Revealing the Risk Perception of Investors using Machine Learning

Marina Koelbl[1], Ralf Laschinger[2], Bertram I. Steininger[3], and Wolfgang Schaefers[4]

26 August 2021

Abstract

Text in corporate disclosures conveys important information to financial market participants. If incorporated in quantitative models, the intended meaning of text is often hidden by the use of idiosyncratic terminology within an industry-specific vocabulary. This study uses an unsupervised machine learning algorithm, the Structural Topic Model, to overcome this issue. It illustrates the connection between machine-extracted risk factors discussed in corporate disclosures (10-Ks) and the corresponding pricing behavior of investors for a not yet investigated US REIT sample from 2005 to 2019. When disclosed, most risk factors counterintuitively decrease stock return volatility and are therefore beneficiary for the pricing process on financial markets.

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

JEL Classification: C45; C80; G14; G18; K22; K40; M41; M48; R30

Acknowledgements: The authors wish to thank Thies Lindenthal, Masaki Mori, McKay Price, Erik Devos, Erkan Yönder, and the participants of the AREUEA-ASSA Conference 2021, Cambridge Real Estate Seminar 2021, ARES and ERES Meeting 2021, and the AREUEA International Conference 2021 for their insightful comments and suggestions on earlier versions of this paper. This paper circulated previously under the title: "Can Risks be Good News? Revealing Risk Perception of Real Estate Investors using Machine Learning".

[1]University of Regensburg
IREBS International Real Estate Business School
Universitätstr. 31
93040 Regensburg, Germany
marina.koelbl@wiwi.uni-regensburg.de

[2]University of Regensburg
Department of Finance
Universitätstr. 31
93040 Regensburg, Germany
ralf.laschinger@wiwi.uni-regensburg.de

[3]KTH Royal Institute of Technology
Banking and Finance
Teknikringen 10 B
100 44 Stockholm, Sweden
bertram.steininger@abe.kth.se

[4]University of Regensburg
IREBS International Real Estate Business School
Universitätstr. 31
93040 Regensburg, Germany
wolfgang.schaefers@wiwi.uni-regensburg.de

Revealing the Risk Perception of Investors using Machine Learning

26 August 2021

Abstract

Text in corporate disclosures conveys important information to financial market participants. If incorporated in quantitative models, the intended meaning of text is often hidden by the use of idiosyncratic terminology within an industry-specific vocabulary. This study uses an unsupervised machine learning algorithm, the Structural Topic Model, to overcome this issue. It illustrates the connection between machine-extracted risk factors discussed in corporate disclosures (10-Ks) and the corresponding pricing behavior of investors for a not yet investigated US REIT sample from 2005 to 2019. When disclosed, most risk factors counterintuitively decrease stock return volatility and are therefore beneficiary for the pricing process on financial markets.

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

JEL Classification: C45; C80; G14; G18; K22; K40; M41; M48; R30

1. Introduction

It is still a matter of academic debate, whether markets efficiently incorporate information into prices. In financial markets, pricing is a continuous process of investors' reactions to new information (Fama, 1970). The resulting outcome is thereby often characterized by the first two moments of distribution: the expected return and the standard deviation of returns. Since the seminal work of Markowitz (1952), the expected return is perceived as the desirable property of an asset, while its standard deviation is the undesirable one. A low standard deviation of returns does not imply stable prices, but rather a low fluctuation around the mean of returns. Consequently, low volatility is a sign of consistent expectations across investors regarding returns when new information emerges. Contrary, high volatility indicates dissent about how to value and incorporate new information. By revealing information, a new pricing process begins on their release date resulting in three possible outcomes: no price reaction if the information is irrelevant or already known among the investors, increasing volatility if the investors are in disagreement with the pricing outcome of the information, or decreasing volatility if the investors coincide about the informational impact on the firm's future prospect. From a theoretical perspective, new information can increase or decrease investors' risk perception. In line with this ambiguity, empirical research identifies information factors increasing as well as decreasing the volatility; whereas the latter finding is in the majority. We propose a new method to identify which information factors are positive or negative linked with risk to dissolve these mixed empirical findings.

Previous studies about market efficiency show theoretically and empirically that information asymmetry reduces market efficiency and increases stock misvaluation (e.g., Ross, 1973; Myers 1984; Myers and Majluf, 1984; Miller and Rock, 1985). An effective tool to overcome this asymmetry is to inform the public of any relevant news helping them to make the right decision and thereby finding the right price. For the US, the Securities and Exchange Commission (SEC) demands various standardized disclosures of publicly listed firms to establish and maintain efficient markets. For that, firms are mandated to discuss the factors which make a firm speculative or risky in their 10-Ks (see 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 (i.e. Item 1A – a section describing risk factors in 10-K filings) are less explored than in quantitative form (e.g., stock volatility).

To process text in corporate disclosures, the Latent Dirichlet Allocation (LDA), a topic model approach based on machine learning, has become predominant in economics and finance. The advantage of LDA is that it does not require predefined rules (i.e. *a priori* determined keywords) to quantify latent topics within a huge amount of documents. The disadvantage is that LDA tends to identify already known topics since the Dirichlet distribution assumes almost uncorrelated topics and ignores the existence of idiosyncratic language (covariate words) within a subset of the documents; for example, firms within the same sector often use similar terminology. These methodical drawbacks are partly solved with its technical successor, the Correlated Topic Model (CTM, see Blei and Lafferty, 2007) which has so far not been used empirically. Surprisingly, even more sophisticated approaches are not yet used in the financial and accounting domain. To overcome the problems encountered in the quantitative analysis of textual disclosures, we propose the application of the Structural Topic Model (STM), which includes covariates (idiosyncratic language) and covariances between topics (see Roberts et al., 2019). Figure 1 highlights the formulized problem of the LDA as well as the proposed solution by the STM.

<<< Insert Figure 1 here. >>>

The text corpora (corpus A and B) in Figure 1 illustrate examples of our later-used data set. The identified words defining the topics by 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 not identified by LDA but 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.

Our study contributes to the literature in various ways. To the best of our knowledge, this is the first study applying STM to the accounting and finance domain while also benchmarking it with LDA and CTM. We show, that the so-far predominantly used LDA is biased by the used idiosyncratic language within an industry reflecting rather the already known operative line of business or business models than significant topics of a document. This is also true for CTM, the advanced LDA algorithm, which is the most suitable benchmark for STM although it is not used in the economic literature so far. In addition, our analysis provides insights into whether and how information is incorporated into the pricing process. By introducing STM, we apply the algorithm to the important but rather neglected asset class of REITs (Real Estate Investment Trusts). This industry is an appealing testing ground for multiple reasons. First, while the sector is described by relatively homogenous business models and firm characteristics, its firms invest in different property types (e.g., healthcare, residential). Consequently, the industry-specific vocabulary distracts the LDA and CTM from extracting common risk factors. We show that LDA and CTM are distracted from extracting common risk factors and can therefore hardly be linked to the pricing behavior of investors. Contrary, the STM-extracted risk factors are statistically significantly associated with volatility and consequently, with the risk perception of investors.

The remainder of the paper is organized as follows. Section 2 discusses related literature on mandatory risk disclosures and develops hypotheses. Section 3 explains the textual analysis procedures (i.e. LDA, CTM, and STM) and the empirical model, while Section 4 introduces the data used and describes the variables. The empirical results are reported in Section 5, and Section 6 concludes

2. Previous Literature and Hypotheses Development

2.1. 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" by counting specific words. 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 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 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; Jegadeesh and Wu, 2013; Chen et al., 2014), 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 extract topics out of corporate risk disclosures using machine learning approaches.

2.2. Textual Analysis of Risk Disclosures

The literature 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 forward-looking disclosures (e.g., risk disclosures) are linked to an increased trading volume (Kravet and Muslu, 2013), and lower future earnings and stock returns (Li, 2006). The result for stock return volatility is not so clear; the majority find a decreasing effect (e.g., Beyer et al., 2010; Muslu et al., 2015), whereas others an increasing effect (e.g., Kravet and Muslu, 2013; Campbell et al., 2014). Common to the studies using straightforward approaches is that they can process a large number of textual documents which is beyond human capacity, but 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 the majority have no effect at all. Gaulin (2019) 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 (2020) 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 general speaking topics derive naturally from fitting the statistical model to the textual corpus, based on word co-occurrences in the documents.

2.3. Hypotheses

Common to all approaches, whether straightforward or sophisticated, is that they attempt to quantify qualitative information in disclosures without the need for a human being to read them. However, quantifying risk disclosures 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, forward-looking 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 10-K filings. In fact, previous research (e.g., Kravet and Muslu, 2013; Bao and Datta, 2014) finds a stock market reaction of risk disclosures confirming its informativeness.

Recognizing that management's discretion entails considerable leeway in deciding which information about a risk factor is disclosed and how much of the filing is allocated to a particular risk-factor topic, we assume that these probabilities of topics provide valuable information on how companies assess the extent of the risks. Accordingly, the topic probabilities in the filings could serve as a proxy for risk beyond the level of previous straight-forwarded proxies (e.g., word count, text length), allowing investors to quantify the information provided in narrative form.

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

The nature of risk disclosures is that it explains but does not necessarily resolve uncertainties. Thus, theoretic models (e.g., Kim and Verrecchia, 1994; Cready, 2007) see the possibility that risk disclosures increase or decrease investors' risk perceptions. Kravet and Muslu (2013) define three opposing arguments. The first argument suggests that investor risk perceptions remain unaffected since risk disclosures are vague and use boilerplates because managers are likely to report 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). This ambiguity is supported by the mixed results in empirical research (see previous subsection), whereas the majority find resolved uncertainties (lower volatility) in response to corporates' disclosures. Since we are able to extract risk topics at a higher level of granularity than previous straight-forwarded risk proxies, we assume that we find all three risk perceptions (null, convergence, and divergence argument). Knowing that the annual frequency of 10-Ks is from the legal and practical perspective inappropriate to discuss new risks, we assume that the majority of disclosures resolve known risk factors and contingencies and formulate our next hypothesis as follows.

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

3. Model Design

3.1. Textual Analysis with Machine Learning: Topic Models

Topics derive naturally from fitting the statistical model to the textual corpus based on word co-occurrences in the documents. Thus, this procedure eliminates subjectivity that would otherwise be introduced by predefined wordlists, and yet provides more informative results than straight-forwarded approaches, which can still be interpreted economically. 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). It is a mixture model, 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 used Dirichlet distribution assumes almost uncorrelated topics. However, they are likely correlated in reality since particular topics occur at the same time. For an illustration, see Figure 1 in our Introduction. 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. They find that CTM is always superior to LDA for altering the number of topics from 5 to 120.[1]

The Structural Topic Model (STM) by Roberts et al. (2019) goes even one step further and incorporates metadata of pre-specified covariates (industry-specific vocabulary), not only covariances; see Figure 1 in our Introduction. Again, it remains a mixture model based on a logistic normal distribution, so that it corresponds to CTM if covariates are ignored. More details on the STM algorithm are given in the next subsections, technical details are in Appendix A, and the use of covariates is explained in Subsection B.1 and Table B.2 in Appendix B.

We run various tests checking whether the higher flexibility of STM corresponds to a better fitting among the approaches. The better the topic identification works the higher the probability that the topics may help to explain the investors' risk perception. In a pre-test, we run a technical comparison for CTM and STM similar to Blei and Lafferty's (2007) comparison for LDA and CTM. We fit a smaller collection of documents of our later-used dataset to a varying number of topics (between 10 and 25) and calculate the residuals, lower bounds, and log likelihoods of the held-out data. The better a model fits the lower are the residuals and the higher are the lower bounds as well as the probability of the held-out data. All three measures indicate a better fit for STM for the full range of topic numbers (see Figure 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 disclosures.

<<< Insert Figure 2 here. >>>

Based on the superiority of CTM over LDA (see Blei and Lafferty, 2007) and STM over CTM as well as LDA (see Roberts et al., 2014 and our pre-test), we assume that STM is most suitable to extract

[1] 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).

topics explaining the investors' risk perception. In our later analysis (Subsection 5.5), we compare the explanatory power of all three approaches to explain the investors' risk perception.

3.2. Topic Identifications: Pre-steps

To apply topic models, we use the programming language R (version 4.0.2) and the corresponding packages topicmodels and STM, authored by Grün and Hornik (2011) and Roberts et al. (2019). Several preprocessing steps are necessary before running the topic models. First, we parse with the edgarWebR package the downloaded 10-K filings to extract the risk report part from the entire document. In addition, we clean the data by removing 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 words appearing in fewer than 20 disclosures to avoid their influence. On the one hand, this threshold (20 occurrences) rules out words occurring solely in 10-K filings of one particular firm (e.g., the firm names), since we have 14 years of observations. 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.

3.3. Topic Identifications: Risk Factors Labeling

Although topic-modeling approaches classify textual data without further instruction by the user, the topics created by the algorithms (LDA, CTM, and STM) do require an 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. We discuss the associated words selected labels in Subsection 5.3. Table B.1 in Appendix B presents the full list of the 20 highest associated words for each risk factor topic for STM and the corresponding name; Table C.1 in Appendix C does it for LDA.

3.4. Risk Model Specification

To assess whether the probabilities of appearance of the extracted risk factors helps to explain the perceived risk on the stock market, we regress whose frequencies ($Freq\_Topics$) on the firms' stock return volatility ($Vola$) by using the following two-way fixed-effects regression model:

$$Vola_{it} = \beta_0 + \beta_1 Freq\_Topics_{it} + \beta_2 Controls_{it} + a_i + \lambda_t + u_{it} \, , \qquad (1)$$

where i denotes the firm, and t the year. In addition to the vector of the distribution 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 incorporates 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.

4. Data

4.1. Data Source and Sample

To test our hypotheses, we combine multiple datasets: (1) investors' risk perception proxied by stock volatility from CRSP, (2) the text corpus given by the Risk Factor report (Item 1A) of the annual 10-Ks obtained from the Electronic Data Gathering and Retrieval (EDGAR) database, and (3) firms' financial and accounting fundamentals obtained from Compustat or Thomson Reuters.

Our sample begins with the earliest date when Item 1A. Risk Factors was available (December 1, 2005) and extends through the fiscal year-end 2019.[2] In contrast to other studies focusing on the entire firm-year sample available from EDGAR database, we limit our examination to a single industry, namely the REIT industry, for multiple reasons. First, while the sector is characterized by relatively homogenous business models and firm characteristics, different investment foci in property types (e.g., healthcare, residential) are salient and distract the LDA from extracting common risk factors (see Figure 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 well-known business model – high investments in fixed assets generate relatively constant cash flow for their

[2] Actually, there is a second risk section in the 10-K. Item 7A lists "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.

investors. This property is attractive for institutional investors since the early 1990s as shown by others (e.g., Ling and Rygaert, 1997; Lee et al., 2008). This type of investor can process lengthy and complex disclosures easier, so it is reasonable to assume that can observe stock market reactions based on the disclosed information. Furthermore, investors must intensively monitor this type of industry for adverse information and outcome (risk) since their capital is tied in fixed assets, which do not have high future expectancies regarding new technologies or where losses can be compensated by new exceptional growth opportunities. In addition, institutional investors are rarely driven by noise trading or herding behavior, which irrationally influence the stock prices. However, institutional investors apply often passive investment styles with a buy-and-hold strategy and a long-term horizon (see e.g., Chung et al., 2012; Devos et al., 2013). Consequently, positive news keeps the ownership of institutional investors constant whereas negative news may not lead to a direct divestment, if they are not severe.

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. Figure 3 displays the sample composition of the 10-Ks over years; it mostly follows the number of REITs included in the FTSE NAREIT All REITs Index over the same time period. For some years, the observations exceed the number of index constituents, since we include a firm in our sample if it was a constituent at any point during the period. We thus address survivorship bias and index effects such as greater investor attention to firms listed in an index. 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 obtained from CRSP and Compustat and not the risk factors extracted from the 10-K filings (see Table 1 for more details about *N*).

<<< Insert Figure 3 here. >>>

4.2. Investors' Risk Perception

The dependent variable of interest is the perceived risk on the stock market measured by the stock volatility after the filing date using the daily closing prices from CRSP. 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. The 5 trading-day period gives investors 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 quarterly reports (10-Qs).[3] We calculate volatility as the standard

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

deviation of daily log returns extrapolated to the 5, 40, and 60 trading-day periods after the 10-K filing day.

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

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

In contrast to the common approach using a 252 trading-day volatility, our procedure concentrates on the volatility induced by the information released in the 10-K. A 252 trading-day window may be 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 our testing period. Additionally, the standard deviation over a 252 trading-day window would cause autocorrelation problems after adding a control variable for the lag volatility for the days before the 10-K filing date, since the majority of the time window overlap. We illustrate this in Figure 4, Panel A.

By contrast, our method 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 4, Panel B.

<<< Insert Figure 4 here. >>>

4.3. Independent Variables

Our primary influencing variables of interest are the frequencies of the machine learning-extracted risk factors discussed in corporate disclosures ($Freq\_Topic$). We start with the STM and verify our results using CTM and LDA; their calculations are described in Section 3. To control for information beyond the risk factors, 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 B.3 in Appendix B. We cluster the controls into two subsets: 1) accounting-based/market-based and 2) 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. We calculate FFO by following NAREIT's guideline: the sum of net income, amortization & depreciation, and the difference of 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, controls for Fama and French's (1993) finding that small firms are more volatile than large firms; we expect its coefficient to be negative. $Leverage$ 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 (ΔREV) as well as sales growth ($Sales\_Growth$). ΔREV 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. Among the market-based controls, $Beta$ proxies the firm risk similar to $Leverage$, so that we expect a positive nexus to volatility. Book-to-Market (BTM) is calculated as the book value of equity, scaled by the market capitalization of equity. Our expectations of 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 the corresponding T trading-days before the 10-K filing date, to control for positive volatility 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 add 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 the general market 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 ($\Delta\, Volume$), 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 Thomson 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.

For the second subset of controls, we include straight-forwarded textual content measures of previous research. In line with Campbell et al., (2014) who show that the number of words is positively related to stock return volatility, we incorporate the natural logarithm of the total text length of the risk sections ($Text\_Length$). Additionally, we follow Li (2008) and Lehavy et al. (2011) and incorporate the readability measured by the Gunning fog index (FOG) to account for higher information-processing costs of complex language.

4.4. Descriptive Statistics

Table 1 presents descriptive statistics for all variables. The STM's frequencies for the risk factor topics ($Freq\_Topic$) sum to 1 within each document but not over all documents. We observe rather small topic frequencies for Item 1A by looking at their means; the highest is around 7.6% for Topic #16 "Property", the lowest for Topic #14 "REIT Status" 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 distribution of all topics is extremely skewed so that we use a log transformation of these factors in our later regressions. 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 B.4 in Appendix B). Thus, the topics have no direct linear relationship, but as shown in Section 3, 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.

<<< Insert Table 1 here. >>>

The classical fundamentals in the control set show the common 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 B.5 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 B.5 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).

5. Results

5.1. Topic Models and Investor Risk Perception

To test whether the probabilities of risk topics help to explain investor risk perception (Hypothesis 1), we regress those probabilities 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.

<<< Insert Table 2 here. >>>

After controlling for firm-level characteristics and other textual measures 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 for all three model specifications. The risk factors are statistically more relevant (12 of 16 topics) in the short run (Model 1) than in the long run (Model 2 and 3) with 6 and 7 topics, respectively. Not only the number but also the magnitudes of the risk factors' coefficients decrease from Model 1 to Model 3 with the exemption of Topic #1 "Transaction" and #15 "Single Tenant Risk". The number of significant firm fundamentals does not vary among the three model specifications, but their magnitude increases on average in the long run, which is in line with the efficient market hypothesis. The results for the other topic model approaches (LDA and CTM) have similar results for the fundamentals (significance and magnitude). However, the majority of whose risk topics are insignificant which is in line with Bao and Datta (2014). We compare all approaches in more detail in Subsection 5.5 and use STM for the next analyses since it is more efficient to extract topics explaining the investors' risk perception.

Some fundamentals are never relevant ($FFO/Share$, ΔREV, and $Sales\_Growth$), others increase their impact over the time horizons 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 ($\Delta Volume$) also increase their impact over the models with a longer time window. The two alternative textual variables ($Text\_Length$ and FOG) are never relevant so that the risk factors convey the information. The goodness of fit (R^2) decreases from Model 1 to Model 2 and 3 (32% vs. 18% and 27%) due to the lower importance of the risk factors but improves from Model 2 and 3. This latter effect is mostly driven by the higher importance of few controls (IO, $Beta$, and Lag_Vola) in the long run.

5.2. Risk Disclosures resolve Uncertainties

To test Hypothesis 2, which predicts a risk-reducing effect for the majority of risk factors, we evaluate the coefficient signs of the extracted risk factors. Consistent with Bao and Datta (2014), our results provide support for all three influencing effects. Contrary to those who find that the majority of their LDA-extracted risk factors carry no relevant information for the market, the majority of our STM-extracted risk factors reduce significantly the volatility and follow therefore the convergence argument.

In Model 1 (5-day window), four risk factor topics #6, #9, #12, and #20[4], have an insignificant coefficient, 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). The convergence factors are in the majority (topics #1, #3, #8, #10, #13, #15, #16, #17, and #19), which is in line with the assumption that firms use 10-Ks to resolve known risk factors or give more facts about known risk factors and thus, reduce risk perceptions among investors. These values are economically significant, too. For example, the standardized beta of topics #1, #3, and #13; if we increase the risk topic by one standard deviation, the volatility decreases by -17%, -24%, and -53% of its standard deviation. The economic impact for the divergence topics is on average greater with 91%, 107%, and 23%. Overall and on average, the risk topics' impact is on the same scale as those of the traditional

[4] We describe the topic labels in the next subsection.

fundamental variables (e.g., $Size$ 6%, Lag_Vola 37%, or BTM -137%). The results for the longer time windows (Model 2 and 3) are the same as discussed in the previous subsection: the risk factors are more relevant in the short run (Model 1) than in the long run (Model 2 and 3) and most fundamentals increase their impact in the long run.

Based on the statistical and economic significance of the convergence factors, we conclude that executives use this type of disclosure (Item 1A in 10-K) mainly to resolve risk instead of presenting new risk factors so that risk disclosures may even be seen as ‘good news’ as long as they clarify the impact of already known factors. This is in line with the majority of the previous literature of a volatility reducing effect of risk disclosures even if they are not or only to a limited extent able to explain why this happens (e.g., Huang and Li, 2011). Common to most of the so-far used measures (e.g., text length or number of keywords) is that they do not allow a deeper look (i.e. semantic) into the risk-reducing drivers of their – mostly – single risk factor model. Our proposed solution instead allows to combine risk increasing and reducing effects in a single model.

5.3. Semantic and Economic Interpretation

Topic modeling has the advantage that it delivers more risk factors with a higher granularity which can be interpreted economically (e.g., Bao and Datta, 2014). For example, STM does 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 (see Model 1 of Table 2). 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, these topic labels give only a first insight. Topic modeling provides the set of words (e.g., top 20) representing the risk factor while researchers choose the label. Therefore, labels may not describe topics entirely. Israelsen (2014) gets to the heart of this dilemma by stating that “it is the words that define the topics, not the title”. For example, the convergence factor #1 “Transaction” includes words such as ‘unenforceable’, ‘origination’, ‘repurchases’, and ‘sale-leaseback’. The frequent appearance of phrases such as ‘plaintiffs’, ‘defendant’, ‘supreme’, and ‘prejudice’ suggests that the corresponding topic #13 is related to “Legal & Litigation risk”. For other 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 first glance, the words 'migration' and 'recycling' do not fit with the other words in the divergence topic #5 (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.

5.4. 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 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 word-count 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 and the results of the regression model which follows Equation (1) are in Table 4.

<<< Insert Table 3 here. >>>

<<< Insert Table 4 here. >>>

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 (Subsection 5.1), we observe lower significant coefficients for the risk factors if we move to 40 trading days (8 risk factors instead of 7) or to 60 trading days (8 risk factors instead of 7). In comparison to the probability model, the absolute allocation of words model explains the variations better; the R^2 is on average 2 percentage points greater 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 $Freq\_Topics$.

Based on the comparable coefficients and the higher explanatory power for the $Abs\_Allocation$ model, we evaluate this hybrid model as a good instance 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 firm-specific accounting and market control variables.

5.5. Alternative of Risk Perception and Alternative Topic Models

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. They 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 B.6 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 CTM and LDA extracted risk factor topics over the 5 trading-day and 60 trading-day periods and compare them with STM. 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 within the economic field (see Subsection 3.1) as the advanced approach. The results are presented in Table B.8 in Appendix B.

5.6. Validity of the STM to capture Changes in Reporting Behavior

The lessons of the subprime crises (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 5 illustrates these growth rates for selected topics whose reporting certainly changed during or after the crisis: Regulation (#2), Rating (#5), and Single Tenant Risk (#15).

<<< Insert Figure 5 here. >>>

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 subprime 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 (#5) 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., subprime crisis) occur. Thus, disclosure frequencies reflect changes in firms' reporting behavior caused by specific events, confirming the validity of the STM.

6. 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 the so-far most used LDA is limited when identifying common risk factors across industries or sectors, we extend the applied toolbox with the advanced topic modeling approaches (STM and CTM) and are the first who apply these techniques in the accounting and finance domain. We are able to confirm the theoretical and previously shown superiority of STM over CTM and LDA in an economic application.

To assess whether our machine-assisted topic 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 in general. Indeed, we find that the majority of risk topics are significantly associated with volatility, confirming the effectiveness of our model in comparison to LDA-focused studies which find for example mostly insignificant results (Bao and Datta, 2014). Furthermore, we allow our fine-grained risk topics to carry all three types of risk perception (null argument, divergence argument, and convergence argument, see Kravet and Muslu, 2013). This helps us to resolve contradicting results in the literature by our way of addressing a problem.

We find evidence supporting all three types of price reactions to information. 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 (nine risk factors) decrease risk perceptions (convergence argument). The predominance of risk-reducing risk factors is in line with the majority of the previous literature using more straight forwarded measures. In addition to their method of measuring qualitative textual information by counting words, we can combine this idea of an impact by quantity with our measure of probability. This hybrid model – combining machine learning with the word counting factor – confirms our previous finding and explains best the variations within our dataset. This achieved finding would not be possible by the so-far used approaches. Thus, we conclude that a combination of the classical word count and our machine-assisted topic modeling helps to explain investor risk perceptions most efficiently. This is our contribution from the technical part.

From the practical part, we contribute the finding that Item 1A in the 10-K filings primarily provides essential information on risk factors resolving uncertainties instead of disclosing new risk factors. 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.

Our findings support the pursue to reduce information asymmetry by regulators (e.g., SEC) since both firms and shareholders benefit from reduced volatility showing that markets efficiently incorporate information into prices. In addition, our idea combining machine learning/topic modeling with a classical and straight forwarded word counting method as well as state-of-the-art econometric models may help to pave the way for more applications of natural language processing since previous methods were not able to give a deeper understanding of whether and which risk topics influence investors' risk perception.

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.

Beyer, A., Cohen, D.A., Lys, T.Z. and Walther, B.R. (2010), “The financial reporting environment: Review of the recent literature”, *Journal of Accounting and Economics,* Vol. 50 No. 2-3, pp. 296–343.

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. (2012), “Probabilistic topic models”, *Communications of the ACM,* Vol. 55 No. 4, pp. 77–84.

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 No. 1, 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.

Chung, R., Fung, S. and Hung, S.Y.K. (2012), “Institutional Investors and Firm Efficiency of Real Estate Investment Trusts”, *Journal of Real Estate Finance and Economics,* Vol. 45 No. 1, pp. 171–211.

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

Cready, W.M. (2007), “Understanding rational expectations models of financial markets: A guide for the analytically challenged”, available at: http://dx.doi.org/10.2139/ssrn.999409.

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.

Devos, E., Ong, S.E, Spieler, A.C. and Tsang, D. (2013), “REIT Institutional Ownership Dynamics and the Financial Crisis”, *Journal of Real Estate Finance and Economics,* Vol. 47 No. 2, pp. 266–288.

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.

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

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. (2019), “Risk Fact or Fiction: The information content of risk factor disclosures”, Working Paper, Rice University.

Grün, B. and Hornik, K. (2011), “topicmodels: An R package for fitting topic models”, *Journal of Statistical Software,* Vol. 40 No. 1, pp. 1–30.

Huang, K.W. and Li, Z.L. (2011), “A multilabel text classification algorithm for labeling risk factors in SEC form 10-K”, *ACM Transactions on Information Systems*, Vol. 2 No. 3, pp. 1–19.

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.

Kim, O. and Verrecchia, R.E. (1994), “Market liquidity and volume around earnings announcements”, *Journal of Accounting and Economics*, Vol. 17 No. 1-2, pp. 41–67.

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.

Lee, M.-L., Lee, M.-T. and Chiang, K.C.H. (2008), ”Real Estate Risk Exposure of Equity Real Estate Investment Trusts”, *Journal of Real Estate Finance and Economics*, Vol. 36 No. 165, pp. 165–181.

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.

Ling, D. 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.

Lopez-Lira, A. (2020), “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.

Markowitz, H.M. (1952), “Portfolio Selection”, *Journal of Finance*, Vol. 7 No. 1, pp. 77–91.

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

Miller, M.H. and Rock, K. (1985), “Dividend Policy under Asymmetric Information”, *Journal of Finance*, Vol. 40 No. 4, pp. 1031–1051.

Muslu, V., Radhakrishnan, S., Subramanyam, K.R. and Lim, D. (2015), “Forward-Looking MD&A Disclosures and the Information Environment”, *Management Science,* Vol. 61 No. 5, pp. 931–948.

Myers, S.C. (1984), “The Capital Structure Puzzle”, *Journal of Finance,* Vol. 39 No. 3, pp. 574–592.

Myers, S.C. and Majluf, N.S. (1984), "Corporate Financing and Investment Decisions When Firms Have Information That Investors Do Not Have", *Journal of Financial Economics,* Vol. 13 No. 2, pp. 187–221.

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. and 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.

Ross, S.A. (1973), "The Economic Theory of Agency: The Principal's Problem", *American Economic Review,* Vol. 63 No. 2, pp. 134–139.

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

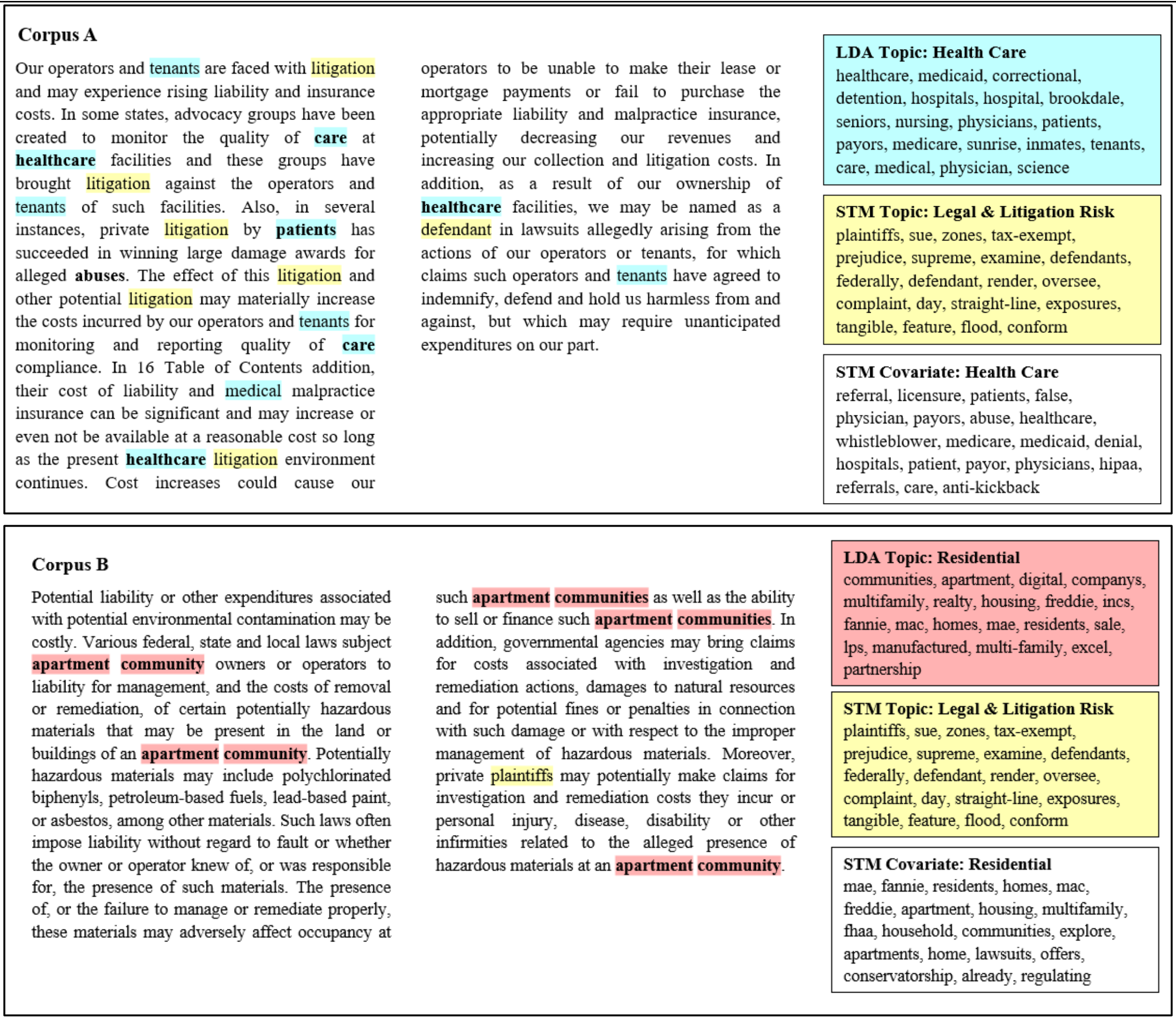
Tetlock, P.C. (2007), "Giving content to investor sentiment: the role of media in the stock market", *Journal of Finance,* Vol. 62 No. 3, pp. 1139–1168.

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.

Figures

Figure 1



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

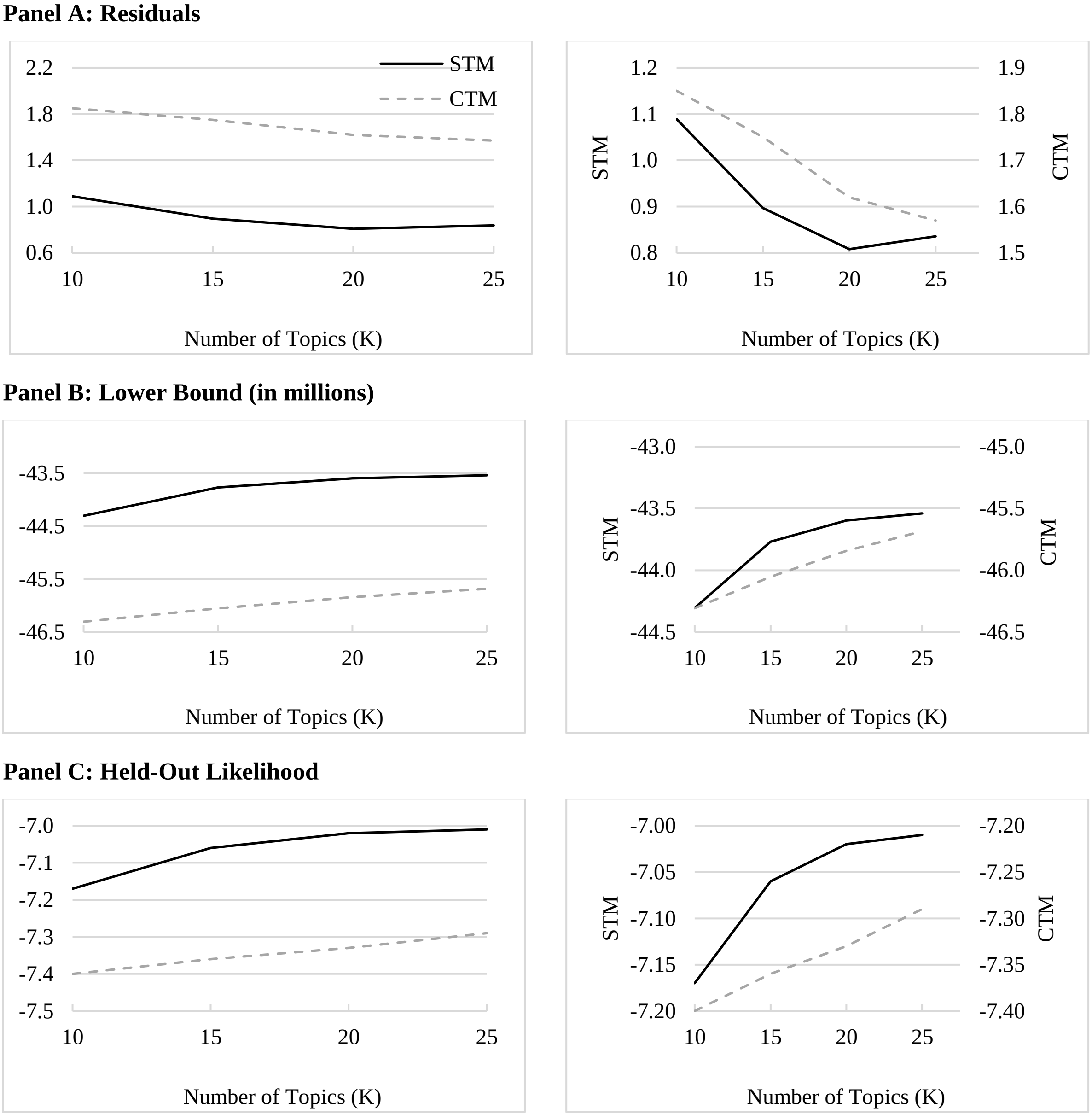
STM Topic: Legal & Litigation Risk
plaintiffs, sue, zones, tax-exempt, prejudice, supreme, examine, defendants, federally, defendant, render, oversee, complaint, day, straight-line, exposures, tangible, feature, flood, conform

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.

Figure 1: Stylized Illustration of LDA and STM

Figure 2



Panel A: Residuals

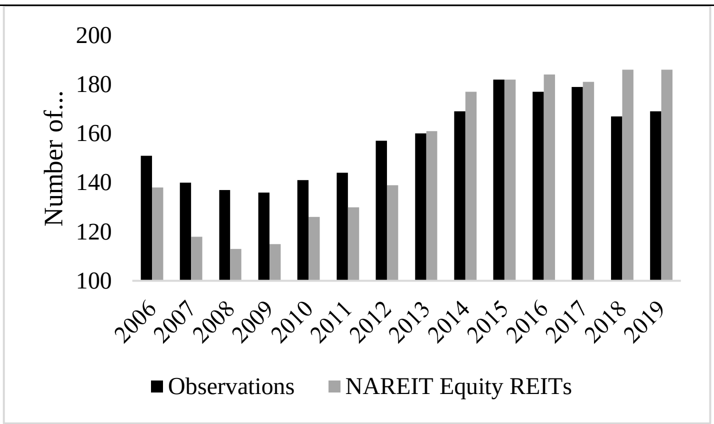
Panel B: Lower Bound (in millions)

Panel C: Held-Out Likelihood

This figure shows the standard criteria for comparing different topic models, namely residuals regarding the text corpora (Panel A), lower bound (Panel B), and held-out likelihood with a standard of 20 percent (Panel C). 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.

Figure 2: Comparison of CTM and STM

Figure 3

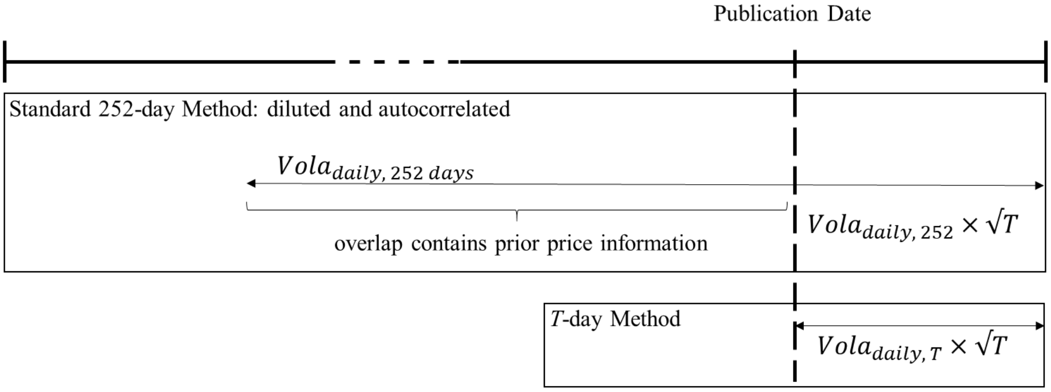


This figure shows the number of observations included in the sample and the number of Equity REITs present in the FTSE NAREIT All REITs Index over years.

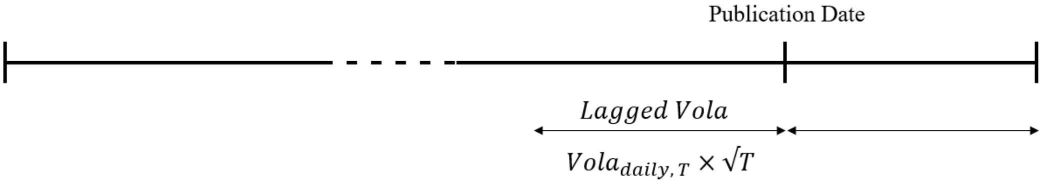
Figure 3: Sample Distribution over Years

Figure 4

Panel A: *T*-day Method


Publication Date
Standard 252-day Method: diluted and autocorrelated
$Vola_{daily,\,252\ days}$
overlap contains prior price information
$Vola_{daily,\,252} \times \sqrt{T}$
T-day Method
$Vola_{daily,\,T} \times \sqrt{T}$

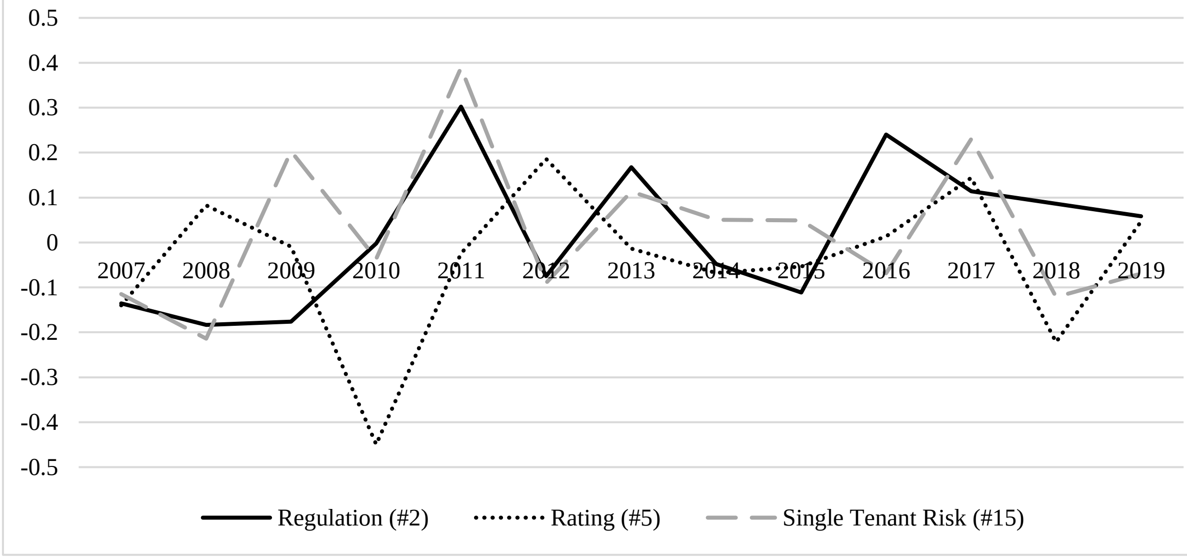
Panel B: Lagged Volatility



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.

Figure 4: Volatility around Publication Date

Figure 5



This figure shows yearly growth rates of the probability of appearance for the topics Regulation (#2), Single Tenant Risk (#15), and Rating (#5).

Figure 5: Yearly Growth Rate of the Probability of Appearance

Tables

Table 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
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	1,861	1.986	4,114	-18.258	0.593	1.385	2.579	127.368
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
Δ*REV*	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
Δ*Volume* (0, 5 days)	1,543	0.119	0.893	-4.306	-0.049	0.025	0.183	20.333
Δ*Volume* (0, 40 days)	1,529	0.052	0.545	-2.601	-0.085	0.001	0.095	7.790
Δ*Volume* (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 B.3 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.

Table 2: Probability of Appearance – Risk Perception

	Model 1	**Model 2**	**Model 3**
	(0, 5 days)	(0, 40 days)	(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
Location	(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)
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)

see next page

Table 2: continued

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 B.3 in Appendix B.

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

Table 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, 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 4: Absolute Allocation of Words – Risk Perception

	Model 1	**Model 2**	**Model 3**
	(0, 5 days)	(0, 40 days)	(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)
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)

see next page

Table 4: continued

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)
Δ*Volume*	0.007***	0.020***	0.020***
	(0.002)	(0.004)	(0.004)
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 B.3 in Appendix B.

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

Online Appendices

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 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:

$$\vec{\theta}_d \mid X_d\gamma, \Sigma \sim LogisticNormal(\mu = X_d\gamma, \Sigma)\,, \tag{A.1}$$

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:

$$\beta_{d,k} \propto exp(m + K_k + K_{y_d} + K_{y_d,k}) \tag{A.2}$$

Figure A.1 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)

<<< Insert Figure A.1 here. >>>

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.

Figure A.1

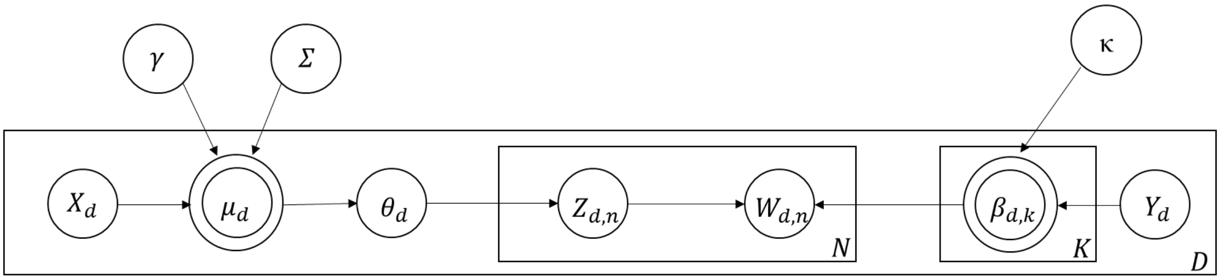


Figure A.2: Structural Topic Modeling, in plate notation (following Roberts et al., 2019)

Appendix B

Table B.1: 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
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
see next page

Table B.1: continued

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: Location
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.

B.1. Metadata Covariates for STM

To apply 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 B.2 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 C.1 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 C.1 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 C.2 in Appendix C).

Table B.2: 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
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 B.3: 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; $T \in [5, 40, 60]$.
Δ*Vola*	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
Δ*REV*	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.
Δ*Volume*	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 B.4: 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 B.5: 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 B.6: 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)
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
Location	(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)

see next page

Table B.6: continued			
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)
Δ*Volume*	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

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 B.7 in Appendix B. The definition of all variables is presented in Table B.3.

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

Table B.7: 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 B.8: 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)
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)

see next page

Table B.8: continued

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)
Δ*REV*	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)
Δ*Volume*	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

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 B.3 in Appendix B.

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

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 C.1 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 C.2 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 C.1: 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
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

see next page

Table C.1: continued

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 C.2: 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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 D.1 in Appendix D). The descriptive statistics of *Abs_Allocation* are given in Table D.2.

In the second step, we apply the fixed-effect panel regression model as stated in Section 4 to Item 7A, to address Hypotheses 1 and 2. The results are given in Table D.3 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 D.1: 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
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 D.2: 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, 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 D.3: 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
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)
Δ*Volume*	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 B.3 in Appendix B.

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