_Allocation* are given in Table 3.16.

In the second step, we apply the fixed-effect panel regression model as stated in Section 3.5 to Item 7A, to address Hypotheses 1 and 2. The results are given in Table 3.17 in Appendix D. Our results suggest that the extracted risk factors are less informative for this item than those identified in Item 1A – none of the 5 factors is significant for the short-term (5 day) window. If we change to longer windows, three risk topics become significant. We conclude that this goes in line with the more long-term nature of the risk factors described in Item 7A. The goodness of fit is for all windows smaller than for Item 1A – ranging from 14% to 21% instead of 21% to 35%. This can be explained by the composition of Item 7A, since this section not only names but additionally quantifies the impact of the individual risk factors on future firm performance. Thus, managers usually use numbers to describe how risk factors affect firms' filings in this section. However, our method focuses on textual data i.e., the words used to qualitatively describe relevant risks and topic models cannot take numbers into account. In addition, with an average length of only 6,680 words, Item 7A is just a tenth of the average length of Item 1A. As explained by Papilloud and Hinneburg (2018), shorter documents decrease the robustness of the topic model, because it "learns" less from the data. Third, many documents have (almost) the same content, which further distorts the topic model (Papilloud and Hinneburg, 2018).

Table 3.15: STM Top Word Lists for Item 7A

Item 7A
Topic 1: Contractual Risks
discounted, excluding, one-month, fix, agreements, policy, notional, maturities, effectively, contractual, techniques, weighted-average, corresponding, giving, reflects, rating, transactions, fixes, discount, fees

see next page

Table 3.15: continued

Topic 2: Accounting
liability, direct, eliminated, actively, stock, accrued, amounted, plan, relating, carried, years, recognized, sale, liquidation, statements, statement, investing, accounts, permanent, carrying
Topic 3: Capital
segments, redeemable, capitalized, section, venture, immediately, regarding, act, joint, redemption, acquired, discussions, consolidation, disclosure, projects, iii, general, reference, receivable, common
Topic 4: Politics and Regions
refers, political, monetary, domestic, international, structure, considering, beyond, governmental, considerations, factors, many, economic, prices, event, financings, take, unable, high, dependent
Topic 5: (Re-)financing
flexibility, refinance, opportunity, issue, change, present, matures, unsecured, although, refinancing, assuming, principal, respect, near, term, revolving, exceeds, premiums, mitigate, time
This table shows the top 20 words for each of the topics.

Table 3.16: Descriptive Statistics – Absolute Allocation of Words for Item 7A

	N	Mean	StDev	Min	Q1	Median	Q3	Max
				Item 1A				
Abs_Allocation 1	2,514	1,237.987	2,987.863	0.965	12.295	51.106	1,302.653	42,934.940
Abs_Allocation 2	2,514	2,075.822	23,277.310	0.573	3.710	10.673	63.392	436,479.300
Abs_Allocation 3	2,514	1,101.688	12,891.150	0.958	5.486	13.465	175.900	373,974.400
Abs_Allocation 4	2,514	1,048.147	2,079.992	3.164	14.860	79.551	1,166.431	37,108.090
Abs_Allocation 5	2,514	1,198.234	3,656.759	0.418	6.859	31.509	1,058.209	94,708.970

This table shows the descriptive statistics for the frequencies (in %) for the risk factor topics multiplied by the total length of the corresponding disclosure (*Abs_Allocation*). *N* is the number of observations, StDev stands for standard deviation, Q1 is the first and Q3 the third quartile of the distribution, and Min is the minimum and Max the maximum of each variable. *N* is set to the maximal available number of observations for each variable.

Table 3.17: Absolute Allocation of Words – Risk Perception for Item 7A

	Model 1	**Model 2**	**Model 3**
	(0, 5 days)	(0, 40 days)	(0, 60 days)
Abs_Allocation 1	-0.001	-0.011***	-0.009**
Contractual Risks	(0.002)	(0.004)	(0.004)
Abs_Allocation 2	-0.0001	-0.001	-0.001
Accounting	(0.001)	(0.003)	(0.003)
Abs_Allocation 3	-0.00001	-0.002	-0.003
Capital	(0.001)	(0.003)	(0.003)
Abs_Allocation 4	0.001	0.013***	0.012***
Politics and Regions	(0.002)	(0.004)	(0.004)
Abs_Allocation 5	-0.002	-0.008**	-0.009**
(Re-)financing	(0.002)	(0.004)	0.0001
FFO/Share	0.00005	0.001	(0.002)
	(0.001)	(0.002)	0.010
Size	-0.001	0.011	(0.008)
	(0.003)	(0.008)	-0.025
Leverage	0.021	-0.013	(0.030)
	(0.013)	(0.031)	-0.00001
ΔREV	0.00000	-0.00001	(0.00002)
	(0.00001)	(0.00002)	0.003

see next page

Table 3.17: continued

Sales_Growth	0.005	0.005	(0.010)
	(0.004)	(0.010)	0.027***
Beta	0.008**	0.033***	(0.008)
	(0.004)	(0.008)	0.0001
BTM	-0.015***	0.005	0.017***
	(0.002)	(0.006)	(0.006)
IO	-0.019***	-0.051***	-0.045***
	(0.006)	(0.015)	(0.015)
Lag_Vola	0.315***	0.355***	0.503***
	(0.041)	(0.066)	(0.049)
$Vola^{S\&P}$	0.954***	1.540***	1.228***
	(0.148)	(0.342)	(0.354)
ΔVolume	0.006***	0.021***	0.021***
	(0.002)	(0.005)	(0.005)
Text_Length	0.002	0.030***	0.029***
	(0.003)	(0.007)	(0.007)
FOG	-0.0001	-0.004**	-0.004**
	(0.001)	(0.002)	(0.002)
N	1,209	1,205	1,204
R^2	0.195	0.144	0.211

This table presents the results of fixed-effect models controlling for unobserved firm and time effects for Item 7A. The table reports panel regression results of fixed effects models, which include coefficients and standard errors (in parentheses) of determinants affecting investor's risk perception. The dependent variable (*Vola*) takes a different number of trading days after the 10-K filing date into account – 5 trading days (Model 1), 40 trading days (Model 2), and 60 trading days (Model 3). The definition of all variables is presented in Table 3.7 in Appendix B.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4 Can Textual Analysis solve the Underpricing Puzzle? A US REIT Study

4.1 Abstract

Purpose: Although many theories aim to explain IPO underpricing, initial-day returns of US REIT IPOs remain a "puzzle". This study proposes textual analysis to exploit the qualitative information revealed through one of the most important documents during the IPO process – Form S-11.

Design/methodology/approach: This study determines the level of uncertain language in the prospectus, as well as its similarity to recently filed registration statements, to assess whether textual features can solve the underpricing puzzle. It assumes that uncertain language makes it more difficult for potential investors to price the issue and thus increases underpricing. Furthermore, it is hypothesized that a higher similarity to previous filings indicates that the prospectus provides little useful information and thus does not resolve existing information asymmetries, leading to increased underpricing.

Findings: Contrary to expectations, this research does not find a statistically significant association between uncertain language in Form S-11 and initial-day returns. This result is interpreted as suggesting that uncertain language in the prospectus does not reflect the issuer's expectations about the company's future prospects, but rather is necessary because of forecasting difficulties and litigation risk. Analyzing disclosure similarity instead, this study finds a statistically and economically significant impact of qualitative information on initial-day returns. Thus, REIT managers may reduce underpricing by voluntarily providing more information to potential investors in Form S-11.

Practical implications: The results demonstrate that textual analysis can in fact help to explain underpricing of US REIT IPOs.

Originality/value: This is the first study which applies textual analysis to corporate disclosures of US REITs in order to explain IPO underpricing.

Keywords: Initial public offering, Underpricing, Information asymmetry, Textual analysis, IPO Prospectus, Form S-11

4.2 Introduction

Underpricing, the phenomenon that an Initial Public Offerings' (IPOs') closing price at the end of the first trading-day is higher than its offer price, has attracted considerable attention in the financial literature. Researchers have proposed numerous theories to explain why IPOs have positive initial returns. Whereas the vast majority of theories revolve around asymmetric information between the parties involved in the IPO, others entail for example, litigation and behavioral theories. However, despite the substantial efforts to explain why issuers accept the large amount of "money left on the table", unresolved questions remain as to how IPO shares are priced.

Only with the increasing online availability of texts and rising computational power in recent decades, have researchers resumed their investigations, relying increasingly on qualitative data to solve the underpricing puzzle. For example, Ferris et al. (2013) and Loughran and McDonald (2013) depart from the numerical data that is typically considered to analyze the narratives that contextualize the numbers in IPO prospectuses. Linking the use of language in the disclosure to underpricing, the authors emphasize the importance of qualitative information. However, the focus on quantitative data has predominated so far, presumably because processing textual data is complex and demanding. Nonetheless, it would be negligent to ignore the increasing amount of qualitative information, especially since most information disseminated to potential investors during the IPO process comes in textual form, either spoken or written. For instance, media attention increases when companies announce plans to go public, managers talk to potential investors during the roadshow to market the IPO, and the Security and Exchange Commission (SEC) requires companies to provide information to investors, also in narrative form, to ensure they have the information needed to make well-founded investment decisions. Thus, with the ever-advancing possibilities of natural language processing, the analysis of qualitative data is indispensable and the question inevitably arises as to whether the qualitative information provided contributes to our understanding of IPO underpricing.

This paper contributes to answering this question by analyzing whether and how qualitative information impacts initial-day returns of US Real Estate Investment Trusts (REITs). Prior studies investigating the impact of qualitative information on IPO underpricing exclude the REIT sector due to its financial characteristics, legal requirements, and differing performance.[16] Nonetheless, as evidenced by its recent market capitalization

[16] While the underpricing discount for non-REIT IPOs is consensual, REIT IPOs can historically be either overpriced (e.g. Wang et al., 1992) or underpriced (e.g., Ling and Ryngaert, 1997). However, after the Tax Reform Act of 1986, allowing REITs to be internally managed, studies generally find significant underpricing for REIT IPOs, albeit lower than for industrial firm IPOs (e.g., Dolvin and Pyles, 2009; Gokkaya et al., 2015).

of $1,249,186.3 million as of December 31, 2020 for 223 listed REITs, the US REIT sector is significant to the overall economy and certainly worthy of consideration. In addition, the high dividend payout requirement of at least 90% of their taxable earnings renders REITs dependent on external equity capital in order to take advantage of growth opportunities. As a result, US REIT managers have an unusually strong incentive to reduce underpricing. Likewise, they have a high motivation to maintain investor trust (e.g., Price et al., 2017), which ensures high disclosure quality.

To assess whether qualitative information helps to explain initial-day returns of US REIT IPOs, we analyze the IPO prospectus, or more accurately, the registration statement which must be filed with the SEC before stock offerings, to provide investors with crucial information. For REITs or companies whose primary business is acquiring and holding real estate for investment purposes, the registration statement is presented in Form S-11. Given that the filing must pass the SEC's critical review in order to proceed with the IPO, it is reasonable to assume that Form S-11 discloses valuable information about the issue. We expect this information to be of major importance in solving the underpricing puzzle, as theories attributing the phenomenon to asymmetric information are widely accepted. The variety of theories revolving around asymmetric information thereby explain the underpricing phenomenon by differences in the level of information between issuers and investors or between different groups of investors (i.e., informed vs. uninformed). Clearly, a filing required by the SEC with the intention to provide investors with substantial information about the IPO, should reduce information asymmetry. Therefore, we expect companies filing registration statements that reveal useful and explicit information to have lower underpricing.

Analyzing the level of uncertain language in a prospectus, as well as its similarity to registration forms filed in the past six months, we indeed find that qualitative information in Form S-11 contributes to the understanding of US REIT IPOs. However, against our initial expectation that cautious wording makes it more difficult for investors to value the issue and thus increases initial-day returns, we do not find a statistically significant impact of the level of uncertain language on underpricing. We conclude that cautious wording is induced by forecasting difficulties and litigation risks, rather than reflecting issuer confidence in the future prospects of the firm. By contrast, document similarity is statistically and economically significant related to underpricing of US REIT IPOs. We assume that higher similarity to prior filings indicates that the filing contains a relatively high proportion of standard content or boilerplates, but little useful information. Thus, information asymmetries persist, leading to increased underpricing. In sum, our results show that qualitative information, just like quantitative information, conveys valuable insights, impacts investors' capability to price the issue and thus affects underpricing.

To the best of our knowledge, this is the first study which applies textual analysis to corporate disclosures of US REITs in order to explain IPO underpricing. We demonstrate that textual analysis can indeed help to explain initial-day returns of US REIT IPOs, as qualitative information in Forms S-11 decreases information asymmetries between US REIT managers and investors, thus reducing underpricing.

The remainder of the paper is organized as follows. Section 4.3 presents common theories explaining the underpricing phenomenon, discusses related literature analyzing how textual features affect initial-day returns of industrial firms, summarizes studies examining REIT underpricing, and finally, develops hypotheses. The textual analysis procedure and data used for this study are described in Sections 4.4 and 4.6, while the empirical methods for the analysis are presented in Section 4.5. Section 4.7 discusses the empirical results, and Section 4.8 concludes.

4.3 Previous Literature and Hypotheses Development

4.3.1 Common Theories to explain IPO Underpricing

Since Ibbotson (1975) provided evidence of underpricing, researchers have proposed numerous theories to explain why IPOs have positive initial returns, mostly citing litigation risk, information asymmetries, or behavioral aspects as potential causes.

Ibbotson (1975) originally suggested that underpricing serves as insurance against lawsuits, opening up a whole strand of litigation-based literature. Since the Securities Act of 1933, each party signing the registration statement in the US is liable for its contents, which is the SEC's way of ensuring adequate pre-market due diligence and investor protection (Tinic, 1988). According to Lowry and Shu (2002), issuers try to minimize the risk of being sued and to avoid associated costs, by underpricing their IPOs. Underpricing is intended to prevent financial damages from unforeseen risks that serve as the basis for a lawsuit, and thus reduce investor propensity to sue. While litigation theories help us understand the underpricing puzzle for countries with strong investor protection, such as the US, they have minor importance in other markets.

By contrast, information-asymmetry-based theories, as initially proposed by Ritter (1984) and Rock (1986), find a wider application. According to the authors, underpricing arises due to asymmetric information between investors, whereby informed investors are assumed to be the best-informed stakeholders, as they collectively even know more than the issuer. Once shares are evenly distributed among all participants and informed investors only place orders for underpriced IPOs, uninformed investors suffer from their order being allocated partially to underpriced IPOs, but fully to overpriced IPOs. Hence, IPOs must on average be underpriced, so that uninformed investors generate a positive return and keep participating in the IPO market, despite the predominant allocation of less attractive IPOs. Beatty and Ritter (1986) further find that underpricing increases with rising *ex-ante* uncertainty about an IPO. Firms with higher *ex-ante* uncertainty promise higher returns, which leads to more investors becoming informed. Consequently, uninformed investors are allocated even fewer shares, the more attractive or uncertain an IPO, requiring higher underpricing to participate in the offering. The book-building theory of Benveniste and Spindt (1989) modifies these theories by assuming that shares are distributed with priority to regular investors. Since investors in their entirety are the best-informed stakeholder group, the issuer relies on their premarket indications of interest when pricing an IPO. However, the issuer faces investors who are unwilling to reveal any information, as this would increase the proportion of informed investors and decrease their returns. Using investors' premarket indications of interest in share allocation, issuers force investors to disclose their interest, otherwise investors would jeopardize their own interests by indicating for something other than actual interest. Underpricing hereby acts as an incentive for investors to reveal their information. Sherman and Titman (2002) further emphasize the value of information by arguing that information-generation is costly and can either be conducted by investors or the issuer. According to the authors, it depends on the relative costs and benefits of issuers and investors, regarding who acquires the information. If investors are to participate in the information gathering process, their effort must be compensated for through higher initial returns.

While litigation and information asymmetry theories support the existence of underpricing, they fail to explain oscillations in hot market periods, such as the dot-com bubble. For those periods, behavioral theories, assuming irrational stakeholder behavior, come into play. Miller (1977) proposed that a small group of excessively optimistic investors, who are characteristic of hot markets, can drive up first-day closing prices if they can clear the market. Welch (1992) adds that the sequential sale of the shares creates a cascade effect, where later investors solely rely on prior sales as a buying motive instead of their own private information. Therefore, issuers try to attract early investors through underpricing their IPOs. In contrast, Loughran and Ritter (2002) state that issuers are the ones behaving irrationally, as they do not care about the level of wealth, but about the change in wealth. Issuers most of the time, and especially during hot market phases, experience a positive change in wealth, as the increase in share prices during the IPO usually outweighs the losses due to undervaluation.

Despite major efforts to explain why issuers accept the large amount of "money left on the table", none of the above theories can fully explain the phenomenon. The theories are thus not mutually exclusive and seem to have varying importance during selected time periods and market phases. While behavioral theories help to understand oscillations in hot market periods, litigation based and information asymmetry theories justify the general existence of underpricing. However, the vast majority of theories center on asymmetric information, as this seems to best explain underpricing over time and space.

4.3.2 Textual Analysis and Underpricing of IPOs

Fueled by the increasing availability of textual information and advances in the processing of this complex data, an emerging body of accounting and finance literature has attempted to explain initial-day returns of industrial firms through the use of textual analysis. Thereby, researchers have most frequently relied on news articles or corporate disclosures to examine the impact of qualitative information on IPO pricing.

The literature investigating news articles has mainly focused on news coverage (e.g., Schrand and Verrecchia, 2005; Cook et al., 2006; Chen et al., 2020) or sentiment (e.g., Bajo and Raimondo, 2017; Zou et al., 2020) to assess whether and how textual analysis contributes to solving the underpricing puzzle. News coverage, that is the number of articles mentioning the issuing company before the IPO day, has thereby provided ambivalent evidence on how it affects underpricing. Schrand and Verrecchia (2005) and Chen et al. (2020) show, for a US and global sample respectively, that greater pre-IPO disclosure frequency reduces information asymmetry and hence underpricing. By contrast, Cook et al. (2006) present supporting evidence of a positive correlation between pre-IPO media coverage and underpricing of US IPOs. Using news coverage as a proxy for investment banker promotional activity to attract sentiment investors, this study postulates that higher pre-offer publicity leads to more sentiment investors and higher underpricing. Sentiment, in the sense of the tone conveyed by the media, has been examined, for example, by Bajo and Raimondo (2017) and Zou et al. (2020). While the former claim that a positive media tone increases investor interest and demand for shares, and find that positive sentiment increases first-day returns, the latter state that both positive and negative words contain useful information, thus reducing the degree of information asymmetry between investors and issuer, and thus decreasing underpricing.

Studies focusing on corporate disclosures have investigated, for example, the industrial counterpart to Form S-11, the initial document for registering a non-REIT stock offering with the SEC (Form S-1), or the final IPO prospectus filed at the time of or a few days after the IPO (Form 424). Again, sentiment is a common technique to process qualitative information. For example, Ferris et al. (2013) and Loughran and McDonald (2013) analyze the tone conveyed in certain sections of Form 424 and the entire Form S-1, respectively, and find that conservatism and uncertainty revealed by the issuing firm increase underpricing. According to the authors, conservative or uncertain language reflects issuer confidence in the company's future prospects and makes it more difficult for investors to price the issue, so that companies must set a lower price to attract investors. Conducting similar analyses for the Chinese market, Yan et al. (2019) confirm that uncertain or negative tone is positively associated with first-day returns. In contrast to previously mentioned studies, which focus on financial words, Brau et al. (2016) analyze strategic words and find a more positive strategic tone to be associated with higher underpricing. Brau et al. (2016) interpret their results as suggesting that investors misprice soft information, i.e., the tone in the registration statement. To utilize the qualitative information in Form S-1, Hanley and Hoberg (2010) determine the informativeness of disclosures. Specifically, they split prospectuses into their standard and informative components, where standard content is the part of the prospectus that was already included in recent and past industry IPOs, while informative content is the residual that is not explained by these two sources. In line with Benveniste and Spindt’s (1989) book-building theory, Hanley and Hoberg (2010) find that more informative disclosures signaling greater effort by issuers during premarket are associated with lower underpricing. Instead of analyzing corporate disclosures in their entirety, some studies assume that a specific section of the prospectus is particularly important for pricing, and limit themselves to parts of the disclosure such as the use-of-proceeds or the risk-factor section. Beatty and Ritter (1986), who were among the first to analyze IPO prospectuses, restrict their analysis to the use-of-proceeds section as an example, and find a positive relationship between the number of uses, which serves as a proxy for *ex-ante* uncertainty and underpricing. Similarly, Leone et al. (2007) show that firms that are more specific in the use-of-proceeds section reduce *ex-ante* uncertainty and therefore have lower underpricing. Examining the risk factor section instead, Beatty and Welch (1996) and Arnold et al. (2010) link greater disclosure to higher first-day returns. Beatty and Welch (1996) employ the number of risk factors declared in the prospectus to proxy for *ex-ante* uncertainty, while Arnold et al. (2010) use the section to measure ambiguity, claiming that risk information is *per se* soft and ambiguous, and that investors demand compensation for ambiguous prospectuses.

4.3.3 Underpricing of US REIT IPOs

Whereas the non-REIT literature has analyzed several disclosure-media and textual features to solve the underpricing puzzle, the literature on REIT initial-day returns has so far only relied on quantitative factors to explain the phenomenon.

One of the first studies to analyze REIT IPOs was Wang et al. (1992), who find, in contrast to general evidence from industrial firm IPOs, a return of -2.82% on the first trading day for 87 REITs going public between 1971 and 1988. REIT overpricing was invariant to several influencing factors such as offer price, issue size, distribution method, offer period, and underwriter reputation. Ling and Ryngaert (1997), by contrast, report significant average underpricing of 3.6% for a sample of equity REIT IPOs issued during the period 1991-1994. The authors attribute this turnaround from over- to underpricing to the regime change that occurred with US equity REITs in the late 1980s, and the subsequent increased interest of institutional investors. The evolvement of the REIT industry from being externally managed to internally managed has increased uncertainty regarding the issuing firms' value. Additionally, a larger share of institutional, presumably better-informed investors, has made REIT IPOs more susceptible to the winner's curse (Ling and Ryngaert, 1997). Indeed, analyzing both direct and indirect costs of REIT IPOs, Chen and Lu (2006) find higher gross spreads, overpricing, and high frequency of integer offer prices in the 1980s, which confirms that REIT IPOs were at that time mainly marketed to less informed individual investors. In line with previous literature, initial-day returns shifted from -1.30% in the 1980s to 4.30% in the 1990s for their sample of 197 US REITs. The researchers further identify determinants of the indirect costs of going public (i.e., underpricing). Whereas gross spreads are negatively associated with initial-day returns, underpricing is higher for self-managed REITs, internally advised REITs, UPREITs, and REITs with higher institutional holdings. Bairagi and Dimovski (2011) add to the findings of Chen and Lu (2006), by revealing a negative relation between underpricing and underwriter reputation, industry differentiated auditors, and REITs' post-offer ownership structure. Their study also documents an average underpricing of 3.18% and a value-weighted underpricing of 4.76% for a sample of 123 US REIT IPOs issued between 1996 and 2010, suggesting that offer size affects initial-day returns of US REITs. Buttimer et al. (2005) and Hartzell et al. (2005) analyze the presence of cycles or waves for the REIT industry. The former define a wave (i.e., hot market) as any year with 10 or more REIT IPOs, while the latter use a supply- and demand-side definition represented by the IPO volume and initial day returns, respectively. Both studies find positive initial returns of US REIT IPOs (2.47% and 0.27%), although the stock price performance is the same in both hot and cold markets.

Further adding to the understanding of initial-day returns of US REIT IPOs, Dolvin and Pyles (2009) and Gokkaya et al. (2015) compare underpricing of REIT and non-REIT IPOs. Although both studies find significant underpricing for REIT issuances, initial-day returns are lower than that observed for a matching industrial sample. The lower underpricing of REITs has thereby been attributed to their unique nature and specific regulatory requirements, making them more transparent than general stocks, and thus easier to value for potential investors (Buttimer et al., 2005; Hartzell et al., 2005; and Dolvin and Pyles, 2009). However, despite being considered more transparent than their industrial counterparts, REITs suffer from information asymmetry. Measuring the net asset value of REITs is difficult for investors, as the underlying assets are heterogenous and illiquid. In addition, REITs are actively managed, further complicating valuation due to the growth options developed by managers (Ghosh et al., 2000). Therefore, incorporating the qualitative information from corporate disclosures using textual analysis may also be valuable in solving the underpricing puzzle of US REIT IPOs. REITs thereby offer a unique and useful setting for conducting textual analysis, as they are strongly incentivized to minimize "money left on the table" when going public, and thus guarantee high disclosure quality. Given the high distribution requirement of at least 90% of their taxable earnings, REITs have a very limited cash reserve by regulation, and are highly dependent on external equity capital in order to take advantage of growth opportunities. Thus, consistent with Beatty and Ritter's (1986) asymmetric information hypothesis, REIT managers might voluntarily reveal valuable insights into the upcoming IPO through corporate disclosures, in order to reduce information asymmetry and underpricing.

4.3.4 Hypothesis Development

Recognizing that asymmetric information between the parties involved in the IPO can be reduced by quantitative and qualitative information alike, this study complements earlier studies by investigating whether and how qualitative data in Form S-11 affects initial-day returns of US REIT IPOs. Informing investors about the firms' business model, financial situation, potential problems or risks, and other important information, Form S-11 presents one of the most important documents during the IPO process. Before it is filed, IPO firms are typically without public record. As such, we not only expect Form S-11 to contain valuable insights, but also that sophisticated investors process the filing to use the information in order to value the offer. Under these circumstances, Form S-11 could help to reduce information asymmetry between the parties involved in the IPO. However, the language used in Form S-11 can impact on investors capability to price the issue (Loughran and McDonald, 2013; Ferris et al., 2013). Whereas explicit language would facilitate the inclusion of value-related information, executives tend to use uncertain language, frequently using words like "may," "could," "depends," and "approximately". Although necessary to reduce litigation risk, as the future prospects of the issuing firm are uncertain, cautious wording makes it difficult for investors to evaluate the IPO. Furthermore, word choice may reflect issuer confidence in projecting financial outcomes, so that extreme linguistic caution could even imply that executives themselves are not confident about the company's prospects. Hence, the level of uncertain language in Form S-11 might serve as a proxy for *ex-ante* uncertainty.

In line with Beatty and Ritter (1986), who demonstrate a positive link between *ex-ante* uncertainty about an IPO’s value and its initial return, we expect that firms providing filings with uncertain language need to underprice more, to attract investors, and we thus formulate the following hypothesis:

Hypothesis 1: *Higher levels of uncertain language in Form S-11 make it more difficult for investors to price the issue and thus increase initial-day returns of US REIT IPOs.*

Although Form S-11 is the major document providing information to investors during the IPO process, it is also a standard document required by the SEC. Therefore, Form S-11 contains a relatively high proportion of standardized content. In particular, registration statements filed at the same time, subject to the same regulatory requirements, and exposed to the same market environment, may be similar. Clearly, registration statements which are by large boilerplate do not help investors to determine the true value of the firm. By contrast, unique and specific disclosures that are different from registration statements filed by firms that recently brought an IPO to the market, are likely to provide valuable insights. Thus, we develop a similarity measure indicating the informativeness of a disclosure by comparing it to registration statements filed up to six months prior to the document in question. Assuming that disclosures which are similar to past prospectuses provide little useful information and thus not resolve information asymmetry, we formulate our second hypothesis as follows:

Hypothesis 2: *A higher level of similarity of Form S-11 to previously filed registration statements indicates low information content, persistent information asymmetry, and thus higher underpricing of US REIT IPOs.*

4.4 Textual Analysis Procedure

To test our hypotheses, we extract textual features from Form S-11. Specifically, we determine the level of uncertain language in the prospectus, as well as its similarity to recently filed registration statements.

4.4.1 Disclosure Tone

To determine the level of uncertain language, we rely on sentiment dictionaries which classify the words in the prospectus into pre-defined categories (e.g., positive, negative). Tetlock (2007) provides what is probably the pioneering study applying the dictionary-based approach to extract sentiment in the financial domain. Examining news articles from *The Wall Street Journal*, Tetlock emphasizes that high values of media pessimism induce downward pressure on market prices, and that unusually high or low pessimism predicts temporarily high market-trading volumes. Subsequently, several papers followed his example and relied on dictionary-based approaches to extract qualitative information from text (e.g., Feldman et al., 2010; Loughran and McDonald, 2011; Garcia, 2013). Accordingly, researchers have employed a variety of dictionaries, including general English language dictionaries such as the Diction or the Harvard GI wordlists (e.g., Tetlock, 2007; Davis and Tama-Sweet, 2012), as well as dictionaries created specifically for financial text (e.g., Henry, 2008; Loughran and McDonald, 2011). Since domain-specific wordlists have proven superior (Henry and Leone, 2016; Doran et al., 2012), a financial dictionary created by Loughran and McDonald (2011), by examining word usage in 10-Ks, has become popular. In line with previous literature, we measure *Uncertainty* using the Loughran and McDonald (2011) sentiment wordlists.[17] We thereby follow Loughran and McDonald's (2013) recommendation and use an aggregate uncertainty measure, which comprises words from the uncertain, weak modal, and negative wordlists to determine the level of uncertain language in Form S-11 (*Uncertainty*).

Assigning all words in the prospectus to the respective categories of the LM dictionary, we obtain the raw count of uncertainty words included in the prospectus. Since this number is, however, strongly tied to document length, we use the ratio of uncertain words to the total number of words in the prospectus as a measure of disclosure uncertainty:

[17] The wordlists are available at https://sraf.nd.edu/textual-analysis/resources/.

$$Uncertainty = \frac{Uncertainty\ Words}{Total\ Number\ of\ Words} \tag{4.1}$$

Before applying the financial dictionary of Loughran and McDonald (2011), we clean the data to reduce linguistic complexity and facilitate the textual analysis procedure. Specifically, we remove appendices and html formatting, transfer all text to lower case and eliminate white spaces, numbers, and punctuation.

4.4.2 Disclosure Similarity

Given that we expect prospectuses filed at the same time to be particularly similar, we determine a similarity score (*Cosine*) for each disclosure, by comparing it to registration statements filed up to six months prior to the document in question. If there are not at least two filings available for that time period, we go back further in time to obtain two documents for comparison. Similarity is thus determined using one of the most popular measures for identifying the similarity of two documents – the Cosine Similarity. This technique was used, for example, by Brown and Tucker (2011), who measure year-over-year modifications in MD&A disclosures; Lang and Stice-Lawrence (2015) who examine annual report similarity; Peterson et al. (2015) who compare accounting policy footnotes in 10-K filings; Hoberg and Phillips (2010; 2016) who focus on product descriptions in 10-K filings; and Hanley and Hoberg (2010) who decompose IPO prospectuses into their standard and informative components.

The basic idea of applying cosine similarity to text classification problems is to map documents onto a vector space model (VSM), which enables measuring the similarity between two vectors or textual documents by computing the dot product. We thus create a VSM by representing each IPO prospectus *i* as an *N*-dimensional vector V_i summarizing its word usage. Thereby, each element *n* of the vector V_i corresponds to a specific word that is present in the textual corpus and denotes the frequency of appearance of the respective word in prospectus *i*. For example, the vectors v_1 and v_2 for two documents in the corpus are as follows:

$$v_1 = (w_1, w_2, \dots, w_{N-1}, w_N) \text{ and } v_2 = (\psi_1, \psi_2, \dots, \psi_{N-1}, \psi_N), \tag{4.2}$$

where w_n and ψ_n specify the frequency of each word $n \in [1, N]$. *N* comprises only essential words, i.e., no numbers, punctuation, and stopwords. Additionally, vectors are based on word roots identified using the Porter (1980) stemming algorithm, rather than explicit word inflections.

Having stored all prospectuses *i* in the sample into vectors V_i, we determine the similarity of any two prospectuses by computing the cosine similarity between the corresponding vectors:

$$Cosine\ Similarity\ (v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\|\ \|v_2\|}, \tag{4.3}$$

where $v_1 \cdot v_2$ represents the dot product of the two vectors, $\| v_1\|$ is the length of vector v_1, and $\| v_2\|$ is the length of vector v_2. Thus, the cosine similarity is geometrically the cosine of the angle between the two vectors, normalized to their vector length. Although the dot product itself provides a measure of similarity, normalization avoids over-scoring larger documents. To determine the similarity of three documents (i.e., the prospectus in question and two prospectuses filed up to six months earlier), we determine the mean vector of the comparative documents, and calculate *Cosine* using the mean vector along with the vector of the document of interest. For non-negative vectors, the measure is bounded between 0 and 1, where two prospectuses using exactly the same words have the same orientation, and thus, a cosine similarity of 1, whereas prospectuses with no words in common are orthogonal and have a similarity of 0.

Figure 4.1 provides a simplified example illustrating the textual analysis procedure for determining disclosure similarity. Each of the three introductory statements from Form S-11, A, B, and C, is represented as a vector, with each value indicating the number of occurrences of the corresponding word in the sentence. The resulting VSM enables calculating the level of similarity of the respective statements.

Figure 4.1: Stylized Illustration of Cosine Similarity

Panel A

A is a REIT that focuses on the ownership of community shopping centers.
B is a REIT that focuses on the ownership of regional shopping centers.
C is a REIT that focuses on the acquisition of suburban office properties.

Panel B

	REIT	focuses	ownership	community	shopping centers	regional	acquisition	suburban	office properties
A	1	1	1	1	1	0	0	0	0
B	1	1	1	0	1	1	0	0	0
C	1	1	0	0	0	0	1	1	1

Panel C

	A and B	**A and C**	**B and C**
Vector	v_A = (1,1,1,1,1,0,0,0,0) v_B = (1,1,1,0,1,1,0,0,0)	v_A = (1,1,1,1,1,0,0,0,0) v_C = (1,1,0,0,0,0,1,1,1)	v_B = (1,1,1,0,1,1,0,0,0) v_C = (1,1,0,0,0,0,1,1,1)
Length of vector	$\| v_A\| = \sqrt{5}$, $\| v_B\| = \sqrt{5}$	$\| v_A\| = \sqrt{5}$, $\| v_C\| = \sqrt{5}$	$\| v_B\| = \sqrt{5}$, $\| v_C\| = \sqrt{5}$
Dot Product	4	2	2
Similarity	0.8	0.4	0.4

This figure illustrates the textual analysis procedure for determining cosine similarity. Panel A presents three introductory statements from Form S-11 of US REITs, stopwords are highlighted in grey. Panel B shows tokens in a document frequency matrix. Panel C presents the determination of the similarity score.

With a similarity score (*Cosine*) of 0.8 for sentences A and B and 0.4 for sentences A and C and B and C, respectively, Figure 4.1 shows that sentences A and B are similar, while sentences A and C and B and C are rather different.

4.5 Model Specification

To assess whether qualitative information in Form S-11 helps to explain the underpricing of US REIT IPOs, we regress the tone revealed in Form S-11 (*Uncertainty*) and the similarity score (*Cosine*) on the initial-day returns of the IPO (*IR*), respectively. Specifically, the ordinary least squares (OLS) multiple regression model is as follows:

$$IR_i = \beta_0 + \beta_1 Textual\ Feature_i + \beta_2 Controls_i + \varepsilon_i\ , \tag{4.4}$$

where ε_i denotes the error term and we assume $\varepsilon_i \sim N(0, \sigma^2)$. In addition to the vector of textual features comprising tone (*Uncertainty*) and similarity (*Cosine*), the regression equation includes a vector of control variables (*Controls*), which are described in Section 4.6.

4.6 Data

For the purpose of our analysis, we combine multiple datasets: (1) the text corpus given by Form S-11 obtained from EDGAR, (2) the IPO first-day return calculated by using the first closing price from CRSP and the initial offer price, as well as (3) IPO specific control variables comprising firm characteristics, offering characteristics, third-party certification, and market conditions. Control variables are derived from CRSP, Compustat, Thompson Reuters, and Jay R Ritter's website.

4.6.1 Textual Corpus

Our initial sample includes US Equity REITs that have completed an IPO between January 1996 and December 2019, as reported by the National Association of Real Estate Investment Trusts (NAREIT). Mortgage REITs are excluded from the analysis, because they differ in characteristics (e.g., underlying asset, risk structure), and are recognized as less transparent to value for external investors (Buttimer et al., 2005). For each IPO in the sample, we employ a web-crawling algorithm to download Form S-11 from the SEC's Electronic Data Gathering and Retrieval (EDGAR) database. If there are multiple filings available, we select the closest to the IPO date, to ensure that we include all information available prior to the day of going public. However, we do not intend to download amended filings, as amendments are typically shorter, include more numbers and legal notes. Likewise, the final IPO prospectus (Form 424) is not the subject of this study, as it is filed on or shortly after the day of going public, so that the qualitative information in this disclosure is unlikely to affect IPO underpricing. Although part of the initial sample, prospectuses filed in 1996 are only used to calculate the similarity (*Cosine*) of subsequent IPOs. Since no filings are available to determine *Cosine* for REITs that went public in 1996, these observations are dropped. Furthermore, IPOs that do not have a Form S-11 filing available online, lack the necessary control variables or initial-day returns are eliminated, resulting in an overall sample of 114 US REIT IPOs.

4.6.2 Underpricing

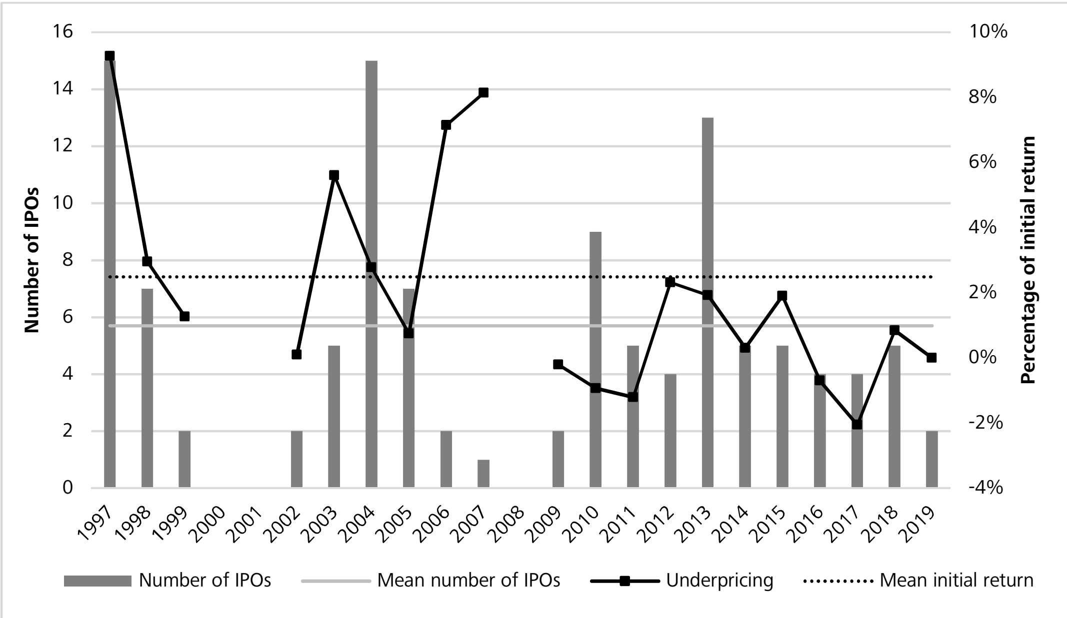
In order to calculate underpricing, we match each REIT that issued an IPO during the sample period with data from CRSP. In accordance with prior studies (e.g., Brobert, 2016; Chen et al., 2020), the initial-day return represents the return enjoyed by IPO investors on the first day of trading of a new issue, and is defined as follows:

$$IR_{i,t} = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}}, \tag{4.5}$$

where $P_{i,t}$ is the closing price on the first day of trading from CRSP, and $P_{i,t-1}$ is the initial offer price as recorded by NAREIT.

Figure 4.2 displays the number of IPOs, as well as average underpricing on a yearly basis. The sample of IPOs is not distributed equally across time; instead Figure 4.2 suggests the prevalence of hot and cold market periods. For example, 2000/2001 and 2008 are cold market periods that correspond with stock market declines, commonly referred to as the end of the dot-com era and the global financial crisis. Similarly, average underpricing varies over time. During the hottest market period (1997-1999), underpricing is high, suggesting that irrationally optimistic sentiment investors were active in the REIT IPO market at that time. Likewise, high initial-day returns are observed in the years before the global financial crisis, while the post-crisis period is characterized by low underpricing.

Figure 4.2: Number of IPOs and average underpricing over Years



This figure shows the number of IPOs, as well as average underpricing on a yearly basis between 1997 and 2019.

4.6.3 Control Variables

To control for information beyond textual clues revealed in Form S-11 that might affect initial-day returns of REITs, a set of control variables is included. We describe all control variables below, and provide more specific definitions in Table 4.5 in the Appendix. We cluster the controls into four subsets: firm characteristics, offering characteristics, third-party certification, and market conditions.

For the first of the four, we include the natural logarithm of the age of the firm (*Age*), and the natural logarithm of the firm's valuation (*Size*) as a measure of *ex-ante* uncertainty. More mature firms provide a longer track record and operating history, while larger offerings are typically issued by well-established firms that have more publicly available information. Thus, the information-gathering process of potential investors is easier for both older and larger firms, which results in a lower level of underpricing. Because we expect more profitable companies to require less underpricing to attract investors, we include a dummy variable equal to one if the firms' earnings per share is positive in the IPO year (*EPS*), as a measure of profitability. Furthermore, we incorporate a common proxy for firm risk, *Leverage*, which we expect to be positively related to underpricing. Recognizing that this study is based on a REIT sample, we additionally include a dummy variable (*Property_Type*) which indicates whether the firm invests in one or more property types (e.g., residential, office, retail). According to Freybote et al. (2008), it is more difficult for investors to value a firm that is invested in multiple property types, so that the issuer must compensate for the complexity by higher initial-day returns.

Controls belonging to the subset of offering characteristics are *Offer_Price*, *Up_Revision*, and *Share_Overhang*. We take into consideration whether the *Offer_Price* is an integer or not, since integer values might signal valuation uncertainty (Bradley et al., 2004). Underwriter and issuing firm dispense with a holistic valuation in order to precisely calculate the price and agree on an overall price, so as to save time and costs (Harris, 1991). Thus, integer values are associated with higher initial-day returns (Bradley et al., 2004). Another control variable related to the offer price is *Up_Revision,* the percentage upward revision from the mid-point of the initial filing range. The offer price is usually raised by the underwriter when the issue is in high demand. However, investors require for compensation to truthfully disclose their demand during book-building. Thus, the offer price is only partially adjusted, with the remainder of the adjustment taking on the form of underpricing to compensate investors (Bradley and Jordan, 2002). This causes a positive relation between *Up_Revision* and underpricing. It is important to note, however, that firms often sell only a fraction of their outstanding shares in the IPO. The unsold shares, or more precisely the number of shares retained divided by the number of shares in the IPO, is usually referred to as *Share_Overhang*. Since the cost of underpricing for the issuer decreases as *Share_Overhang* increases, a larger *Share_Overhang* is associated with greater underpricing (Bradley and Jordan, 2002).

Third-party certification is represented by a dummy variable (*Underwriter*) equal to one if the IPO's lead underwriter was ranked as top tier (value of nine) according to the Carter and Manaster (1990) ranking, as updated by Loughran and Ritter (2004). The underwriter serves a certification role, with underwriters of high reputation guaranteeing a more accurate premarket valuation of the issuing company. Additionally, prestigious underwriters tend to underwrite less risky IPOs in order to protect their reputation. Therefore, top-tier underwriters signal high quality, in addition to increasing the transparency and credibility of the information. Consequently, underwriter prestige is expected to have a negative effect on underpricing.

The final subset of controls addresses market conditions at the time of the IPO and comprises a dummy variable for a *Hot_Market*, and the standard deviation of stock returns for the first 20 days after the IPO (*Volatility*). Following Buttimer et al. (2005), *Hot_Market* equals one in years experiencing ten or more IPOs. Many new IPO issuances indicate positive sentiment in the capital market, making investors more receptive to further offerings and expecting lower compensation (Ljungqvist and Wilhelm, 2003). *Volatility*, as a common measure of risk, shows whether market participants agreed with the pricing before and after the IPO. If this is the case, volatility and underpricing are low (Ascherl et al, 2018).

4.6.4 Descriptive Statistics

Table 4.1 presents descriptive statistics of the characteristics of the 114 REIT IPOs, and textual features of their registration statements. The average underpricing for our sample is a statistically and economically significant 2.49%. This is relatively low compared to non-REITs, but consistent with previous research on US REITs (e.g., Buttimer et al., 2005; Chan et al., 2013). On average, the offer prices are revised upwards by 4.83% from the mid-point of the initial filing range. Hereby, 78.1% of the final offer prices are integers. The average size of a REIT at the time of the IPO is \$665.9 million, with a maximum of \$7,215.6 million and a minimum of only \$12.0 million. However, only 29.94% of the outstanding shares are sold to the market. As evidenced by an average score of 7.7 on the Carter and Manaster (1990) scale and 44.7% selecting an underwriter ranked 9, REITs appear to prefer underwriters with excellent reputation. The average standard deviation of daily stock returns for the first 20 days after the IPO is 1.25%. The age of the issuing company is on average 6 years. In the IPO year, 47.4% of the issuing firms have positive earnings per share and are on average 43.3% debt financed. While 76.3% of the firms were invested in a single property type, the remaining 23.7% held multiple types of real estate assets.

Table 4.1: Descriptive Statistics

	Mean	StDev	Min	5th	Median	95th	Max
			Form S-11				
Uncertainty	0.0256	0.0031	0.0179	0.0202	0.0263	0.0303	0.0325
Uncertainty$_{LM}$	0.0272	0.0032	0.0197	0.0219	0.0277	0.0319	0.0341
Cosine	0.8805	0.0547	0.6650	0.7814	0.8888	0.9456	0.9700
Cosine$_{LM}$	0.8665	0.0562	0.6445	0.7791	0.8754	0.9348	0.9643
			Control Variables				
Age	6.3684	12.0568	1	1	2	26	80
Size (mm)	665.8947	1,091.5417	12.0000	58.6946	32.8077	2,558.1013	7,215.5800
EPS	0.4737	0.5015	0	0	0	1	1
Leverage	0.4333	0.1970	0.0030	0.04087	0.4554	0.7457	0.8352
Property_Type	0.2368	0.4270	0	0	0	1	1
Offer_Price	0.7807	0.4156	0	0	1	1	1
Up_Revision	-0.0483	0.0806	-0.2727	-0.1938	-0.0426	0.0531	0.1500
Share_Overhang	0.2994	0.2848	0.0000	0.0000	0.2084	0.8209	0.9032
Underwriter	0.4474	0.4994	0	0	0	1	1
Hot_Market	0.3772	0.4868	0	0	0	1	1
Volatility	0.0125	0.0078	0.0044	0.0056	0.0114	0.0219	0.0658

see next page

Table 4.1: continued

Dependent Variable							
IR	0.0249	0.0676	-0.1338	-0.0540	0.0017	0.1300	0.3750

This table shows the descriptive statistics for the textual features extracted from Form S-11 (*Uncertainty* and *Cosine*), further control variables, and the dependent variables (*IR*). The definition of all variables is presented in Table 4.5 in the Appendix. The sample consists of 114 US REIT IPOs from 1996 to 2019.

The textual features extracted for the analysis suggest that REIT IPO prospectuses display a low level of *Uncertainty* and are relatively similar to each other. The mean of *Uncertainty* is 0.0256, indicating that on average, 2.56% of the total words in disclosures are uncertainty words. The observed standard deviation of 0.0031 is quite small, so that the uncertainty levels of the individual prospectuses are close to the overall mean. The similarity score, which ranges between 0 and 1 by construction, averages 0.88, and varies from a *5th* percentile of 0.78 to a *95th* percentile of 0.95. The high similarity of documents can be attributed to the comparison within a single industry, but is certainly also induced by the strict regulatory requirements for SEC registration statements and the specific requirements for US REITs.

4.7 Results

4.7.1 Disclosure Tone is not informative

To analyze Hypothesis 1, which predicts that a higher level of uncertain language in Form S-11 is related to increased underpricing, we regress the disclosure tone on the level of underpricing. We run two model specifications, the first of which only includes quantitative factors influencing initial-day returns (Model 1), whereas the second contains traditional factors along with the level of uncertain language (Model 2).

Table 4.2 column one presents regression results with the traditional IPO control variables. Seven of the eleven independent variables are statistically significantly associated with initial-day returns. The insignificance of the variables *Offer_Price*, *Underwriter*, *Leverage*, and *Property_Type* could be due to the strong influence of the other variables. The goodness of fit of this regression is 39.5%. Contradicting our hypothesis, we do not find a statistically significant relation between *Uncertainty* and IPO underpricing (*IR*), when incorporating our measure of uncertain language into Model 2.

Table 4.2: Initial-Day Return – Disclosure Tone

	Model 1 (*Quantitative Factors*)		Model 2 (*Uncertainty*)	
	Regression Results	Economic Significance	Regression Results	Economic Significance
Intercept	-0.344***		-0.270**	
	(-2.776)		(-2.021)	
Age	0.011**	1.22%	0.010**	1.11%
	(2.117)		(2.005)	
Size	0.019***	2.22%	0.019***	2.22%
	(2.955)		(2.981)	
EPS	0.028**	1.40%	0.024**	1.20%
	(2.601)		(2.138)	
Leverage	-0.015	-0.30%	-0.015	-0.30%
	(-0.553)		(-0.530)	
Property_Type	-0.003	-0.13%	-0.001	-0.04%
	(-0.255)		(-0.102)	
Offer_Price	-0.002	-0.08%	-0.004	-0.17%
	(-0.128)		(-0.336)	
Up_Revision	0.283***	2.28%	0.297***	2.39%
	(4.047)		(4.227)	
Share_Overhang	-0.040*	-1.14%	-0.036	-1.03%
	(-1.662)		(-1.480)	
Underwriter	-0.019	-0.95%	-0.019	-0.95%
	(-1.540)		(-1.582)	
Hot Market	0.029**	1.41%	0.026**	1.27%
	(2.619)		(2.327)	
Volatility	0.641	0.50%	0.390	0.30%
	(0.829)		(0.494)	
Uncertainty			-2.635	-0.82%
			(-1.442)	
N	114		114	
R^2	0.395		0.407	
Adj. R^2	0.329		0.336	

This table presents the results of the ordinary least squares (OLS) multiple regression model using first-day IPO returns as the dependent variable. The table reports coefficients and standard errors (in parentheses) of determinants affecting initial-day returns of US REIT IPOs. Economic significance is defined as the coefficient multiplied by the standard deviation. The definition of all variables is presented in Table 4.5 in the Appendix.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This result contrasts with Loughran and McDonald (2013), who find that uncertain language produces higher initial-day returns. However, the authors put their results into perspective, providing a possible explanation for the diverging results. While Loughran and McDonald (2013) analyze a sample of firms characterized by low sales and negative earnings, they clarify that it is unclear whether prospectus wording is as important for large, established firms with highly profitable businesses. With a positive earnings ratio of 48% compared to 37% of firms in the Loughran and McDonald (2013) sample which shows positive earnings in the IPO year, we conclude that our US REIT sample presents precisely this exception. In addition, soft information like language has proven to exert a greater impact on prices when hard information is noisier (Dye and Sridhar, 2004). For example, tech IPOs are more susceptible to soft information (i.e., sentiment) because they are more likely to be driven by hype or buzz caused by fads, news coverage, rumors, or speculation, than by factual information (e.g., historical and current key figures). Accordingly, Ferris et al*.* (2013) find that textual tone in terms of its conservatism (as measured by negative words) affects initial-day returns of tech IPOs, but not those of non-tech IPOs. Compared to a sample of industrial firms that includes tech firms, the US REIT sector can certainly be considered less noisy. While tech companies, especially during the dot-com bubble, did not have a long history and benefited from speculative investments and overly optimistic markets, the real estate sector in general and the US REIT sector in particular have proven to be stable and lucrative investments that guarantee a high level of hard information through specific regulatory requirements. Thus, the insignificant coefficient on *Uncertainty* is plausible, although we initially expected a positive correlation between *Uncertainty* and underpricing.

We conclude that the tone of Form S-11 in terms of its *Uncertainty* does not present a suitable proxy for *ex-ante* uncertainty about an US REIT IPO’s valuation. This is supported by the low standard deviation of our measure of uncertain language. While it was hypothesized that *Uncertainty* might reflect issuer confidence in the future prospects of the firm, the low variation suggests that managers use cautious wording only because the future prospects of the issuing firm are uncertain. If uncertain language provided information about the issuer's expectations, one would expect stronger variations in *Uncertainty* between prospectuses of companies with good future prospects and those about which the issuer itself is less confident.

4.7.2 Disclosure Similarity is informative

Although Form S-11 is the major document providing information to investors during the IPO process, it is also a standard document required by the SEC. As such, we expect the disclosure to include standardized content; particularly prospectuses of companies bringing an IPO to market at the same time could be similar. We hypothesize that disclosures with a higher similarity to previously filed registration statements reveal little useful information to potential investors, thereby preserving information asymmetry and leading to greater underpricing. To test Hypothesis 2, we regress the similarity measure for each registration form on the corresponding first-day return of the IPO. After controlling for firm and offering characteristics, third-party certification, and market conditions, all of which have shown to be associated with initial-day returns in prior literature, we find that the qualitative information revealed in the Form S-11 indeed helps to explain the underpricing of US REIT IPOs (see Table 4.3). Specifically, a one-standard deviation increase in the similarity score (*Cosine*) leads to a 1,61% increase in first-day returns (0.294 coefficient value multiplied by standard deviation of 0.0547). The economic significance of *Cosine* thus lies within the range of widely accepted determinants of IPO underpricing, namely *Size* and *EPS.* Only *Up_Revision*, a variable that can only be measured after the S-11 filing date, specifically on the day the shares are sold, has a higher economic significance of 2,15% (0.267 coefficient value multiplied by standard deviation of 0.0806). The key point to note here is that among the variables with the greatest economic significance, *Cosine* is the only factor that the firm can influence directly before going public. While *Up_Revision* and the firms' earnings per share in the IPO year (*EPS*) can only be determined after the IPO, *Size* cannot be changed by the company at short notice. Further emphasizing the importance of qualitative information, the goodness of fit of the regression improves with R^2 increasing from 39.5% to 44.5% when *Cosine* is added as an independent variable.

Table 4.3: Initial-Day Return – Disclosure Similarity

	Model 1 (*Quantitative Factors*)		**Model 2** (*Cosine*)	
	Regression Results	Economic Significance	Regression Results	Economic Significance
Intercept	-0.344***		-0.559***	
	(-2.776)		(-4.029)	
Age	0.011**	1.22%	0.012**	1.33%
	(2.117)		(2.414)	
Size	0.019***	2.22%	0.016***	1.87%
	(2.955)		(2.670)	
EPS	0.028**	1.40%	0.036***	1.81%
	(2.601)		(3.337)	
Leverage	-0.015	-0.30%	-0.011	-0.22%
	(-0.553)		(-0.394)	
Property_Type	-0.003	-0.13%	-0.004	-0.17%
	(-0.255)		(-0.307)	
Offer_Price	-0.002	-0.08%	-0.004	-0.17%
	(-0.128)		(-0.287)	
Up_Revision	0.283***	2.28%	0.267***	2.15%
	(4.047)		(3.949)	
Share_Overhang	-0.040*	-1.14%	-0.043*	-1.22%
	(-1.662)		(-1.830)	
Underwriter	-0.019	-0.95%	-0.020*	-1.00%
	(-1.540)		(-1.718)	
Hot Market	0.029**	1.41%	0.027**	1.31%
	(2.619)		(2.533)	

see next page

Table 4.3: continued

Volatility	0.641	0.50%	0.484	0.38%
	(0.829)		(0.649)	
Cosine			0.294***	1.61%
			(3.027)	
N	114		114	
R^2	0.395		0.445	
Adj. R^2	0.329		0.379	

This table presents the results of the ordinary least squares (OLS) multiple regression model using first-day IPO returns as the dependent variable. The table reports coefficients and standard errors (in parentheses) of determinants affecting initial-day returns of US REIT IPOs. Economic significance is defined as the coefficient multiplied by the standard deviation. The definition of all variables is presented in Table 4.5 in the Appendix.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Demonstrating that the informativeness of corporate disclosures as measured by document similarity is statistically and economically significant related to underpricing of US REIT IPOs, our results confirm, that qualitative information is as important as quantitative information in solving the underpricing puzzle. Instead of merely restating qualitative data, the narratives in corporate disclosures add context to numerical disclosures, which provides additional insights to investors and thus impacts their capability to price the issue. Accordingly, REIT managers can reduce underpricing by voluntarily providing more information in Form S-11. We acknowledge, however, that there is a trade-off between acquiring costly information in the premarket, so as to provide informative disclosures that simplify the evaluation of the offering, and gathering information from investors during book-building (Sherman and Titman, 2002). In order to provide valuable insights in the prospectus and determine an accurate offer price that potential investors agree with, companies have to bear high costs for information procurement in the premarket. If the company deems this too expensive, it can rely on investors themselves to gather information and value the IPO during book-building. However, investors must be compensated for their efforts through higher initial-day returns.

4.7.3 Robustness

To examine the robustness of our findings, we additionally run the analysis when limiting the total number of words in the prospectus to the number of words that also appear in the Loughran and McDonald (2011) master dictionary. This adjustment ensures that all words included in the textual analysis procedure contain meaningful information, thus reducing the impact of noise caused by irrelevant words.

Uncertainty is thus defined as the ratio of words assigned to the uncertain wordlist, to the number of words in the prospectus also appearing in the Loughran and McDonald (2011) master dictionary. For *Cosine*, the maximum number of elements N of each vector V_i

representing a document is limited to the total number of words that compose the Loughran and McDonald (2011) master dictionary. Furthermore, we now use specific word inflections instead of word roots. Our results are robust to these alternate main independent variables, since all coefficient signs are the same and their magnitudes have a comparable size (see Table 4.4). Again, there is no association between the level of uncertain language (*Uncertainty*) and a REIT's initial-day return, while a higher similarity score (*Cosine*) is associated with increased underpricing. These results survive several other robustness checks which are not tabulated separately. For example, we omit control variables and winsorize all variables to common levels.

Table 4.4: Initial-Day Return – Robustness

	Model 1 (*Quantitative Factors*)		**Model 2** (*Uncertainty*)		**Model 3** (*Cosine*)	
	Regression Results	Economic Significance	Regression Results	Economic Significance	Regression Results	Economic Significance
Intercept	-0.344***		-0.265*		-0.537***	
	(-2.776)		(-1.963)		(-3.889)	
Age	0.011**	1.22%	0.010**	1.11%	0.012**	1.33%
	(2.117)		(1.988)		(2.461)	
Size	0.019***	2.22%	0.019***	2.22%	0.017***	1.98%
	(2.955)		(2.967)		(2.716)	
EPS	0.028**	1.40%	0.024**	1.20%	0.034***	1.71%
	(2.601)		(2.144)		(3.161)	
Leverage	-0.015	-0.30%	-0.014	-0.28%	-0.012	-0.24%
	(-0.553)		(-0.518)		(-0.438)	
Property_Type	-0.003	-0.13%	-0.001	-0.04%	-0.004	-0.17%
	(-0.255)		(-0.087)		(-0.355)	
Offer_Price	-0.002	-0.08%	-0.004	-0.17%	-0.003	-0.12%
	(-0.128)		(-0.335)		(-0.234)	
Up_Revision	0.283***	2.28%	0.298***	2.40%	0.266***	2.14%
	(4.047)		(4.239)		(3.917)	
Share_Overhang	-0.040*	-1.14%	-0.035	-1.00%	-0.044*	-1.25%
	(-1.662)		(-1.461)		(-1.865)	
Underwriter	-0.019	-0.95%	-0.019	-0.95%	-0.020*	-1.00%
	(-1.540)		(-1.586)		(-1.718)	
Hot Market	0.029**	1.41%	0.026**	1.27%	0.028**	1.36%
	(2.619)		(2.315)		(2.561)	
Volatility	0.641	0.50%	0.375	0.29%	0.513	0.40%
	(0.829)		(0.473)		(0.684)	
$Uncertainty_{LM}$			-2.592	-0.83%		
			(-1.435)			
$Cosine_{LM}$					0.266***	1.49%
					(2.817)	
N	114		114		114	
R^2	0.395		0.407		0.439	

see next page

Table 4.4: continued

Adj. R^2	0.329	0.336	0.372

This table presents the results of the ordinary least squares (OLS) multiple regression model using first-day IPO returns as the dependent variable. The table reports coefficients and standard errors (in parentheses) of determinants affecting initial-day returns of US REIT IPOs. Economic significance is defined as the coefficient multiplied by the standard deviation. The definition of all variables is presented in Table 4.5 in the Appendix.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

As our empirical findings persist when using a different approach to quantify textual data and altering the set of controls, we conclude that qualitative information in Form S-11 in fact helps to explain initial-day returns of REIT IPOs.

4.8 Conclusion

REIT IPO candidates provide investors with information about the business model, financial situation, potential problems or risks, and other important information in Form S-11, the initial filing for registering stock offerings with the SEC. However, while quantitative factors have been well studied, the impact of qualitative information on the underpricing of US REIT IPOs has been neglected so far. Recognizing that the vast majority of theories attribute underpricing to asymmetric information between the parties involved in the IPO, which can be reduced by both quantitative and qualitative information, this study examines whether textual features extracted from Form S-11 contribute to the understanding of US REIT IPOs.

To assess whether and how textual features help to explain initial-day returns of US REITs, we determine the level of uncertain language in the prospectus, as well as its similarity to recently filed registration statements. We assume that cautious language makes it more difficult for investors to value the firm, and we thus expect a positive relation between uncertain language and underpricing. Higher similarity to past disclosures suggests that the prospectus provides little useful information, does not resolve information asymmetry, and is therefore associated with increased underpricing.

Contrary to our initial expectation that a higher level of uncertain language is associated with higher initial-day returns, we find an insignificant coefficient for *Uncertainty*. However, this can be explained by the characteristics of the US REIT sample. Being subject to strict regulatory requirements, US REITs are mandated to provide a variety of information on a regular basis. Moreover, they often represent established firms with highly profitable businesses. Thus, the information environment is dominated by facts and figures, known as hard information. According to Dye and Sridhar (2004) US REITS are therefore less prone to soft information such as language. Furthermore, the low standard deviation observed for the measure of uncertain language suggests that cautious wording is induced by forecasting difficulties and litigation risks, rather than reflecting a lack of confidence in the firm's prospects. Therefore, the level of uncertain language in Form S-11 is not suitable as a proxy for *ex-ante* uncertainty. Analyzing the similarity of disclosures brought to the market at the same time, we find a statistically and economically significant impact of qualitative data on initial-day returns of US REIT IPOs. As hypothesized, US REITs providing disclosures that are more similar to previously filed prospectuses do not resolve information asymmetry and thus suffer from increased underpricing. Our findings demonstrate that qualitative information, just like quantitative information, conveys valuable insights and impacts on investor ability to price the issue. Thus, it is up to the REIT managers to reduce underpricing by providing more information to potential investors in Form S-1. We conclude that analyzing qualitative information in corporate disclosures of US REITs offers a new perspective on IPO pricing. Furthermore, our results show that textual analysis can in fact contribute to solving the underpricing puzzle of US REIT IPOs.

To the best of our knowledge this is the first study to analyze the impact of corporate disclosures on underpricing of US REIT IPOs. Clearly, understanding the market reaction to corporate disclosures is essential for REIT managers, investors and regulators alike. Given their high dependency on external capital, REITs are incentivized to provide as much information as possible to reduce underpricing. Investors interested in buying new shares benefit from the opportunity to use textual analysis to identify offerings that promise the highest returns on the first day of trading. Ultimately, regulators can use disclosure similarity to measure the informativeness of required filings and decide on the need for further guidance on that basis.

Despite these many applications, our analysis is subject to certain limitations. The relatively small number of US REITs that went public during the sample period limits ways to process textual data to bag-of-words measures. More sophisticated textual analysis procedures such as machine learning are not applicable. Nonetheless, all contemporary approaches can capture only a small portion of the narratives, which are complex by nature. Further research should therefore aim to expand the information that can be gained from Form S-11 by investigating multiple textual features, for example, by incorporating readability or key word counts. Additionally, it is worth examining whether qualitative information in Form S-11 provides signals regarding other quantitative measures, such as the firms' post-IPO operating performance and volatility.

4.9 References

Arnold, T., Fishe, R.P.H. and North, D. (2010), "The Effects of Ambiguous Information on Initial and Subsequent IPO Returns", *Financial Management*, Vol. 39 No. 4, pp. 1497–1519.

Ascherl, C. and Schaefers, W. (2018), "REITs and REOCs and their initial stock market performance: a European perspective", *Journal of European Real Estate Research*, Vol. 11 No. 1, pp. 4–27.

Bairagi, R.K. and Dimovski, W. (2011), "The underpricing of US REIT IPOs: 1996-2010", *Journal of Property Research*, Vol. 28 No. 3, pp. 233–248.

Bajo, E. and Raimondo, C. (2017), "Media sentiment and IPO underpricing", *Journal of Corporate Finance*, Vol. 46, pp. 139–153.

Beatty, R.P. and Ritter, J.R. (1986), "Investment banking, reputation, and the underpricing of initial public offerings", *Journal of Financial Economics*, Vol. 15 No. 1–2, pp. 213–232.

Beatty, R. P. and Welch, I. (1996), "Issuer Expenses and Legal Liability in Initial Public Offerings." *Journal of Law and Economics*, Vol. 39, pp. 545–602.

Benveniste, L.M. and Spindt, P.A. (1989), "How investment bankers determine the offer price and allocation of new issues", *Journal of Financial Economics*, Vol. 24 No. 2, pp. 343–361.

Bradley, D.J., Cooney, J.W., Jordan, B.D. and Singh, A.K. (2004), "Negotiation and the IPO Offer Price: A Comparison of Integer vs. Non-Integer IPOs", *Journal of Financial and Quantitative Analysis,* Vol. 39 No. 3, pp. 517–540.

Bradley, D.J. and Jordan, B.D. (2002), "Partial Adjustment to Public Information and IPO Underpricing", *Journal of Financial and Quantitative Analysis*, Vol. 37 No. 4, pp. 595–616.

Brau, J.C., Cicon, J. and McQueen, G. (2016), "Soft Strategic Information and IPO Underpricing", *Journal of Behavioral Finance*, Vol. 17 No. 1, pp. 1–17.

Brobert, G. (2016), "The global REIT market: initial-day performance of IPOs", *Journal of European Real Estate Research*, Vol. 9 No. 3, pp. 231–249.

Brown, S. V. and Tucker, J.W. (2011), "Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications", *Journal of Accounting Research*, Vol. 49 No. 2, pp. 309–346.

Buttimer, R.J., Hyland, D.C. and Sanders, A.B. (2005), "Real Estate REITs, IPO Waves and Long-Run Performance", *Real Estate Economics*, Vol. 33 No. 1, pp. 51–87.

Carter, R. and Manaster, S. (1990), "Initial Public Offerings and Underwriter Reputation", *Journal of Finance*, Vol. 45 No. 4, pp. 1045–1067.

Chan, S.H., Chen, J. and Wang, K. (2013), "Are REIT IPOs Unique? The Global Evidence", *Journal of Real Estate Finance and Economics*, Vol. 47 No. 4, pp. 719–759.

Chen, H.C. and Lu, C. (2006), "How much do REITs pay for their IPOs?", *Journal of Real Estate Finance and Economics*, Vol. 33 No. 2, pp. 105–125.

Chen, Y., Goyal, A., Veeraraghavan, M. and Zolotoy, L. (2020), "Media coverage and IPO pricing around the World", *Journal of Financial and Quantitative Analysis*, Vol. 55 No. 5, pp. 1515–1553.

Cook, D.O., Kieschnick, R. and Van Ness, R.A. (2006), "On the marketing of IPOs", *Journal of Financial Economics*, Vol. 82 No. 1, pp. 35–61.

Davis, A.K. and Tama-Sweet, I. (2012), "Managers' use of language across alternative disclosure outlets: earnings press releases versus MD&A", *Contemporary Accounting Research*, Vol. 29 No. 3, pp. 804-837.

Dolvin, S.D. and Pyles, M.K. (2009), "REIT IPOs and the cost of going public", *Journal of Real Estate Finance and Economics*, Vol. 39 No. 1, pp. 92–106.

Doran, J.S., Peterson, D.R. and Price, S.M. (2012), "Earnings conference call content and stock price: the case of REITs", *Journal of Real Estate Finance and Economics*, Vol. 45 No. 2, pp. 402-434.

Dye, R.A. and Sridhar, S.S. (2004), "Reliability-relevance trade-offs and the efficiency of aggregation", *Journal of Accounting Research*, Vol. 42 No. 1, pp. 51–88.

Feldman, R., Govindaraj, S., Livnat, J. and Segal, B. (2010), "Management's tone change, post earnings announcement drift and accruals", *Review of Accounting Studies*, Vol. 15 No. 4, pp. 915–953.

Ferris, S.P., Hao, Q. and Liao, M.Y. (2013), "The Effect of Issuer Conservatism on IPO Pricing and Performance", *Review of Finance*, Vol. 17 No. 3, pp. 993–1027.

Field, L.C. and Karpoff, J.M. (2002), "Takeover Defenses of IPO Firms", *Journal of Finance*, Vol. 57 No. 5, pp. 1857–1889.

Freybote, T., Rottke, N. and Schiereck, D. (2008), "Underpricing of European property companies and the IPO cycle: A note", *Journal of Property Investment & Finance*, Vol. 26 No. 5, pp. 376–387.

Garcia, D. (2013), "Sentiment during Recessions", *Journal of Finance*, Vol. 68 No. 3, pp. 1267–1300.

Ghosh, C., Nag, R. and Sirmans, C.F. (2000), "A Test of the Signaling Value of IPO Underpricing with REIT IPO-SEO Pairs", *Journal of Real Estate Finance and Economics*, Vol. 20 No. 2, pp. 137–154.

Gokkaya, S., Highfield, M.J., Roskelley, K. and Steele, D.F. (2015), "An Examination of Primary and Secondary Market Returns in Equity REIT IPOs", *Journal of Real Estate Research*, Vol. 37 No. 1, pp. 23–64.

Hanley, K.W. and Hoberg, G. (2010), "The information content of IPO prospectuses", *Review of Financial Studies*, Vol. 23 No. 7, pp. 2821–2864.

Harris, L. (1991), "Stock Price Clustering and Discreteness", *Review of Financial Studies*, Vol. 4 No. 3, pp. 389–415.

Hartzell, J.C., Kallberg, J.G. and Liu, C.H. (2005), "The Role of the Underlying Real Asset Market in REIT IPOs", *Real Estate Economics*, Vol. 33 No. 1, pp. 27–50.

Henry, E. (2008), "Are investors influenced by how earnings press releases are written?", *Journal of Business Communication*, Vol. 45 No. 4, pp. 363-407.

Henry, E. and Leone, J.A. (2016), "Measuring qualitative information in capital markets research: comparison of alternative methodologies to measure disclosure tone", *Accounting Review*, Vol. 91 No. 1, pp. 153-178.

Hoberg, G. and Phillips, G. (2010), "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis", *Review of Financial Studies*, Vol. 23 No. 10, pp. 3773–3811.

Hoberg, G. and Phillips, G. (2016), "Text-based network industry classifications and endogenous product differentiation", *Journal of Political Economy*, Vol. 124 No. 5, pp. 1423–65.

Ibbotson, R.G. (1975), "Price performance of common stock new issues", *Journal of Financial Economics*, Vol. 2 No. 3, pp. 235–272.

Lang, M. and Stice-Lawrence, L. (2015), "Textual analysis and international financial reporting: Large sample evidence", *Journal of Accounting and Economics*, Vol. 60 No. 2–3, pp. 110–135.

Leone, A.J., Rock, S. and Willenborg, M. (2007), "Disclosure of Intended Use of Proceeds and Underpricing in Initial Public Offerings", *Journal of Accounting Research*, Vol. 45 No. 1, pp. 111–153.

Ling, D.C. and Ryngaert, M. (1997), "Valuation uncertainty, institutional involvement, and the underpricing of IPOs: The case of REITs", *Journal of Financial Economics*, Vol. 43 No. 3, pp. 433–456.

Ljungqvist, A. and Wilhelm, W.J. (2003), "IPO Pricing in the Dot-Com Bubble", *Journal of Finance*, Vol. 58 No. 2, pp. 723–752.

Loughran, T. and McDonald, B. (2011), "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks", *Journal of Finance*, Vol. 66 No. 1, pp. 35–65.

Loughran, T. and McDonald, B. (2013), "IPO first-day returns, offer price revisions, volatility, and form S-1 language", *Journal of Financial Economics*, Vol. 109 No. 2, pp. 307–326.

Loughran, T. and Ritter, J.R. (2002), "Why Don't Issuers Get Upset About Leaving Money on the Table in IPOs?", *Review of Financial Studies*, Vol. 15 No. 2, pp. 413–444.

Loughran, T. and Ritter, J.R. (2004), "Why has IPO Underpricing Changed over Time?", *Financial Management*, Vol. 33 No. 3, pp. 5–37.

Lowry, M. and Shu, S. (2002), "Litigation risk and IPO underpricing", *Journal of Financial Economics*, Vol. 65 No. 3, pp. 309–335.

Miller, E.M. (1977), "Risk, Uncertainty, and Divergence of Opinion", *Journal of Finance*, Vol. 32 No. 4, p. 1151.

Peterson, K., Schmardebeck, R. and Wilks, T.J. (2015), "The Earnings Quality and Information Processing Effects of Accounting Consistency", *Accounting Review*, Vol. 90 No. 6, pp. 2483–2514

Porter, M.F. (1980), "An algorithm for suffix stripping", *Program*, Vol. 14 No. 3, pp. 130–137.

Price, S.M., Seiler, M.J. and Shen, J. (2017), "Do investors infer Vocal cues from CEOs during quarterly REIT conference calls?", *Journal of Real Estate Finance and Economics*, Vol. 54 No. 4, pp. 515-557.

Ritter, J.R. (1984), "The "Hot Issue" Market of 1980", *Journal of Business*, Vol. 57 No. 2, p. 215.

Rock, K. (1986), "Why new issues are underpriced", *Journal of Financial Economics*, Vol. 15 No. 1-2, pp. 187–212.

Sherman, A.E. and Titman, S. (2002), "Building the IPO order book: underpricing and participation limits with costly information", *Journal of Financial Economics*, Vol. 65 No. 1, pp. 3–29.

Schrand, C.M. and Verrecchia, R.E. (2005), "Information Disclosure and Adverse Selection Explanations for IPO Underpricing", available at: http://dx.doi.org/10.2139/ssrn.316824.

Tetlock, P.C. (2007), "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", *Journal of Finance*, Vol. 62 No. 3, pp. 1139–1168.

Tinic, S.M. (1988), "Anatomy of Initial Public Offerings of Common Stock", *Journal of Finance*, Vol. 43 No. 4, pp. 789–822.

Wang, K., Chan, S.H. and Gau, G.W. (1992), "Initial public offerings of equity securities. Anomalous evidence using REITs", *Journal of Financial Economics*, Vol. 31 No. 3, pp. 381–410.

Welch, I. (1992), "Sequential Sales, Learning, and Cascades", *Journal of Finance*, Vol. 47 No. 2, pp. 695–732.

Yan, Y., Xiong, X., Meng, J.G. and Zou, G. (2019), "Uncertainty and IPO initial returns: Evidence from the Tone Analysis of China's IPO Prospectuses", *Pacific Basin Finance Journal*, Vol. 57, 101075.

Zou, G., Li, H., Meng, J.G. and Wu, C. (2020), "Asymmetric Effect of Media Tone on IPO Underpricing and Volatility", *Emerging Markets Finance and Trade*, Vol. 56 No. 11, pp. 2474–2490.

4.10 Appendix

Table 4.5: Description of Control Variables

Variable	Description
Age	The natural logarithm of a firm's age in years. Age is thereby defined as the period of time between a firm's founding date and its IPO in years. Founding dates are obtained from Datastream and the Field-Ritter dataset, as used in Field and Karpoff (2002) and Loughran and Ritter (2004).
Size	The natural logarithm of of the firm's valuation, measured as the product of the offer price and the number of shares outstanding.
EPS	Dummy variable equal to one if the firm's earnings per share is positive in the IPO year.
Leverage	Ratio of total liabilities to total assets; LT/AT
Property_Type	Dummy variable equal to one if the issuing firm invests in multiple property types.
Offer_Price	Dummy variable equal to one if the IPO offer price is an integer.
Up_Revision	The percentage upward revision from the mid-point of the filing range.
Share_Overhang	The number of shares retained, divided by the number of shares in the IPO.
Underwriter	Dummy variable equal to one if the IPO's lead underwriter was ranked as top tier (value of nine) according to the Carter and Manaster (1990) ranking, updated by Loughran and Ritter (2004).
Hot Market	Dummy variable equal to one for observations in years experiencing ten or more IPOs.
Volatility	The standard deviation of stock returns for the first 20 days after the IPO.

This table describes the control variables used.

5 Conclusion

This dissertation examines whether qualitative data revealed in corporate disclosures of US REITs conveys useful information to market participants, helping them to make well-informed investment decisions. The three papers comprising this thesis hereby differ in the textual analysis procedure, disclosure media, methodologies, textual features, and outcomes investigated. The following section summarizes the motivation, research design, and key findings for each article and concludes with final remarks, along with suggestions for further research.

5.1 Executive Summary

Paper 1 | Is the MD&A of US REITs informative? A Textual Sentiment Study

Although the "MD&A is intended to give the investor an opportunity to look at the company through the eyes of management" (SEC, 1987) the informativeness of the MD&A has frequently been criticized (e.g., Pava and Epstein, 1993). With the aim of assessing the informativeness of US REIT disclosures, this study examines the language disclosed in approximately 6,500 MD&As filed by the SEC between 2003 and 2018. Specifically, this study investigates whether textual sentiment conveyed through the MD&A of US REITs provides signals regarding future firm performance, and thus generates a market response.

To determine the overall sentiment inherent in each filing, that is, the level of pessimistic or optimistic language, the Loughran and McDonald (2011) financial dictionary and a custom wordlist for the real estate industry created by Ruscheinsky et al. (2018) are employed. Thereafter, a panel fixed effects regression enables investigating the relationship between sentiment and future firm performance, as well as the market reaction.

Confirming the informativeness of the MD&A, this study finds that higher levels of pessimistic (optimistic) language in the MD&A are indeed associated with lower (higher) future firm performance. This holds even after controlling for the information released in other concurrent disclosures that may predict future performance. In accordance with prior literature from Rogers et al. (2011), Doran et al. (2012), and Henry and Leone (2016), the use of a domain-specific real estate dictionary, namely that developed by Ruscheinsky et al. (2018), leads to superior results. Furthermore, this study finds a significant market response to pessimistic language in the MD&A at the time of the SEC filing. However, corresponding to the notion that the human psyche is affected more strongly by negative than positive news (Rozin and Royzman, 2001), the impact of optimistic language is insignificant.

This is the first study to provide evidence that the use of language in the MD&A reveals US REIT manager expectations regarding future firm performance and that the market responds to this information. Given that investigating the language in the MD&A decreases information asymmetries between US REIT managers and investors, this study demonstrates that the market can indeed benefit from textual analysis.

Paper 2 | Can Risks be Good News? Revealing Risk Perception of Real Estate Investors using Machine Learning

The SEC mandates firms to discuss the most relevant factors that may entail speculative or risky aspects for the firm in their 10-Ks, so that investors are able to monitor the current and future risk factors a firm is facing and integrate them into their decision-making analysis. However lengthy and complex disclosures – mostly for dozens of firms in an investor's portfolio – can barely be processed by a human being. To cope with the flood of information, this study proposes using unsupervised machine-learning approaches (STM, CTM, and LDA) to identify and quantify the risk factors discussed in REITs' 10-Ks.

Assuming that documents are characterized as a collection of topics, and topics as a collection of words, STM, CTM, and LDA automatically cluster words around topics and thus enable uncovering latent topics in a relatively short time period. However, since LDA is limited when identifying common risk factors across industries (i.e., property types of REITs), this study focusses on the advanced topic modeling approaches (STM and CTM). Hereby, STM has demonstrated superiority over CTM.

To assess whether the STM presents a valid approach to quantifying risk in narrative form, this study examines whether the machine-extracted risk factors help to explain the perceived risk on the stock market. Analyzing a US REIT sample between 2005 and 2019, this study finds that the majority of the STM-extracted risk topics is significantly associated with volatility, confirming the effectiveness of the model. Furthermore, this study disentangles how the identified risk topics explain investor risk perception, in addition to traditional firm characteristics. Although initial intuition suggests that risks are *per se* negative, a risk-reducing effect of risk disclosures is expected, since most risk factors are revealed in a timely manner through press releases or Form 8-K. Thus, instead of disclosing new risk factors, Item 1A primarily provides essential information on risk factors that have already been communicated to investors using more frequent channels, thereby reducing uncertainty and risk perception. Accordingly, most of the STM-identified factors follow the convergence argument, indicating a risk-reducing effect.

To the best of our knowledge, this is the first study employing STM in the accounting and finance domain. By analyzing the topics' generated vocabulary, the nature and scope of the disclosed risk factors can be identified. This allows assessing the stock market reaction to each risk factor. Therefore, investors would benefit from using machine-learning techniques which enable them to process a huge amount of company information simultaneously, just as the risk items in the 10-Ks.

Paper 3 | Can Textual Analysis solve the Underpricing Puzzle? A US REIT Study

REIT IPOs provide investors with information about the business model, financial situation, potential problems or risks, and other important information in Form S-11, the initial filing for registering stock offerings with the SEC. However, while quantitative factors have been well studied, the impact of qualitative information on underpricing of US REIT IPOs has been neglected so far. Recognizing that the vast majority of theories attribute underpricing to asymmetric information between the parties involved in the IPO, which can be reduced by both quantitative and qualitative factors, this study examines whether textual features extracted from Form S-11 contribute to solving the underpricing puzzle of US REIT IPOs. In particular, this study determines the level of uncertain language in the prospectus, using the Loughran and McDonald (2011) wordlists, as well as its similarity to recently filed registration statements employing one of the most popular measures to identify the similarity of two documents – the Cosine Similarity. To assess whether and how textual features help to explain initial-day returns of US REITs, this study regresses the level of uncertain language and document similarity on the initial-day returns of a sample of 114 US Equity REITs that completed an IPO between January 1996 and December 2019.

Contrary to the initial expectation that a higher level of uncertain language makes it more difficult for investors to price the issue and thus increases underpricing, this study does not find a statistically significant association between uncertain language and initial-day returns. The study concludes that cautious wording is induced by forecasting difficulties and litigation risks, rather than reflecting issuer confidence in the future prospects of the firm. Analyzing the similarity of disclosures brought to the market at the same time, this study finds a statistically and economically significant impact of qualitative data on initial-day returns of US REIT IPOs. Specifically, a higher similarity to previous filings suggests that the prospectus provides little useful information and thus does not resolve existing information asymmetries, leading to increased underpricing. Overall, the results confirm that qualitative information, just like quantitative information, conveys valuable insights and impacts on investor capability to price the issue. Thus, textual analysis can in fact contribute to solving the underpricing puzzle of US REIT IPOs.

This is the first study to analyze the impact of corporate disclosures on initial-day returns of US REIT IPOs. Yet, understanding the market reaction to corporate disclosures is essential for REIT managers, investors and regulators alike.

5.2 Final Remarks and Suggestions for Further Research

While academics and practitioners have traditionally relied on quantitative data as the basis for economic decision-making, the qualitative data contextualizing numbers in corporate disclosures has lately become increasingly valuable to the financial sector. However, the US REIT market and its narratives have been largely neglected, despite its emergence as an attractive asset class that offers investors an efficient way to diversify their investments, reduce risk and enhance long-term returns. The present dissertation addresses this important omission in the literature by investigating whether qualitative data revealed in corporate disclosures of US REITs indeed conveys valuable insights to market participants.

Specifically, the first paper examines whether language disclosed in the MD&A of US REITs provides signals regarding future firm performance and generates a market response. The empirical results suggest that sentiment, that is, the level of pessimistic or optimistic language in the MD&A, is indeed associated with future firm performance. Moreover, this study finds a significant market response to pessimistic language in the MD&A at the time of the SEC filing. Corresponding to the notion that the human psyche is affected more strongly by negative than positive news (Rozin and Royzman, 2001), the market does not respond to optimistic language in the MD&A. The second paper complements the analysis on annual reports and exploits an unsupervised machine learning algorithm to identify the specific risk-factor topics discussed in 10-Ks. To evaluate the validity of this approach, the study investigates whether the probability of appearance of the extracted risk factors helps to explain the perceived stock market risk. The results indicate that the majority of risk factors is indeed significantly associated with volatility confirming the effectiveness of the machine-assisted modeling. The third paper determines the level of uncertain language in IPO prospectuses, as well as their similarity to prospectuses filed up to six months prior to the document in question, to assess whether textual features can solve the underpricing puzzle. Contrary to expectations, this study does not find a statistically significant association between uncertain language in Form S-11 and initial-day returns. This finding disproves the initial assumption that uncertain language makes it more difficult for potential investors to price the issue and thus increases underpricing. On the other hand, document similarity is statistically and economically significantly related to the underpricing of US REIT IPOs. Thus, the hypothesis that a prospectus showing high similarity to previous filings provides little useful information and does not resolve existing information asymmetries is confirmed.

In their entirety, the three papers forming this dissertation show that the market responds to various textual features (i.e., sentiment, risk factor topics, disclosure similarity) revealed through corporate disclosures of US REITs, thus, providing evidence that market participants incorporate more than just quantitative data into their decision making. Rather, qualitative information appears to constitute an essential component of the information set that financial market participants use for economic decision making. Furthermore, the results of all three studies demonstrate that US REITs reveal material information in their narratives, indicating that qualitative information in corporate disclosures is truly informative. In this respect, the SEC appears to be successfully pursuing its mission of maintaining efficient markets. More importantly however, this dissertation demonstrates the tremendous necessity of analyzing qualitative information. Thereby, the use of computer-based techniques to process textual data can deliver significant competitive advantage. The consequent ability of processing textual data more quickly and accurately equips users of textual analysis for faster and better-informed decisions. For example, textual analysis allows traders to identify market-moving events at an early stage, providing them with enough time to initiate respective steps to capitalize on changes, or at least avoid losses. Thus, investors can utilize textual analysis to enhance returns, reduce risk or increase efficiency.

Although demonstrating the importance of qualitative information and the potential of conducting textual analysis on corporate disclosures, this dissertation is subject to certain limitations. While representing the most comprehensive analysis, it is still restricted to a subset of the enormous volume of disclosures available. To further evaluate the importance of qualitative data, subsequent research should investigate other disclosure outlets such as press releases, internet postings, and news media. Similarly, future research could be devoted to other aspects of the unstructured world of big data such as imagery, recordings, audio, video. Moreover, textual analysis is often criticized for being imprecise due to its inability to identify the full information content of a document. However, the findings in this dissertation show that such lack of precision is not something that precludes using this technique. More so, the given limitations of the technology should motivate researchers to approach this shortcoming and develop new methodologies that enable to extract more information from texts. In this regard, this study aims to serve as a starting point to a generation of researchers who analyze qualitative information to gain insights as valuable as their quantitative predecessors.

5.3 References

Doran, J.S., Peterson, D.R. and Price, S.M. (2012), "Earnings Conference Call Content and Stock Price: The Case of REITs", *Journal of Real Estate Finance and Economics*, Vol. 45 No. 2, pp. 402–434.

Henry, E. and Leone, J.A. (2016), "Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone", *Accounting Review*, Vol. 91 No. 1, pp. 153–178.

Loughran, T. and McDonald, B. (2011), "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks", *Journal of Finance*, Vol. 66 No. 1, pp. 35–65.

Pava, M. and Epstein, M. (1993), "How good is MD&A as an investment tool?", *Journal of Accountancy*, Vol. 175 No. 3, pp. 51–54.

Rogers, J.L., Van Buskirk, A. and Zechman, S.L.C. (2011), "Disclosure tone and shareholder litigation", *Accounting Review*, Vol. 86 No. 6, pp. 2155–2183.

Rozin, P. and Royzman, E.B. (2001), "Negativity Bias, Negativity Dominance, and Contagion", *Personality and Social Psychology Review*, Vol. 5 No. 4, pp. 296–320.

Ruscheinsky, J.R., Lang, M. and Schäfers, W. (2018), "Real estate media sentiment through textual analysis", *Journal of Property Investment and Finance*, Vol. 36 No. 5, pp. 410–428.

SEC. Securities and Exchange Commission (1987), Securities Act Release No. 6711, available at: https://www.sec.gov/rules/interp/33-6835.htm.

Universität Regensburg
IRE|BS Institut für Immobilienwirtschaft
Fakultät für Wirtschaftswissenschaften