From Corporate Disclosure to Social Media – Understanding Real Estate Markets with Textual Analysis

Dissertation zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft

eingereicht an der Fakultät für Wirtschaftswissenschaften der Universität Regensburg

vorgelegt von:

NINO PAULUS

Berichterstatter: Prof. Dr. Wolfgang Schäfers

Prof. Dr. Bertram Steininger

Dissertation zur Erlangung des Grades eines

Doktors der Wirtschaftswissenschaft

Eingereicht an der Fakultät für Wirtschaftswissenschaften der Universität Regensburg

Vorgelegt von:

Nino Paulus
M.Sc. Immobilienwirtschaft
(Universität Regensburg)
M.Sc. Real Estate
(University of Hong Kong)

Tag der Disputation:
19. Juli 2023

Berichterstatter:
Prof. Dr. Wolfgang Schäfers
Prof. Dr. Bertram Steininger

From Corporate Disclosure to Social Media – Understanding Real Estate Markets with Textual Analysis

Nino Paulus

Table of Contents

List of Tables .... **VII**

List of Figures .... **VIII**

1 Introduction **1**

1.1 Motivation and Background 1

1.2 Research Questions 4

1.3 Co-Authors, Submissions and Conference Presentations 7

1.4 References 9

2 Can Textual Analysis solve the Underpricing Puzzle? A US REIT Study **11**

2.1 Abstract 11

2.2 Introduction 12

2.3 Previous Literature and Hypothesis Development 14

2.3.1 Common Theories to explain IPO Underpricing 14
2.3.2 Textual Analysis and Underpricing of IPOs 16
2.3.3 Underpricing of US REIT IPOs 18
2.3.4 Hypothesis Development 20

2.4 Textual Analysis Procedure 21

2.4.1 Disclosure Tone 21
2.4.2 Disclosure Similarity 22

2.5 Model Specification 24

2.6 Data 25

2.6.1 Textual Corpus 25
2.6.2 Underpricing 25
2.6.3 Control Variables 26
2.6.4 Descriptive Statistics 28

2.7 Results 29

2.7.1 Disclosure Tone is not informative 29
2.7.2 Disclosure Similarity is informative 32
2.7.3 Robustness 34

2.8 Conclusion 35

2.9 Appendix 38

2.10 References 39

3 **Trade vs. Daily Press: The Role of News Coverage and Sentiment in Real Estate Market Analysis ... 44**

3.1 Abstract ... 44

3.2 Introduction ... 45

3.3 Literature Review and Hypothesis Development ... 47

3.3.1 Sentiment Analysis ... 47
3.3.2 Topic Modelling ... 49
3.3.3 Hypothesis Development ... 50

3.4 Data ... 52

3.4.1 Textual Corpus ... 52
3.4.2 INREV Returns ... 53
3.4.3 Macroeconomic and Real Estate-Specific Controls ... 54

3.5 Methodology ... 55

3.5.1 News Coverage – sLDA and GloVe Model ... 55
3.5.2 News Sentiment – Unbounded Polarity Score ... 58
3.5.3 Model Specification ... 59

3.6 Results ... 60

3.6.1 News Coverage ... 60
3.6.2 News Sentiment ... 63
3.6.3 Vector Autoregression ... 66
3.6.4 Robustness Tests ... 71

3.7 Conclusion ... 73

3.8 Appendix ... 76

3.9 References ... 79

4 **Social Media and Real Estate: Do Twitter Users predict REIT Performance? 84**

4.1 Abstract ... 84

4.2 Introduction ... 85

4.3 Literature Review and Hypothesis Development ... 87

4.3.1 Sentiment in Real Estate ... 87
4.3.2 Social Media Sentiment in Finance ... 88
4.3.3 Complexity of Social Media Data ... 90
4.3.4 Hypothesis ... 91

4.4 Data ... 92

4.4.1 Social Media Data ... 92

4.4.2 Dependent Variable 93
4.4.3 Economic Control Variables 93
4.5 Methodology 94
4.5.1 Tweet Selection and Data Cleaning 94
4.5.2 Sentiment Analysis 97
4.5.3 Econometric Approaches 101
4.6 Results 103
4.6.1 Tweet Selection 103
4.6.2 Model Accuracy 104
4.6.3 Sentiment Indices 104
4.6.4 Vector Autoregression 106
4.6.5 Granger Causality and Impulse Response Function 110
4.7 Conclusion 111
4.8 Appendix 114
4.9 References 120
5 Conclusion 126
5.1 Executive Summary 126
5.2 Final Remarks 132
5.3 References 134

List of Tables

Table 2.1 - Descriptive Statistics 29

Table 2.2 - Initial-Day Return – Disclosure Tone 30

Table 2.3 - Initial-Day Return – Disclosure Similarity 33

Table 2.4 - Initial-Day Return – Robustness 35

Table 2.5 - Description of Control Variables 38

Table 3.1 - Descriptive Statistics 54

Table 3.2 - Residential: VAR Equation on Returns and Granger Causality Test Results.... 67

Table 3.3 - Office: VAR Equation on Returns and Granger Causality Test Results 68

Table 3.4 - Retail: VAR Equation on Returns and Granger Causality Test Results............ 69

Table 3.5 - Robustness Tests: Hypo Index – VAR and Granger Causality Test Results 71

Table 3.6 - Robustness Tests: INREV Capital Growth Component – VAR and Granger Causality Test Results 72

Table 3.7 - Seed Words selected by GloVe Model 77

Table 3.8 - Descriptive Statistics of Sentiment Indicators 78

Table 4.1 - Descriptive Statistics 93

Table 4.2 - VAR Results monthly Lags Optimism Indicator.. 107

Table 4.3 - VAR Results monthly Lags Pessimism Indicator.. 108

Table 4.4 - VAR Results monthly Lags Sentiment Quotient.. 109

Table 4.5 - Signal Words 117

Table 4.6 - VAR Results quarterly Lags Optimism Indicator... 117

Table 4.7 - VAR Results quarterly Lags Pessimism Indicator.. 118

Table 4.8 - VAR Results quarterly Lags Sentiment Quotient.. 119

List of Figures

Figure 2.1 - Stylized Illustration of Cosine Similarity 24

Figure 2.2 - Number of IPOs and Average Underpricing over Years 26

Figure 3.1 - Number of Articles over Time by Newspaper 52

Figure 3.2 - Percentage Total Return over Time by Asset class per Quarter 53

Figure 3.3 - Residential News Coverage over Time 62

Figure 3.4 - Office News Coverage over Time 62

Figure 3.5 - Retail News Coverage over Time 63

Figure 3.6 - Residential News Sentiment over Time 64

Figure 3.7 - Office News Sentiment over Time 64

Figure 3.8 - Retail News Sentiment over Time 65

Figure 4.1 - Unfiltered Tweet Distribution over Time 92

Figure 4.2 - NLP Process 94

Figure 4.3 - LSTM Model 100

Figure 4.4 - Filtered Tweet Distribution over Time 103

Figure 4.5 - Optimism Indicators and monthly Total Returns over Time 105

Figure 4.6 - Pessimism Indicators and monthly Total Returns over Time 105

Figure 4.7 - Sentiment Quotient and monthly Total Returns over Time 106

Figure 4.8 - Impulse Response Functions 111

1 Introduction

1.1 Motivation and Background

The phenomenon of sentiment and its influence on real estate markets is a widely explored topic in empirical studies for more than a decade. Within this context, sentiment is defined as the beliefs regarding future cash flows and risks that cannot be explained solely by market fundamentals (Baker and Wurgler, 2007). However, according to the theories of efficient markets, every market participant acts rationally and unemotionally at every point in time (Fama, 1970). Therefore, the influence of sentiment on any market or the presence of market participants not adhering to these assumptions contradicts these theories. Researchers have analyzed this contradiction and shown that sentiment can indeed have a significant impact on markets, especially when "hard" information is ambiguous or even absent. Sentiment, hence, can act as a substitute for missing or inadequate information.

While stock markets are generally considered highly efficient and the sentiment is therefore of secondary importance, the characteristics of real estate markets abet the impact of sentiment (Gallimore and Gray, 2002). The immobility, heterogeneity, and low divisibility of real estate lead to segmented submarkets, intransparency, and information asymmetries among stakeholders and are thus major drivers for sentiment-induced trading behavior. The first evidence of a positive relationship between sentiment and direct real estate markets was provided by Clayton et al. in 2009 and by Lin et al. (2009), who discovered a similar relation for indirect real estate markets.

Since, plenty of studies have demonstrated the significant impact of sentiment caused by the "animal spirit" of investors (Keynes, 1936) and "noise traders" (Black, 1986) in real estate markets. With the influence of sentiment being proven and the augmentation of sentiment proxies in fundamental market models enhancing explanatory power, the literature focused on how to improve the measurement of market sentiment. Previously mentioned studies of Clayton et al. (2009) and Lin et al. (2009) employed traditional survey-based (direct) sentiment indicators, which is one of two methods for capturing sentiment and has often been adopted (e.g. Ling et al., 2014; Das et al., 2015; Freybote, 2016). However, direct indicators, which are mainly based on interviews and surveys, are often criticized for being expensive, time-consuming, and prone to selection bias. The second method, namely indirect indicators, derive sentiment from market fundamentals, operating on the premise that sentiment corresponds to what cannot be explained by market data (Baker and Wurgler, 2007). Indirect indicators were also often used in the literature (e.g. Marcato and Nanda, 2016; Freybote and Seagraves, 2017; Heinig et al., 2020), but again face criticism due to the dependence on hard and past market data rather than soft sentiment components.

However, there exists another source of sentiment that can potentially address these issues, yet has been relatively overlooked until recently: text. Thanks to advances in natural language processing, any textual corpus can now be analyzed for its sentiment. With all types of texts, such as corporate disclosure, news articles, or social media messages, available digitally and online, there are no particular restrictions to time, topic, or people participating. This offers tremendous potential to improve the measurement of sentiment and consequently gain a deeper understanding of its influence. The potential has been demonstrated in initial studies by Walker (2014) and Soo (2015), and subsequently confirmed across various real estate submarkets, including different use types and direct as well as indirect markets. Hereby, the most recent studies have focused on enhancing the performance of sentiment classification by employing more sophisticated algorithms (e.g. Hausler 2018, Beracha 2019).

Although textual sentiment analysis has shown to contribute to a better understanding of real estate markets, it is important to note that sentiment only captures a specific type of qualitative information conveyed in text. The real estate literature has largely focused on this type of qualitative information due to technical limitations. However, the results of the sentiment literature already suggest that other qualitative information could be of considerable value. Furthermore, when other hard information is ambiguous or missing, the proportion of information transmitted via text, whether spoken or written, is even greater. Hence, there is a need for further research into different types of qualitative information. New natural language processing techniques, such as similarity analysis and topic modelling, have evolved and can be used to gain further insights into text. Therefore, the primary objective of this dissertation is to explore the impact of various forms of qualitative information, extracted through natural language processing, on real estate markets.

Sentiment analysis is useful to show how information is presented, but it does not provide insights into what the information is or whether it is valuable. Often, stakeholders may place more importance on the content rather than its sentiment. Thus, it is the content that leads to different actions and affects market outcomes. However, it is difficult to quantify content and predict stakeholders' reactions in order to incorporate it into a quantitative model, which is why there is a lack of research. Therefore, the first paper proposes an approach to quantify the amount of new and valuable information provided in corporate disclosures and investigates how the real estate market reacts to this information.

Another potential way to advance the utility of textual analysis is to refine sentiment analysis. Historically, in news media sentiment analysis, all articles in the corpus were assigned equal weights without regard for the subject matter being studied. Even if a pre-selection of the corpus has taken place, it cannot be assumed that all articles are of equal relevance. However, modern techniques permit researchers to classify the content of texts, gauge their relevance to specific topics, and get rid of noise from irrelevant texts. The question that consequently arises, and that will be investigated in the second paper, is whether tailoring the corpus to suit the research objectives captures the nuances of sentiment more effectively and improves our understanding of the market.

One step further as tailoring corpora, is the examination of less structured corpora. The real estate literature has focused on well-structured sources such as newspaper articles or corporate publications, which are typically written in formal language and often are categorized by topics. However, there are other less structured sources, such as social media, that may offer valuable insights into real estate markets. Due to the complexity of analyzing unstructured social media data, there is a lack of research and only a few small-scale studies. Hence, the impact of social media sentiment on real estate markets remains unclear. Therefore, paper three, by combining the techniques presented in paper one and two, will structure social media data and examine its relationship with real estate markets.

To sum up, the progress made in natural language processing has given us the capability to extract information from textual data and explore new types of text. This is particularly relevant in the real estate market, where information is often scarce, and any text can be a valuable source of insights that can enhance our understanding of the industry. While previous literature has mainly focused on sentiment classification, this dissertation aims to expand on that by exploring the content of texts and leveraging it to improve our understanding of real estate markets.

1.2 Research Questions

This section outlines the progression of this dissertation and introduces the major research questions that will be addressed. While the impact of qualitative information and textual sentiment in real estate markets are the unifying elements of all papers, each paper uses a different type of text as a source of information and applies different methods to enhance information extraction. Going from insider to outsider information, the first paper examines corporate disclosure, the second paper investigates news articles, and the third paper concentrates on social media. Thus, the following research questions can be raised:

Paper 1: Can Textual Analysis solve the Underpricing Puzzle? A US REIT Study

Paper 1 concentrates on the sentiment and informativeness of Form S-11, the initial document for registering US REIT initial public offerings with the Securities and Exchange Commission, using natural language processing techniques. The document is intended to provide potential investors with initial information about the company's business and financial condition, but it is not clear to what extent Form S-11 meets this requirement. Based on Beatty and Ritter's (1986) asymmetric information hypothesis, it is expected that informative content in the S-11 is associated with REIT underpricing. The central research questions can hence be summarized as follows:

- Does the qualitative information, extracted from Form S-11, help to explain the underpricing of US REIT IPOs?
- Can the language used in Form S-11 affect the pricing accuracy of US REIT IPOs by investors?
- Does the uncertain language in Form S-11 yield difficulties for investors to price the issue and thus increase initial-day returns of US REIT IPOs?
- Does the level of uncertain language in Form S-11 serve as a proxy for ex ante uncertainty?
- Does a greater similarity to previous filings indicate that the filing contains a relatively high proportion of platitudes but little useful information?
- Do higher levels of similarity to prior filings suggest that information asymmetries persist and lead to increased underpricing in US REIT IPOs?

Paper 2: Trade vs. Daily Press: The Role of News Coverage and Sentiment in Real Estate Market Analysis

Paper 2 examines the impact of newspapers on direct real estate markets. Newspapers serve as an important source of information for market participants and thus contribute to the spread of a narrative (Shiller 2017). Using textual analysis, this paper separates newspaper articles by asset class in order to derive asset class-specific sentiment indices, which are compared to the total returns of these asset classes. Using different types of newspapers from Germany that have broad market coverage, it is expected that those contribute to the spread of narratives and hence significantly impact total returns. The research questions are therefore as follows:

- Do daily newspapers differ in their reporting style from trade newspapers?
- Are different types of newspapers equally informative and thus equally suitable for market forecasts?
- Which real estate asset classes are particularly prone to news reporting?
- Do asset class-specific sentiment indices add to the understanding of real estate markets or is a general index sufficient?
- Can relative news coverage serve as another indicator for future market development?

Paper 3: Social Media and Real Estate: Do Twitter Users predict REIT Performance?

Paper 3 uses multiple textual sentiment classifiers to extract a market sentiment from real estate-related social media posts. Social media platforms have become vibrant online platforms where all kinds of market participants share their opinions and thoughts on equity markets (Yadav and Vishwakarma, 2020). This paper, therefore, identifies relevant posts and measures their sentiment to determine the predictive power of social media sentiment on REIT returns. It is anticipated that there exists a correlation between REIT returns and social media sentiment, as the sentiment expressed on social media is likely to mirror the beliefs and actions of investors. The central considerations can be stated as such:

- Is it possible to determine social media messages relevant to a real estate context to create a valuable sentiment indicator?
- Do more sophisticated sentiment classifiers yield higher classification accuracy?
- Does social media sentiment represent the beliefs and opinions of real estate investors?
- Does social media sentiment predict REIT returns?
- Do more sophisticated sentiment classifiers yield a higher prediction accuracy and are hence better predictors?

1.3 Co-Authors, Submissions and Conference Presentations

The following overview provides information about co-authors, journal submissions, publication status and conferences presentations.

Paper 1: **Can Textual Analysis solve the Underpricing Puzzle? A US REIT Study**

Authors:

Nino Paulus, Dr. Marina Koelbl, Prof. Dr. Wolfgang Schaefers

Submission Details:

Journal: Journal of Property Investment & Finance
Current Status: accepted (10/10/2021) and published in Volume 40, Issue 6 (11/11/2021)

Conference Presentations:

This paper was presented at:

- Center of Finance Workshop "Artificial Intelligence and Finance" at the University of Regensburg in July 2021
- the 38th Annual Conference of the American Real Estate Society (ARES) in Bonita Springs, Florida, USA (2022)
- the 28th Annual Conference of the European Real Estate Society (ERES) in Milan, Italy (2022)

Paper 2: Trade vs. Daily Press: The Role of News Coverage and Sentiment in Real Estate Market Analysis

Authors:

Dr. Franziska Ploessl, Nino Paulus, Prof. Dr. Tobias Just

Submission Details:

Journal: Journal of Property Research
Current Status: accepted for publication (03/04/2023)

Conference Presentations:

This paper was presented at:

- Center of Finance Workshop "Artificial Intelligence and Finance" at the University of Regensburg in July 2022
- the 39th Annual Conference of the American Real Estate Society (ARES) in San Antonio, Texas, USA (2023)

This paper will be presented at:

- the 29th Annual Conference of the European Real Estate Society (ERES) in London, England (2023) (submission accepted)

Paper 3: Social Media and Real Estate: Do Twitter Users predict REIT Performance?

Authors:

Nino Paulus, Lukas Lautenschlaeger, Prof. Dr. Wolfgang Schaefers

Submission Details:

Journal: Journal of Real Estate Research
Current Status: Under review (04/29/2023)

Conference Presentations:

This paper was presented at:

- the 39th Annual Conference of the American Real Estate Society (ARES) in San Antonio, Texas, USA (2023)

This paper will be presented at:

- Center of Finance Workshop "Artificial Intelligence and Finance" at the University of Regensburg in June 2023 (submission accepted)
- the 29th Annual Conference of the European Real Estate Society (ERES) in London, England (2023) (submission accepted)

1.4 References

Baker, M. and Wurgler, J. (2007), "Investor Sentiment in the Stock Market", Journal of Economic Perspectives, Vol. 21 No. 2, pp. 129–151.

Beatty, R.P. and Ritter, J.R. (1986), "Investment banking, reputation, and the underpricing of initial public offerings", Journal of Financial Economics, Vol. 15 1-2, pp. 213–232.

Beracha, E., Lang, M. and Hausler, J. (2019), "On the Relationship between Market Sentiment and Commercial Real Estate Performance—A Textual Analysis Examination", Journal of Real Estate Research, Vol. 41 No. 4, pp. 605–638.

Black, F. (1986), "Noise", The Journal of Finance, Vol. 41 No. 3, pp. 528–543.

Clayton, J., Ling, D.C. and Naranjo, A. (2009), "Commercial Real Estate Valuation: Fundamentals Versus Investor Sentiment", The Journal of Real Estate Finance and Economics, Vol. 38 No. 1, pp. 5–37.

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

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

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

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

Gallimore, P. and Gray, A. (2002), "The role of investor sentiment in property investment decisions", Journal of Property Research, Vol. 19 No. 2, pp. 111–120.

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

Heinig, S., Nanda, A. and Tsolacos, S. (2020), "Which Sentiment Indicators Matter? Evidence from the European Commercial Real Estate Market", Journal of Real Estate Research, Vol. 42 No. 4, pp. 499–530.

Keynes, J.M. (Ed.) (1936), The General Theory of Employment, Interest, and Money, Palgrave Macmillan, Cham.

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

Ling, D.C., Naranjo, A. and Scheick, B. (2014), "Investor Sentiment, Limits to Arbitrage and Private Market Returns", Real Estate Economics, Vol. 42 No. 3, pp. 531–577.

Marcato, G. and Nanda, A. (2016), "Information Content and Forecasting Ability of Sentiment Indicators: Case of Real Estate Market", Journal of Real Estate Research, Vol. 38 No. 2, pp. 165–204.

Shiller, R.J. (2017), "Narrative Economics", American Economic Review, Vol. 107 No. 4, pp. 967–1004.

Soo, C.K. (2015), "Quantifying Animal Spirits: News Media and Sentiment in the Housing Market", Ross School of Business Working Paper No. 1200.

Walker, C.B. (2014), "Housing booms and media coverage", Applied Economics, Vol. 46 No. 32, pp. 3954–3967.

Yadav, A. and Vishwakarma, D.K. (2020), "Sentiment analysis using deep learning architectures: a review", Artificial Intelligence Review, Vol. 53 No. 6, pp. 4335–4385.

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

2.1 Abstract

Purpose: Although many theories aim to explain initial public offering (IPO) underpricing, initial-day returns of US Real Estate Investment Trust (REIT) IPOs remain a "puzzle". The literature on REIT IPOs has focused on indirect quantitative proxies for information asymmetries between REITs and investors to determine IPO underpricing. This study, however, proposes textual analysis to exploit the qualitative information, revealed through one of the most important documents during the IPO process – Form S-11 – as a direct measure of information asymmetries.

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

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

Practical implications: The results demonstrate that textual analysis can in fact help to explain underpricing of US REIT IPOs, as qualitative information in Forms S-11 decreases information asymmetries between US REIT managers and investors, thus reducing underpricing. Consequently, REIT managers are incentivized to provide as much information as possible to reduce underpricing, while investors could use textual analysis to identify offerings that promise the highest returns.

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

Keywords: Initial public offering, Information asymmetry, Textual analysis, Underpricing

2.2 Introduction

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

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

This paper contributes to answering this question by analyzing whether and how qualitative information impacts initial-day returns of US Real Estate Investment Trusts (REITs). Prior studies investigating the impact of qualitative information on IPO underpricing exclude the REIT sector due to its financial characteristics, legal requirements and differing performance.[1] Nonetheless, as evidenced by its recent market capitalization of $1,249,186.3 million as of December 31, 2020 for 223 listed REITs, the US REIT sector is significant to the overall economy and certainly worthy of consideration. In addition, the high dividend payout requirement of at least 90% of their taxable earnings renders REITs dependent on external equity capital in order to take advantage of growth opportunities. As a result, US REIT managers have an unusually strong incentive to reduce underpricing. Likewise, they have a high motivation to maintain investor trust (e.g. Price et al., 2017), which ensures high disclosure quality.

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

Analyzing the level of uncertain language in a prospectus, as well as its similarity to registration forms filed in the past six months, we indeed find that qualitative information in Form S-11 contributes to the understanding of US REIT IPOs. However, against our initial expectation that cautious wording makes it more difficult for investors to value the issue and thus increases initial-day returns, we do not find a statistically significant impact of the level of uncertain language on underpricing. We conclude that cautious wording is induced by forecasting difficulties and litigation risks, rather than reflecting issuer confidence in the future prospects of the firm. By contrast, document similarity is statistically and

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

economically significant related to underpricing of US REIT IPOs. We assume that higher similarity to prior filings indicates that the filing contains a relatively high proportion of standard content or boilerplates, but little useful information. Thus, information asymmetries persist, leading to increased underpricing. In sum, our results show that qualitative information, just like quantitative information, conveys valuable insights, impacts investors' capability to price the issue and thus affects underpricing.

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

The remainder of the paper is organized as follows. Section 2.3 presents common theories explaining the underpricing phenomenon, discusses the related literature analyzing how textual features affect initial-day returns of industrial firms, summarizes studies examining REIT underpricing, and finally, develops hypotheses. The textual analysis procedure and data used for this study are described in sections 2.4 and 2.5, while the empirical methods for the analysis are presented in section 2.6. Section 2.7 discusses the empirical results, and section 2.8 concludes.

2.3 Previous Literature and Hypothesis Development

2.3.1 Common Theories to explain IPO Underpricing

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

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

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

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

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

2.3.2 Textual Analysis and Underpricing of IPOs

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

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

2.3.3 Underpricing of US REIT IPOs

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

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

Further adding to the understanding of initial-day returns of US REIT IPOs, Dolvin and Pyles (2009) and Gokkaya et al. (2015) compare underpricing of REIT and non-REIT IPOs. Although both studies find significant underpricing for REIT issuances, initial-day returns are lower than that observed for a matching industrial sample. The lower underpricing of REITs has thereby been attributed to their unique nature and specific regulatory requirements, making them more transparent than general stocks and thus easier to value for potential investors (Buttimer et al., 2005; Hartzell et al., 2005; Dolvin and Pyles, 2009; Lorenz, 2020). However, despite being considered more transparent than their industrial counterparts, REITs suffer from information asymmetry. Measuring the net asset value of REITs is difficult for investors, as the underlying assets are heterogenous and illiquid. In addition, REITs are actively managed, further complicating valuation due to the growth options developed by managers (Ghosh et al., 2000). Therefore, incorporating the qualitative information from corporate disclosures using textual analysis may also be valuable in solving the underpricing puzzle of US REIT IPOs. REITs thereby offer a unique and useful setting for conducting textual analysis, as they are strongly incentivized to minimize “money left on the table” when going public and thus guarantee high disclosure quality. Given the high distribution requirement of at least 90% of their taxable earnings, REITs have a very limited cash reserve by regulation and are highly dependent on external equity capital in order to take advantage of growth opportunities. Thus, consistent with Beatty and Ritter’s (1986) asymmetric information hypothesis, REIT managers might voluntarily reveal valuable insights into the upcoming IPO through corporate disclosures, in order to reduce information asymmetry and underpricing.

In fact, the importance of textual information in the context of real estate has already been demonstrated in the literature. By analyzing earnings conference calls for a REIT sample, Doran et al. (2012) and Price et al. (2017) show that sentiment impacts initial reaction-window abnormal returns. Ruscheinsky et al. (2018) reveal a leading relationship of media expressed sentiment to future REIT market movements; while, Beracha et al. (2019) show that news based sentiment has predictive power for the direct commercial real estate market in the US. Hausler et al. (2018) confirms that news-based sentiment leads both the direct as well as the securitized real estate market in the US. Focusing on “abnormally” positive tone in annual and quarterly reports of US REITs, Carstens and Freybote (2019) find that institutional investors react positively (negatively) to an abnormally positive tone and behave as net buyers (net sellers) in periods of institutional investor optimism (pessimism).

2.3.4 Hypothesis Development

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

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

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

Although Form S-11 is the major document providing information to investors during the IPO process, it is also a standard document required by the SEC. Therefore, Form S-11 contains a relatively high proportion of standardized content. In particular, registration statements filed at the same time, subject to the same regulatory requirements and exposed to the same market environment, may be similar. Clearly, registration statements that merely repeat statements disclosed by other companies or that consist largely of standard phrases do not help investors to determine the true value of the firm. By contrast, unique and specific disclosures that are different from registration statements filed by firms that recently brought an IPO to the market are likely to provide valuable guidance to investors in their search for the right price. As such, unique and informative disclosures can present an effective tool to overcome information asymmetries between the parties involved in the IPO – one of the main reasons for underpricing. To determine the informativeness or uniqueness of a disclosure as opposed to standard phrases, following Hanley and Hoberg (2010), we develop a similarity measure by comparing the document in question to registration statements filed up to six months earlier. Assuming that disclosures which are more similar to past prospectuses provide less useful information and thus not resolve information asymmetry, we formulate our second hypothesis as follows:

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

2.4 Textual Analysis Procedure

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

2.4.1 Disclosure Tone

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

In line with previous literature, we measure *Uncertainty* using the Loughran and McDonald (2011) sentiment wordlists.[2] We thereby follow Loughran and McDonald's (2013) recommendation and use an aggregate uncertainty measure, which comprises words from the uncertain, weak modal and negative wordlists to determine the level of uncertain language in Form S-11 (*Uncertainty*).

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

$$Uncertainty = \frac{Uncertainty\ Words}{Total\ Number\ of\ Words} \tag{2.1}$$

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

2.4.2 Disclosure Similarity

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

The basic idea of applying cosine similarity to text classification problems is to map documents onto a vector space model (VSM), which enables measuring the similarity between two vectors or textual documents by computing the dot product. We thus create a VSM by representing each IPO prospectus *i* as an *N*-dimensional vector V_i summarizing

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

its word usage. Thereby, each element n of the vector V_i corresponds to a specific word that is present in the textual corpus and denotes the frequency of appearance of the respective word in prospectus i. For example, the vectors v_1 and v_2 for two documents in the corpus are as follows:

$$v_1 = (w_1, w_2, \dots, w_{N-1}, w_N) \text{ and } v_2 = (\psi_1, \psi_2, \dots, \psi_{N-1}, \psi_N), \tag{2.2}$$

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

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

$$Cosine\ Similarity\ (v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\|\ \|v_2\|}, \tag{2.3}$$

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

Figure 2.1 provides a simplified example illustrating the textual analysis procedure for determining disclosure similarity. Each of the three introductory statements from Form S-11, A, B and C, is represented as a vector, with each value indicating the number of occurrences of the corresponding word in the sentence. The resulting VSM enables calculating the level of similarity of the respective statements.

Figure 2.1 - Stylized Illustration of Cosine Similarity

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

Panel B									
	REIT	focuses	owner-ship	commu-nity	shoppin g centers	regional	acquisi-tion	suburba n	office proper-ties
A	1	1	1	1	1	0	0	0	0
B	1	1	1	0	1	1	0	0	0
C	1	1	0	0	0	0	1	1	1

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

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

With a similarity score (*Cosine*) of 0.8 for sentences A and B and 0.4 for sentences A and C and B and C, respectively, Figure 2.1 shows that sentences A and B are similar, while sentences A and C and B and C are rather different. According to our hypothesis, a higher similarity (dummy = 1) implies a higher information asymmetry, therefore we expect a positive sign. Consequently, if this variable takes a value closer to one the underpricing, initial returns respectively, are greater.

2.5 Model Specification

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

$$IR_i = \beta_0 + \beta_1 Textual\ Feature_i + \beta_2 Controls_i + \varepsilon_i\ , \tag{2.4}$$

where ε_i denotes the error term and we assume $\varepsilon_i \sim N(0, \sigma^2)$. In addition to the vector of textual features comprising tone (*Uncertainty*) and similarity (*Cosine*), the regression equation includes a vector of control variables (Controls), which are described in Section 2.6.

2.6 Data

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

2.6.1 Textual Corpus

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

2.6.2 Underpricing

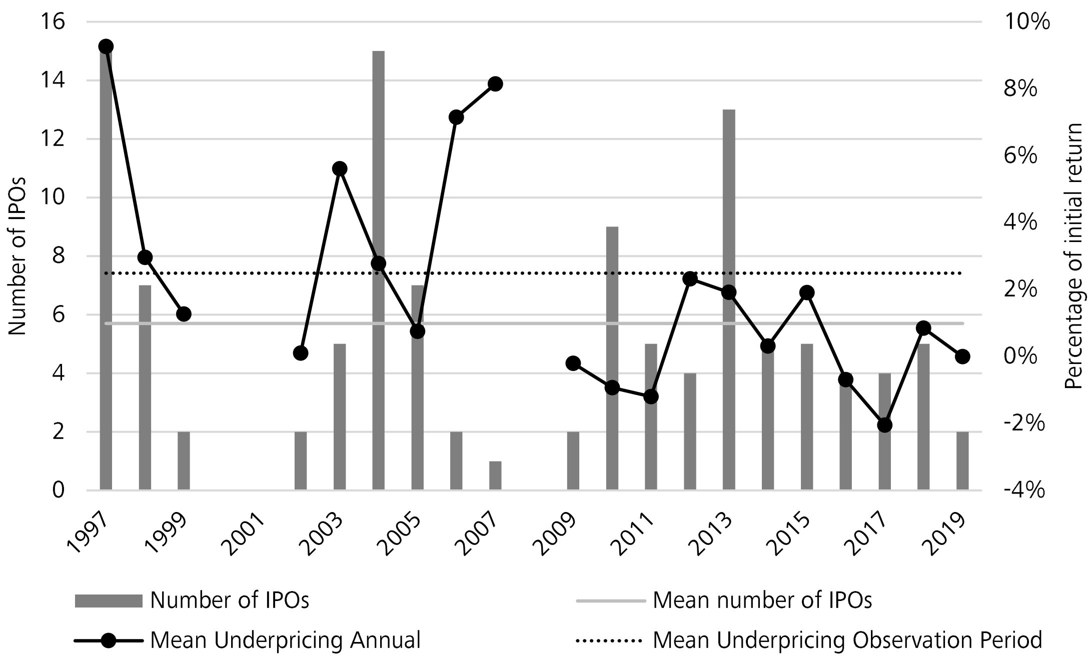
In order to calculate underpricing, we match each REIT that issued an IPO during the sample period with data from *CRSP*. In accordance with prior studies (e.g. Brobert, 2016; Chen et al., 2020), the initial-day return represents the return enjoyed by IPO investors on the first day of trading of a new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

$$IR_{i,t} = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}}\,, \tag{2.5}$$

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

Figure 2.2 displays the number of IPOs, as well as average underpricing on a yearly basis. The sample of IPOs is not distributed equally across time; instead Figure 2.2 suggests the prevalence of hot and cold market periods. For example, 2000/2001 and 2008 are cold market periods that correspond with stock market declines, commonly referred to as the end of the dot-com era and the global financial crisis. Similarly, average underpricing varies over time. During the hottest market period (1997–1999), underpricing is high, suggesting that irrationally optimistic sentiment investors were active in the REIT IPO market at that time. Likewise, high initial-day returns are observed in the years before the global financial crisis, while the post-crisis period is characterized by low underpricing.

Figure 2.2 - Number of IPOs and Average Underpricing over Years



2.6.3 Control Variables

To control for information beyond textual clues revealed in Form S-11 that might affect initial-day returns of REITs, a set of control variables is included. We describe all control variables below and provide more specific definitions in Table 2.5 in the Appendix. We cluster the controls into four subsets: firm characteristics, offering characteristics, third-party certification and market conditions.

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

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

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

Therefore, top-tier underwriters signal high quality, in addition to increasing the transparency and credibility of the information. Consequently, underwriter prestige is expected to have a negative effect on underpricing.

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

2.6.4 Descriptive Statistics

Table 2.1 presents descriptive statistics of the characteristics of the 114 REIT IPOs and textual features of their registration statements. The average underpricing for our sample is a statistically and economically significant 2.49%, visualized by the dotted line in Figure 2.2. This is relatively low compared to non-REITs, but consistent with previous research on US REITs (e.g. Buttimer et al., 2005; Chan et al., 2013). Overall, 61 IPOs are undervalued, while 29 are overvalued and 24 are priced at market value. On average, the offer prices are revised upwards by 4.83% from the mid-point of the initial filing range. Hereby, 78.1% of the final offer prices are integers. The average size of a REIT at the time of the IPO is $665.9 million, with a maximum of $7,215.6 million and a minimum of only $12.0 million. However, only 29.94% of the outstanding shares are sold to the market. As evidenced by an average score of 7.7 on the Carter and Manaster (1990) scale and 44.7% selecting an underwriter ranked 9, REITs appear to prefer underwriters with excellent reputation. The average standard deviation of daily stock returns for the first 20 days after the IPO is 1.25%. The age of the issuing company is on average 6 years. In the IPO year, 47.4% of the issuing firms have positive earnings per share and are on average 43.3% debt financed. While 76.3% of the firms were invested in a single property type, the remaining 23.7% held multiple types of real estate assets.

The textual features extracted for the analysis suggest that REIT IPO prospectuses display a low level of *Uncertainty* and are relatively similar to each other. The mean of *Uncertainty* is 0.0256, indicating that on average, 2.56% of the total words in disclosures are uncertainty words. The observed standard deviation of 0.0031 is quite small, so that the

Table 2.1 - Descriptive Statistics

	Mean	StDev	Min	5th	Median	95th	Max
				Form S-11			
Uncertainty	0.0256	0.0031	0.0179	0.0202	0.0263	0.0303	0.0325
Uncertainty$_{LM}$	0.0272	0.0032	0.0197	0.0219	0.0277	0.0319	0.0341
Cosine	0.8805	0.0547	0.6650	0.7814	0.8888	0.9456	0.9700
Cosine$_{LM}$	0.8665	0.0562	0.6445	0.7791	0.8754	0.9348	0.9643
				Control Variables			
Age	6.3684	12.0568	1	1	2	26	80
Size (mm)	665.895	1,091.542	12.000	58.695	32.808	2,558.101	7,215.580
EPS	0.4737	0.5015	0	0	0	1	1
Leverage	0.4333	0.1970	0.0030	0.04087	0.4554	0.7457	0.8352
Property_Type	0.2368	0.4270	0	0	0	1	1
Offer_Price	0.7807	0.4156	0	0	1	1	1
Up_Revision	-0.0483	0.0806	-0.2727	-0.1938	-0.0426	0.0531	0.1500
Share_Overhang	0.2994	0.2848	0.0000	0.0000	0.2084	0.8209	0.9032
Underwriter	0.4474	0.4994	0	0	0	1	1
Hot_Market	0.3772	0.4868	0	0	0	1	1
Volatility	0.0125	0.0078	0.0044	0.0056	0.0114	0.0219	0.0658
				Dependent Variable			
IR	0.0249	0.0676	-0.1338	-0.0540	0.0017	0.1300	0.3750

Notes: This table shows the descriptive statistics for the textual features extracted from Form S-11 (*Uncertainty* and *Cosine*), further control variables, and the dependent variables (*IR*). The definition of all variables is presented in Table 2.5 in the Appendix. The sample consists of 114 US REIT IPOs from 1996 to 2019.

uncertainty levels of the individual prospectuses are close to the overall mean. The similarity score, which ranges between 0 and 1 by construction, averages 0.88 and varies from a 5th percentile of 0.78 to a 95th percentile of 0.95. The high similarity of documents can be attributed to the comparison within a single industry, but is certainly also induced by the strict regulatory requirements for SEC registration statements and the specific requirements for US REITs.

2.7 Results

2.7.1 Disclosure Tone is not informative

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

Table 2.2 column one presents regression results with the traditional IPO control variables. Seven of the eleven independent variables are statistically significantly associated with initial-day returns. The insignificance of the variables *Offer_Price*, *Underwriter*, *Leverage* and *Property_Type* could be due to the strong influence of the other variables. The control variables have the expected signs, except for *Age*. The variable has a positive signs, which could be due to the transparency of REITs. The goodness of fit of this regression is 39.5%. Contradicting our hypothesis, we do not find a statistically significant relation between *Uncertainty* and IPO underpricing (*IR*), when incorporating our measure of uncertain language into Model 2.

Table 2.2 - Initial-Day Return – Disclosure Tone

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

Notes: This table presents the results of the ordinary least squares (OLS) multiple regression model using first-day IPO returns as the dependent variable. The table reports coefficients and standard errors (in parentheses) of determinants affecting initial-day returns of US REIT IPOs. Economic significance is defined as the coefficient multiplied by the standard deviation. The definition of all variables is presented in Table 2.5 in the Appendix.

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

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

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

In light of the risk-return relationship on which investors typically base their investment decisions, our results show that the level of uncertain language in Form S-11 is not appropriate for quantifying risk. Thus, analyzing IPO prospectuses with regard to *Uncertainty* does not help investors in their investment decisions.

2.7.2 Disclosure Similarity is informative

Although Form S-11 is the major document providing information to investors during the IPO process, it is also a standard document required by the SEC. As such, we expect the disclosure to include standardized content; particularly prospectuses of companies bringing an IPO to market at the same time could be similar. We hypothesize that disclosures with a higher similarity to previously filed registration statements reveal little useful information to potential investors, thereby preserving information asymmetry and leading to greater underpricing. To test the second hypothesis, we regress the similarity measure for each registration form on the corresponding first-day return of the IPO. After controlling for firm and offering characteristics, third-party certification and market conditions, all of which have shown to be associated with initial-day returns in prior literature, we find that the qualitative information revealed in the Form S-11 indeed helps to explain the underpricing of US REIT IPOs (see Table 2.3). Specifically, a one-standard deviation increase in the similarity score (*Cosine*) leads to a 1.61% increase in first-day returns (0.294 coefficient value multiplied by standard deviation of 0.0547). The economic significance of *Cosine* thus lies within the range of widely accepted determinants of IPO underpricing, namely *Size* and *EPS*. Only *Up_Revision*, a variable that can only be measured after the S-11 filing date, specifically on the day the shares are sold, has a higher economic significance of 2.15% (0.267 coefficient value multiplied by standard deviation of 0.0806). The key point to note here is that among the variables with the greatest economic significance, *Cosine* is the only factor that the firm can influence directly before going public. While *Up_Revision* and the firms' earnings per share in the IPO year (*EPS*) can only be determined after the IPO, *Size* cannot be changed by the company at short notice. Further emphasizing the importance of qualitative information, the goodness of fit of the regression improves with R^2 increasing from 39.5% to 44.5% when *Cosine* is added as an independent variable.

Demonstrating that the informativeness of corporate disclosures as measured by document similarity is statistically and economically significant related to underpricing of US REIT IPOs, our results confirm, that qualitative information is as important as quantitative information in solving the underpricing puzzle. Instead of merely restating qualitative data, the narratives in corporate disclosures add context to numerical disclosures, which provides additional insights to investors and thus impacts their capability to price the issue. Being aware of the relationship between *Cosine* and underpricing, investors interested in buying new shares can use textual analysis to identify offerings that

Table 2.3 - Initial-Day Return – Disclosure Similarity

	Model 1 (*Quantitative Factors*)		Model 2 (*Cosine*)	
	Regression Results	Economic Significance	Regression Results	Economic Significance
Intercept	-0.344*** (-2.776)		-0.559*** (-4.029)	
Age	0.011** (2.117)	1.22%	0.012** (2.414)	1.33%
Size	0.019*** (2.955)	2.22%	0.016*** (2.670)	1.87%
EPS	0.028** (2.601)	1.40%	0.036*** (3.337)	1.81%
Leverage	-0.015 (-0.553)	-0.30%	-0.011 (-0.394)	-0.22%
Property_Type	-0.003 (-0.255)	-0.13%	-0.004 (-0.307)	-0.17%
Offer_Price	-0.002 (-0.128)	-0.08%	-0.004 (-0.287)	-0.17%
Up_Revision	0.283*** (4.047)	2.28%	0.267*** (3.949)	2.15%
Share_Overhang	-0.040* (-1.662)	-1.14%	-0.043* (-1.830)	-1.22%
Underwriter	-0.019 (-1.540)	-0.95%	-0.020* (-1.718)	-1.00%
Hot Market	0.029** (2.619)	1.41%	0.027** (2.533)	1.31%
Volatility	0.641 (0.829)	0.50%	0.484 (0.649)	0.38%
Cosine			0.294*** (3.027)	1.61%
N	114		114	
R^2	0.395		0.445	
Adj. R^2	0.329		0.379	

Notes: This table presents the results of the ordinary least squares (OLS) multiple regression model using first-day IPO returns as the dependent variable. The table reports coefficients and standard errors (in parentheses) of determinants affecting initial-day returns of US REIT IPOs. Economic significance is defined as the coefficient multiplied by the standard deviation. The definition of all variables is presented in Table 2.5 in the Appendix.

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

have low similarity scores and, in turn, promise the highest returns on the first day of trading. Likewise, REIT managers can reduce underpricing by voluntarily providing more information in Form S-11. We acknowledge, however, that there is a trade-off between acquiring costly information in the premarket, so as to provide informative disclosures that simplify the evaluation of the offering and gathering information from investors during book-building (Sherman and Titman, 2002). In order to provide valuable insights in the prospectus and determine an accurate offer price that potential investors agree with, companies have to bear high costs for information procurement in the premarket. If the company deems this too expensive, it can rely on investors themselves to gather information and value the IPO during book-building. However, investors must be compensated for their efforts through higher initial-day returns.

2.7.3 Robustness

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

Uncertainty is thus defined as the ratio of words assigned to the uncertain wordlist, to the number of words in the prospectus also appearing in the Loughran and McDonald (2011) master dictionary. For *Cosine*, the maximum number of elements N of each vector v_i representing a document is limited to the total number of words that compose the Loughran and McDonald (2011) master dictionary. Furthermore, we now use specific word inflections instead of word roots. Our results are robust to these alternate main independent variables, since all coefficient signs are the same and their magnitudes have a comparable size (see Table 2.4). Again, there is no association between the level of uncertain language (*Uncertainty*) and a REIT's initial-day return, while a higher similarity score (*Cosine*) is associated with increased underpricing. These results survive several other robustness checks which are not tabulated separately. For example, we omit control variables and winsorize all variables to common levels.

All models in Tables 2.2 to 2.4 are tested for multicollinearity and heteroscedasticity. We do not find any multicollinearity issues applying a variance inflation factor (VIF) test with the common threshold, as suggested by Chatterjee and Price (1995), of 10. Besides, correlations range between 0.196 and 0.549, which is below the critical level of 0.80 (Kennedy, 2003). For testing for heteroscedasticity, a Breusch-Pagan test was applied. The null hypothesis of homoscedasticity has not been rejected for the models in Tables 2.2 to 2.4.

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

Table 2.4 - Initial-Day Return – Robustness

	Model 1 (*Quantitative Factors*)		Model 2 (*Uncertainty*)		Model 3 (*Cosine*)	
	Regression Results	Economic Significance	Regression Results	Economic Significance	Regression Results	Economic Significance
Intercept	-0.344***		-0.265*		-0.537***	
	(-2.776)		(-1.963)		(-3.889)	
Age	0.011**	1.22%	0.010**	1.11%	0.012**	1.33%
	(2.117)		(1.988)		(2.461)	
Size	0.019***	2.22%	0.019***	2.22%	0.017***	1.98%
	(2.955)		(2.967)		(2.716)	
EPS	0.028**	1.40%	0.024**	1.20%	0.034***	1.71%
	(2.601)		(2.144)		(3.161)	
Leverage	-0.015	-0.30%	-0.014	-0.28%	-0.012	-0.24%
	(-0.553)		(-0.518)		(-0.438)	
Property_Type	-0.003	-0.13%	-0.001	-0.04%	-0.004	-0.17%
	(-0.255)		(-0.087)		(-0.355)	
Offer_Price	-0.002	-0.08%	-0.004	-0.17%	-0.003	-0.12%
	(-0.128)		(-0.335)		(-0.234)	
Up_Revision	0.283***	2.28%	0.298***	2.40%	0.266***	2.14%
	(4.047)		(4.239)		(3.917)	
Share_Overhang	-0.040*	-1.14%	-0.035	-1.00%	-0.044*	-1.25%
	(-1.662)		(-1.461)		(-1.865)	
Underwriter	-0.019	-0.95%	-0.019	-0.95%	-0.020*	-1.00%
	(-1.540)		(-1.586)		(-1.718)	
Hot Market	0.029**	1.41%	0.026**	1.27%	0.028**	1.36%
	(2.619)		(2.315)		(2.561)	
Volatility	0.641	0.50%	0.375	0.29%	0.513	0.40%
	(0.829)		(0.473)		(0.684)	
$Uncertainty_{LM}$			-2.592	-0.83%		
			(-1.435)			
$Cosine_{LM}$					0.266***	1.49%
					(2.817)	
N	114		114		114	
R^2	0.395		0.407		0.439	
Adj. R^2	0.329		0.336		0.372	

Notes: This table presents the results of the ordinary least squares (OLS) multiple regression model using first-day IPO returns as the dependent variable. The table reports coefficients and standard errors (in parentheses) of determinants affecting initial-day returns of US REIT IPOs. Economic significance is defined as the coefficient multiplied by the standard deviation. The definition of all variables is presented in Table 2.5 in the Appendix.

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

2.8 Conclusion

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

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

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

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

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

2.9 Appendix

Table 2.5 - Description of Control Variables

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

Notes: This table describes the control variables used.

2.10 References

Arnold, T., Fishe, R.P.H. and North, D. (2010), “The effects of ambiguous information on initial and subsequent IPO returns”, Financial Management, Vol. 39 No. 4, pp. 1497–1519.

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

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

Bajo, E. and Raimondo, C. (2017), “Media sentiment and IPO underpricing”, Journal of Corporate Finance, Vol. 46, pp. 139–53.

Beatty, R.P. and Ritter, J.R. (1986), “Investment banking, reputation, and the underpricing of initial public offerings”, Journal of Financial Economics, Vol. 15 Nos 1–2, pp. 213–232.

Beatty, R.P. and Welch, I. (1996), “Issuer expenses and legal liability in initial public offerings”, Journal of Law and Economics, Vol. 39, pp. 545–602.

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

Beracha, E., Lang, M. and Hausler, J. (2019), “On the relationship between market sentiment and commercial real estate performance - a textual analysis examination”, Journal of Real Estate Research, Vol. 41 No. 4, pp. 605–637.

Bradley, D.J. and Jordan, B.D. (2002), “Partial adjustment to public information and IPO underpricing”, Journal of Financial and Quantitative Analysis, Vol. 37 No. 4, pp. 595–616.

Bradley, D.J., Cooney, J.W., Jordan, B.D. and Singh, A.K. (2004), “Negotiation and the IPO offer price: a comparison of integer vs. non-integer IPOs”, Journal of Financial and Quantitative Analysis, Vol. 39 No. 3, pp. 517–540.

Brau, J.C., Cicon, J. and McQueen, G. (2016), “Soft strategic information and IPO underpricing”, Journal of Behavioral Finance, Vol. 17 No. 1, pp. 1–17.

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

Brown, S.V. and Tucker, J.W. (2011), “Large-sample evidence on firms’ year-over-year MD&A modifications”, Journal of Accounting Research, Vol. 49 No. 2, pp. 309–346.

Buttimer, R.J., Hyland, D.C. and Sanders, A.B. (2005), “Real Estate REITs, IPO waves and long-run performance”, Real Estate Economics, Vol. 33 No. 1, pp. 51–87.

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

Carter, R. and Manaster, S. (1990), "Initial public offerings and underwriter reputation", The Journal of Finance, Vol. 45 No. 4, pp. 1045–1067.

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

Chatterjee, S. and Price, B. (1995), Praxis Der Regressionsanalys, 2nd ed., Oldenburg Verlag, München.

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

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

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

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

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

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

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

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

Ferris, S.P., Hao, Q. and Liao, M.Y. (2013), "The effect of issuer conservatism on IPO pricing and performance", Review of Finance, Vol. 17 No. 3, pp. 993–1027.

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

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

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

Ghosh, C., Nag, R. and Sirmans, C.F. (2000), "A test of the signaling value of IPO underpricing with REIT IPO-SEO pairs", Journal of Real Estate Finance and Economics, Vol. 20 No. 2, pp. 137–154.

Gokkaya, S., Highfield, M.J., Roskelley, K. and Steele, D.F. (2015), "An examination of primary and secondary market returns in equity REIT IPOs", Journal of Real Estate Research, Vol. 37 No. 1, pp. 23–64.

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

Harris, L. (1991), "Stock price clustering and discreteness", Review of Financial Studies, Vol. 4 No. 3, pp. 389–415.

Hartzell, J.C., Kallberg, J.G. and Liu, C.H. (2005), "The role of the underlying real asset market in REIT IPOs", Real Estate Economics, Vol. 33 No. 1, pp. 27–50.

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

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

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

Hoberg, G. and Phillips, G. (2010), "Product market synergies and competition in mergers and acquisitions: a text-based analysis", Review of Financial Studies, Vol. 23 No. 10, pp. 3773–3811.

Hoberg, G. and Phillips, G. (2016), "Text-based network industry classifications and endogenous product differentiation", Journal of Political Economy, Vol. 124 No. 5, pp. 1423–1465.

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

Kennedy, P. (2003), A Guide to Econometrics, 5th ed., John Wiley and Sons, Cornwall.

Lang, M. and Stice-Lawrence, L. (2015), "Textual analysis and international financial reporting: large sample evidence", Journal of Accounting and Economics, Vol. 60 Nos 2–3, pp. 110–135.

Leone, A.J., Rock, S. and Willenborg, M. (2007), "Disclosure of intended use of proceeds and underpricing in initial public offerings", Journal of Accounting Research, Vol. 45 No. 1, pp. 111–153.

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

Ljungqvist, A. and Wilhelm, W.J. (2003), "IPO pricing in the dot-com bubble", The Journal of Finance, Vol. 58 No. 2, pp. 723–752.

Lorenz, F. (2020), "Underpricing and market timing in SEOs of European REITs and REOCs", Journal of Property Investment and Finance, Vol. 38 No. 3, pp. 163–180.

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

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

Loughran, T. and McDonald, B. (2020), "Textual analysis in finance", Annual Review of Financial Economics, Vol. 12 No. 1, pp. 357–375.

Loughran, T. and Ritter, J.R. (2002), "Why don't issuers get upset about leaving money on the table in IPOs?", Review of Financial Studies, Vol. 15 No. 2, pp. 413–444.

Loughran, T. and Ritter, J.R. (2004), "Why has IPO underpricing changed over time?", Financial Management, Vol. 33 No. 3, pp. 5–37.

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

Miller, E.M. (1977), "Risk, uncertainty, and divergence of opinion", The Journal of Finance, Vol. 32 No. 4, p. 1151.

Peterson, K., Schmardebeck, R. and Wilks, T.J. (2015), "The earnings quality and information processing effects of accounting consistency", The Accounting Review, Vol. 90 No. 6, pp. 2483–2514.

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

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

Ritter, J.R. (1984), "The 'hot issue' market of 1980", The Journal of Business, Vol. 57 No. 2, p. 215.

Rock, K. (1986), “Why new issues are underpriced”, Journal of Financial Economics, Vol. 15 Nos 1-2, pp. 187–212.

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

Schrand, C.M. and Verrecchia, R.E. (2005), “Information disclosure and adverse selection explanations for IPO underpricing”, Working Paper, University of Pennsylvania.

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

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

Tinic, S.M. (1988), “Anatomy of initial public offerings of common stock”, The Journal of Finance, Vol. 43 No. 4, pp. 789–822.

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

Welch, I. (1992), “Sequential sales, learning, and cascades”, The Journal of Finance, Vol. 47 No. 2, pp. 695–732.

Yan, Y., Xiong, X., Meng, J.G. and Zou, G. (2019), “Uncertainty and IPO initial returns: evidence from the tone analysis of China’s IPO prospectuses”, Pacific Basin Finance Journal, Vol. 57, 101075.

Zou, G., Li, H., Meng, J.G. and Wu, C. (2020), “Asymmetric effect of media tone on IPO underpricing and volatility”, Emerging Markets Finance and Trade, Vol. 56 No. 11, pp. 2474–2490.

3 Trade vs. Daily Press: The Role of News Coverage and Sentiment in Real Estate Market Analysis

3.1 Abstract

Each week, thousands of newspaper articles on real estate topics are read by market participants. While the market is comparatively intransparent, readers hope to find valuable information. This raises the question of whether this investment of time pays off and whether different types of newspapers are an equivalent source of information. This paper examines the relationship between news-based coverage of real estate topics respectively news-based market sentiment and total returns of the asset classes of residential, office and retail. Using methods of natural language processing, including word embedding, topic modelling and sentiment analysis, three sentiment indicators for each asset class can be derived from 137,000 articles of two trade and two daily newspapers. Our results suggest that trade newspapers outperform daily newspapers in the prediction of future total returns and that the generated sentiment indicators Granger-cause total returns. Moreover, the results indicate that daily newspapers report more negatively on rising returns in the residential market than the trade press. To the best knowledge of the authors, this is the first study to quantify news coverage and sentiment for the main real estate asset classes through means of textual analysis, and to assess different sentiments in trade and daily press.

Keywords: News Sentiment, Textual Analysis, Real Estate Asset Classes, GloVe, Seeded LDA

Acknowledgments: The authors especially thank KPMG Germany for contributing to this study. All statements of opinions are those of the authors and do not necessarily reflect the opinion of KPMG Germany or its associated companies.

3.2 Introduction

A key assumption of modern finance is that a stock price reacts reasonably and promptly to new information (Fama, 1991). This, however, does not exclude the fact that soft factors such as emotions or rules of thumb can impact investor behaviour where hard data is lacking. Investor sentiment plays an important role in financial markets. Keynes (1936) argued that markets can fluctuate under the influence of investors' 'animal spirit', implying that changes in asset prices may be driven by more than just changes in market fundamentals. Revisited by Tversky and Kahneman (1974) in recognition of the bounded rationality and psychological biases of investors who often rely on simplistic models and assumptions, and formalised by Black (1986), who showed that noise traders can increase volatility of stock prices, behavioural finance has become elementary in understanding market pricing. Since, a substantial body of literature, quantifying the stock market-related effects of investor sentiment, has evolved.

In recent years, comprehensive research has expounded the importance of investor sentiment in real estate markets. Real estate investors are particularly susceptible to sentiment and the opinion of others in their decision making (Gallimore and Gray, 2002) due to the informational inefficiencies of real estate markets (Fisher et al., 2003; Heinig and Nanda, 2018). The effects of sentiment are manifold as investor sentiment affects for example real estate market returns (Ling et al., 2014; Heinig et al., 2020), market liquidity (Freybote and Seagraves, 2018), cap rates (Clayton et al., 2009) and the ex ante risk premium required by investors (Beracha et al., 2019b). These effects in turn have implications for securitised real estate markets, where investor sentiment also influences REIT returns (Lin et al., 2009; Das et al., 2015; Carstens and Freybote, 2021) and REIT SEOs (Deng et al., 2014). Hence, it is crucial to understand the effects of investor sentiment for analysts, policy makers and market participants in order to improve investment decisions and policy formulation.

Despite the importance of investor sentiment, above mentioned studies have significant shortcomings in determining sentiment. Investor sentiment is either measured by surveys, which are subject to selection biases and often not available in a timely manner or based on market fundamentals, which is at most an indirect measure of investor sentiment (e.g. Clayton et al., 2009; Ling et al., 2014). Therefore, text-based sentiment indicators have come into focus, providing more flexibility when it comes to temporal aggregation periods and transferability to other key figures (Hausler, 2019). Consequently, Beracha et al. (2019a), Hausler et al. (2018), and Ruscheinsky et al. (2018b) brought together textual-sentiment analysis and real estate markets and found significant explanatory power of those indicators in the commercial and securitised real estate sector. However, those studies do not consider the findings of Marcato and Nanda (2016), who have demonstrated varying importance of survey-based as well as market fundamental-based sentiment indicators for residential and commercial real estate markets. Hence, deriving general market sentiment from newspaper headlines does not meet the requirements for asset class-specific real estate cycles. Contributing to this body of literature, this paper uses textual analysis to examine the impact of newspapers sentiment on investors and thus on real estate markets. Since, it is to be presumed that news-based sentiment indicators should be asset class-specific, this paper supplements the existing literature by creating asset class-specific textual sentiment indicators.

When analysing news, it should be considered that publisher meet varying informational needs and therefore the selection of articles and the tone can change depending on the targeted readership. Thus, the market sentiment would be reflected incompletely by the analysis of one newspaper only. For this reason, this study is the first to apply a comprehensive approach facilitating the examination of the different publishing priorities of leading trade newspapers and daily newspapers in Germany. This approach also allows us to go beyond textual sentiment already explored by prior research, by addressing the concept of the economic narratives by Shiller (2017), in which a narrative is donated by a collective and meaningful story-like interpretation of specific information. According to Shiller, these narratives, if they reach a sufficiently large number of market players, can have an impact on economic outcomes and, as a consequence, also help to explain and predict future economic events. Therefore, we quantify the news coverage of newspapers in relation to each asset class. The insights gained from this approach can be crucial for investors in several respects. First, since news coverage and news sentiment affect real estate markets, our asset class-specific indicators could improve their understanding of the market mechanisms. Second, since investors have to decide based on the limited information available, it is even more important to understand the information objectives of their sources. Last, our approach enables them to process plenty textual information in a short time.

We collected a unique set of 136,548 real estate-related articles, containing approximately 39 million words, from the two leading German real estate newspapers and two major daily newspapers for the time period between 2010 and 2020. The articles are then classified to the asset classes residential, office, and retail and the underlying textual sentiment is calibrated for each article. More precisely, in the first step we applied a seeded Latent Dirichlet Allocation algorithm, enriched with asset class-specific seed words, which assigns articles to an asset class. This step allows the analysis of the respective newspapers' focus with regards to content by asset class. In the second step, we used a combination of two dictionaries to determine the sentiment of each article and thus for each asset class over time. Ultimately, the intertemporal links between reporting intensity and sentiment, on the one hand, and the total returns in the residential, office, and retail market, on the other hand, are examined. Hereby the unique dataset allowed us to shed light on the similarities, but also, and more importantly, on the disparities between trade newspapers and daily newspapers. Recognising the varying level of responsiveness of certain markets to shifts and shocks in investor sentiment, we indeed find that news coverage and sentiment are Granger-causal for future asset class total returns. Both indicators are up to two quarters ahead of returns, allowing investors to react in time to changes in the market environment and to predict returns. Further, trade newspapers and daily newspapers differ in their reporting, which can be at least partially explained by considering their contrasting target groups. More specifically, we find that trade newspapers are a more reliable source for investors as they share the investors' perspective.

The remainder of this paper is organised as follows: Section 3.3 presents applications of news classification in real estate, the importance of sentiment indicators in real estate markets, and lastly develops the hypotheses for this paper. The dataset is presented in section 3.4., while both the textual analysis procedure and methodology regarding the analysis of our hypotheses are described in section 3.5. Section 3.6 discusses the empirical results, which then lead to the conclusion in section 3.7.

3.3 Literature Review and Hypothesis Development

3.3.1 Sentiment Analysis

Textual analysis is the process of analysing textual data to extract meaningful information and is used to quantify of large amounts of unstructured data. Various techniques and algorithms enable computers to understand human language and convert it into structured data, which is key to uncover trends, identify patterns, and make predictions. It is a rapidly growing area of research as there are still challenges to overcome in textual analysis. This also applies to one of the most important branches of textual analysis: sentiment analysis. Researchers have to consider that language respectively text is ambivalent and vague, and therefore can be disparate to the information contained within it as well as the intentions of authors (Loughran and McDonald, 2020).

There is a wide range of both financial and real estate literature in which sentiment indicators have been utilised. Traditionally, investor sentiment has been measured by surveys or by indicators derived from capital market data (e.g., Clayton et al., 2009;

Freybote and Seagraves, 2017). While the former usually suffer from selection bias and of interviewees lacking either the incentive or ability to reveal their true preferences, the latter overstretches the concept of sentiment as it is difficult to disentangle the relative impact of soft and hard data, both of which is often comprised in capital market indicators. Tetlock (2007) quantifies the sentiment of news articles from *The Wall Street Journal* by assigning words to various sentiment dictionaries. Hereby, each dictionary comprises words of a certain sentiment, e.g., positive or negative. In contrast to the survey and capital market methods, this approach enables the examination of a larger variety of sources as well as the representation of qualitative, soft information. Through use of the textual sentiment approach, Tetlock discovered that elevated media pessimism leads to downward movements on stock markets and that sentiment is therefore an important leading indicator for financial markets. Subsequent to Tetlock's findings, several dictionary-based papers followed (e.g., Feldman et al., 2010; Loughran and McDonald, 2011; García, 2013). One of the biggest criticisms of the dictionary approach is its dependence on predefined dictionaries, which, in addition, should be domain-specifically adjusted as the meaning of a word is sometimes inextricably tied to its context (Henry and Leone, 2016). Another stream of literature attempts to tackle this criticism by utilising machine learning (ML)-based approaches such as support vector machines and Naïve Bayes classifiers to extract textual sentiment (e.g., Li, 2010; Hausler et al., 2018). Nonetheless, due to its traceability and replicability, the dictionary-based approach remains the dominant form of methodology, in particular when it comes to sentiment analysis (Loughran and McDonald, 2020).

Regardless of which sentiment index is used, various studies have found significant relationships between market sentiment and real estate market developments. In particular, when examining securitised real estate markets, it becomes clear that sentiment can have a significant impact on the demand for and prices of these assets. Studies, like Clayton et al. (2009), Ling et al. (2014) and Ruscheinsky et al. (2018b), have shown that positive survey-based investor sentiment indicators can lead to increased market returns in subsequent periods. Sentiment has an impact not only on returns, but also on the trading behaviour of institutional investors (Das et al., 2015) and the liquidity of real estate markets (Freybote and Seagraves, 2018). Therefore, the inclusion of sentiment indicators in traditional forecasting models is a logical step to improve the accuracy of these models (Heinig and Nanda, 2018). In addition, there are some studies that have found comparable results in the direct real estate market. Hausler et al. (2018) show that an ML-based sentiment obtained from professional financial news leads the US securitised and commercial real estate markets. Using a different type of textual corpus, REIT financial statements, Carstens and Freybote (2021) similarly find that that the sentiment could be used to predict commercial real estate total returns in the next quarter. These findings can also be confirmed with regards to the residential real estate market. For example, Soo (2018) explains that changes in housing media sentiment have significant predictive power for future changes in house prices that lead by nearly two years. Focusing on the German residential market, Ruscheinsky et al. (2018a) created a German domain-specific dictionary and revealed a statistically significant relationship between the sentiment of the most important German real estate newspaper and the development of residential real estate prices. Marcato and Nanda (2016) are the only researchers to analyse both residential and commercial real estate markets. Despite the significant impact of sentiment in both markets, they found that survey-based sentiment indices vary in importance depending on the market, which may reflect a greater level of responsiveness of certain markets to shifts and shocks in sentiment.

3.3.2 Topic Modelling

Another important branch of textual analysis is topic modelling, a technique used to automatically discover the topics present in a corpus of text. This is done by using algorithms that identify and group together words that frequently occur together in the text, and then assign a label to each group of words, called a "topic." Through this process researchers can quickly and easily analyse large amounts of textual data and gain insights into the underlying topics in the data. The flexibility of topic modelling has made it applicable to a wide range of academic subjects and has been applied in areas such as Social Media Analytics or Finance (e.g., Jianfeng Si et al., 2013; Nguyen and Shirai, 2015; Nordheim et al., 2018). The predominant topic modelling approach, Latent Dirichlet Allocation (LDA), from Blei et al. (2003) is a generative, unsupervised method employed for the identification of latent attributes producing topics, i.e., word groups with a common context. LDA and Naïve Bayes, a similar classification algorithm, have often been used in finance to classify news, e.g., from newspapers or social media, as either buy, sell or hold recommendations (similar to the sentiment indicator described above). For example, Antweiler and Frank (2004), Das and Chen (2007), and Sprenger et al. (2014) examine the relationship between social media content and market movements as well as the performance of individual stocks, finding significant correlations. In the context of real estate research, topic modelling has been applied in a few studies: Winson-Geideman and Evangelopoulos (2013) used Latent Semantic Analyses (LSA)[3] to structure the body of

[3] LSA aims to reduce matrix dimension in contrast to LDA, which is focused on solving the topic modelling issue.

published research in four real estate related journals to identify the main topics in real estate research and Evangelopoulos et al. (2015) presented methods and applications of LSA in order to investigate unstructured data in real estate research to better understand real estate issues. Koelbl et al. (2020) are the only ones to apply structural topic modelling (STM) with regard to financial markets, by determining risk-factor topics discussed in the 10-K filings of US REITs and by examining whether and how these topics affect the risk perceptions of investors. In order to proceed, the algorithm extracts word clusters from a corpus and indicates for each filing which cluster is predominant. Hereby, the user has no impact on the detected word clusters. This, however, is a major limitation of the LDA algorithm, especially when the goal is to identify specific topics of interest in the text, or to classify text into pre-defined categories. Jagarlamudi et al. (2012) addressed this problem and developed a semi-supervised LDA algorithm. The so-called seeded Latent Dirichlet Allocation (sLDA) determines the probability of a document discussing a certain topic. Hereby, the topics are defined ex ante and the algorithm is enhanced with seed words for each topic. By using sLDA, Mai (2021), for example, extracts narratives from the New York Times and ascertains that these extracted narratives can serve as strong market predictors. Likewise, Antweiler and Frank (2006) cluster *The Wall Street Journal* corporate news stories into business incidents and identify an overreaction to news on the American stock markets. The only application of sLDA in the context of real estate is in the form of a recent paper of Ploessl et al. (2021), in which the authors use sLDA to assign almost 170,000 German real estate news articles to six real estate-related trends to quantify their relevance over time. It is revealed that both the news coverage and sentiment of these putatively stable trends have cyclical elements throughout the examined 20-year period.

3.3.3 Hypothesis Development

Given that real estate asset classes are not moving in sync, we revisit in the findings of Marcato and Nanda (2016) and develop asset class-specific sentiment indicators. However, our analysis goes one step further than the existing literature as we divide the text corpus not only according to its relevance for residential and commercial markets, but by subdividing the corpus for the commercial market into retail and office, reflecting the recently diverging market drivers for these two submarkets. In addition, our analyses allow us to shed light on the reporting intensity for each asset class, namely another indicator of the degree of presence of a particular topic, which turn having an impact on economic outcomes, according to Shiller (2017).

So far, sentiment indicators, based on text-based measured tonality changes in social media or news media, implicitly hypothesise that any of these calculated media sentiments reflect the market sentiment. However, given that trade and daily press have two diverging readerships, they may address various topics differently. Thus, in this study we utilise a unique dataset, which enables the comparison of differences between reporting intensity and sentiment indicators in a text corpus aimed at professional and non-professional readers, i.e., indicators that are derived from the trade press and daily press, respectively.

Prima facie, it is not clear whether trade and daily newspapers should share a common understanding concerning real estate markets, due to their divergent target groups: daily newspapers aim to inform the general public and trade newspapers serve professional clients. Two aspects could result from these varying target groups. Firstly, trade newspapers can be expected to put more emphasis on market-related topics and thus be ahead and more precisely informed as they need to provide information for investment decisions to reach their target readership. The daily press, however, may address more multi-layered objectives, since their readers could mirror a broader set of stakeholders in the real estate industry, i.e., non-profit-oriented actors, occupiers, or political players. Secondly, one can expect both types of newspapers to interpret changing market conditions differently, depending on their readership. Therefore, the first hypothesis can be formulated as follows:

Hypothesis 1: *There are significant differences between news coverage and sentiment obtained from either trade newspapers or daily newspapers.*

Despite these differences, the indicators are expected to have a similar relationship towards the performance (total return) of asset classes. More precisely, we expect news coverage and sentiment to be positively related to performance, as newspapers generally report both more frequently and more optimistically upon well performing asset classes, resulting in the following hypothesis:

Hypothesis 2: *News coverage and sentiment are positively correlated to performance for each asset class.*

Following Shiller's (2017, 2020) argument, being that the narrative style, i.e., sentiment, can influence individual and collective investor decision-making if continuously and consistently disseminated across information sources, we hypothesise that it is primarily the sentiment indicators that lead real estate market movements. Assuming that newspapers are an important source of information for investors, we expect that neither indicator to be solely positively correlated with performance, but rather affect the formation of investors' opinion. Furthermore, real estate markets are not fully efficient and are slow to react to new information (Baum et al., 1996; Clayton et al., 2009).

Consequently, the sentiment indices should have a predictive power over the future performance of an asset class, leading to the third hypothesis:

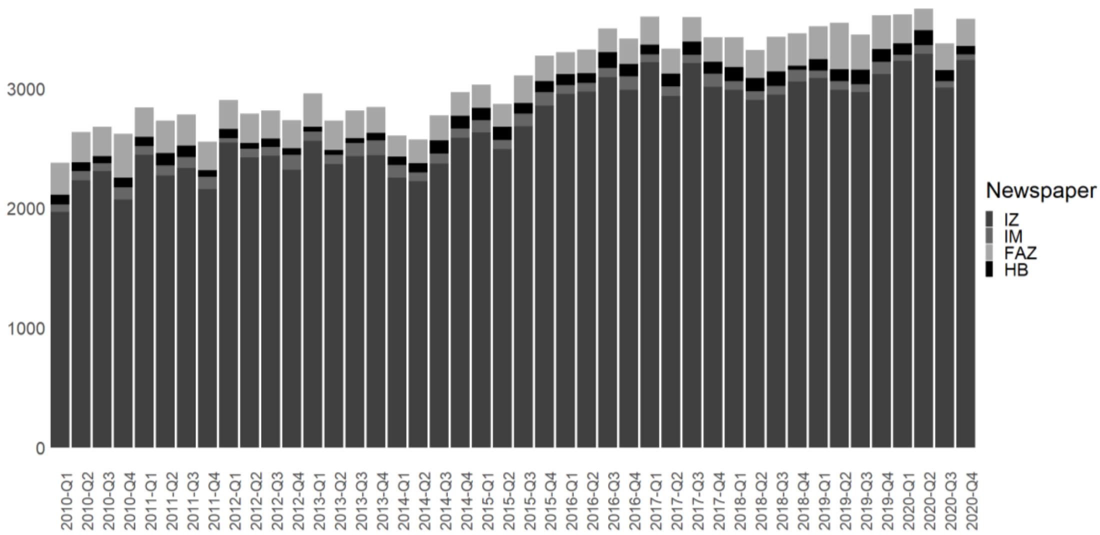
Hypothesis 3: *News Coverage and sentiment have predictive power over future performance for each asset class.*

3.4 Data

3.4.1 Textual Corpus

The corpus of this study consists of a unique dataset of 136,548 newspaper articles published between Q1-2010 and Q4-2020 by four renowned German newspapers. These newspapers comprise two real estate-specific trade newspapers, *Immobilien Zeitung* (*IZ*) and *Immobilien Manager* (*IM*), and two general interest daily newspapers, *Frankfurter Allgemeine Zeitung* (*FAZ*) and *Handelsblatt* (*HB*). *IZ* is the leading newspaper in the real estate industry, which has a weekly circulation of 11,000 copies; the *IM* is published monthly with a circulation of approximately 14,000 copies. In contrast, the *FAZ* and *HB* have a significantly greater daily circulation of almost 210,000 and 135,000 copies, respectively.[4] However, since these general interest daily newspapers cover diverse sections such as politics, economics, finance, sports, etc., the number of real estate-related articles is relatively low and stronger outliers due to more randomly selected topics therefore become possible. Figure 3.1 reinforces the disparity between trade and daily newspapers, since the daily *FAZ* and *HB* only slightly exceed the number of real estate-related articles from *IM*. The articles of the *IZ* have an average length of 229 words and

Figure 3.1 - Number of Articles over Time by Newspaper



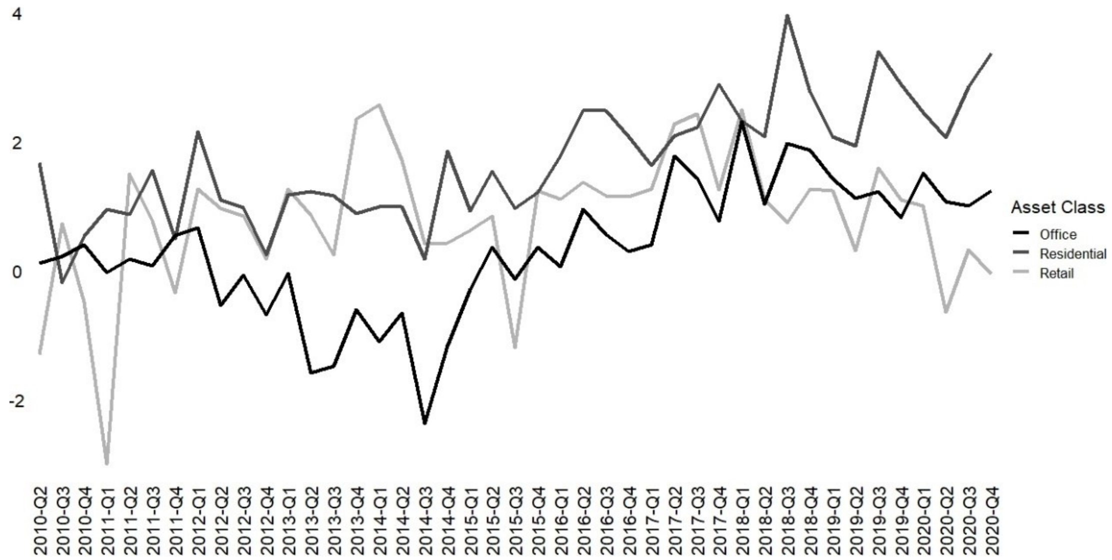
[4] Newspapers' circulation in Q4 2021 according to *IVW* (*Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e. V.*)

the *IM* of 998 words; i.e., the *IZ* provides compact sector-related news for professionals and investors, while more detailed information can be found in the *IM*. The articles of the daily newspapers have similar article lengths (*FAZ*: 565, *HB*: 612)[5] and offer detailed information to the reader. In order to reduce the impact of potential random outliers in the daily newspapers and the impact of longer articles in *IM*, the results for *IM*, *FAZ* and *HB* are smoothed in the following analysis.[6]

3.4.2 INREV Returns

The dependent performance variable is based on total return of the *German Vehicles Index* extracted from the *European Association for Investors in Non-Listed Real Estate Vehicles* (*INREV*) database. *INREV*'s *German Vehicles Index* comprises the data of 49 German vehicles, invested in German retail, residential or office properties, which provide quarterly total return data for each asset class since 2003. During the observation period, the vehicles hold, on average, properties in Germany totalling a gross asset value of 14.4 billion Euro, whereby the allocation amongst residential, office and retail is almost equal. However, this index is merely an approximation regarding the actual performance of the asset classes, as vehicle returns are distorted by company-specific factors. Nonetheless, in the absence of quarterly data for the broader *MSCI Germany Annual Property Index* or the *Bulwiengesa Property Index*, the *INREV* index remains the best alternative with which to capture general and short-term market movements.

Figure 3.2 - Percentage Total Return over Time by Asset class per Quarter



[5] The standard deviation of the article length for the *IZ* (*IM/FAZ/HB*) is 238 (431/391/469).

[6] The total number of articles is 118,645 (3,601/10,497/3,805) of *IZ* (*IM/FAZ/HB*).

Figure 3.2 shows that the total returns for residential and office properties increased in a similar pattern; nevertheless, office returns declined more strongly between 2012 and 2014. All in all, residential properties manifest the highest average quarterly total returns for the observation period (1.71%) and outperform the office market (0.36%). Total returns of retail properties initially show solid growth but subsequently decline. However, due to the early growth, retail real estate still yields, on average, the second highest total returns (0.82%) while simultaneously remaining the most volatile market with a standard deviation of 0.0106.

3.4.3 Macroeconomic and Real Estate-Specific Controls

Since real estate market returns are driven by many factors, a set of macroeconomic and real estate-specific control variables is included. The control variables have been selected following Clayton et al. (2009), Ling et al. (2014), Walker (2014), Das et al. (2015), and Freybote and Seagraves (2017). To account for the performance of the general stock market, we incorporate the return of the *DAX* index (*DAX*), which is the German stock index consisting of the 30 major German blue-chip companies during the observation period and is obtained from *boerse.de Finanzportal*. Secondly, the model includes the Gross Domestic Product of Germany (*GDP*), which is published by the Federal Statistical Office of Germany. Further, to account for asset class-specific factors causing variations in returns, two additional variables are included: the trading volume of each asset class (*TV*), which is an indicator for the liquidity risk of the submarkets (Franzoni et al., 2012), and the average lending rates (*INT*), as another market fundamental. While data on *TV* has been obtained from multiple sources (residential: *JLL*; office: *BNP Paribas Real Estate*; retail: *NAI*

Table 3.1 - Descriptive Statistics

	Mean	Median	SD	Min	Max
Return Residential	0.0171	0.0167	0.0092	-0.0018	0.0396
Return Retail	0.0036	0.0038	0.0100	-0.0237	0.0231
Return Office	0.0082	0.0101	0.0105	-0.0299	0.0258
GDP	0.0087	0.0127	0.0290	-0.0997	0.1025
DAX	0.0075	0.0123	0.0333	-0.0901	0.0742
TV Residential	0.0568	0.0321	0.1376	-0.2539	0.4009
TV Retail	0.0271	0.0410	0.0739	-0.1662	0.1612
TV Office	0.0186	0.0179	0.0952	-0.1803	0.3015
INT Residential	-0.0003	-0.0002	0.0007	-0.0018	0.0013
INT Retail	0.0106	0.0060	0.0131	-0.0110	0.0370
INT Office	-0.0014	-0.0010	0.0022	-0.0060	0.0030

Notes: Return is the total return as provided by *INREV*. Macroeconomic control variables are the growth rate of the German Gross Domestic Product (*GDP*) and the growth rate of the *DAX* index (*DAX*). The asset class-specific controls comprise the growth rate of the trading volume (*TV*) and the growth rate of the interest rate (*INT*).

Global), *INT* was provided by *Deutsche Bundesbank*. Each control variable was converted into growth rates in order to avoid non-stationarity, and data only available on an annual basis was temporally disaggregated.[7] Descriptive statistics for *INREV* returns and control variables are presented in Table 3.1.

3.5 Methodology

3.5.1 News Coverage – sLDA and GloVe Model

In this study we apply a twostep natural language processing procedure to derive a news coverage (*NC*) and a news sentiment (*NS*) index for three real estate asset classes (residential, office and retail) from German newspapers articles. The first step, assigning pre-processed articles[8] to an asset class, is required as the articles are not tagged by keywords that allowed for inference to the asset class discussed in an article. Therefore, a sLDA algorithm is used, which is based on the state-of-the-art LDA algorithm for topic modelling in the finance literature and which has been developed by Jagarlamudi et al. (2012). The sLDA assigns the articles to one of the ex ante defined topics k, in this case the three asset classes, which the algorithm creates around certain seed words s, provided by the researcher.

Accordingly, before applying the sLDA algorithm, a list of suitable seed words must be created for each asset class, i.e., words that best describe the respective topics 'residential', 'office' and 'retail'. Following Watanabe and Zhou (2020), seed words should be both knowledge-based and frequency-based. Regarding knowledge-based seed words, King et al. (2017) show that even expert humans perform poorly and are unreliable on seed word-selection. To automatically identify 80 suitable seed words for each asset class from the corpus, a Global Vector model (GloVe), a word embedding approach introduced by

[7] Temporal disaggregation is referred to as the process of deriving high frequency data from that of low frequency. Furthermore, macroeconomic and real estate-specific control variables have either been tested (e.g., unemployment rate, wages, population growth, building permits, construction cost indices) but did not lead to the improvement of estimation results, or data was not available on asset class level (e.g., construction turnover).

[8] Pre-processing of the articles involves the removal of punctuation marks, numbers, nonalphabetical and special characters, and stop words. "Stop words" relate to frequently occurring words that have no relevance to the content of a text, such as "the" or "and". For this large text corpus, a general German stop word list is extended by frequent words in the real estate industry context. Furthermore, illustrations, tables, English articles, and editorial shortcuts are also excluded. The data is tokenised for the ensuing tasks. This process divides the text into units (tokens) such as phrases, words, and other meaningful entities. In this case, the text corpus is segmented by words.

Pennington et al. (2014), is applied. GloVe creates a word vector for each word in the corpus, which then can be used to identify relationships between words. By measuring the similarity of these vectors using cosine similarity, words that are synonyms to 'residential', 'office' and 'retail' or at least used in the same context can be identified and used as seed words. In this study the similarity is measured using cosine similarity, which means that synonyms and words used in the same context have a higher similarity. For instance, the algorithm can reveal, considering the two sentences 'Company X bought a property in London.' and 'Corporation Y purchased a building in Berlin.', that the tokens 'property' and 'building' or 'bought' and 'purchased' have similar meaning. To generate these vectors for word representation, the algorithm combines the advantages of global matrix factorisation and local context window methods.[9] For this reason, the training of the algorithm is conducted on only non-zero elements in a global word-word co-occurrence statistic X, rather than on the entire sparse matrix or the local context windows within a large text corpus, and therefore generates a vector space of meaningful substructures. The objective function J of the weighted least squares regression model is specified as follows:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \widetilde{w}_j + b_i + \tilde{b}_j - logX_{ij})^2 \tag{3.1}$$

where V is the size of vocabulary; X_{ji} lists the number of times word j occurs in the context of word i; w $(\widetilde{w})$ represents the word vector for a main (context) word and b $(\tilde{b})$ describes the scalar biases for main (context) words; w^T indicates a transposed word vector. In order to prevent the learning of simple common word pairs, a weighting function readjusts the cost for each word pair because word pairs exhibit different occurrence frequencies in the text corpus. Given a co-occurrence count higher than or equal to a certain threshold, the weight is set to 1. Otherwise, the weight is based on the co-occurrence count (see Pennington et al. (2014) for a detailed description).

Using the resulting 80 seed words for each asset class as an input, the newspaper articles can then be assigned using semi-supervised topic modelling. Unsupervised topic models implicitly use document-level co-occurrence information to group semantically related words into a single topic. However, since the goal of these models is to maximise the likelihood of the observed data, they tend to identify only the most obvious and superficial topics within a corpus. To avoid this and to identify ex ante selected topics, a semi-

[9] Matrix factorisation methods decompose large matrices that capture statistical information of a text corpus and generate word representations of low-dimensional latent space in order to reduce computation time, while the concept of context window methods is to predict linguistic patterns as linear relationships between the word vectors based on local context windows and to perform better on word analogy tasks.

supervised sLDA algorithm is applied, which helps the model to better capture the topics of interest by providing some initial guidance. Guidance is provided in the form of the seed words, which in turn inform the model about which words should be associated with each topic. Technically, in sLDA each topic k is defined by two distributions, the regular topic distribution $\emptyset_k^r$, and an additional seed topic distribution $\emptyset_k^s$, where parameter π controls the probability from which of the two distributions a word is drawn. In this way, topic-word distributions and document-topic distributions can be biased towards the starting words and topics respectively. The generative story of the seeded LDA by Jagarlamudi et al. (2012) is as follows:

(1) For each $k = 1, \ldots, T$,
- Choose regular topic $\emptyset_k^r \sim$ Dirichlet (β_r)
- Choose seed topic $\emptyset_k^s \sim$ Dirichlet (β_s)
- Choose $\pi_k \sim$ Beta (1,1)

(2) For each seed set $s = 1, \ldots, S$,
- Choose group-topic distribution $\psi_s \sim$ Dirichlet (α)

(3) For each document d,
- Choose a binary vector $\vec{b}$ of length S
- Choose a document-group distribution $\zeta^d \sim$ Dirichlet ($\tau\vec{b}$)
- Choose a group variable $g \sim$ Multinomial (ζ^d)
- Choose $\theta_d \sim$ Dirichlet (ψ_g) // of length T
- For each token $i = 1, \ldots, N_d$:
  - Select a topic $z_i \sim$ Multinomial (θ_d)
  - Select a indicator $x_i \sim$ Bernoulli (π_{z_i})
  - if x_i is 0: select a word…
  - Select a word $w_i \sim$ Multinomial ($\emptyset_{z_i}^r$)
  - if x_i is 1: select a word…
  - Select a word $w_i \sim$ Multinomial ($\emptyset_{z_i}^s$)

where T is the number of topics; S represents the number of seed sets; N_d illustrates a token; d stands for a document; w is the observed word; z denotes the topic assignment; $x(\vec{b})$ indicates a binary variable (vector); g is a group variable; ψ means the group-topic distribution; $\theta(\zeta)$ describes the document-topic (document-group) distribution and α, β, τ are hyperparameters that are used to control the learning process (see Jagarlamudi et al. (2012) or the appendix in Ploessl et al. (2021) for a detailed description). Once articles are assigned to an asset class, the news coverage index is calculated as follows:

$$NC_{i,k,t} = \frac{k_{i,t}}{n_{i,t}} \tag{3.2}$$

where $NC_{i,k,t}$ is the news coverage of topic (asset class) *k* in newspaper *i* in quarter *t* and $k_{i,t}$ the number of articles in newspaper *i* regarding an asset class *k* in quarter *t,* with *NC* ∈ [0; 1], with a higher value meaning that an asset class is discussed more frequently.

3.5.2 News Sentiment – Unbounded Polarity Score

In the second step of the natural language processing procedure, articles' polarities are determined using a dictionary approach. Sentiment dictionaries classify words within a corpus into pre-defined categories (e.g., positive, negative) (Tetlock, 2007), which possesses the advantage of leading to reduced researcher subjectivity and high replicability (Loughran and McDonald, 2016). However, given that domain-specific idiosyncratic word meanings exist, the use of domain-specific dictionaries, as Doran et al. (2012) and Henry and Leone (2016) have shown, is to be recommended. Ruscheinsky et al. (2018a) developed a German real estate-specific dictionary containing 8,144 negative and 5,745 positive words, validated by a representative survey of German real estate professionals. Initially, each word *w*[10] is compared to the polarised words of the respective sentiment dictionary, which are tagged with +1 or -1. Then, a polarised context cluster $c_{i,j,l}$ is defined with *n* words before and after each polarised word by the researcher. In addition to conventional sentiment dictionaries, a dictionary containing 76 valence shifters, such as negators (e.g., not) $w_{i,j,k}^{n}$, amplifiers (e.g., very) $w_{i,j,k}^{a}$, de-amplifiers (e.g., less) $w_{i,j,k}^{d}$ or adversative conjunctions (e.g., but) $w_{i,j,k}^{\mathrm{ac}}$, is incorporated, since valence shifters can strengthen, weaken, or even reverse the polarity of a word where the researcher provides the weight *z*. Once a valence shifter has been identified, it affects the score of the polarised context cluster; if not, the cluster is tagged as neutral $w_{i,j,k}^{0}$. Since the number of positive and negative words is bound to the article length, a weighted function is used instead of raw counts for each article. Accordingly, to calculate the unbounded polarity score δ by Rinker (2019), the weighted context clusters $c_{i,j,l}$ are summed to $c'_{i,j}$ and divided by the square root of the word count $\sqrt{w_{i,\, jn}}$ (see Appendix 3.1 for a detailed description):

$$\delta = \frac{c'_{i,j}}{\sqrt{w_{i,jn}}} \tag{3.3}$$

After determining the polarity of each article, all polarity scores for each quarter and for each of the three asset classes are aggregated to a mean representing the overall

[10] For instance, $w_{1,2,3}$ represents the third word of the second sentence of the first paragraph.

sentiment of a specific newspaper towards an asset class. Hence, news sentiment is determined as:

$$NS_{i,k,t} = \frac{\sum_{1}^{n} \delta_{i,k,t,n}}{n_{i,k,t}} \tag{3.4}$$

where $NS_{i,k,t}$ is the sentiment of newspaper *i* in quarter *t* towards asset class *k* and $\delta_{i,k,t,n}$ the polarity of article *n* about asset class *k* in newspaper *i* in quarter *t*. The sentiment scores are determined using article length weighted polarity scores, with a higher (lower) or above 0 score indicating a preponderance of positive (negative) indicators, respectively.

Since it is conceivable that markets react differently in times when both *NC* and *NS* are high compared with times when only one indicator is high, a joint analysis of both variables is conducted in addition to the isolated analysis, which is why we combine both variables by adding them together:

$$NCS_{i,k,t} = NC_{i,k,t} + NS_{i,k,t} \tag{3.5}$$

A high *NCS* score therefore indicates that *NC* and *NS* are high, which means that an asset class is in the media spotlight and is being positively promoted.[11]

3.5.3 Model Specification

To analyse whether newspapers' sentiment leads or lags in terms of the performance of assets classes, a vector autoregression (VAR) framework is applied. A VAR model does not require a priori assumptions about existing causalities and allows estimation of the intertemporal relationship between sentiment and performance while controlling for possible endogeneity. Each variable is a linear function of its own lags and of other variables' lags. This also enables us to control for any momentum in the dependent variable (Beracha and Downs, 2015). In general, the conventional bivariate VAR model is specified as follows:

$$y_{1t} = \beta_{10} + \sum_{i=1}^{l} \beta_{1i}\, y_{1t-i} + \sum_{i=1}^{l} \alpha_{1i}\, y_{2t-i} + u_{1t} \tag{3.6}$$

$$y_{2t} = \beta_{20} + \sum_{i=1}^{l} \beta_{2i}\, y_{2t-i} + \sum_{i=1}^{l} \alpha_{2i}\, y_{1t-i} + u_{2t} \tag{3.7}$$

[11] Despite the fact that the standard deviations of *NC* and *NS* differ, the differences are minor and do not call for a normalisation of the indices. Also, through the multiplicative linking of the indices, the values would become relatively small and could thus limit the informative power.

where *l* denotes the number of lags and u_{it} a white noise disturbance term with $E(u_{it}) = 0, (i = 1,2), E(u_{1t}, u_{2t}) = 0$. However, based on economic theory, exogenous controls should be included in the VAR framework. In matrix notation, where *X* denotes a matrix of exogenous variables and *B* a matrix of coefficients, the model can be written as follows:

$$y_t = A_0 + \sum_{i=1}^{l} A_i \, y_{t-i} + BX_t + u_{1t} \tag{3.8}$$

The components of the VAR are tested for stationarity through the use of an Augmented Dickey-Fuller Test (ADF). When the null hypothesis and hence the existence of a unit root is rejected, variables are differenced to ensure statistical appropriateness. Additionally, the optimal lag is determined applying Akaike information criterion (AIC) as selection criteria. Accordingly, the lag length, which minimises the value of the information criteria is chosen (Brooks and Tsolacos, 2010). Furthermore, a set of diagnostic tests is performed to ensure robustness of the results. The Breusch-Godfrey Lagrange Multiplier test is applied in order to ensure that residuals are not serially correlated. In addition, normality and heteroscedasticity tests were conducted to ascertain statistical appropriateness.

The relationship between sentiment and performance is further examined by economic significance and Granger causality tests (Granger, 1969). High and statical significant coefficients themselves do not necessarily implicate that their impact on the dependent variable is of economic relevance. If a high coefficient faces a low volatility of the corresponding variable, its real-world impact is still of minor importance. Therefore, the coefficients are multiplied with the standard deviation to determine the economic significance. Further, Granger causality helps us to better understand the lead-lag relationships between sentiment and asset class performance and to eventually determine whether newspapers' sentiment has predictive power and vice versa.

3.6 Results

3.6.1 News Coverage

Following the approach of Watanabe and Zhou (2020) our selection of reliable knowledge- and frequency-based seed words results in 80 seed words for each asset class (see Table 3.7 in the Appendix). For instance, the results of the GloVe model show: The word vector 'residential property' provides the highest correspondence level of 27.93% for the vector 'condominium' or 'housing' (26.97%) in the data set of the *IZ*, while for the daily newspapers 'apartment buildings' (21.00%, *FAZ*) and 'home buyers' (19.30%, *HB*) provide high matches. For 'office property', word vectors such as 'office building' (67.40%) and 'office space' (66.17%), and for 'retail property', vectors such as 'retail park' (23.56%), 'grocery store' (23.24%) or 'drugstore' (22.82%) are detected by the algorithm.[12]

After assigning the articles to the corresponding asset class through the sLDA algorithm, the *NC* indices, as shown in Figures 3.3 to 3.5, can be computed.[13] Hereby, retail is the most frequently reported market, with a mean share of 27.63% of all articles, followed by office (23.48%), and then residential (20.74%).[14] However, during the observation period one can observe how the media's attention shifts away from retail, which could be interpreted as a result of the ongoing transition of the retail sector (e.g., Kaiser and Freybote, 2021), towards residential and office real estate. What is more, when comparing trade newspapers to daily newspapers, one can see that trade newspapers focus more on office. Residential real estate is almost equally represented in both types of newspapers and having become a more frequent item of discussion by both since 2015. Retail is initially more present in daily newspapers, which may be traced back to its presence in everyday life. Moreover, it is the asset class, throughout which the *NC* indices of both types of newspapers move the least synchronously. Even though both indices are positively correlated, as for all asset classes, the reporting intensity of the daily newspapers has decreased significantly since 2015, while it has remained stable in the trade papers.

An analysis of the bivariate relationship between the reporting intensity and the performance reveals that *NC* was able to be related to total return. In particular, the *NC* and performance of residential real estate are highly correlated, with the Pearson correlation coefficient standing at 0.78.[15] The same applies to the office market, in which correlations are highest using one to four period lagged *NC*s (between 0.49 and 0.56).[16] Hence, reporting intensity could be considered as being a good predictor for future returns, with more media attention being indicative of higher total returns. Regarding the retail market, the picture is less clear: Correlation between the *NC* of retail and its total return is even negative (-0.18) and highest (-0.32) when total return leads *NC* by two quarters.

[12] See Table 3.8 in the Appendix for the descriptive statistics of the generated indices. All results in this paper were estimated in German and have been translated. Expressions consisting of two or more words in English appeared as single words in German.

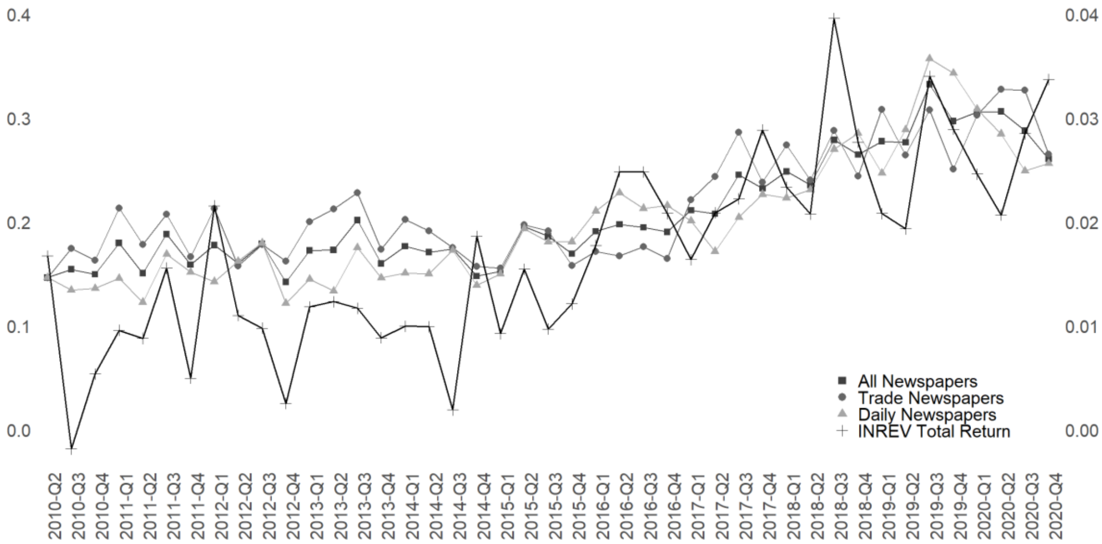
[13] The time series of *IM*, *FAZ* and *HB* have been exponentially smoothed, to reduce their higher variance which are due to the lower circulation and longer articles compared with the *IZ*. Furthermore, indices for trade and daily newspapers are based on the mean of the underlying newspapers.

[14] The remaining 28.06% is not classified to one of the proposed asset classes.

[15] Significant at 1% level.

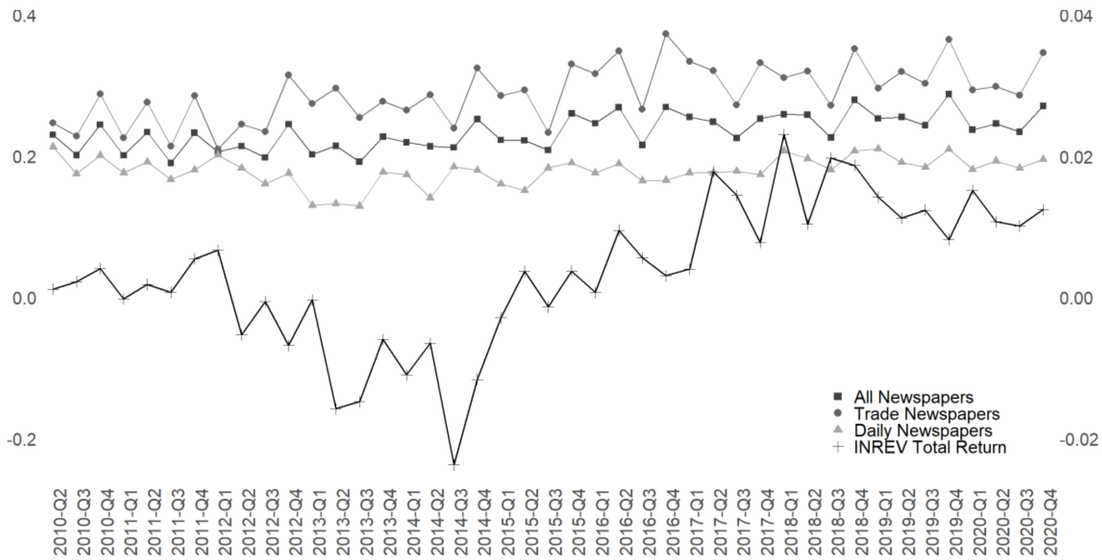
[16] All correlations are at least significant at 5% level.

Figure 3.3 - Residential News Coverage over Time



Notes: News Coverage (left), quarterly *INREV* Index Return (right)

Figure 3.4 - Office News Coverage over Time



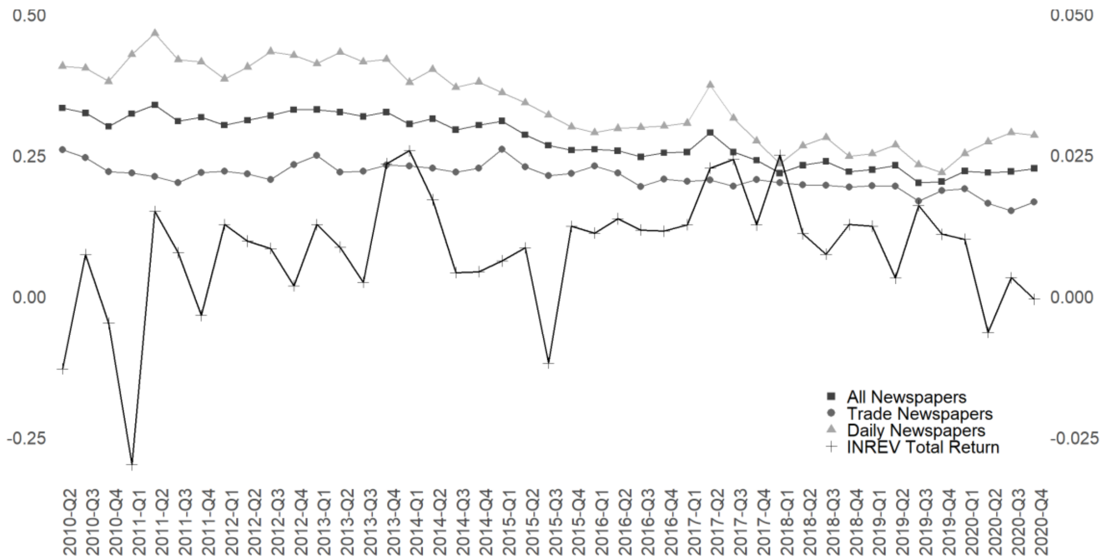
Notes: News Coverage (left), quarterly *INREV* Index Return (right)

These observations are consistent when disaggregating *NC* indices to both types of newspapers and they are validated through a set of bivariate VARs: For residential and office, a diminishing relationship between *NC* and total returns can be identified. For retail, in contrast, no pattern is discernible.

Upon summarising these preliminary results briefly, one can see that there are differences in the news coverage between trade and daily newspapers[17], which seem to be driven by

[17] See Table 3.8 in the Appendix for the descriptive statistics of the generated indices.

Figure 3.5 - Retail News Coverage over Time



Notes: News Coverage (left), quarterly *INREV* Index Return (right)

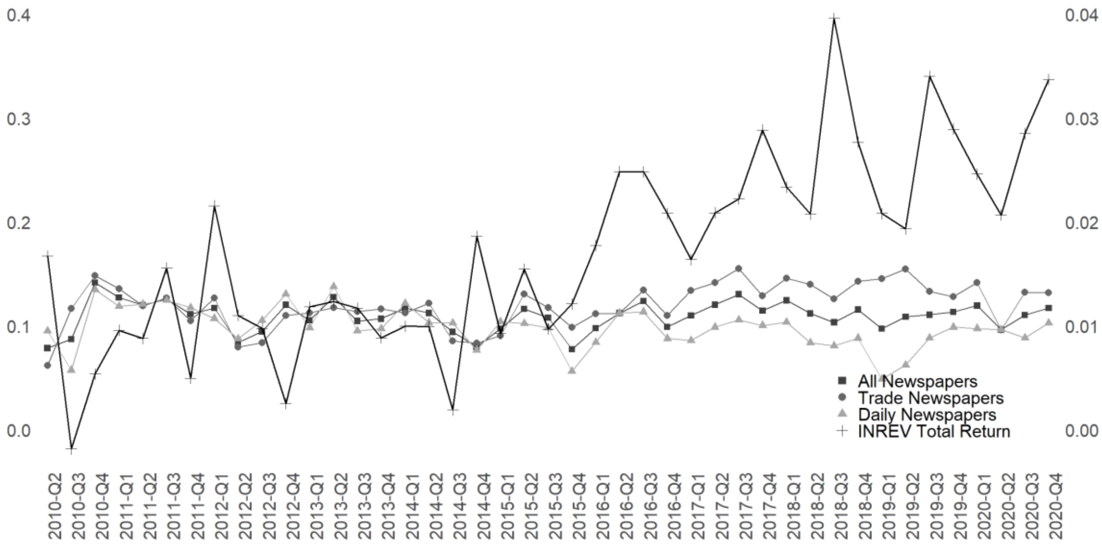
newspaper type applying degrees of varying importance of the changing market conditions. Therefore, regarding our first hypothesis, we conclude that there are in fact differences in news coverage between trade and daily newspapers. Besides, it appears that news coverage precedes the total return indices and may therefore be an indicator of upcoming performance, which, in turn, supports the second hypothesis.

3.6.2 News Sentiment

When turning to the *NS* indices, we observe that *IM* is the most optimistic newspaper concerning each asset class, while *FAZ* is the most pessimistic newspaper. When *IZ* and *HB* are included, however, the tonality of trade newspapers is only slightly more positive than that of daily newspapers, as revealed by Figures 3.6 to 3.8.[18] While residential and retail real estate are assessed similarly optimistically, with a mean *NS* over all newspapers of 0.11, the tone regarding the office market (0.08) is slightly more pessimistic. However, there is a pronounced difference in tonality between trade and daily newspapers within the articles on office real estate: Office is the most positively discussed asset class (0.14) in trade newspapers, whereas in daily newspapers, its mean comprises a mere 0.03. This difference is driven by negative *NS* scores obtained from the *FAZ* (the only negative mean score in the sample). In combination with the low standard deviations of the *NS* indices, we interpret these results as being indicative of a strong opinion or a prejudice of the newspapers toward asset classes. Therefore, it is even more striking that the sentiment

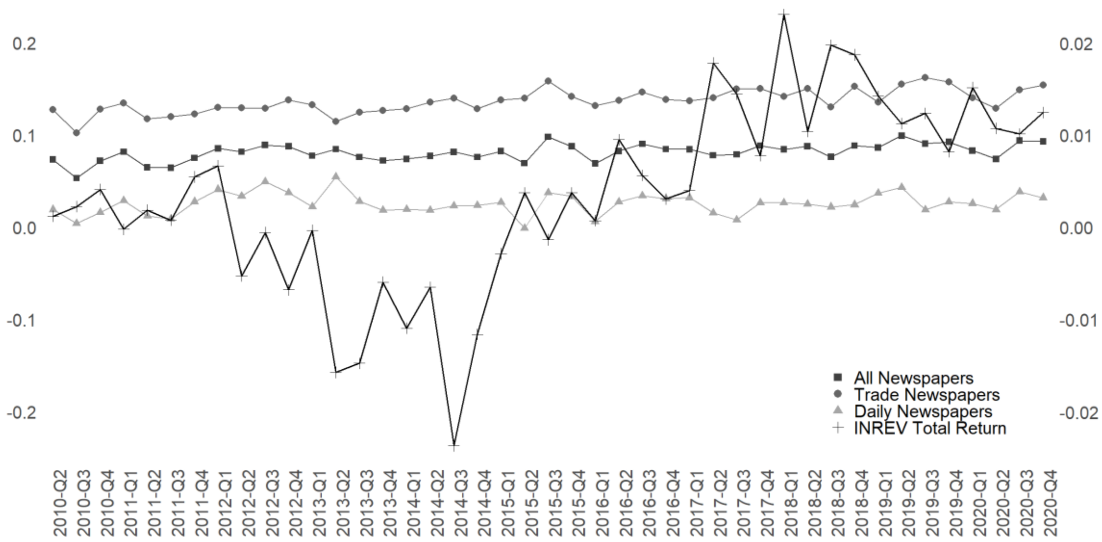
[18] See Table 3.8 in the Appendix for the descriptive statistics of the generated indices.

Figure 3.6 - Residential News Sentiment over Time



Notes: News Sentiment (left), quarterly *INREV* Index Return (right)

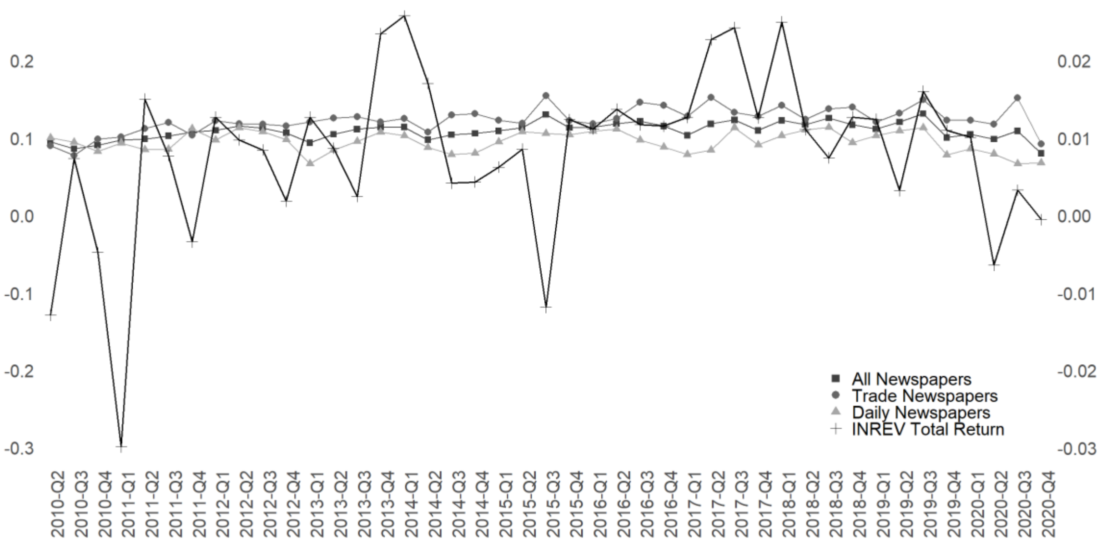
Figure 3.7 - Office News Sentiment over Time



Notes: News Sentiment (left), quarterly *INREV* Index Return (right)

indices of both types of newspapers referring to the residential market diverge. Prior to 2015, both exhibit similar *NS* values, but from 2015 onwards, trade newspapers have been significantly more optimistically written in their articles on the residential market than the daily newspapers. This divergence thus begins simultaneously with the previously described increase in residential reporting intensity, which can possibly be seen as an indicator of changing market conditions.

Figure 3.8 - Retail News Sentiment over Time



Notes: News Sentiment (left), quarterly *INREV* Index Return (right)

To see whether this applies, we compare the *NS* indices to the corresponding total returns. The results are ambiguous: Correlations between sentiment and performance are between 0.14 and 0.36, even for different lags of the *NS* indices, with only office being significant at 1%-level. Additionally, it seems that, as expected, more optimistic reporting is related to higher returns. The use of a set of bivariate VARs largely confirms these findings. However, a breakdown of the *NS* indices by newspaper type reveals one peculiarity: Only for the residential market is there a negative correlation between sentiment, obtained from daily newspapers, and total returns. This is not surprising and finally provides an explanation for the previously described rise in residential *NC* and the accompanying divergence between daily and trade newspaper sentiment. Since 2015 total returns of residential properties have risen sharply, which has led to an increase in the reporting intensity in both types of newspapers. However, while increasing total returns are positively received by the professional clientele of trade newspapers, the readership of daily newspapers, suffers from the implied higher rental and purchase prices. Daily newspapers therefore adopt the perspective of residents rather than investors when reporting on the residential market, rejecting the rising returns.

Overall, it can be stated that there are striking differences between the *NS* indices obtained from trade and daily newspapers. Despite the direction of the sentiment indices for each asset class being largely identical, the tone regarding residential real estate has undergone a significant change since 2015 onwards. Therefore, it can be concluded that newspapers interpret changing market conditions differently. For daily newspapers 'bad news' is 'good news' as real estate does not primarily concern earning money when concerns residential real estate. Raising awareness could be more seen as being important and then 'bad news' becomes more valuable than 'good news'. By contrast, for trade newspapers good information is 'good news', as return is paramount, i.e., readers value any information more symmetrically, as long as it proves to be valuable for investment decisions. Regarding the second hypothesis, it can be concluded that although not unequivocal for residential real estate, there remains a positive relationship between news sentiment and total returns.

3.6.3 Vector Autoregression

The objective of this analysis is to examine the ability of news coverage and sentiment measures to predict the total returns per asset class, despite possible momentum behaviour in total returns and subsequent to controlling for further exogenous effects. Tables 3.2 to 3.4 exhibit the results of the VAR estimations of the return equation, as specified in equation 3.8, for each asset class and newspaper type.

Focusing on the *NC*, the significant effect of the one period lagged *NC* for the office market stands out. The coefficients are significant at the 1% level for the indicator of all newspapers and trade newspapers, which means that there is a lead time, but the comparatively low transparency of the market only allows a prediction of one quarter in advance. Nevertheless, once newspapers anticipate returns in the short-term, they tend to publish more articles upon the office sector (positive coefficient). A similar pattern can be observed for the retail market, where one period lagged coefficients of *NC* are significant. Hereby, the four period lagged coefficients of *NC* are negative which might indicate a potential reversal or correction in reporting intensity, as previously observed by Tetlock (2007), and Antweiler and Frank (2006), with respect to news sentiment and the general stock market. Lastly, for the residential market trade newspapers clearly outperform daily newspaper as coefficients yield higher levels of statistical significance. Even if the direct relationship between the yield component and the reporting intensity appears weaker, the reason for this could be that the reporting of daily newspapers is also more likely than that of trade newspapers to deal with social and political issues in housing. From an economic perspective, it can be observed that *NC* is of significant economic importance, especially for the residential and retail markets. *NC* is hence a valuable indicator to potential investors.

In terms of *NS*, one can ascertain a statistically significant relationship between sentiment and office and retail markets, respectively, but this does not apply to the same degree to the residential market. This contrasts somewhat to the findings of Marcato and Nanda (2016), who found statistically significant effects of a survey-based sentiment indicator on residential real estate returns but not on commercial property returns. Nonetheless, for the

Table 3.2 - Residential: VAR Equation on Returns and Granger Causality Test Results

	All Newspapers			Trade Newspapers			Daily Newspapers		
INREV (-1)	0.019	0.132	0.063	0.138	0.194	0.107 ***	-0.085	0.068	-0.032
INREV (-2)	0.159	0.273 **	0.162	0.136	0.114	0.175 ***	0.137	0.125	-0.019
INREV (-3)	-0.196	-0.087	-0.143	-0.031	-0.080	0.030 ***	-0.342 **	-0.105	-0.243
INREV (-4)	0.452 ***	0.582 ***	0.605 ***	0.722 ***	0.503 ***	0.578 **	0.353 **	0.479 ***	0.486 ***
NC (-1)	-0.002			-0.155 *			0.051		
NC (-2)	-0.008			0.128 **			-0.081 *		
NC (-3)	0.066			0.120 *			0.146 **		
NC (-4)	0.014			-0.133 **			0.012		
NS (-1)		-0.186 *			-0.038			-0.057	
NS (-2)		0.047			0.090			-0.024	
NS (-3)		0.021			-0.054			-0.005	
NS (-4)		-0.100			0.082			-0.111 ***	
NCS (-1)			-0.035			-0.065 **			0.063 **
NCS (-2)			0.032			0.048			-0.037
NCS (-3)			0.068			0.038			0.125 ***
NCS (-4)			-0.049			-0.036			-0.036
Intercept	0.003	0.028 *	0.002	0.012 *	-0.005	0.010 *	-0.008	0.030 **	-0.019 ***
Adj. R^2	0.41	0.43	0.42	0.48	0.42	0.42	0.59	0.49	0.57
χ^2 *NC/NS/NCS*	1.19	1.72	1.03	2.08	1.00	1.04 **	6.11 ***	3.24 **	8.79 ***
χ^2 *INREV*	6.50 ***	3.32 **	4.29 **	9.06 ***	5.32 ***	6.12	2.57 *	5.98 ***	9.29 ***

Notes: Table 3.2 shows the estimated coefficients from the VAR models with quarterly asset class-specific *INREV* total returns (*INREV*) and News Coverage (*NC*), News Sentiment (*NS*) or the combination of both indices (*NCS*). The set of the macroeconomic control variables includes the growth rate of the German Gross Domestic Product (GDP), the growth rate of the *DAX* index (*DAX*), while the asset class-specific real estate controls comprise the growth rate of the trading volume (*TV*) and the growth rate of the interest rate (*INT*). Newey and West (1987) standard errors, which are robust to heteroscedasticity and autocorrelation, are used. * denotes significance at the 10 % level, ** at the 5 % level, and *** at the 1 % level. The sample period is Q1-2010 to Q4-2020.

Table 3.3 - Office: VAR Equation on Returns and Granger Causality Test Results

	All Newspapers			Trade Newspapers			Daily Newspapers		
INREV (-1)	0.161	0.258	0.231	0.295	0.168	0.260	0.246 *	0.077	0.308 *
INREV (-2)	0.481 **	0.516 *	0.316	0.460 **	0.459 *	0.382 *	0.471 **	0.597 **	0.537 *
INREV (-3)	0.013	0.066	-0.059	-0.163	0.062	-0.100	0.041	0.226	-0.089
INREV (-4)	0.172	0.068	0.207	0.273	0.118	0.361 *	0.202	0.370	0.172
NC (-1)	0.247 ***			0.114 ***			0.110 *		
NC (-2)	0.061			-0.025			0.038		
NC (-3)	-0.078			-0.034			-0.001		
NC (-4)	-0.035			0.031			-0.053		
NS (-1)		-0.055			0.068			-0.321 *	
NS (-2)		0.030			-0.121			-0.058	
NS (-3)		0.376 ***			0.402 ***			0.036	
NS (-4)		-0.023			0.140			-0.034	
NCS (-1)			0.156 ***			0.110 **			0.035
NCS (-2)			0.028			-0.056			0.047
NCS (-3)			0.012			-0.019			0.002
NCS (-4)			-0.003			0.061			-0.057
Intercept	-0.038 **	-0.024	-0.055 **	-0.020 **	-0.063 ***	-0.037 ***	-0.013	0.013	-0.003
Adj. R^2	0.52	0.43	0.45	0.46	0.57	0.48	0.36	0.40	0.34
χ^2 *NC/NS/NCS*	17.38 ***	3.70 **	7.42 ***	5.71 ***	11.33 ***	6.31 ***	1.56	2.27	0.75
χ^2 *INREV*	0.93	8.95 ***	1.30	1.75	7.04 ***	0.95	5.55 ***	3.61 **	4.00 **

Notes: Table 3.3 shows the estimated coefficients from the VAR models with quarterly asset class-specific INREV total returns (INREV) and News Coverage (*NC*), News Sentimen (*NS*) or the combination of both indices (*NCS*). The set of the macroeconomic control variables includes the growth rate of the German Gross Domestic Product (GDP), th growth rate of the DAX index (DAX), while the asset class-specific real estate controls comprise the growth rate of the trading volume (*TV*) and the growth rate of the interes rate (*INT*). Newey and West (1987) standard errors, which are robust to heteroscedasticity and autocorrelation, are used. * denotes significance at the 10 % level ** at the 5 % level, and *** at the 1 % level. The sample period is Q1-2010 to Q4-2020.

Table 3.4 - Retail: VAR Equation on Returns and Granger Causality Test Results

	All Newspapers			Trade Newspapers			Daily Newspapers		
INREV (-1)	-0.363 **	0.029	-0.326 ***	-0.225	-0.312 **	0.011	-0.241 **	-0.318 ***	-0.321 ***
INREV (-2)	0.144	0.248 **	0.139	0.225 *	0.003	0.305 ***	0.230 **	0.152	0.095
INREV (-3)	0.150	-0.206	0.116	0.119	-0.111	0.007	0.172	-0.158	0.074
INREV (-4)	0.422 ***	0.292 ***	0.386 ***	0.254	0.082	0.188 *	0.386 ***	0.244 ***	0.327 ***
NC (-1)	0.196 **			0.127			0.094 **		
NC (-2)	-0.012			-0.310 *			0.024		
NC (-3)	-0.093			0.036			-0.044		
NC (-4)	-0.230 *			0.102			-0.112 **		
NS (-1)		-0.036			0.152			-0.143 *	
NS (-2)		-0.637 **			-0.001			-0.263 ***	
NS (-3)		-0.210			0.092			-0.299 ***	
NS (-4)		0.306 **			0.036			-0.039	
NCS (-1)			0.155 **			0.098			0.051
NCS (-2)			-0.084			-0.378 ***			-0.007
NCS (-3)			-0.094			0.034			-0.037
NCS (-4)			-0.170 *			0.254 ***			-0.087 *
Intercept	0.037 *	0.070 **	0.073 **	0.011	-0.027	-0.002	0.012	0.080 ***	0.036 **
Adj. R^2	0.26	0.32	0.26	0.16	0.14	0.31	0.27	0.54	0.22
χ^2 *NC/NS/NCS*	1.64	4.02 **	2.95 *	1.52	1.48	5.15 ***	3.58 **	12.22 ***	3.03 *
χ^2 *INREV*	12.74 ***	7.77 ***	4.41 **	6.74 ***	2.09	2.29	7.43 ***	12.85 ***	3.62 **

Notes: Table 3.4 shows the estimated coefficients from the VAR models with quarterly asset class-specific *INREV* total returns (*INREV*) and News Coverage (*NC*), News Sentimen (*NS*) or the combination of both indices (*NCS*). The set of the macroeconomic control variables includes the growth rate of the German Gross Domestic Product (GDP), th growth rate of the *DAX* index (*DAX*), while the asset class-specific real estate controls comprise the growth rate of the trading volume (*TV*) and the growth rate of the interes rate (*INT*). Newey and West (1987) standard errors, which are robust to heteroscedasticity and autocorrelation, are used. * denotes significance at the 10 % level, ** at the 5 % level, and *** at the 1 % level. The sample period is Q1-2010 to Q4-2020.

residential market it should be noted that signs of the *NS* coefficients are negative for daily newspapers and positive for trade newspapers, which gives proof to the different perspectives of newspapers previously discussed. Regarding office and retail markets, the sentiment indicator is significant further in time, implying that newspapers in general are better at predicting long-term returns. This contrast the combined news coverage and sentiment indicator *NCS*: Significant coefficients can be observed for each asset class using trade newspaper articles, mainly in closer temporal proximity (one and two period lags). These results underline Shiller's (2017) findings by showing that once a topic is discussed more frequently and positively, it has an impact on total returns through the effect it has on readers` mindsets, initially. Further, the positive sign of the statistically significant coefficients shows that the attention created by the newspaper drives readers towards those asset classes and hence returns rise. When comparing the results by type of newspaper it becomes apparent, that trade newspapers outperform daily newspapers in terms of significance and adjusted R^2 for office markets. The picture is less clear for the retail and residential markets. Overall, it can be stated, trade newspapers seem to be better suited for understanding the investment motive of real estate markets and yield higher economic significance, which may be due to their focus on the interests of investors and real estate professionals. However, the investment motive plays a lesser role for residential markets because of the consumer goods component and the high tenant ratio in Germany, which is why daily newspapers have their raison d'être. Similar applies for retail markets, where the direct link to urban topics and consequently external effects is likely to play an additional role.

Tables 3.2 to 3.4 also present the results of the Granger causality tests, where the null hypothesis assumes that neither news coverage nor sentiment Granger-causes total returns and vice versa. χ^2 *NC/NS/NCS* reveals that the rejection of the null hypothesis depends on the type of newspaper, which aligns with above-described focus of the newspapers. For residential and retail the null hypothesis is rejected by daily newspapers, whereas for office it is rejected by trade newspapers. In these cases, then, there is a statistically significant Granger causality which spans the area of news coverage and sentiment to total returns and indicators hence Granger-cause total returns. However, results of χ^2 *INREV* are less unequivocal, since the null hypothesis cannot be rejected for daily newspapers, whereas for trade newspapers it can at least be partially rejected. The reporting in the trade newspapers is therefore less driven by past returns, so there is less of a feedback loop.

3.6.4 Robustness Tests

In order to check the robustness of the results, two further models were fitted in addition to the statistical robustness tests. Firstly, instead of the textual analysis-based sentiment indicator, a survey-based sentiment index for the German real estate market, *Deutsche Hypo Immobilienklima* (*HYPO*) as provided by *bulwiengesa AG*, was used as a sentiment indicator for explaining market returns (Table 3.5). This sentiment indicator is compiled monthly by means of a survey of 1,200 real estate experts for six different asset classes. This indicator may provide insight into whether survey- or news-based sentiment indicators are more suitable for measuring sentiment and for explaining movements in real estate returns. Secondly, the dependent variable *INREV* was replaced by its capital growth component (Table 3.6, results displayed for each asset class at the aggregated level of all newspapers) as we expect that the generated sentiment indicators affect appreciation returns in particular because income returns are more stable overall, being defined by long-term rental contracts. Due to this, income returns are often less suitable for fitting statistical models. Furthermore, as we try to capture investor sentiment, this could be better reflected in changes in cap rates, i.e., in changes in the capital growth component of the total return.

According to Table 3.5, the changes in the *NS* indicator are more closely related to the development of real estate returns than the survey-based sentiment indicator. Only the

Table 3.5 - Robustness Tests: Hypo Index – VAR and Granger Causality Test Results

	Residential-Hypo	Office-Hypo	Retail-Hypo
INREV (-1)	-1.055 ***	0.696 **	-1.017 ***
INREV (-2)	-0.845 **	0.051	-0.569 *
INREV (-3)	-0.955 ***	0.179	-0.500
INREV (-4)	-0.420	0.184	0.010
HYPO (-1)	-0.006	-0.080 **	0.008
HYPO (-2)	0.025	-0.048	-0.026
HYPO (-3)	0.033	0.045	-0.025
HYPO (-4)	0.049	-0.002	0.060
Intercept	0.003	-0.012 *	-0.002
Adj. R^2	0.36	0.23	0.36
χ^2 *HYPO*	14.91 **	13.14 **	14.83 **
χ^2 *INREV*	9.25 *	4.45	15.14 ***

Notes: Table 3.5 shows the estimated coefficients from the VAR models with quarterly asset class-specific *INREV* total returns (*INREV*) and Hypo Index (*HYPO*), which is a survey-based sentiment indicator. The set of the macroeconomic control variables includes the growth rate of the German gross domestic product (*GDP*), the growth rate of the *DAX* index (*DAX*), while the asset class-specific real estate controls comprise the growth rate of the trading volume (*TV*) and the growth rate of the interest rate (*INT*). Newey and West (1987) standard errors, which are robust to heteroscedasticity and autocorrelation, are used. * denotes significance at the 10 % level, ** at the 5 % level, and *** at the 1 % level. The sample period is Q1-2010 to Q4-2020.

Table 3.6 - Robustness Tests: INREV Capital Growth Component – VAR and Granger Causality Test Results

	Residential - All Newspapers			Office - All Newspapers			Retail - All Newspapers		
INREV (-1)	0,057	-0,025	0,069	-0,030	-0,179	-0,042	-0,440 ***	-0,555 ***	-0,502 ***
INREV (-2)	0,123	0,383 **	0,167	0,155	-0,022	0,026	-0,262 **	-0,182	-0,213 *
INREV (-3)	0,267	0,340 **	0,264	-0,169	-0,178	-0,143	-0,042	-0,470	-0,115
INREV (-4)	-0,060	0,300 *	0,079	0,374 ***	0,626 ***	0,347 *	0,219 *	0,062	0,208
NC (-1)	-0,035			0,339 ***			0,156 *		
NC (-2)	0,085			-0,091			-0,109		
NC (-3)	-0,079			-0,242 **			0,318 **		
NC (-4)	0,123			0,034			-0,605 ***		
NS (-1)		-0,060 *			-0,063			-0,024	
NS (-2)		0,054			0,064			-0,182	
NS (-3)		-0,094			0,055			0,499 **	
NS (-4)		-0,057			0,048			-0,342	
NCS (-1)			-0,008			0,273 **			0,089
NCS (-2)			0,099 *			-0,115			-0,086
NCS (-3)			-0,086 *			-0,112			0,246 **
NCS (-4)			0,049			0,017			-0,499 ***
Intercept	-0,014	0,016	-0,014	-0,016	-0,018	-0,029	0,053 ***	0,000	0,083 ***
Adj. R^2	0,22	0,12	0,21	0,56	0,41	0,51	0,66	0,33	0,68
χ^2 *NC/NS/NCS*	2,17	1,55	2,46 *	4,73 **	0,12	2,81 *	18,01 ***	1,93	23,24 ***
χ^2 *INREV*	7,17 ***	7,32 ***	8,21 ***	1,93	1,52	4,25 **	4,36 **	4,61 **	2,19

Notes: Table 3.6 shows the estimated coefficients from the VAR models with quarterly asset class-specific *INREV* capital growth components (*INREV*) and News Coverage (*NC*), News Sentiment (*NS*) or the combination of both indices (*NCS*). The set of the macroeconomic control variables includes the growth rate of the German Gross Domestic Product (*GDP*), the growth rate of the *DAX* index (*DAX*), while the asset class-specific real estate controls comprise the growth rate of the trading volume (*TV*) and the growth rate of the interest rate (*INT*). Newey and West (1987) standard errors, which are robust to heteroscedasticity and autocorrelation, are used. * denotes significance at the 10 % level, ** at the 5 % level, and *** at the 1 % level. The sample period is Q1-2010 to Q4-2020.

survey-based sentiment coefficient for the office market is statistically significant, and also exhibits higher explanatory power than the indicator based on textual analyses. But this does not hold for residential or retail markets. As stated above, the standard deviations of the *NS* indices are low, particularly when compared with the *HYPO* indices. Evidently, the survey-based indicators are subject to more pronounced short-term changes than the extracted tone from journalists in news articles. Hence, news-based sentiment may be more suitable for catching undistorted market sentiment and therefore better to predict future returns.

Limiting the dependent variable to the capital growth component of the total return, which is generally the more volatile component, provides evidence of the fact that information obtained from newspapers still has explanatory power when omitting the stable component. As displayed in Table 3.6, there are significant results for each of our textual analysis-based indicators, even though slight deviations can be observed. For office and retail adjusted R^2 increased substantially only for the residential markets the explanatory power decreases. Overall, both models provide convincing evidence of the fact that, as expected, news-based indicators indeed affect real estate returns and due to higher explanatory power of the models, the indices capture better investor sentiment than the more stable total returns. This result implies that valuable additional information can be extracted from news articles, especially for professional market participants.

3.7 Conclusion

In this article we analyse a broad data set consisting of real estate articles that were published by four leading German newspapers, covering the period from Q1-2010 to Q4-2020. Two of these newspapers are general-interest daily newspapers, while two newspapers are specialised trade newspapers. To the best of our knowledge this is the first study to cluster newspaper articles by asset class in order to derive news coverage and also news sentiment for the analysis of real estate markets. The compiled unique dataset allows us not only to capture the general consensus, but at the same time to distinguish between trade newspapers and daily newspapers.

The aim of this article is to build on the existing literature regarding the relevance of market sentiment indicators that are derived from a large text corpus and not from surveys. This is done in the following two ways: First, following Marcato and Nanda (2016), we distinguish sentiment indicators by asset class, hypothesising that real estate markets are so diverse that such differentiation is necessary. Second, we wish to understand whether trade and daily newspapers present different market narratives, given that their readership varies. This implicitly addresses a third issue, namely, whether the narratives in the newspapers either lead or lag regarding real estate cycles. In particular, regarding professional market participants, this also raises the question of whether reading the news can either be considered primarily as an activity or an investment in the acquisition of relevant information.

With regard to these questions, we formulate three hypotheses: Our first hypotheses is that there are measurable differences between the news coverage and sentiment of trade and daily newspapers. We find an indication that this is the case with regards to residential real estate and, to a lesser extent, with regards to commercial real estate. Given the different readerships of the two newspaper types, with more professional investors reading trade newspapers and a far broader audience reading daily papers, daily newspapers report in a significantly more negative tonality during market phases with rising residential returns. Secondly, we hypothesise that there is a correlation between changes in news sentiment and coverage and the development of total returns within the respective asset classes. Here, we find ample indication of the fact that this is the case, particularly for the capital growth component. It also appears that the news-based sentiment indicator is more suitable for explaining future market developments compared with survey-based indicators. This finding leads to the necessity of further research: survey-based indicators appear to be more volatile than our newspaper sentiment indicator. If this becomes the case, the survey-based indicators may perform better for short-term directional changes, such as real estate stock movements, while news media indicators prove more resilient to short-term changes and thus perform better for more slowly moving direct investments. Our third hypothesis is that the sentiment indicators tend to lead the real estate cycle rather than to follow it. We also find an indication of this being the case, as well as finding that trade newspapers outperform daily newspapers in 'predicting' future return shifts.

In this sense, trade newspapers contain significant market-relevant information even beyond the hard facts that are presented in both newspaper types. The understanding of this information revealed in textual data can enable investors to better comprehend asset class-specific market narratives, and given that daily newspapers appear to follow different narratives for residential real estate, it would be wise to also learn about these, even though they possess less predictive power. Thus, the information derived from newspapers could be valuable to investors and influence both opinion-formation and decision-making. In essence, it could help to increase the understanding of the asset class-specific narratives and sentiment in real estate markets.

Our approach bears several strengths: Qualitative information from newspapers is available digitally and in real-time. Hence, information can be extracted in a standardised, timely and replicable way through the algorithms and methods presented in this paper. This approach can be extended to new text corpora (different newspapers in other countries, social media content, corporate publications, professional research publications, etc.) and employed for new research questions, e.g., whether the findings also hold for other real estate asset classes and for different response variables such as price indices, or whether survey-based sentiment indicators perform better in explaining more volatile time series like real estate stock indices. With future research it may be possible to go beyond the limitations of this study, i.e., the sample of only one decade for only one country and only four newspapers. Additionally, the chosen dictionary-approach may be challenged by employing either different dictionaries or less supervised ML-algorithms for measuring sentiment within the corpora or for detecting suitable seed words.

3.8 Appendix

Appendix 3.1 – Unbounded Polarity Score

$$\delta = \frac{c'_{i,j}}{\sqrt{w_{i,jn}}} \tag{A3}$$

where:

$$c'_{i,j} = \sum((1 + w_{amp} + w_{deamp}) \cdot w^{p}_{i,j,k}(-1)^{2+w_{neg}}) \tag{A3.1}$$

$$w_{amp} = (w_b > 1) + \sum(w_{neg} \cdot (z \cdot w^{a}_{i,j,k})) \tag{A3.2}$$

$$w_{deamp} = max(w_{deamp'}, 1) \tag{A3.3}$$

$$w_{deamp'} = (w_b < 1) + \sum(z(-w_{neg} \cdot w^{a}_{i,j,k} + w^{d}_{i,j,k})) \tag{A3.4}$$

$$w_b = 1 + z_2 * w_{b'} \tag{A3.5}$$

$$w_{b'} = \sum(|w_{adversative\ conjunction}|, \ \ldots, w^{p}_{i,j,k}, w^{p}_{i,j,k}, \ \ldots, |w_{adversative\ conjunction}| \cdot -1) \tag{A3.6}$$

$$w_{neg} = \left(\sum w^{n}_{i,j,k}\right) mod\ 2 \tag{A3.7}$$

Table 3.7 - Seed Words selected by GloVe Model

Seed Words	
Residential	flat; existing flat; resident; two-family house; two-family house half; owner-occupied home; homeowner subsidy; owner-occupied flat; one-family house; one-bedroom flat; holiday flat; total living space; multi-storey building; household; multi-family house; rent cap; rent control; rental flat; new-build flat; new-build apartment; penthouse; townhouse; residential; senior housing; senior living; single household; social housing; student residence; student flat; housing estate; housing area; housing project; housing development; housing construction; housing entitlement certificate; housing stock; housing ownership; home ownership rate; housing unit; housing; residential area; type of housing; residential building; residential site; residential house; residential property; residential property market; residential complex; residential location; residential use; residential object; residential portfolio; residential project; residential quarter; residential space; residential support; residential development; residential unit; residential supply; residential construction; residential building society; residential property ownership; housing shortage; housing vacancy; housing lack; housing market; housing market report; housing rent; housing demand; new housing construction; housing need; housing users; housing policy; housing prices; housing sector; housing companies; housing industry; housing economy; housing district; housing value
Office	work space; work spaces; working space; working place; working places; working environment; work environment; working world; meeting rooms; office supply; office construction; office stock; office unit; office ensemble; office floor; office spaces; office space supply; office space stock; office space shortage; office space market; office space demand; office space turnover; office buildings; office building; office tower; office towers; office property; office property market; office complex; office occupancy; office vacancy; office suite; office market; office rents; office tenants; office demand; office new construction; office occupants; office use; office object; office project; office quarter; office space; office bar; office room; office sector; office top rent; office location; office rooftop; office rental; office centre; business centre; business park; coworking; coworking provider; coworking offices; coworking space; coworking spaces; single office; single offices; remote work; company headquarters; headquarters; large office; large offices; company offices; working from home; home office; conference rooms; conference offices; office; offices; conference room; teleworking; head office; event rooms; management office; work; working; cellular offices
Retail	anchor; anchor tenant; construction market; bookstore; centre management; centre manager; discounter; drugstore; ecommerce; shopping opportunity; shopping mall; shopping street; shopping area; retailer; retail trade; retail space; retail property; retail investment; retail market; retail object; retail project; retail location; retail turnover; retail park; franchise store; franchise store concept; retail turnover; flagship store; food court; frequented; frequency; catering space; stores; trade; commercial space; retail building; retail company; retail shop; retail shops; department store; department store chain; shop; shop unit; shop space; shops; shop spaces; food industry; food discounter; food retailing; food market; mall; market hall; local supplier; local supply; local supply centres; local supply centre; online retailer; online trade; online shop; outlet; footfall; products; restaurant; retail; retailer; shop; shopping; shopping centre; shopping centre operator; range; store; supermarket; supermarket chain; retail turnover rent; sales level; sales area; sales channel; warehouse; warehouse manager; warehouse chain

Table 3.8 - Descriptive Statistics of Sentiment Indicators

	Mean	Median	SD	Min	Max
NC_residential_trade	0.2159	0.2024	0.0522	0.1465	0.3279
NC_residential_daily	0.1990	0.1810	0.0600	0.1220	0.3575
NC_office_trade	0.2901	0.2885	0.0406	0.2104	0.3735
NC_office_daily	0.1795	0.1808	0.0203	0.1297	0.2135
NC_retail_trade	0.2112	0.2136	0.0232	0.1519	0.2603
NC_retail_daily	0.3415	0.3428	0.0689	0.2193	0.4661
NS_residential_trade	0.1202	0.1218	0.0213	0.0622	0.1554
NS_residential_daily	0.0984	0.0990	0.0195	0.0491	0.1379
NS_office_trade	0.1370	0.1371	0.0125	0.1028	0.1628
NS_office_daily	0.0260	0.0269	0.0115	-0.0008	0.0551
NS_retail_trade	0.1237	0.1231	0.0165	0.0778	0.1549
NS_retail_daily	0.0952	0.0957	0.0135	0.0668	0.1140

3.9 References

Antweiler, W. and Frank, M.Z. (2004), "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards", The Journal of Finance, Vol. 59 No. 3, pp. 1259–1294.

Antweiler, W. and Frank, M.Z. (2006), "Do US Stock Markets Typically Overreact to Corporate News Stories?", SSRN Electronic Journal.

Baum, A., Crosby, N. and MacGregor, B. (1996), "Price formation, mispricing and investment analysis in the property market", Journal of Property Valuation and Investment, Vol. 14 No. 1, pp. 36–49.

Beracha, E. and Downs, D.H. (2015), "Value and Momentum in Commercial Real Estate: A Market-Level Analysis", The Journal of Portfolio Management, Vol. 41 No. 6, pp. 48–61.

Beracha, E., Lang, M. and Hausler, J. (2019a), "On the Relationship between Market Sentiment and Commercial Real Estate Performance—A Textual Analysis Examination", Journal of Real Estate Research, Vol. 41 No. 4, pp. 605–638.

Beracha, E.i., Freybote, J. and Lin, Z. (2019b), "The Determinants of the Ex Ante Risk Premiumin Commercial Real Estate", Journal of Real Estate Research, Vol. 41 No. 3, pp. 411–442.

Black, F. (1986), "Noise", The Journal of Finance, Vol. 41 No. 3, pp. 528–543.

Blei, D.M., Andrew, Y.N. and Michael, I.J. (2003), "Latent dirichlet allocation", The Journal of Machine Learning Research, Vol. 3, pp. 993–1022.

Brooks, C. and Tsolacos, S. (2010), Real Estate Modelling and Forecasting, Cambridge University Press, Cambridge.

Carstens, R. and Freybote, J. (2021), "Can the Textual Tone in REIT Financial Statements Improve the Information Environment for Commercial Real Estate Investors? An Investigation", Journal of Real Estate Research, Vol. 43 No. 3, pp. 335–354.

Clayton, J., Ling, D.C. and Naranjo, A. (2009), "Commercial Real Estate Valuation: Fundamentals Versus Investor Sentiment", The Journal of Real Estate Finance and Economics, Vol. 38 No. 1, pp. 5–37.

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

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

Deng, X., Hrnjić, E. and Ong, S.E. (2014), "Investor sentiment and the SEO pricing process: Evidence from REITs", Journal of Real Estate Literature, Vol. 20 No. 2, pp. 85–109.

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

Evangelopoulos, N., Ashton, T., Winson-Geideman, K. and Roulac, S. (2015), “Latent Semantic Analysis and Real Estate Research: Methods and Applications”, Journal of Real Estate Literature, Vol. 23 No. 2, pp. 353–380.

Fama, E.F. (1991), “Time, Salary, and Incentive Payoffs in Labor Contracts”, Journal of Labor Economics, Vol. 9 No. 1, pp. 25–44.

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.

Fisher, J., Gatzlaff, D., Geltner, D. and Haurin, D. (2003), “Controlling for the Impact of Variable Liquidity in Commercial Real Estate Price Indices”, Real Estate Economics, Vol. 31 No. 2, pp. 269–303.

Franzoni, F., Nowak, E. and Phalippou, L. (2012), “Private Equity Performance and Liquidity Risk”, The Journal of Finance, Vol. 67 No. 6, pp. 2341–2373.

Freybote, J. and Seagraves, P. (2018), “The Impact of Investor Sentiment on Commercial Real Estate Market Liquidity”, Journal of Real Estate Research, Vol. 40 No. 4, pp. 597–628.

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

Gallimore, P. and Gray, A. (2002), “The role of investor sentiment in property investment decisions”, Journal of Property Research, Vol. 19 No. 2, pp. 111–120.

García, D. (2013), “Sentiment during Recessions”, The Journal of Finance, Vol. 68 No. 3, pp. 1267–1300.

Granger, C.W.J. (1969), “Investigating Causal Relations by Econometric Models and Cross-spectral Methods”, Econometrica, Vol. 37 No. 3, pp. 424–438.

Hausler, J. (2019), Essays on the influence of textual sentiment in real estate markets, Schriften zu Immobilienökonomie and Immobilienrecht, Band 96, Regensburg.

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

Heinig, S. and Nanda, A. (2018), “Measuring sentiment in real estate – a comparison study”, Journal of Property Investment & Finance, Vol. 36 No. 3, pp. 248–258.

Heinig, S., Nanda, A. and Tsolacos, S. (2020), “Which Sentiment Indicators Matter? Evidence from the European Commercial Real Estate Market”, Journal of Real Estate Research, Vol. 42 No. 4, pp. 499–530.

Henry, E. and Leone, A.J. (2016), “Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone”, The Accounting Review, Vol. 91 No. 1, pp. 153–178.

Jagarlamudi, J., Daumé, H. and Udupa, R. (2012), “Incorporating Lexical Priors into Topic Models”, Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 204–213.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li and Xiaotie Deng (2013), “Exploiting Topic based Twitter Sentiment for Stock Prediction”, in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 24–29.

Kaiser, C. and Freybote, J. (2021), “Is e-commerce an investment risk priced by retail real estate investors? An investigation”, Journal of Property Research, pp. 1–18.

Keynes, J.M. (Ed.) (1936), The General Theory of Employment, Interest, and Money, Palgrave Macmillan, Cham.

King, G., Lam, P. and Roberts, M.E. (2017), “Computer-Assisted Keyword and Document Set Discovery from Unstructured Text”, American Journal of Political Science, Vol. 61 No. 4, pp. 971–988.

Koelbl, M., Schuierer, R., Steininger, B.I. and Schäfers, W. (2020), “Is Mandatory Risk Reporting Informative? Evidence from US REITS Using Machine Learning for Text Analysis”, SSRN Electronic Journal.

Li, F. (2010), “The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach”, Journal of Accounting Research, Vol. 48 No. 5, pp. 1049–1102.

Lin, C.Y., Rahman, H. and Yung, K. (2009), “Investor Sentiment and REIT Returns”, The Journal of Real Estate Finance and Economics, Vol. 39 No. 4, pp. 450–471.

Ling, D.C., Naranjo, A. and Scheick, B. (2014), “Investor Sentiment, Limits to Arbitrage and Private Market Returns”, Real Estate Economics, Vol. 42 No. 3, pp. 531–577.

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

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

Loughran, T. and McDonald, B. (2020), “Textual Analysis in Finance”, Annual Review of Financial Economics, Vol. 12 No. 1, pp. 357–375.

Mai, D. (2021), "Economic Narratives and Market Outcomes: A Semi-supervised Topic Modeling Approach", SSRN Electronic Journal.

Marcato, G. and Nanda, A. (2016), "Information Content and Forecasting Ability of Sentiment Indicators: Case of Real Estate Market", Journal of Real Estate Research, Vol. 38 No. 2, pp. 165–204.

Newey, W.K. and West, K.D. (1987), "Hypothesis Testing with Efficient Method of Moments Estimation", International Economic Review, Vol. 28 No. 3, pp. 777–787.

Nguyen, T.H. and Shirai, K. (2015), "Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction", in Zong, C. and Strube, M. (Eds.), Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1354–1364.

Nordheim, G. von, Boczek, K. and Koppers, L. (2018), "Sourcing the Sources", Digital Journalism, Vol. 6 No. 7, pp. 807–828.

Pennington, J., Socher, R. and Manning, C. (2014), "Glove: Global Vectors for Word Representation", in Alessandro Moschitti, Q.C.R.I., Bo Pang, G. and Walter Daelemans, U.o.A. (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1532–1543.

Ploessl, F., Just, T. and Wehrheim, L. (2021), "Cyclicity of real estate-related trends: topic modelling and sentiment analysis on German real estate news", Journal of European Real Estate Research, ahead-of-print ahead-of-print.

Rinker, T. (2019), "Package 'sentimentr': calculate text polarity sentiment, R 3.4.0", Vol. 2019.

Ruscheinsky, J., Kandlbinder, K., Alexander, M., Dietzel, K. and Schaefers, W. (2018a), "Forecasting the German Real Estate Market with Dictionary-based Textual Analysis", in 25th Annual European Real Estate Society Conference, Reading, United Kingdom, European Real Estate Society.

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

Shiller, R.J. (2017), "Narrative Economics", American Economic Review, Vol. 107 No. 4, pp. 967–1004.

Shiller, R.J. (2020), Narrative economics: How stories go viral and drive major economic events, Princeton University Press.

Soo, C.K. (2018), "Quantifying Sentiment with News Media across Local Housing Markets", Review of Financial Studies, Vol. 31 No. 10, pp. 3689–3719.

Sprenger, T.O., Tumasjan, A., Sandner, P.G. and Welpe, I.M. (2014), "Tweets and Trades: the Information Content of Stock Microblogs", European Financial Management, Vol. 20 No. 5, pp. 926–957.

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

Tversky, A. and Kahneman, D. (1974), "Judgment under Uncertainty: Heuristics and Biases", Science (New York, N.Y.), Vol. 185 No. 4157, pp. 1124–1131.

Walker, C.B. (2014), "Housing booms and media coverage", Applied Economics, Vol. 46 No. 32, pp. 3954–3967.

Watanabe, K. and Zhou, Y. (2020), "Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches", Social Science Computer Review, Vol. 40 No. 2, 346-366.

Winson-Geideman, K. and Evangelopoulos, N. (2013), "Topics in Real Estate Research, 1973-2010: A Latent Semantic Analysis", Journal of Real Estate Literature, Vol. 21 No. 1, pp. 59–76.

4 Social Media and Real Estate: Do Twitter Users predict REIT Performance?

4.1 Abstract

This study investigates the impact of social media sentiment on indirect real estate market returns by utilizing a comprehensive natural language processing approach to identify relevant *Twitter* posts and extract sentiment from them. To handle the complex linguistic features inherent in social media messages, three different sentiment classifiers are compared. The findings suggest a significant relationship between monthly sentiment and REIT returns, which occurs in two phases: a short-term speculative reaction and a greater longer-term reaction related to actual changes in the real estate market. The study also highlights that while the conventional dictionary approach can identify this relationship, more sophisticated classifiers can achieve higher accuracy. Overall, the results demonstrate the valuable insights that can be gained from analyzing social media data and its potential impact on the real estate market.

Keywords: Social Media Sentiment, Textual Analysis, Real Estate Investment Trust, Machine Learning

4.2 Introduction

Social media has emerged as an increasingly important source of information in today's digital age. Since millions of users share their ideas, opinions and experiences online, the platforms have become essential for businesses and individuals to gain information and communicate with their audience (Yadav and Vishwakarma, 2020). Thanks to widespread Internet access, social media platforms have become incredibly convenient and accessible. As a result, users are immediately and always provided with multiple perspectives about any event worldwide. These technical advancements have fundamentally changed how people consume information. Even journalists increasingly rely on social media as a primary source for reporting (e.g. Broersma and Graham, 2013; Paulussen and Harder, 2014), highlighting its growing influence on today's society.

Consequently, social media provides insights into the opinions of those who actively share their thoughts online and reaches people who consume information directly on those platforms or indirectly via traditional sources. This wide reach of social media, in combination with a significantly growing exchange of stock market-related information on social media, has attracted the interest of researchers and business organizations in recent years. Sentiment analysis has found that social media's effect on stock markets is becoming more salient (Li et al., 2018) and utilized to predict stock market trends. In the finance literature, as early as 2011, Bollen et al. found evidence for public sentiment, obtained from social media, being a valuable predictor of stock market performance. Since then, many more researchers have found significant relationships between social media and stock markets (e.g. Sprenger et al., 2014; Xu and Cohen, 2018).

However, these findings have limited implications for real estate markets. Unlike the efficient stock markets, real estate markets are characterized by their heterogeneity, low divisibility, and high investment volumes, leading to segmented submarkets and long transaction periods. Therefore, the short-term social media analysis typically applied to stock markets has limited value for protracted real estate markets. Despite these limitations, there is little research on social media in the real estate literature and none which deals with it at a greater level, caused by the complexity of social media data. Two significant issues can describe this complexity: First, researchers must identify relevant tweets, which is difficult due to the lack of a consistent classification by topic of social media messages and the existence of non-user-generated content. Second, researchers must handle the linguistic complexity of social media messages, characterized by the use of informal language, shortened words and hashtags, and sarcasm and irony. These factors make it difficult to extract sentiment accurately, as traditional sentiment analysis techniques developed for news articles and corporate disclosures may not be suitable. Therefore, this study provides a comprehensive approach to social media analysis, which includes an adoptable selection process and the comparison of three different sentiment classifiers.

To simplify the selection of tweets in the first place, this paper investigates their relationship with real estate investment trusts (REIT), which can be traced using firm identifiers. As with previous social media sentiment literature, we expect social media can affect people’s behavior and, therefore, shape economic outcomes. Consequently, we believe social media helps predict future REIT returns, and that prediction accuracy can be increased using more sophisticated sentiment classifiers. In our study, we find that social media provides valuable insights into real estate market performance.

The results show that changes in social media sentiment cause a significant effect, which can be divided into two phases. The short-term reaction could be related to changes in the stock market and the more substantial long-term effect is probably related to actual real estate market changes. Overall, the observed lag structure is similar to the relationship between sentiment expressed in the media and real estate markets as Hausler et al. (2018) and Braun et al. (2020) observed. However, in line with our argumentation, the results indicate that social media sentiment is an even earlier indicator, which adds to our market understanding. This holds for all sentiment classifiers applied, even though we observe that more sophisticated approaches yield better classification performance and are better predictors.

In summary, similar to the findings of Li et al. (2018), we believe social media sentiment indicators can supplement traditional sentiment indicators. The accessibility, reach and timeliness of the information on social media make it a powerful tool for predicting real estate market trends. Being able to quantify the sentiment expressed on social media therefore provides valuable insights to stakeholders in the real estate industry. The following chapter presents an overview of the sentiment literature focusing on real estate and social media. It also introduces the complexity of social media data and derives the hypotheses. The subsequent chapters show the data utilized and explain the methodology. The sixth chapter presents the empirical analysis results, and the conclusion summarizes the study's most significant findings, their implications for practice, and provides suggestions for future research.

4.3 Literature Review and Hypothesis Development

4.3.1 Sentiment in Real Estate

The impact of sentiment on real estate markets has been the subject of empirical studies for more than a decade, with a growing body of evidence suggesting a correlation between sentiment and market performance. As far back as 1936, Keynes observed that a lack of information leads market participants to supplement the limited information available with their subjective perceptions, which he called the "animal spirit". Given that the real estate industry is frequently deemed inefficient owing to fragmented markets, property illiquidity, and a high information asymmetry (Gallimore and Gray, 2002), sentiment should have an influence on market performance. Many studies show a significant impact of sentiment on direct and indirect real estate markets. Some studies use traditional survey-based (direct) sentiment indicators (e.g. Clayton et al., 2009; Lin et al., 2009; Ling et al., 2014; Das et al., 2015; Freybote, 2016), while others use (indirect) sentiment indicators derived from market fundamentals (e.g. Marcato and Nanda, 2016; Freybote and Seagraves, 2017; Heinig et al., 2020).

With both types of sentiment indicators having limitations, a new category of sentiment measure has become increasingly widespread. In 2007 Tetlock pioneered textual sentiment analysis in the finance literature, deriving a sentiment measure from qualitative information from news articles in *The Wall Street Journal*. He used a dictionary approach, identifying positive and negative words using predefined word lists and determines the article's overall sentiment based on the ratio. Subsequently, Walker (2014) was among the first researchers to utilize textual sentiment analysis in the context of real estate housing markets, examining news media articles concerning the UK housing market. Walker's finding that media Granger-causes house price changes has been confirmed by Soo (2015) using a simple frequency-based method on news media articles and by Ruscheinsky et al. (2018a), who created a domain-specific dictionary for the German market. Nowak and Smith (2017) took it a step further by incorporating text data from *MLS property listings* into a hedonic pricing model for house price valuation, which significantly enhanced the accuracy of their pricing model.

Textual analysis has been applied to housing markets and to commercial real estate markets. Beracha et al. (2019) demonstrated that sentiment within news abstracts of *The Wall Street Journal* can forecast the performance of the US private commercial real estate market. Using a more advanced machine learning technique to measure sentiment, Hausler et al. (2018) revealed that sentiment extracted from professional financial news leads the performance of the US securitized and commercial real estate markets. While previous studies focused on direct real estate markets, Ruscheinsky et al. (2018b) turned to indirect real estate markets. Using news media headlines, they found a significant leading relationship between sentiment and future REIT market movements. However, compared to direct real estate markets, indirect real estate markets with corporate disclosure offer an additional potential source for textual analysis. Therefore, Carstens and Freybote (2019) investigated the response of institutional REIT investors to abnormally positive sentiment in REIT financial statements. In a later study, Carstens and Freybote (2021) used those statements to predict commercial real estate total returns in the next quarter. Paulus et al. (2022) extended the research on textual sentiment by utilizing IPO filings to predict the underpricing of REIT IPOs. Alternatively, Price et al. (2017) investigated the vocal cues of CEOs during quarterly REIT conference calls to predict investors' reactions.

To summarize, previous studies indicate that real estate markets are influenced by human behavior and that textual analysis is a valuable tool for measuring sentiment. Nevertheless, despite analyzing different textual sources, such as newspapers and corporate disclosure, there is a scarcity of research on the role of social media sentiment in real estate markets. There are only a few studies that look at its impact, all of them using direct housing real estate markets. Zamani and Schwartz (2017) collected 131 million tweets from US counties based on self-reported locations to improve the prediction of foreclosure rates and price increases compared to traditional socioeconomic factors. Similarly, Hannum et al. (2019) explored the connection between microblogging sentiment and property prices in Istanbul by analyzing two million tweets marked with either specific geotags or spatial lexical terms. They discovered a negative correlation between *Twitter* sentiment and property prices, including price appreciation. Recently, Tan and Guan (2021) investigated nine million tweets from Manhattan, New York City, and found that high sentiment levels aligned with high-frequency spatial clusters and higher housing prices. These studies highlight the predictive power of sentiment expressed in social media on local housing markets on a long-term perspective. However, they are limited in scope, as they focus solely on local markets and have yet to be applied to larger markets or explored broader market sentiment.

4.3.2 Social Media Sentiment in Finance

While the real estate literature has yet to explore the impact of social media extensively, there is a rapidly growing body of research in finance dedicated to understanding its influence. Social media sentiment analysis has emerged as a valuable source due to its real-time availability and increasing message volume related to stock markets. Studies have analyzed various online sources, including *Twitter*, *StockTwits*, and other message boards, with Antweiler and Frank (2004) being among the first to do so. Their study utilized a Naive Bayes classifier and a Support Vector Machines (SVM) to analyze over 1.5 million message board postings on *Yahoo! Finance* and *Raging Bull* of 45 companies, revealing a link between disagreement in sentiment and increased trading volume. Additionally, the study found predictive power of their sentiment measure on next-day returns and volatility.

These early findings align with Bollen et al. (2011), who investigated whether collective mood states derived from large-scale *Twitter* feeds correlate with the value of the *Dow Jones Industrial Average*. They used two mood tracking tools to measure positive vs. negative mood and mood in six dimensions. The results indicate that specific public mood dimensions can significantly improve the accuracy of their predictions. Chen et al. (2011) found evidence for a relationship between social media sentiment and stock markets and compared the effect with that of news media. Comparing *The Wall Street Journal* and *SeekingAlpha*, they state that there is a significant relationship in both cases, but social media has an even stronger and longer-lasting effect on the stock market.

Sprenger et al. (2014) diverged from previous research that concentrated on investigating the connection between social media sentiment and market indices, as they examine the correlation to individual stocks. Using a dictionary-based sentiment classifier, they found relationships between the bullishness in tweet sentiment and stock returns, message volume and trading volume, and disagreement and volatility. Especially, their findings about bullishness in social media have since been revisited. Li et al. (2018), for example, used a more sophisticated SVM and found that the bullishness in stock-related microblog messages had an impact on contemporaneous abnormal returns. Besides, they confirm that message volume significantly correlates with trading volume and volatility. Overall, the literature suggests that there are significant short-term relationships between stock markets and social media sentiment.

In recent years, there has been a surge of research aimed at improving the handling of complex linguistic features in social media messages to enhance classification and prediction accuracy. This trend is attributable to the increased computing capacity now available, which enables the application of more sophisticated classifiers to larger datasets, ultimately resulting in better prediction accuracy. Therefore, Yildirim et al. (2019) compared deep learning and machine learning methods based on their classification performance of *StockTwits* messages and found that deep learning approaches outperform classic machine learning. These findings are supported by Shi et al. (2021) and

Wang et al. (2020), which imply that different deep learning models outperformed other tested classical machine learning models. A recent study by Gupta and Chen (2020) examined the influence of financial-related social media messages on major companies' stock prices. They were able to improve prediction accuracy when including sentiment data. In conclusion, the literature shows that sentiment expressed on different social media platforms can be used to predict stock returns, volatility, and trading volume. Besides, the results suggest that deep learning models outperform classical machine learning models in analyzing social media sentiment.

4.3.3 Complexity of Social Media Data

Substantial evidence suggests the utility of social media sentiment in comprehending financial markets. However, the limited literature on real estate sentiment may be due to the multiple aspects of deriving sentiment from social media. Maynard et al. (2012) highlight significant challenges accompanying social media-based texts, which include issues of relevance, target identification, negations, contextual information, and volatility over time.

The issue of relevance arises when attempting to extract sentiments regarding a specific topic from social media texts. The corpus provided by social media platforms is enormous. However not all of the messages are relevant for the investigation. Therefore, selecting the right messages is crucial since false selection will produce no or inaccurate results. Closely related to relevance is the issue of target identification. Ensuring that the sentiment is expressed about the analyzed topic can be difficult, especially when the topic is discussed in a broader context. This challenge requires the development of algorithms and tools that can identify the specific target topic of interest and exclude irrelevant discussions within the previously selected corpus. Besides those message selection issues, negations pose a significant challenge in textual analysis, particularly in dictionary-based sentiment analysis.

In addition, social media messages are characterized by opinions expressed through irony, sarcasm, and other semantic feature, making it even more difficult for conventional sentiment analysis to classify messages accurately. Hence, contextual information is required to understand those linguistic features and give a correct prediction for the expressed opinion. And not least, social media sentiment is exposed to strong temporal dynamics, resulting in volatility over time. Opinions expressed on a social media platform like *Twitter* can change radically from positive to negative and vice versa in a short period, which is why the time dimension is crucial for further analysis.

4.3.4 Hypothesis

Following Maynard et al. (2012), who emphasize the importance of relevance and target identification in social media analysis, we have chosen to focus on indirect real estate markets as they present a more manageable task of identifying relevant social media messages. Companies are often tagged using the company name or firm identifiers, which enables us to retrieve those messages. Besides, we chose to investigate the long-term relationship between social media sentiment and real estate markets rather than short-term relationships, due to the perception that real estate markets are inefficient and slow moving compared to stock markets.

This approach provides three additional benefits, such as comparing results to the abovementioned real estate studies, which typically examine monthly or quarterly relationships. Second, it aligns with Shiller's (2017, 2020) narrative economics, which suggests that beliefs and expectations about the future take time to form. Accordingly, the behavior of social media users or investors, even those who only monitor or indirectly consume social media via news media, could be influenced if the change in sentiment persists for an extended period. Third, investigating longer periods, might reduce the high volatility issue in daily sentiment (Maynard et al. 2012). Despite that, we assume social media can be a valuable sentiment indicator since it more directly represents people's opinion compared to traditional news sources. Therefore, our first hypothesis is:

Hypothesis 1: *The sentiment expressed in social media affects future REIT market returns.*

In light of Maynard et al.'s (2012) emphasis on the linguistic features inherent in social media messages and the aim to establish a baseline for future research, this study advocates for the application of more advanced techniques that can handle the challenges presented. Accordingly, the performance of different sentiment classification methods is evaluated. Consistent with prior literature, this study encompasses the full spectrum of classifiers, encompassing a dictionary-based, a machine learning and a deep learning approach. Hence, the second hypothesis can be formulated as:

Hypothesis 2: *The more sophisticated a classifier the more accurate is its classification performance.*

As the accurate classification of tweets provides a foundation for the calculation of sentiment indices, it is essential to achieve a high level of accuracy in sentiment classification for reliable prediction of market performance. Since the sentiment indices are expected to reflect changes in the underlying market performance, a higher classification accuracy will result in more accurate predictions of market trends. Therefore, we hypothesize that higher classification accuracy will lead to better predictions of market performance. We, therefore, expect the following:

Hypothesis 3: *The more sophisticated a sentiment classifier, the better its predictions of future REIT market returns.*

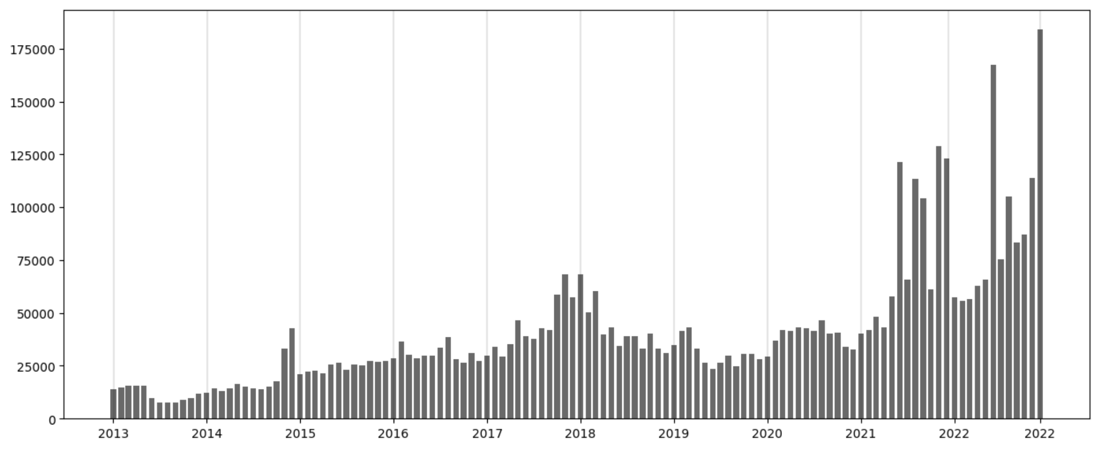
4.4 Data

4.4.1 Social Media Data

The corpus of this study consists of 4.97 million *Twitter* messages spanning a ten-year horizon ending in December 2022. Due to its popularity, we decided to use *Twitter* as our source as it provides a large and diverse sample of data for sentiment analysis. Besides *Twitter* has proven to be a valuable source of sentiment analysis in the finance literature, particularly regarding investor sentiment indicators. Therefore, the tweets were obtained from the *Twitter Developer API* using the name or cashtag (i.e., a "$" sign followed by the ticker symbol) of REITs. Given that we seek to compare sentiment indices with a representative performance index, the sample is restricted to tweets about REITs that have been listed in the performance index for the corresponding year. This approach ensures that the data used for comparison is both relevant and comparable, allowing for a robust examination of the relationship between sentiment and performance.

The monthly distribution of tweets, as depicted in Figure 4.1, reveals a steady rise in the number of messages from 2013 to 2020, followed by exponential growth until the end. This trend is not aligned with *Twitter*'s monthly user base but rather with the onset of the global COVID-19 pandemic, indicating increased interest in REITs since then.

Figure 4.1 - Unfiltered Tweet Distribution over Time



4.4.2 Dependent Variable

We used the US REIT market to investigate the relationship between social media sentiment and indirect real estate markets. Being the largest REIT market in the world diminishes the impact of outliers on overall performance, which makes it more suitable for analysis. The *FTSE/NAREIT All REIT Total Return Index* indicates the market performance (NAREIT, 2022). The *National Association of Real Estate Investment Trusts* (NAREIT, www.reit.com) provides the index, which aggregates the total returns of all US REITs. It represents all REIT types and includes price returns and income returns from dividends. Descriptive statistics of the index can be found in Table 4.1.

Table 4.1 - Descriptive Statistics

	Mean	Median	SD	Min	Max
Total Return	0.0067	0.0112	0.0479	-0.2085	0.1142
Spread	0.0091	0.0089	0.0022	0.0055	0.0170
Term	0.0136	0.0141	0.0082	-0.0055	0.0296
Inflation	0.0248	0.0176	0.0227	-0.0020	0.0906

Notes: Monthly total return as provided by *NAREIT*. Macroeconomic control variables are the spread between Moody's Seasoned Baa- and Aaa-rated corporate Bonds (SPREAD), the difference between the 10-year Treasury bond yield and the 3-month Treasury bill (TERM) and inflation rate of the US (INFLATION).

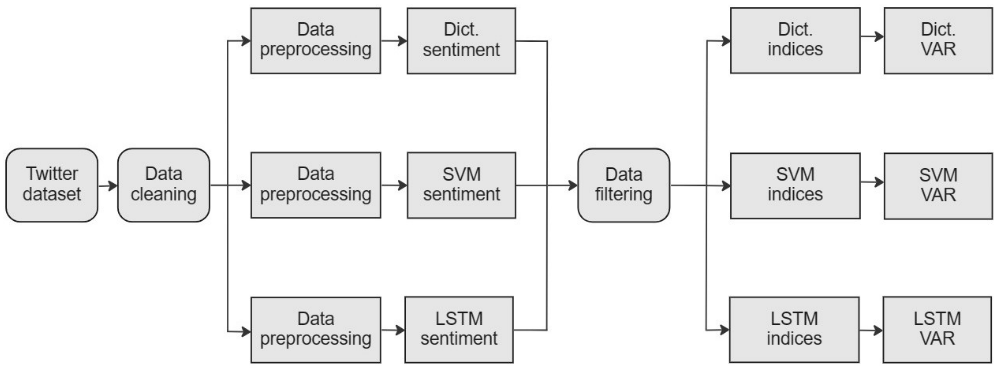
4.4.3 Economic Control Variables

When performing regression analysis for real estate returns, it is crucial to consider other factors that may impact, such as the overall economy and capital markets (Brooks and Tsolacos, 2010). In academic literature, various exogenous variables have been employed to explicate the performance of indirect real estate. For this study, we use the most commonly utilized variables. As a metric to represent the overall economic default risk, this study includes the spread between *Moody*'s Seasoned Baa- and Aaa-rated corporate Bonds (*SPREAD*), as established in the literature by Lin et al. (2009), Ling et al. (2014), Freybote and Seagraves (2017), and Hausler et al. (2018). To account for expectations regarding future economic development, a macroeconomic control in the form of the difference between the 10-year Treasury bond yield and the three-month Treasury bill (*TERM*) is included, as previously utilized by Clayton et al. (2009), Freybote and Seagraves (2017), and Hausler et al. (2018). Lastly, following Hausler et al. (2018), the regression accounts for the inflation rate of the US (*INFLATION*) as real estate is commonly regarded as a possible hedge against the change in the consumer price index (Hoesli et al., 2008). Summary statistics of the described variables can be obtained from Table 4.1.

4.5 Methodology

Examining the relationship between *Twitter* messages and REIT market performance is a process that involves various techniques of natural language processing. This complex process is illustrated in Figure 4.2 and can be divided into three phases. The first phase involves the selection of tweets and begins with the retrieval of messages from *Twitter*. Those messages are then filtered to obtain a corpus of relevant tweets. During this phase, the textual corpus undergoes pre-processing. In the second phase, the filtered tweets are assigned to one of three sentiment categories (positive, negative, or neutral) by applying three classification techniques. Once all of the messages have been classified, monthly and quarterly sentiment indices are generated and, in the final phase, integrated into econometric models to evaluate their relationship with REIT market performance.

Figure 4.2 - NLP Process



4.5.1 Tweet Selection and Data Cleaning

In light of Maynard et al. (2012) research on the importance of relevance and target identification in social media sentiment analysis, this study places substantial emphasis on the selection of tweets about all US REITs. To this end, a three-step procedure is employed for tweet selection: First, English tweets containing the name or cashtag (i.e., a "$" sign followed by the ticker symbol) of a REIT are downloaded. Second, tweets without interactions, such as likes, comments, or retweets, as well as duplicates are removed from the dataset (compare Sprenger et al., 2014). Third, drawing on the methodology proposed by Zhou et al. (2017), a word embeddings approach is utilized to categorize tweets and eliminate irrelevant or non-pertinent topics.

Upon completing the first step and examining the data, it is evident that a substantial proportion of the tweets downloaded are irrelevant to the research question (e.g., tweets in other languages or on unrelated topics) and that further selection is required. Tweets that only include the ticker symbol or appear to be generated by automated systems and typically do not have any interactions were removed from the dataset, as were all duplicates. This decision is in line with the argumentation about relevance put forth by Maynard et al. (2012), that tweets without interactions are likely to be irrelevant to the target audience, and the findings of Sprenger et al. (2014), demonstrating that users providing above average investment advice are retweeted more often amplifying their share of voice. A closer examination of the distribution of tweets among REITs then reveals that some REITs appear to have higher levels of public interest at first glance. Upon further screening of these tweets, it becomes evident that some of them share ticker symbols with other financial products, such as cryptocurrencies, highlighting the importance of further data cleaning and the third step of our procedure.

Aiming to identify those tweets that are really about REITs, we, similar to Zhou et al. (2017), employ a combination of two techniques for word embedding and similarity measurement, namely the Global Vector (GloVe) model and cosine similarity. These algorithms allow tweets to be categorized by topic through signal words and then create a sentiment indicator only for “REIT” related tweets. This automated procedure aligns with the findings of Watanabe and Zhou (2020), who determined that signal words should possess both knowledge-based and frequency-based properties. The GloVe model, introduced by Pennington et al. (2014), is a word embedding technique that generates a vector representation for each word in the corpus, capturing the semantic and syntactic properties of the words. To generate these, the algorithm combines the advantages of global matrix factorization and local context window methods.[19] The training of the algorithm is carried out on only non-zero elements in a global word-word co-occurrence matrix X, rather than on the entire sparse vector space or the local context windows within a large text corpus, and therefore generates a matrix of meaningful substructures. The objective function J of the weighted least squares regression model is specified as follows:

$$J = \sum_{i,j=1}^{V} f(X_{ij})(w_i^T \widetilde{w}_j + b_i + \tilde{b}_j - log X_{ij})^2 \tag{4.1}$$

where V is the size of vocabulary; X_{ji} lists the number of times word j occurs in the context of word i; w ($\widetilde{w}$) represents the word vector for a main (context) word and b ($\tilde{b}$) describes the scalar biases for main (context) words; w^T indicates a transposed word vector. In order to prevent the learning of simple common word pairs, a weighting function readjusts the

[19] Matrix factorization techniques are used to decompose large matrices that contain statistical information about a text corpus, resulting in low-dimensional word representations and reduced computation time. On the other hand, context window methods use linear relationships between word vectors based on local context windows to predict linguistic patterns and improve word analogy tasks.

cost for each word pair because word pairs exhibit different occurrence frequencies in the text corpus. This weight is either set to 1 if the co-occurrence count exceeds a certain threshold or based on the co-occurrence count (see Pennington et al. (2014) for additional information).

The matrix obtained from the GloVe model consists of word vectors that identify relationships between words in the corpus. To measure the similarity between these vectors, we apply the cosine similarity, a nearest neighbor approach, which calculates the cosine of the angle between two vectors. This method identifies synonyms or words used in similar contexts. Cosine similarity is computed as follows:

$$Cosine\ Similarity\ (v_i, v_j) = \frac{v_i \cdot v_j}{\|v_i\| \, \|v_j\|} \tag{4.2}$$

where $v_i \cdot v_j$ is the dot product of two GloVe vectors, $\| v_i\|$ is the length of vector v_i, and $\| v_j\|$ is the length of vector v_j. Cosine similarity is geometrically the cosine of the angle between the two vectors, normalized by the vector length. Since word vectors are non-negative, the measure is bounded between 0 and 1. Besides, a higher value indicates that the words are more semantically similar and tend to occur in the same context.

To create a meaningful dataset of tweets, we filter for tweets that include at least one of the 25 words best describing "REIT".[20] Hence, v_i represents the global vector of "REIT" and v_j is iterated through for all other words, with the 25 words yielding the highest cosine similarity selected as the signal words. Using signal words to filter tweets, instead of applying an entire topic model, as performed by Zhou et al. (2017), is a more efficient and targeted approach. Using signal words allows for a focused identification of relevant tweets, instead of to the broad topic identification provided by a topic model such as the Latent Dirichlet Allocation (LDA) model. Furthermore, using signal words can reduce the computational complexity associated with topic models, making it a more practical solution for large-scale text analysis tasks. Nonetheless, future studies may seek to apply topic modelling for this purpose.

It is essential to mention that pre-processing of tweets was carried out between steps two and three to enhance the accuracy and efficiency of the classification and the upcoming sentiment analysis. This involved lowercasing, removal of punctuation, numbers, non-alphabetic and special characters and stop words. Contractions were expanded, tweets were tokenized, and words ultimately lemmatized. Contractions can cause noise in the

[20] It was determined that filtering for the 25 closest words was optimal as a significant decline in the quality of the word representations occurred beyond this threshold.

analysis since word frequencies are distorted, but their expansion resolves this issue (Dorle and Pise, 2018). Tokenization involves dividing the text into meaningful units such as phrases, n-grams, or, in this case, words. Ultimately, lemmatization (Manning et al., 2008), the conversion of words to their root form, was preferred over stemming as it is a more advanced method.

Besides, the literature suggests additional pre-processing for social media data (e.g., Dorle and Pise, 2018; Ankit and Saleena, 2018; Bahrainian and Dengel, 2013), of which the most commonly used steps were incorporated. These included the removal of @ mentions, the # symbol, URLs, retweets, non-ASCII letters and symbols. The slang commonly used on social media platforms was addressed with by replacing slang terms with their actual phrasal equivalents (Bahrainian and Dengel, 2013). This conversion reduced the vocabulary size, thereby improving performance for dictionary-based approaches, as actual words are more likely to be in the dictionary than slang terms. The Internet & Text Slang Dictionary from www.noslang.com, which contains more than 5000 slang terms and their formal meaning, was provided for converting the slang (Jones, 2022).

4.5.2 Sentiment Analysis

After conducting pre-processing and selecting tweets, sentiment analysis is performed to determine the sentiment for each of the remaining tweets. This involves classifying tweets as positive, negative or neutral and then aggregating those individual scores within specified time frames to derive a sentiment index.

Classification with a Dictionary Approach

In this study, the dictionary approach is employed as the first algorithm and serves as the baseline for evaluating the performance of subsequent methods. The dictionary approach was selected due to its widespread use as the most fundamental and commonly used method for sentiment analysis (e.g., Beracha et al., 2019; Carstens and Freybote, 2019; Paulus et al., 2022). It is a simple yet effective method brought into finance by Tetlock (2007). A predefined list of words, referred to as the sentiment dictionary, is used to assign sentiment polarity to text. Since the outcome of sentiment analysis heavily depends on the dictionary used and, given that domain-specific idiosyncratic word meanings exist, the use of domain-specific dictionaries is recommended. Doran et al. (2012) and Henry and Leone (2016) have shown that domain-specific dictionaries help to account for these nuances and enhance the accuracy of sentiment analysis. Therefore, we apply the real estate-specific dictionary developed by Ruscheinsky et al. (2018b), which is based on the finance-related dictionary by Loughran and McDonald (2011) and has been adjusted for a real estate context.

Further, the dictionary-based approach is also known as a bag-of-words approach because it disregards sentence structure. This simplicity does not make it an adequate baseline for comparison with more advanced algorithms as it is the only one not capturing critical contextual information in sentiment analysis of social media content, as Maynard et al. (2012) described. To address this limitation, valence shifters, such as negators (e.g., not), amplifiers (e.g., very), de-amplifiers (e.g., less), or adversative conjunctions (e.g., but), are controlled for. In doing so, a context cluster is defined with *n* words before and after each polarized word. Within this context cluster valence shifters are identified using an additional dictionary containing 76 valence shifters. Once a valence shifter has been identified, it affects the score of the polarized context cluster by a certain weight (see Rinker (2019) for a detailed description).

Classification with a Support Vector Machines Model

Since machine-learning approaches tend to yield higher classification accuracy than dictionary-based approaches (Li, 2010), we include an SVM approach. The SVM approach was selected due to its effectiveness in handling high-dimensional data and ability to learn complex decision boundaries. Besides, in real estate literature, it is among the commonly used machine learning approaches for textual sentiment classification (e.g., Hausler et al., 2018; Antweiler and Frank, 2004). The SVM, as developed by Cortes and Vapnik (1995), involves training a model using a labeled dataset[21] to predict sentiment polarity. Therefore, it is trained to map tweets, in the form of input vectors, into a high-dimensional space via a non-linear mapping technique chosen a priori. This vector space is subdivided by a hyperplane, which subsequently allows identifying the category of tweets depending on their position in the feature space relative to the surface. However, since SVMs are binary classifiers and cannot classify into three categories, Platt scaling (Platt, 1999) is applied to get a probabilistic output, with *P* interpreted as the probability of a text being positive, we classify the tweets in three sentiments:

$$P \geq 0.667 \rightarrow \text{positive sentiment}$$
$$0.333 < P < 0.667 \rightarrow \text{neutral sentiment}$$
$$P \leq 0.333 \rightarrow \text{negative sentiment}$$

A detailed description of the SVM model can be found in Appendix 4.1.

[21] The training dataset utilized in this study is sentiment140 and is classified by Go et al.2009 using distant supervision. This balanced dataset contains 1.6 million tweets labeled as either positive or negative.

Classification with a Long-Short-Term-Memory Model

The third classification model chosen for this study, is a Long Short-Term Memory (LSTM) model. LSTMs are modified versions of Recurrent Neural Networks (RNN) introduced by Hochreiter and Schmidhuber (1997) and are deep-learning classifiers. They are well suited for sentiment analysis due to their ability to model the contextual dependencies within textual data effectively. The internal memory state of the model allows it to capture both short- and long-term dependencies in the text, improving its ability to accurately predict sentiment. LSTMs have been widely applied in sentiment analysis, demonstrating superior performance compared to traditional models (e.g. Fu et al., 2018; Jin et al., 2020; Wu et al., 2022). With ever more (social media) data available for training[22], they provide better scalability, which further deems them superior to traditional indicators.

A Neural Network is a multi-layer computational model composed of nodes organized into multiple layers. It comprises an input layer, several hidden layers and an output layer (Sarker, 2021). The input is received as a document feature matrix at the input layer, then transformed through the applying weights and biases, after then passed to the first hidden layer. In each hidden layer, the transformed values are further transformed by an activation function. This process is repeated based on the number of hidden layers and nodes until the final transformed values are sent to the output layer. Ultimately, the values from the final hidden layer are aggregated to produce the final output.[23] In RNNs, the output values of a cell can also be added back to the inputs of the current layer. This is a major feature, called a feedback loop, that distinguishes them from other neuronal networks.[24] Additional information on the LSTM model applied in this study is presented in Appendix 4.2.

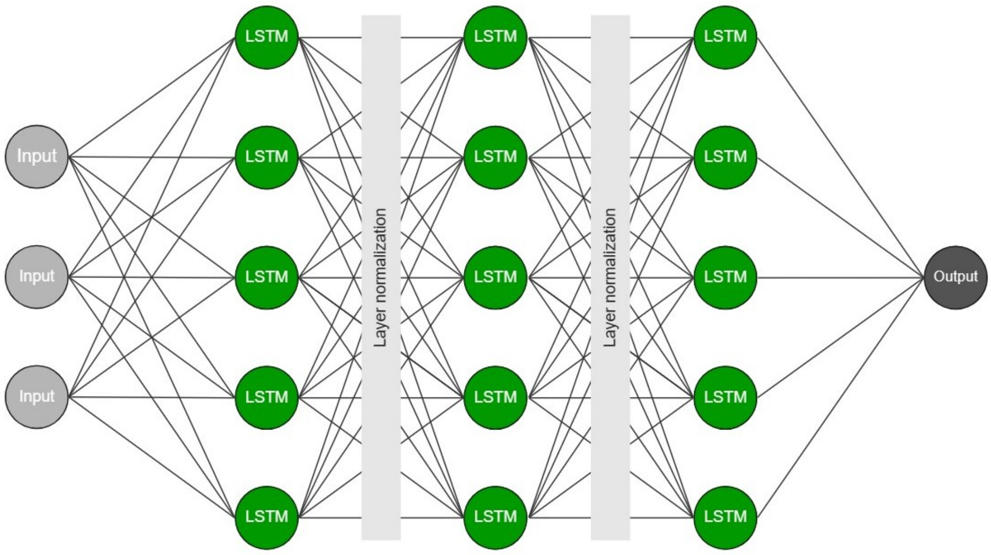
The LSTM network used in this study is illustrated in a simplified version in Figure 4.3. The model consists of an input layer, three hidden LSTM layers with 256 nodes each and an output layer. Additionally, layer normalization is included between each LSTM layer, improving the efficiency of a model (Ba et al., 2016).

[22] The same training data set and pre-processing steps as for the SVM have been applied.

[23] Since the output is again probabilistic, with *P* indicating the probability of an input having positive sentiment, the same classification ranges as for the SVM have been applied.

[24] Due to the feedback loop, RNNs encounter the challenge of vanishing or exploding gradients. As the weights are applied repeatedly to the values in the feedback loop, the magnitude of the input values either decreases (weights < 1) or increases (weights > 1) with every iteration, potentially leading to the input values either dropping to zero or increasing to infinity after a sufficient number of iterations. Vanishing or exploding gradients, however, impede the ability of the network to converge toward optimal weights and biases (Hochreiter, 1991).

Figure 4.3 - LSTM Model



Notes: The LSTM model consists of an input layer, three hidden LSTM layers with 256 nodes each, and an output layer. Between each layer, layer normalization is included.

Model Accuracy

Five independent models are trained for both machine learning approaches. For this process, the training data is randomly split into a training set, containing 80% of the data and a validation or testing set, containing the remaining 20%.[25] The average output over these five independent models will then be used for the final classification. Training five independent models with unique training and test or validation splits improves the robustness of the classification results. After implementing the three sentiment classifiers, their accuracy can be compared by assessing the proportion of correctly classified tweets from the validation or testing split. It should be noted that due to the lack of neutral classification in the training dataset, the accuracy of the classifiers is measured only for a binary classification. In the case of the dictionary approach, this implies that only tweets containing at least one polarized word are considered. For the other two approaches, a text is classified as negative if the probability of this text being positive (P) is below 0.5 and as positive if it is above 0.5.

Index construction

After tweets are classified by all three models, the positive, negative, and neutral tweets are aggregated into sentiment indices. As we seek to understand the long-term relationship and face temporal dynamics in sentiment, we decided to consider different

[25] For the SVM only 25% of the training dataset is used for training and testing, due to computational limitations.

time horizons in the analysis. As such, we have chosen to employ monthly indices for our main analysis and quarterly indices to check the robustness of the results. The first measure following Antweiler and Frank (2004), is called the optimism indicator (*OI*) and shows the level of optimism expressed on social media:

$$OI_t = \frac{\sum_1^I \text{ positive messages }_{i,t}}{\sum \text{ total number of messages }_t} \tag{4.3}$$

The second measure is called the pessimism indicator (*PI*). It is based on Rozin and Royzman (2001), who found that individuals are affected more strongly by negative rather than positive influences and therefore measures the ratio of negative messages:

$$PI_t = \frac{\sum_1^J \text{ negative messages }_{j,t}}{\sum \text{ total number of messages }_t} \tag{4.4}$$

The third measure is the Sentiment Quotient (*SQ*), which is the most used in literature and provides an indicator of the number of positive messages compared to positive and negative messages:

$$SQ_t = \frac{\sum_1^I \text{ positive messages }_{i,t}}{\sum_1^I \text{ positive messages }_{i,t} + \sum_1^J \text{ negative messages }_{j,t}} \tag{4.5}$$

with i as a positive message, j as a negative message, and t states the aggregated time frame.

4.5.3 Econometric Approaches

Vector Autoregression

A Vector Autoregression (VAR) framework is applied to examine the intertemporal relationship between social media sentiment regarding REITs and the performance of the corresponding REIT market. The VAR model offers an advantage in that it does not make a priori assumptions about causal relationships and allows estimating the intertemporal relationship between sentiment and performance while controlling for possible endogeneity. The model specifies each variable as a linear function of its own lags and the lags of other variables, enabling control for momentum in the dependent variable (Beracha and Downs, 2015). In general, the conventional bivariate VAR model is specified as follows:

$$y_{1t} = \beta_{10} + \sum_{i=1}^{l} \beta_{1i}\, y_{1t-i} + \sum_{i=1}^{l} \alpha_{1i}\, y_{2t-i} + u_{1t} \tag{4.6}$$

$$y_{2t} = \beta_{20} + \sum_{i=1}^{l} \beta_{2i}\, y_{2t-i} + \sum_{i=1}^{l} \alpha_{2i}\, y_{1t-i} + u_{2t} \tag{4.7}$$

where *l* denotes the number of lags and u_{it} a white noise disturbance term with $E(u_{it}) = 0$, (i = 1,2), $E(u_{1t}, u_{2t}) = 0$. However, following economic theory, exogenous variables should be incorporated into the VAR framework. In matrix notation, where *X* denotes a matrix of exogenous variables and *B* a matrix of coefficients, the model can be written as follows:

$$y_t = A_0 + \sum_{i=1}^{l} A_i\, y_{t-i} + BX_t + u_{1t} \tag{4.8}$$

In the VAR framework, the stationarity of its components is tested using the Augmented Dickey-Fuller (ADF) Test. Whenever the null hypothesis of unit root presence is rejected, it indicates the need for differencing the variables to ensure statistical validity. Additionally, a suite of diagnostic tests is conducted to confirm the robustness of the results. The Breusch-Godfrey Lagrange Multiplier test is employed to verify that the residuals are not serially correlated. Furthermore, normality and heteroscedasticity tests are performed to guarantee the statistical validity of the results. Newey and West (1987) standard errors, which are robust to heteroscedasticity and autocorrelation (HAC), are applied whenever necessary.

Determining the optimal lag length is a crucial step in configuring the VAR. The optimal lag length is optimized using the dictionary approach, and the results applied to the other models to ensure comparability among the various models. This is selected using the Akaike Information Criterion (AIC) as selection criterion and choosing the lag length that minimizes its value (Brooks and Tsolacos, 2010). The optimal lag length for the dictionary approach with three sentiment indicators (*OI*, *PI*, *SQ*) is determined by calculating the AIC for each indicator and ranking their values. The lag with the lowest mean rank is ultimately selected. Hereby, the minimum lag length is restricted according to recent findings: As Hausler et al. (2018) and Braun et al. (2020) found a significant impact of sentiment on real estate markets of up to eight months, this will serve as our minimum lag length in the monthly analysis.

Granger Causality and Impulse Response Function

To gain a deeper understanding of the causal relationship and the impact of social media sentiment on market performance, Granger causality will be assessed and Impulse Response Functions (IRFs) constructed.

Granger causality is a concept used in time series analysis to determine the causal relationship between two variables (Granger, 1969). The idea behind Granger causality is that if a variable X is said to Granger-cause a variable Y, it means that the past values of X provide information that helps predict future Y values beyond what can be predicted from the past values of Y alone. In other words, X is a useful predictor of Y. Thus, the use of Granger causality aids in comprehending the temporal relationships between social media sentiment and market performance, ultimately determining if sentiment holds predictive power and vice versa.

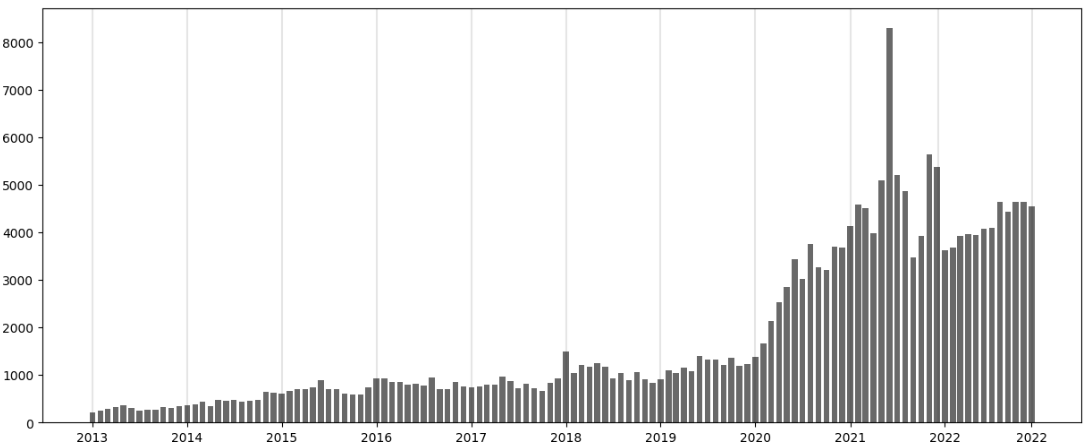
Impulse Response Functions (IRFs) are used to measure the impact of a sudden, short-lived shock to a particular variable (i.e., sentiment) on other variables (i.e., performance) within the system. Therefore, an exogenous event in the form of a one standard deviation shock in the sentiment variable is simulated and, its effect, i.e., the dynamic response of the total return, during the analyzed lags, is graphically represented. This provides further insights into the transmission mechanisms of the interdependent variables and the speed at which the shock decays over time.

4.6 Results

4.6.1 Tweet Selection

Initially, the corpus of tweets consisted of more than 4.97 million tweets, of which a significant proportion was deemed irrelevant to the study. After removing duplicates and tweets with no interaction, the dataset was reduced to approximately 650,000 tweets and used to train the GloVe model. Finally, the corpus was pruned to 208,332 tweets, utilizing the nearest neighbors of "REIT" as signal words to ensure that the remaining messages

Figure 4.4 - Filtered Tweet Distribution over Time



exclusively pertained to REITs.[26] Figure 4.4 shows the distribution of the final selection during the observation period, which is similar to the distribution of all tweets as shown in Figure 4.1. The number of tweets steadily increase until 2020, with a significant surge thereafter. This suggests that our selection process is not biased and selects a constant percentage of available tweets over time, which indicates this approach is generalizable.

4.6.2 Model Accuracy

After tweet selection was accomplished and the sentiment classifiers trained, the accuracy of the three models is determined. According to our findings, the dictionary approach yielded the lowest accuracy of 75.64% in correctly classifying tweets. For the SVM an accuracy of 79.56% was achieved when tested on the unseen test data. For the LSTM an accuracy of 82.79% is reported in the final validation process of the training.[27] Note that this validation accuracy does not precisely reflect the accuracy the model will achieve on unseen data but indicates the generalization ability. As the validation data is assessed multiple times while training, thus being part of the training process, the performance on unseen data is expected to be slightly lower. Further, note that these three accuracies are based on a general *Twitter* dataset; thus, they are not directly representative in the context of finance-related messages. Nevertheless, they give a good impression of the capabilities of the different methods and suggest that more sophisticated models generally lead to better classification performance, supporting the hypothesis that more sophisticated classifiers are more accurate. To further improve the performance, it is important to note that incorporating a neutral category in our actual model mitigates the impact of wrongly classified tweets and tweets that are not polarized.[28]

4.6.3 Sentiment Indices

Figures 4.5 to 4.7 show the *OI*s, *PI*s, *SQ*s and the total return index during the observation period. Upon visual inspection, it is evident that the dictionary approach is the least optimistic, which may be attributed to the limited number of positive words in the dictionary. Besides, the three classifiers exhibit a relatively low standard deviation, which

[26] The list of signal words can be found in Table 4.5 in the Appendix

[27] The model accuracy of the SVM and LSTM increased from 76.22% respectively 78.02% when stop words were excluded, which confirms our assumptions described in the data cleaning section.

[28] Since it is impossible to measure the accuracy of our unlabeled corpus, we compared the percentage of tweets that were classified identically by the models. The LSTM model and dictionary shared 40.44% similarity, the dictionary and SVM shared 48.80%, and LSTM and SVM shared 61.03%. For the dictionary approach significantly more tweets have been classified neutral due to the limited size of the word lists.

for example can be seen at the beginning of the covid pandemic in early 2022, where only a marginal and short-lasting slump in optimism occurs. Moreover, the standard deviation of the *OI*s diminishes over time, which should be related to the increasing number of tweets available. Looking at the development over the entire observation period, it seems there is a correlation between optimism and total returns.

Regarding the *PI*, as indicated by Figure 4.6, the indices are closer together but have a higher standard deviation. For all three models, there was a clear increase in pessimism at the start of the pandemic, but similar to the *OI*, it quickly returned to normal levels. Moreover, the models have identified a rise in pessimism starting from 2017, which has stagnated since the pandemic. This lack of change in recent years is understandable, given the increased volatility of total returns and the challenging economic conditions. The *SQ* is, by definition, connected to the *OI* and the *PI*. Therefore, the previously mentioned findings are reflected in it. Hence, the visual analysis supports our first hypothesis of social media sentiment affecting REIT market returns.

Figure 4.5 - Optimism Indicators and monthly Total Returns over Time

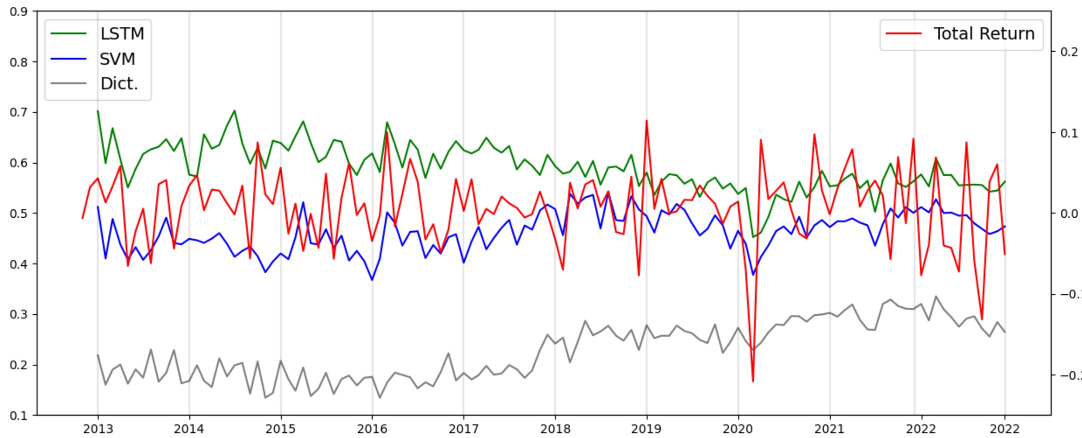


Figure 4.6 - Pessimism Indicators and monthly Total Returns over Time

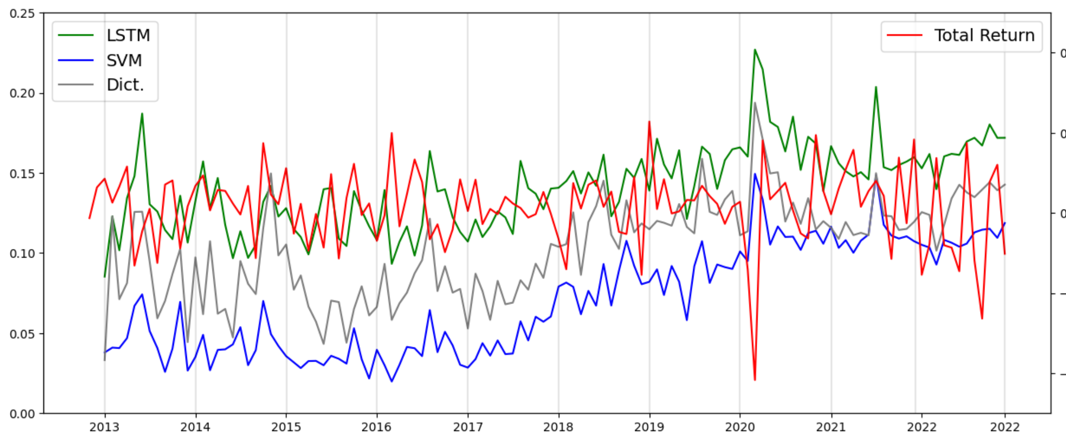
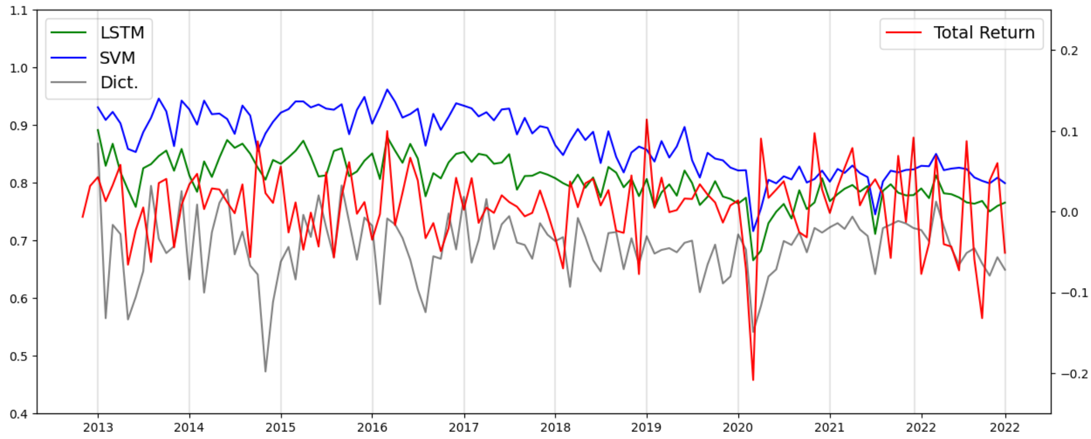


Figure 4.7 - Sentiment Quotient and monthly Total Returns over Time



4.6.4 Vector Autoregression

Table 4.2 presents the results of the regressions, including the *OI* as the endogenous sentiment variable, for the three classifiers. Both the dictionary approach and the LSTM model exhibit a short-term inverse relationship between optimism and total return, which seems counterintuitive at first. An increase in optimism tends to result in a decrease in total return in the following month. Interestingly, this short-term inverse relationship is similar to the relationship observed between total return today and total return in the previous month. This could be interpreted as investors becoming skeptical about the sustainability of the returns and therefore anticipating a downward correction. Despite this short-term reversal effect, the LSTM model detects a long-term relationship significant at 1%-level between optimism and total returns in lag seven. The coefficient has the expected sign, and its economic significance is the highest among the significant coefficients, therefore, having the biggest impact on returns.

As Table 4.3 indicates, the results for the *PI*s are even more promising. Again, the dictionary approach identifies an inverse relationship in the first lag between pessimism and total returns, emphasizing the previously described skepticism among investors and an expectation of a downward correction. The inverse relationship is also consistent with the total return in the previous month. However, in contrast to the *OI*, the LSTM model finds a short-term relationship between pessimism and total returns in lag two and three, which has the expected sign. This indicates that investors might expect returns to decline when pessimism increases. Besides the short-term effects, all models find a long-term relationship between pessimism and total returns. For the dictionary approach and LSTM model, the sixth and seventh lag are significant at 5%, respectively, 1%-level, and for the SVM model, the seventh lag is significant at the 5%-level. These coefficients have the

Table 4.2 - VAR Results monthly Lags Optimism Indicator

	FTSE/NAREIT AllREIT Total Return		
	OI Dict.	**OI SVM**	**OI LSTM**
Total Return $_{t-1}$	-0.253 **	-0.274 **	-0.252 *
Total Return $_{t-2}$	-0.068	-0.046	-0.052
Total Return $_{t-3}$	0.037	0.015	0.058
Total Return $_{t-4}$	0.012	0.068	-0.013
Total Return $_{t-5}$	0.114	0.117	0.096
Total Return $_{t-6}$	-0.097	-0.133	-0.057
Total Return $_{t-7}$	0.029	0.037	-0.027
Total Return $_{t-8}$	0.065	0.129	0.069
OI $_{t-1}$	-0.362 *	0.019	-0.261 **
OI $_{t-2}$	-0.177	-0.061	0.039
OI $_{t-3}$	0.058	0.172	0.132
OI $_{t-4}$	0.171	-0.014	0.100
OI $_{t-5}$	-0.042	-0.056	-0.057
OI $_{t-6}$	0.404	0.198	0.205
OI $_{t-7}$	0.289	0.164	0.400 ***
OI $_{t-8}$	0.281	0.022	0.112
Spread	-17.001 ***	-16.472 ***	-17.068 ***
Term	-5.970 **	-7.215 ***	-7.141 ***
Inflation	1.145	0.996	0.958
Constant	0.006	0.005	0.007
Adj. R^2	0.186	0.149	0.219
Granger Causality			
Sentiment	0.288	0.714	0.088
Total Return	0.667	0.524	0.125

Notes: Table 4.2 shows the estimated coefficients from the VAR models with monthly total returns and the optimism indicator. The set of the macroeconomic control variables includes the spread between Moody's Seasoned Baa- and Aaa-rated corporate Bonds (SPREAD), the difference between the 10-year Treasury bond yield and the 3-month Treasury bill (TERM) and inflation rate of the US (INFLATION). All variables, except for total return, are considered as the first differences. Newey and West (1987) standard errors, which are robust to heteroscedasticity and autocorrelation, are used. * denotes significance at the 10 % level, ** at the 5 % level, and *** at the 1 % level.

expected sign and the largest economic significance in the respective models. Therefore, the *PI* is an even better indicator than the *OI*.

Table 4.3 - VAR Results monthly Lags Pessimism Indicator

	FTSE/NAREIT AllREIT Total Return		
	PI Dict.	**PI SVM**	**PI LSTM**
Total Return $_{t-1}$	-0.285 **	-0.244 *	-0.299 **
Total Return $_{t-2}$	-0.068	-0.069	-0.092
Total Return $_{t-3}$	0.063	0.093	0.044
Total Return $_{t-4}$	0.057	0.070	-0.006
Total Return $_{t-5}$	0.110	0.062	0.040
Total Return $_{t-6}$	-0.111 *	-0.088	-0.091
Total Return $_{t-7}$	0.004	-0.032	-0.060
Total Return $_{t-8}$	0.159 *	0.141	0.133
PI $_{t-1}$	0.385 *	0.136	0.095
PI $_{t-2}$	0.036	-0.199	-0.581 ***
PI $_{t-3}$	-0.023	0.020	-0.434 **
PI $_{t-4}$	-0.063	0.153	-0.281
PI $_{t-5}$	-0.160	0.261	-0.165
PI $_{t-6}$	-0.523 **	-0.170	-0.752 ***
PI $_{t-7}$	-0.425 **	-0.803 **	-0.926 ***
PI $_{t-8}$	0.206	0.081	-0.224
Spread	-19.369 ***	-16.774 ***	-15.995 ***
Term	-7.405 ***	-7.113 ***	-6.642 ***
Inflation	1.331	0.564	1.008
Constant	0.006	0.006	0.009
Adj. R^2	0.243	0.194	0.279
Granger Causality			
Sentiment	0.030	0.224	0.005
Total Return	0.557	0.902	0.491

Notes: Table 4.3 shows the estimated coefficients from the VAR models with monthly total returns and the pessimism indicator. The set of the macroeconomic control variables includes the spread between Moody's Seasoned Baa- and Aaa-rated corporate Bonds (SPREAD), the difference between the 10-year Treasury bond yield and the 3-month Treasury bill (TERM) and inflation rate of the US (INFLATION). All variables, except for total return, are considered as the first differences. Newey and West (1987) standard errors, which are robust to heteroscedasticity and autocorrelation, are used. * denotes significance at the 10 % level, ** at the 5 % level, and *** at the 1 % level.

The regression results for the *SQ*, as presented in Table 4.4, are consistent with the previous sentiment indicators but particularly noteworthy. The short-term and long-term relationships are very similar to the *PI*, with the long-term effects of sentiment again having the highest impact on total returns. Since the results for the three sentiment indicators are

Table 4.4 - VAR Results monthly Lags Sentiment Quotient

	FTSE/NAREIT AllREIT Total Return		
	SQ Dict.	**SQ SVM**	**SQ LSTM**
Total Return $_{t-1}$	-0.303 **	-0.264 **	-0.270 **
Total Return $_{t-2}$	-0.063	-0.060	-0.081
Total Return $_{t-3}$	0.045	0.078	0.042
Total Return $_{t-4}$	0.054	0.053	-0.016
Total Return $_{t-5}$	0.134 *	0.087	0.060
Total Return $_{t-6}$	-0.125 *	-0.100	-0.074
Total Return $_{t-7}$	0.030	-0.033	-0.080
Total Return $_{t-8}$	0.143	0.151	0.111
SQ $_{t-1}$	-0.153 **	-0.037	-0.154
SQ $_{t-2}$	-0.063	0.085	0.329 **
SQ $_{t-3}$	0.028	0.071	0.258 *
SQ $_{t-4}$	0.006	-0.038	0.161
SQ $_{t-5}$	0.029	-0.087	0.035
SQ $_{t-6}$	0.207 **	0.153	0.450 ***
SQ $_{t-7}$	0.111 *	0.426 **	0.617 ***
SQ $_{t-8}$	-0.046	-0.008	0.123
Spread	-20.507 ***	-17.205 ***	-16.309 ***
Term	-6.810 ***	-7.353 ***	-6.874 ***
Inflation	1.308	0.552	0.940
Constant	0.006	0.006	0.008
Adj. R^2	0.232	0.185	0.282
Granger Causality			
Sentiment	0.051	0.297	0.004
Total Return	0.521	0.822	0.285

Notes: Table 4.4 shows the estimated coefficients from the VAR models with monthly total returns and the sentiment quotient. The set of the macroeconomic control variables includes the spread between Moody's Seasoned Baa- and Aaa-rated corporate Bonds (SPREAD), the difference between the 10-year Treasury bond yield and the 3-month Treasury bill (TERM) and inflation rate of the US (INFLATION). All variables, except for total return, are considered as the first differences. Newey and West (1987) standard errors, which are robust to heteroscedasticity and autocorrelation, are used. * denotes significance at the 10 % level, ** at the 5 % level, and *** at the 1 % level.

consistent and reflected in the quarterly analysis presented in Table 4.6 through 4.8 in the Appendix, it is reasonable to assume that there is a robust and well-founded relationship between them. The lower significance and magnitude of the short-term effects (one to three months ahead) lead us to interpret them as speculative reactions to observed changes in the stock market. Conversely, the more substantial long-term effects (six to seven months ahead) could be seen as investors' reactions to changes in the underlying assets, i.e., the real estate market. The results regarding the long-term effects, are consistent with previous research by Hausler et al. (2018) and Braun et al. (2020), who observed a similar lag-structure between the media-expressed sentiment and direct real estate markets. The same applies, when comparing our results to Ruscheinsky et al. (2018b) on news sentiment and REITs, who even find a similar structure in the coefficients of the sentiment variable. However, we find that social media sentiment is an even earlier indicator, by two months, than news sentiment.

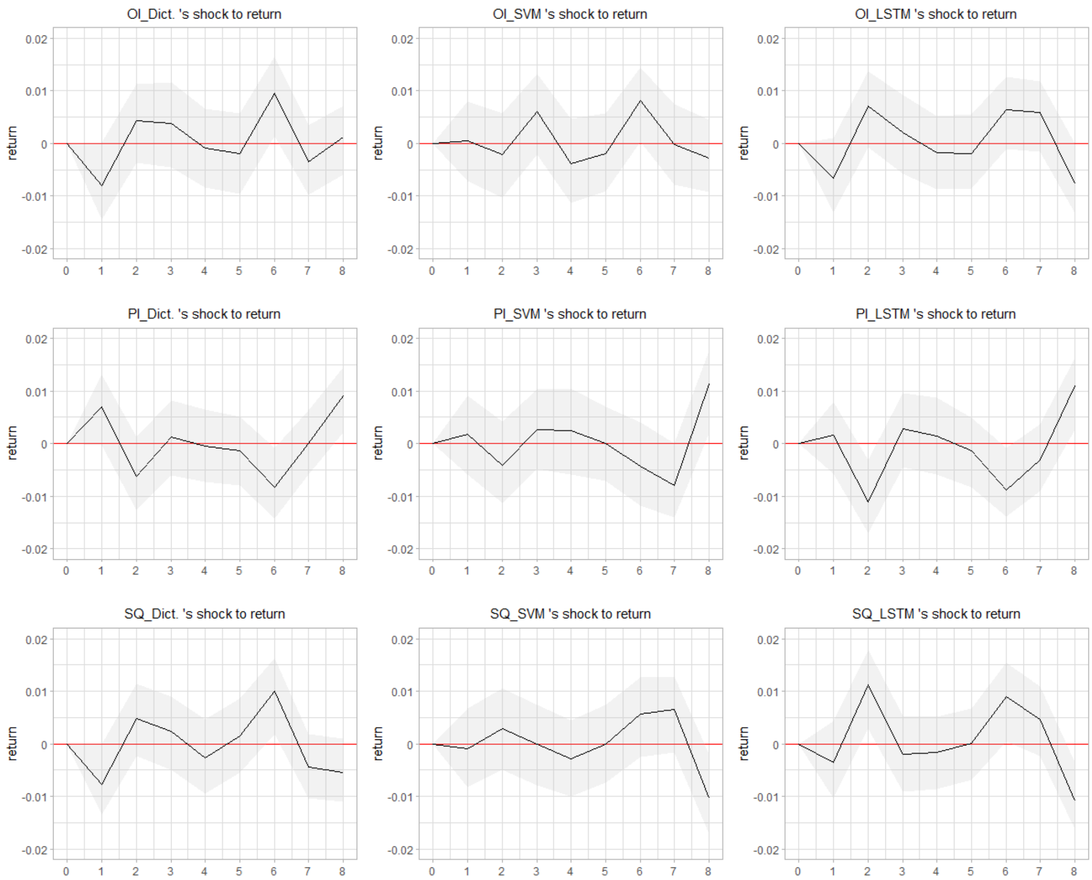
In summary, the VAR results are highly consistent and robust, even when considering the quarterly regression. This supports our hypothesis that sentiment expressed on social media affects the returns of the REIT market. All models demonstrate significant relationships, with the *PI* and *SQ* being the most informative indicators and accounting for a higher adjusted R^2. When comparing the sentiment classifiers, the LSTM model is best at capturing the relationships and yields the highest adjusted R^2 for all sentiment indicators. However, contrary to our expectations, the dictionary approach outperforms the SVM model despite the previously described higher classification accuracy. We think that this is primarily due to two reasons. First, only a subset of the training dataset was used to train the model. Second, the dictionary approach might be better classifying real estate-related messages because of the word lists applied. Hence, utilizing more training data or a real estate-related training dataset might have increased the SVM's performance. Ultimately, this partly supports our hypothesis that the more sophisticated a sentiment classifier is, the better the predictions.

4.6.5 Granger Causality and Impulse Response Function

Looking at the Granger causalities, these findings are confirmed. Tables 4.2 to 4.4 show that the sentiment indicators Granger-cause total returns, but not vice versa. Similar to the VAR results, the *OI* shows the lowest Granger causality, being significant at the 10%-level only for the LSTM model. The dictionary approach and the LSTM model show significant Granger causality for the *PI* and the *SQ*. For both indicators, the LSTM model is significant at 1%-level and is again the best predictor for future REIT market returns.

The results of the impulse response functions for all models are presented in Figure 4.8. Each line represents the outcomes for one of the sentiment indicators. Interestingly, the response of the total return appears to be independent of the classifier utilized. As anticipated, an increase in the *OI* and *SQ* leads to a positive response in total returns, while the opposite is true for the *PI*. These functions further support previous findings that sentiment effects can be categorized into short and long-term effects. The short-term effect occurs within the first three to four months and is typically characterized as a countermovement. Afterward, the long-term effect, which begins after six months, has a more significant and longer-lasting impact on the total return. Ultimately, the effect fades out in month eight.

Figure 4.8 - Impulse Response Functions



Notes: The figure displays the impact of a one-time shock to the sentiment indicators on the total returns over time, as revealed by the impulse response functions (IRFs). Each of the three sentiment indicators is represented by a separate line, while each column provides the results for one of the classification algorithms. The grey area represents the 95% confidence interval.

4.7 Conclusion

This study examines the impact of social media sentiment on REIT market returns from 2013 to 2022. Therefore, a total of 4.97 million tweets related to REITs were collected from *Twitter* and subjected to a rigorous selection process, ensuring the relevance of the remaining tweets and resulting in a final corpus of approximately 210,000 tweets. Those messages were classified as either positive, negative or neutral and comprised in sentiment indices to compare them with the REIT market performance.

The results suggest a significant relationship between social media sentiment and total returns, regardless of whether the focus is on the percentage of optimistic or pessimistic messages or their ratio. As expected, an increase in optimism or the sentiment quotient leads to higher returns in the subsequent months, while the opposite is true for pessimism. In line with prior research using newspapers or corporate disclosure, pessimism and the sentiment quotient have a greater impact than optimism on REIT returns. Nonetheless, the relationship for all sentiment indicators, can be characterized by short-term effects, which are likely to be stock market driven, and long-term effects related to the underlying real estate market, with the latter having a greater impact. Hence, our first hypothesis that social media sentiment affects future REIT market returns is confirmed, which makes social media a valuable source of information for real estate investors. Especially, with the growing public interest and the corresponding increase in message volume in recent years, future social media sentiment can be an even better indicator of public opinion.

This study employed three different sentiment classifiers to test the ability to understand the complex linguistic features inherent to social media messages. We compared a dictionary approach, designed to control for valence shifters and utilizing real estate word lists, with a Support Vector Machine as a machine learning approach and a Long Short-Term Memory as a deep learning approach. Significant relationships of similar lag structure were found by all three models, which provides further evidence for our first hypothesis. Even the conventional and literature-approved dictionary approach seems to handle the linguistic complexity of social media messages and identifies the relationships. The fact that the SVM and LSTM models find similar relationships suggests that it is not a coincidence.

Additionally, as it can be observed that the classification accuracy increases for theses machine and deep learning approaches, our second hypothesis is confirmed. The LSTM approach finds the strongest relationship among all models, likely due to its better classification performance. The relationship detected by the SVM follows the same shape. However, it appears less significant, which might be related to its limited training. Increasing its training could improve its classification accuracy, leading to more significant forecasts. These findings confirm our third hypothesis that more sophisticated sentiment classifiers can better predict future REIT market returns. For real-world applications, the dictionary approach might be sufficient to get an initial understanding of the impact of social media; thereupon, investors may choose to use more sophisticated models that provide more accurate results for investment decisions.

These results are promising. Still, social media analysis is a relatively new field in real estate literature. Therefore, this paper offers a detailed selection process and comparison of sentiment classifiers that can be useful for future research and its application in practice. The selection process can be easily adjusted to identify relevant tweets for different topics using different search words or replacing the GloVe with a topic modelling approach. Future research should focus on improving the selection process, for example by applying different firm identifiers, signal words or topic models. Even though it is more complex, with advances in tweet selection, applying the approach to direct real estate markets is possible. Regardless of the selection process, future research might consider increasing classification performance by using more and recent training data, finance related training data or considering the impact of emojis. This will yield even better classification results and, combined with the increasing number of social media messages in recent years, provides additional possibilities for any analysis. Researchers could for example explore the quantity of messages as another sentiment indicator or change the dependent variable to measures such as stock price volatility or trading volume. In summary, our findings demonstrate the significant potential of social media data as a valuable source of information that merits further exploration.

4.8 Appendix

Appendix 4.1 - SVM

Following Cortes and Vapnik (1995), a set of training vectors $(y_1, x_1), \dots, (y_l, x_l), y_l \in \{-1,1\}$ is separable by a hyperplane if the inequality $y_i(wx_i + b) \geq 1$ with $i = 1, \dots, l$ is fulfilled. The optimal hyperplane, representing the hyperplane that maximizes the distance between the separated points, is given by $w_0 z + b_0 = 0$. The maximized distance between the optimal hyperplane and the data points is given by:

$$p(w_o, b_o) = \frac{2}{|w_0|} = \frac{2}{\sqrt{w_0 * w_0}} \tag{A4.1}$$

The optimal decision surface is determined by w_o, it can be written as a linear combination of training vectors:

$$w_o = \sum_{i=1}^{l} y_i \alpha_i^0 x_i \tag{A4.2}$$

with $\alpha_i^0 \geq 0$. To find α_i, a quadratic programming function:

$$W(\Lambda) = \Lambda^T 1 - \frac{1}{2} \Lambda^T D \Lambda \tag{A4.3}$$

with $\Lambda^T = (\alpha_1, \dots, \alpha_l)$, and the constraints $\Lambda \geq 0; \Lambda^T Y = 0$, where $1^T = (1, \dots, 1)$ is an l-dimensional vector, $Y^T = (y_1, \dots, y_l)$ is an l-dimensional vector of labels, and D is a symmetric lxl matrix $D_i j = y_i y_j x_i x_j$ with $i, j = 1, \dots, l$, needs to be solved. As the formula for w_0 is given, it can be solved for b_0, which provides all parameters for stating the optimal hyperplane.

Cortes and Vapnik (1995) state that the optimal hyperplane may still not be found if the data is not separable without any errors. Therefore, the soft margin classifier is introduced to solve this limitation. A new parameter C is introduced, modifying the quadratic programming function, given in equation (A4.3), to allow for some errors:

$$W(\Lambda) = \Lambda^T 1 - \frac{1}{2} \left[\Lambda^T D \Lambda + \frac{\delta^2}{C} \right] \tag{A4.4}$$

where $\delta = \alpha_{\max} = \max(\alpha_1, \dots, \alpha_l)$. This addition enables the SVM to find the optimal hyperplane for the remaining data.

New data, thus unknown vectors, are classified by applying the sign function. Therefore, the output is binary with a value of $+1$ or -1 for the classification categories. In the context of this study, the SVM outputs a value of $+1$ for positive and -1 for negative messages when classifying the input data. As the output of an SVM hence is not probabilistic, Platt scaling is applied to get probabilistic output values that enable the desired sentiment classification. Platt (1999) uses a parametric form of a sigmoid:

$$P(y = 1 \mid f) = \frac{1}{1 + \exp{(Af + B)}} \tag{A4.5}$$

The variables $A\&B$ are fit by a maximum likelihood estimation based on a training set (f_i, y_i). As stated by Platt (1999), to find $A\&B$, a new training set (f_i, t_i) is defined. With t_i as the target probabilities:

$$t_i = \frac{y_i + 1}{2} \tag{A4.6}$$

Finally, to calculate the parameters $A\&B$, the negative log-likelihood function of the training data must be minimized:

$$min - \sum_i t_i \log{(p_i)} + (1 - t_i) \log{(1 - p_i)} \tag{A4.7}$$

with

$$p_i = \frac{1}{1 + \exp{(Af_i + B)}} \tag{A4.8}$$

After the scaling is applied, the probability of a tweet having a positive sentiment is presented, enabling the desired classification.

Appendix 2 - LSTM

Facing the shortcomings of conventional RNNs, the LSTM network adds a memory cell c_j, which evolves during the constant error carousel (CEC). The CEC is a mechanism where the cell state, the value saved in the memory cell, is multiplied by a weight of 1. The cell state can only be changed by values allowed to pass through the input gate and the output gate regulates access to the memory cell and protects other units from interference (Hochreiter and Schmidhuber, 1997). The output of the memory cell can be stated as follows:

$$y^{c_j}(t) = y^{out\,j}(t)h\left(s_{c_j}(t)\right) \tag{A4.9}$$

y^{out_j} describes the activation of the output gate. The output gate and the input gate y^{in_j} are given as:

$$y^{in_j}(t) = f_{\text{in}\,j}\left(\text{ net }_{\text{in}\,j}(t)\right); y^{\text{out}\,j}(t) = f_{\text{out}\,j}\left(\text{ net }_{\text{out}\,j}(t)\right) \tag{A4.10}$$

where

$$net_{in_j}(t) = \sum_u \omega_{in_j u} y^u\,(t-1);\; net_{out_j}(t) = \sum_u \omega_{out_j u} y^u\,(t-1) \tag{A4.11}$$

with net_j being the current net input for unit j. Modern LSTM implementations add an additional gate to the memory cell called the forget gate. The forget gate was introduced by Gers et al. (2000) as an addition to the memory cell. The authors mention that, despite the presence of the CEC, the internal state can still grow unchecked and pose a threat to the stability of the network, which is countered by an additional gate unit. The forget gate acts as a reset for the LSTM cell, it is forgetting its internal information. As stated in Gers et al. (1999), the activation of the forget gate y^{φ} states as:

$$y^{\varphi j}(t) = f_{\varphi j}\big(net_{\varphi j}(t)\big) \tag{A4.12}$$

with $net_{\varphi j}$ being the net input:

$$net_{\varphi j} = \sum_u w_{\varphi j u} y^u(t-1) \tag{A4.13}$$

Gers et al. (1999) further adjusted the calculation formula for the internal cell state s_{c_j} by adding the forget gate:

$$s_{c_j}(0) = 0; s_{c_j}(t) = y^{\varphi j}(t)s_{c_j}(t-1) + y^{in_j}(t)g\left(\text{ net }_{c_j}(t)\right) \text{ for } t > 0 \tag{A4.14}$$

With the addition of the forget gate, the LSTM unit now includes three gates, determining when different values can flow through it. The output given by the three gates ranges between 0 and 1. Soutner and Müller (2013) describe how the different gates can influence the LSTM cell: The input gates can decide how an input given can influence the cell. With a gate value close to 0, the input does not change the cells' internal values. The forget gate can further decide if stored information in the cell is kept or reset by giving a value close to 1 for keeping it and 0 to reset it. The output gate further decides when the information stored in the unit should be put out.

Table 4.5 - Signal Words

Start word	Signal words
REIT	reit; reits; estate; portfolio; retail; sector; stock; real; property; industrial; dividend; company; yield; focus; business; exposure; favorite; consider; growth; pick; office; healthcare; medical; market; investor

Table 4.6 - VAR Results quarterly Lags Optimism Indicator

	FTSE/NAREIT AllREIT Total Return					
	OI Dict.		OI SVM		OI LSTM	
Total Return $_{t-1}$	-0.618	***	-0.668	***	-0.652	***
Total Return $_{t-2}$	-0.389	*	-0.471	***	-0.469	***
Total Return $_{t-3}$	-0.130		-0.135		-0.196	
Total Return $_{t-4}$	-0.109		-0.162	**	-0.126	
OI $_{t-1}$	-0.257		0.218		0.157	
OI $_{t-2}$	0.078		0.470	**	0.717	
OI $_{t-3}$	-0.896		-0.197		-0.123	
OI $_{t-4}$	-0.818		0.739	**	0.192	
Spread	-26.972	***	-28.952	***	-30.201	***
Term	-5.430		-3.709		-3.279	
Inflation	8.766		11.552		9.411	**
Constant	-0.001		-0.001		-0.002	
Adj. R^2	0.497		0.555		0.532	
Granger Causality						
Sentiment	0.819		0.319		0.490	
Total Return	0.500		0.545		0.002	

Notes: Table 4.6 shows the estimated coefficients from the VAR models with quarterly total returns and the optimism indicator. The set of the macroeconomic control variables includes the spread between Moody's Seasoned Baa- and Aaa-rated corporate Bonds (SPREAD), the difference between the 10-year Treasury bond yield and the 3-month Treasury bill (TERM) and inflation rate of the US (INFLATION). All variables are considered as second differences, except for total returns, which are first differences. Newey and West (1987) standard errors, which are robust to heteroscedasticity and autocorrelation, are used. * denotes significance at the 10 % level, ** at the 5 % level, and *** at the 1 % level.

Table 4.7 - VAR Results quarterly Lags Pessimism Indicator

	FTSE/NAREIT AllREIT Total Return					
	PI Dict.		**PI SVM**		**PI LSTM**	
Total Return $_{t-1}$	-0.646	***	-0.575	***	-0.660	***
Total Return $_{t-2}$	-0.404	**	-0.364	*	-0.563	***
Total Return $_{t-3}$	-0.142		-0.215	*	-0.220	
Total Return $_{t-4}$	-0.085		-0.248		-0.167	
PI $_{t-1}$	-0.258		1.633		-0.296	
PI $_{t-2}$	-1.310	**	1.280		-1.457	***
PI $_{t-3}$	-0.535		0.630		-0.222	
PI $_{t-4}$	-0.422		0.096		-0.753	
Spread	-31.831	**	-22.583	*	-27.135	***
Term	-3.801		-5.658		-2.602	
Inflation	8.419		8.521		11.188	**
Constant	-0.001		-0.002		-0.002	
Adj. R^2	0.513		0.488		0.542	
Granger Causality						
Sentiment	0.669		0.897		0.409	
Total Return	0.362		0.253		0.062	

Notes: Table 4.7 shows the estimated coefficients from the VAR models with quarterly total returns and the pessimism indicator. The set of the macroeconomic control variables includes the spread between Moody's Seasoned Baa- and Aaa-rated corporate Bonds (SPREAD), the difference between the 10-year Treasury bond yield and the 3-month Treasury bill (TERM) and inflation rate of the US (INFLATION). All variables are considered as second differences, except for total returns, which are first differences. Newey and West (1987) standard errors, which are robust to heteroscedasticity and autocorrelation, are used. * denotes significance at the 10 % level, ** at the 5 % level, and *** at the 1 % level.

Table 4.8 - VAR Results quarterly Lags Sentiment Quotient

	FTSE/NAREIT AllREIT Total Return		
	SQ Dict.	**SQ SVM**	**SQ LSTM**
Total Return $_{t-1}$	-0.606 ***	-0.621 ***	-0.649 ***
Total Return $_{t-2}$	-0.354 *	-0.424 **	-0.549 ***
Total Return $_{t-3}$	-0.091	-0.217 *	-0.222
Total Return $_{t-4}$	-0.028	-0.233	-0.165
SQ $_{t-1}$	0.116	-0.586	0.143
SQ $_{t-2}$	0.582 **	-0.270	0.932 **
SQ $_{t-3}$	0.202	-0.262	0.043
SQ $_{t-4}$	0.107	0.061	0.489
Spread	-36.118 ***	-23.082	-27.454 ***
Term	-3.452	-4.747	-2.694
Inflation	7.647	9.992 *	11.169 **
Constant	-0.001	-0.002	-0.002
Adj. R^2	0.543	0.484	0.546
Granger Causality			
Sentiment	0.402	0.921	0.381
Total Return	0.438	0.126	0.028

Notes: Table 4.8 shows the estimated coefficients from the VAR models with quarterly total returns and the sentiment quotient. The set of the macroeconomic control variables includes the spread between Moody's Seasoned Baa- and Aaa-rated corporate Bonds (SPREAD), the difference between the 10-year Treasury bond yield and the 3-month Treasury bill (TERM) and inflation rate of the US (INFLATION). All variables are considered as second differences, except for total returns, which are first differences. Newey and West (1987) standard errors, which are robust to heteroscedasticity and autocorrelation, are used. * denotes significance at the 10 % level, ** at the 5 % level, and *** at the 1 % level.

4.9 References

Ankit and Saleena, N. (2018), "An Ensemble Classification System for Twitter Sentiment Analysis", Procedia Computer Science, Vol. 132, pp. 937–946.

Antweiler, W. and Frank, M.Z. (2004), "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards", The Journal of Finance, Vol. 59 No. 3, pp. 1259–1294.

Ba, J.L., Kiros, J.R. and Hinton, G.E. (2016), Layer Normalization, arXiv.

Bahrainian, S.-A. and Dengel, A. (2013), "Sentiment Analysis Using Sentiment Features", in 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta, GA, USA, 16.11.2013 - 19.11.2013, IEEE, pp. 26–29.

Beracha, E. and Downs, D.H. (2015), "Value and Momentum in Commercial Real Estate: A Market-Level Analysis", The Journal of Portfolio Management, Vol. 41 No. 6, pp. 48–61.

Beracha, E., Lang, M. and Hausler, J. (2019), "On the Relationship between Market Sentiment and Commercial Real Estate Performance—A Textual Analysis Examination", Journal of Real Estate Research, Vol. 41 No. 4, pp. 605–638.

Bollen, J., Mao, H. and Zeng, X. (2011), "Twitter mood predicts the stock market", Journal of Computational Science, Vol. 2 No. 1, pp. 1–8.

Braun, J., Hausler, J. and Schäfers, W. (2020), "Artificial intelligence, news sentiment, and property market liquidity", Journal of Property Investment & Finance, Vol. 38 No. 4, pp. 309–325.

Broersma, M. and Graham, T. (2013), "Twitter as a News Source", Journalism Practice, Vol. 7 No. 4, pp. 446–464.

Brooks, C. and Tsolacos, S. (2010), Real Estate Modelling and Forecasting, Cambridge University Press, Cambridge.

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

Carstens, R. and Freybote, J. (2021), "Can the Textual Tone in REIT Financial Statements Improve the Information Environment for Commercial Real Estate Investors? An Investigation", Journal of Real Estate Research, Vol. 43 No. 3, pp. 335–354.

Chen, H., De, P., Hu, Y. and Hwang, B.-H. (2011), "Sentiment revealed in social media and its effect on the stock market", in 2011 IEEE Statistical Signal Processing Workshop (SSP), Nice, 27.06.2011 - 29.06.2011, IEEE, pp. 25–28.

Clayton, J., Ling, D.C. and Naranjo, A. (2009), "Commercial Real Estate Valuation: Fundamentals Versus Investor Sentiment", The Journal of Real Estate Finance and Economics, Vol. 38 No. 1, pp. 5–37.

Cortes, C. and Vapnik, V. (1995), “Support-vector networks”, Machine Learning, Vol. 20 No. 3, pp. 273–297.

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

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

Dorle, S. and Pise, N. (2018), “Political Sentiment Analysis through Social Media”, in 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, 14.02.2018 - 15.02.2018, IEEE, pp. 869–873.

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

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

Fu, X., Yang, J., Li, J., Fang, M. and Wang, H. (2018), “Lexicon-Enhanced LSTM With Attention for General Sentiment Analysis”, IEEE Access, Vol. 6, pp. 71884–71891.

Gallimore, P. and Gray, A. (2002), “The role of investor sentiment in property investment decisions”, Journal of Property Research, Vol. 19 No. 2, pp. 111–120.

Gers, F.A., Schmidhuber, J. and Cummins, F. (2000), “Learning to forget: continual prediction with LSTM”, Neural computation, Vol. 12 No. 10, pp. 2451–2471.

Go, A., Bhayani, R. and Huang, L. (2009), “Twitter sentiment classification using distant supervision.”, CS224N project report, Stanford, Vol. 1 No. 12.

Granger, C.W.J. (1969), “Investigating Causal Relations by Econometric Models and Cross-spectral Methods”, Econometrica, Vol. 37 No. 3, pp. 424–438.

Gupta, R. and Chen, M. (2020), “Sentiment Analysis for Stock Price Prediction”, in 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Shenzhen, Guangdong, China, 05.08.2020 - 07.08.2020, IEEE, pp. 213–218.

Hannum, C., Arslanli, K.Y. and Kalay, A.F. (2019), “Spatial analysis of Twitter sentiment and district-level housing prices”, Journal of European Real Estate Research, Vol. 12 No. 2, pp. 173–189.

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

Heinig, S., Nanda, A. and Tsolacos, S. (2020), “Which Sentiment Indicators Matter? Evidence from the European Commercial Real Estate Market”, Journal of Real Estate Research, Vol. 42 No. 4, pp. 499–530.

Henry, E. and Leone, A.J. (2016), “Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone”, The Accounting Review, Vol. 91 No. 1, pp. 153–178.

Hochreiter, S. (1991), “Untersuchungen zu dynamischen neuronalen Netzen.”, Diploma, Technische Universität München, Vol. 91 No. 1.

Hochreiter, S. and Schmidhuber, J. (1997), “Long short-term memory”, Neural computation, Vol. 9 No. 8, pp. 1735–1780.

Hoesli, M., Lizieri, C. and MacGregor, B. (2008), “The Inflation Hedging Characteristics of US and UK Investments: A Multi-Factor Error Correction Approach”, The Journal of Real Estate Finance and Economics, Vol. 36 No. 2, pp. 183–206.

Jin, Z., Yang, Y. and Liu, Y. (2020), “Stock closing price prediction based on sentiment analysis and LSTM”, Neural Computing and Applications, Vol. 32 No. 13, pp. 9713–9729.

Jones, R. (2022), “Internet & Text Slang Dictionary”, available at: https://www.noslang.com/dictionary/ (accessed 1 February 2023).

Keynes, J.M. (Ed.) (1936), The General Theory of Employment, Interest, and Money, Palgrave Macmillan, Cham.

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.

Li, T., van Dalen, J. and van Rees, P.J. (2018), “More than just Noise? Examining the Information Content of Stock Microblogs on Financial Markets”, Journal of Information Technology, Vol. 33 No. 1, pp. 50–69.

Lin, C.Y., Rahman, H. and Yung, K. (2009), “Investor Sentiment and REIT Returns”, The Journal of Real Estate Finance and Economics, Vol. 39 No. 4, pp. 450–471.

Ling, D.C., Naranjo, A. and Scheick, B. (2014), “Investor Sentiment, Limits to Arbitrage and Private Market Returns”, Real Estate Economics, Vol. 42 No. 3, pp. 531–577.

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

Manning, C.D., Raghavan, P. and Schuetze, H. (2008), “The term vocabulary and postings lists”, Introduction to information retrieval, pp. 44–78.

Marcato, G. and Nanda, A. (2016), “Information Content and Forecasting Ability of Sentiment Indicators: Case of Real Estate Market”, Journal of Real Estate Research, Vol. 38 No. 2, pp. 165–204.

Maynard, D., Kalina, B. and Dominic, R. (2012), "Challenges in developing opinion mining tools for social media.", Proceedings of the@ NLP can u tag# usergeneratedcontent, pp. 15–22.

NAREIT (2022), "FTSE Nareit All REITs Index [Fact Sheet]", available at: https://www.ftserussell.com/products/indices/nareit (accessed 3 March 2023).

Newey, W.K. and West, K.D. (1987), "Hypothesis Testing with Efficient Method of Moments Estimation", International Economic Review, Vol. 28 No. 3, pp. 777–787.

Nowak, A. and Smith, P. (2017), "Textual Analysis in Real Estate", Journal of Applied Econometrics, Vol. 32 No. 4, pp. 896–918.

Paulus, N.M., Koelbl, M. and Schaefers, W. (2022), "Can textual analysis solve the underpricing puzzle? A US REIT study", Journal of Property Investment & Finance, Vol. 40 No. 6, pp. 548–570.

Paulussen, S. and Harder, R.A. (2014), "Social Media References in Newspapers", Journalism Practice, Vol. 8 No. 5, pp. 542–551.

Pennington, J., Socher, R. and Manning, C. (2014), "Glove: Global Vectors for Word Representation", in Alessandro Moschitti, Q.C.R.I., Bo Pang, G. and Walter Daelemans, U.o.A. (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1532–1543.

Platt, J. (1999), "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods", Advances in large margin classifiers, Vol. 10 No. 3, pp. 61–74.

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

Rinker, T. (2019), "Package 'sentimentr': calculate text polarity sentiment, R 3.4.0", Vol. 2019.

Rozin, P. and Royzman, E.B. (2001), "Negativity bias, negativity dominance, and contagion.", Personality and social psychology review, Vol. 5 No. 4, pp. 296–320.

Ruscheinsky, J., Kandlbinder, K., Alexander, M., Dietzel, K. and Schaefers, W. (2018a), "Forecasting the German Real Estate Market with Dictionary-based Textual Analysis", in 25th Annual European Real Estate Society Conference, Reading, United Kingdom, European Real Estate Society.

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

Sarker, I.H. (2021), "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions", SN computer science, Vol. 2 No. 6, p. 420.

Shi, Y., Zheng, Y., Guo, K. and Ren, X. (2021), "Stock movement prediction with sentiment analysis based on deep learning networks", Concurrency and Computation: Practice and Experience, Vol. 33 No. 6.

Shiller, R.J. (2017), "Narrative Economics", American Economic Review, Vol. 107 No. 4, pp. 967–1004.

Shiller, R.J. (2020), Narrative economics: How stories go viral and drive major economic events, Princeton University Press.

Soo, C.K. (2015), "Quantifying Animal Spirits: News Media and Sentiment in the Housing Market", Ross School of Business Working Paper No. 1200.

Soutner, D. and Müller, L. (2013), "Application of LSTM Neural Networks in Language Modelling", in Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M.Y., Weikum, G., Habernal, I. and Matoušek, V. (Eds.), Text, Speech, and Dialogue, Lecture Notes in Computer Science, Vol. 8082, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 105–112.

Sprenger, T.O., Tumasjan, A., Sandner, P.G. and Welpe, I.M. (2014), "Tweets and Trades: the Information Content of Stock Microblogs", European Financial Management, Vol. 20 No. 5, pp. 926–957.

Tan, M.J. and Guan, C. (2021), "Are people happier in locations of high property value? Spatial temporal analytics of activity frequency, public sentiment and housing price using twitter data", Applied Geography, Vol. 132, p. 102474.

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

Walker, C.B. (2014), "Housing booms and media coverage", Applied Economics, Vol. 46 No. 32, pp. 3954–3967.

Wang, C., Wang, T. and Yuan, C. (2020), "Does Applying Deep Learning in Financial Sentiment Analysis Lead to Better Classification Performance?", Economics Bulletin, Vol. 40 No. 2, pp. 1091–1105.

Watanabe, K. and Zhou, Y. (2020), "Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches", Social Science Computer Review, Vol. 40 No. 2, 346-366.

Wu, S., Liu, Y., Zou, Z. and Weng, T.-H. (2022), "S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis", Connection Science, Vol. 34 No. 1, pp. 44–62.

Xu, Y. and Cohen, S.B. (2018), "Stock Movement Prediction from Tweets and Historical Prices", in Gurevych, I. and Miyao, Y. (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1970–1979.

Yadav, A. and Vishwakarma, D.K. (2020), "Sentiment analysis using deep learning architectures: a review", Artificial Intelligence Review, Vol. 53 No. 6, pp. 4335–4385.

Yildirim, S., Jothimani, D., Kavaklioglu, C. and Basar, A. (2019), "Deep Learning Approaches for Sentiment Analysis on Financial Microblog Dataset", in 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 08.12.2019 - 11.12.2019, IEEE, pp. 5581–5584.

Zamani, M. and Schwartz, H.A. (2017), "Using Twitter Language to Predict the Real Estate Market", in Lapata, M., Blunsom, P. and Koller, A. (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 28–33.

Zhou, X., Tao, X., Rahman, M.M. and Zhang, J. (2017), "Coupling topic modelling in opinion mining for social media analysis", in Sheth, A., Ngonga, A., Wang, y., Chang, E., Ślęzak, D., Franczyk, B., Alt, R. and Tao, X. (Eds.), Proceedings of the International Conference on Web Intelligence, Leipzig Germany, 23 08 2017 26 08 2017, ACM, New York, NY, USA, pp. 533–540.

5 Conclusion

5.1 Executive Summary

This part of the thesis summarizes the content of the three papers. It discusses the objectives of each study, data, and methodologies used, as well as results and implications for science and practice.

Paper 1: Can Textual Analysis solve the Underpricing Puzzle? A US REIT Study

---

Problems and Objective

Although many theories aim to explain initial public offering (IPO) underpricing, unresolved questions remain on how IPOs are priced. Following the predominate asymmetric information hypothesis of Beatty and Ritter (1986), underpricing results from information asymmetry between potential investors and increases with ex ante uncertainty. As property revenues are more certain compared to revenues of a matching industrial sample, the underpricing for REITs is lower than for other industries. However, the underpricing of REITs has already been proven to be significant on several occasions (e.g. Ling and Ryngaert, 1997; Buttimer et al., 2005; Dolvin and Pyles, 2009) and is of particular interest due to the strict regulation and the associated high dependence on external equity of REITs.

Besides the extensive literature analyzing the impact of various quantitative factors, only a few studies have examined the role of qualitative information in understanding underpricing. Loughran and McDonald (2013) argue that IPOs do not have a long history of tangible information to help predict cash flows and investors hence rely on qualitative information. Therefore, Loughran and McDonald (2013) as well as Hanley and Hoberg (2010) used textual analysis to extract qualitative information from the initial document for registering (non-REIT) IPOs with the SEC (Form S-1), which is an important source of information for investors during the IPO process.

Since these studies find a significant impact of qualitative information from Form S-1 on underpricing, the question arises whether the REIT counterpart (Form S-11) helps to explain REIT underpricing. Therefore, the aim of this paper is to provide new insights into the underpricing phenomenon of US REITs, by investigating how qualitative information in

Form S-11 affects initial returns. More precisely, this study hypothesizes that providing an informative Form S-11 could help potential investors to form an individual opinion about the true value of a REIT and thus decreases IPO underpricing.

Methodology and Data

Using cosine similarity, a method to determine the similarity of documents, this paper determines the level of informativeness contained in Form S-11. Hereby, the degree of informativeness is defined by new information not already contained in previous REIT prospectuses. In accordance with Beatty and Ritter's (1986) asymmetric information hypothesis, it is assumed that a higher level of informativeness makes it easier for investors to overcome quantitative information deficits and precisely assimilate value-relevant information resulting in lower underpricing. A similar mechanism applies for the level of uncertain language, as uncertain language leaves more room for speculation, making it more difficult for investors to price the IPO.

The sample includes all US equity REITs that have completed an IPO between 1996 and 2020. For all IPOs in the sample, Form S-11 is downloaded from the SEC's *Electronic Data Gathering, Analysis and Retrieval* (*EDGAR*) database. Underpricing is used as the dependent variable for the analysis. The target variables are the level of informativeness, as well as uncertain language in the S-11, which will be interpreted after controlling for conventional quantitative variables.

Results and their Contribution to Science and Practice

To the best of the authors' knowledge, this is the first study investigating the information content of Form S-11 for a US REIT sample. The results indicate that the tone does not have a decisive impact on underpricing. However, as per our similarity measure, the informativeness of Form S-11 has the expected impact. Hence, Form S-11 is informative, decreases information asymmetries, and improves the investors' ability to value the IPO. However, issuers face a trade-off: Either they let investors value the IPO during book-building and accept a larger IPO discount (Benveniste and Spindt 1989) or they reduce the underpricing by providing costly information. Understanding the impact of qualitative information is therefore valuable for both investors and the issuing company.

Paper 2: Trade vs. Daily Press: The Role of News Coverage and Sentiment in Real Estate Market Analysis

Problems and Objective

Real estate markets are comparatively intransparent, which is why any kind of information can be valuable and market participants are always looking for an information edge. In order to meet the informational needs of their readership, newspaper publishers and especially real estate newspapers must hence inevitably and constantly deal with current topics. Hausler et al. (2018) therefore found a significant relationship between news-based sentiment and returns in the commercial real estate sector by applying textual sentiment analysis to headlines from *The Wall Street Journal*. Since then, textual sentiment has proven superior to survey-based and market fundamental-based sentiment indicators (Beracha et al., 2019; Ruscheinsky et al., 2018b).

However, these studies, focusing on commercial real estate, attempt to measure general market sentiment but not asset class-specific sentiment. Since there are distinct real estate cycles by asset class and since Marcato and Nanda (2016) have demonstrated varying importance of survey-based as well as market fundamental-based indicators for residential and commercial real estate, it is to be presumed that news-based sentiment indicators should be asset class-specific. Besides, as suggested by Hausler et al. (2018) it should be considered that media meets varying informational needs and therefore the selection of articles and the tone can change depending on the targeted readership. Thus, the market sentiment would be reflected incompletely by the analysis of one newspaper only.

For investors, such differences can play an important role as they make their decisions to some extent based on only part of the available information. For this reason, this paper examines whether the investment of time and money of reading newspapers is paying off and whether each type of newspaper is an equally valuable source of information. Therefore, a comprehensive approach was chosen, in order to derive asset class-specific sentiments for each newspaper type. This allows shedding light on the disparities in tone and time lag structure between trade newspapers and daily newspapers.

Methodology and Data

Using word embeddings and seeded topic modelling newspaper articles from two German trade and two German daily newspapers will be assigned to the real estate asset classes of residential, office, and retail. This classification not only allows the creation of asset class-specific sentiment indicators but also further investigations on focal points of reporting. The effect of news coverage and sentiment on German real estate returns will be explored by type of newspaper using a vector autoregressive framework (Brooks and Tsolacos, 2010).

The text corpus comprises almost 137.000 articles from two German daily newspapers (*Frankfurter Allgemeine Zeitung*, *Handelsblatt*) and two German trade newspapers (*Immobilienzeitung*, *Immobilien Manager*) between 2010 and 2020. As the articles are written in German, the dictionary for real estate developed by Ruscheinsky et al. (2018a) is applied. The total return is the dependent variable and news coverage and sentiment, respectively, are the variables of interest after controlling for macroeconomic and property-specific variables.

Results and their Contribution to Science and Practice

To the best of the authors' knowledge, this is the first study to quantify both the news coverage and sentiment for the main real estate asset classes by means of textual analysis and to assess different sentiments in trade and daily newspapers. Moreover, qualitative information from newspapers is available digitally and in real-time and hence can be extracted in a standardized, timely and replicable way. Therefore, information derived from newspapers could be valuable to investors and contribute to the understanding of real estate markets.

The findings of our study indicate that there are disparities in the reporting style of trade and daily newspapers. Specifically, our results suggest that daily newspapers tend to report more negatively on increases in returns in the residential market compared to the trade press, possibly due to their target audience. Although for both types of newspapers news coverage and news sentiment display significant impact on asset class returns, trade newspapers exhibit better performance in the prediction of total returns.

Paper 3: Social Media and Real Estate: Do Twitter Users predict REIT Performance?

Problems and Objective

Social media platforms have become vibrant online platforms where people share their opinions and views on any topic (Yadav and Vishwakarma 2020). With the increasing volume and speed of social media, the exchange of stock market-related information has become more important, which is why the effects of social media information on stock markets are becoming increasingly salient (Li et al., 2018). As social media increasingly influences traditional news media reporting (Broersma and Graham, 2013; Paulussen and Harder, 2014) it is not surprising that there is evidence for public sentiment, obtained from social media, correlating with or even predicting economic indicators (e.g. Bollen et al., 2011; Sprenger et al., 2014; Xu and Cohen, 2018). Therefore, businesses need to comprehend these dynamics and their impact on any market.

Even if the results of the finance literature are promising, they can only be applied to a limited extent to the less efficient and slower reacting real estate markets. However, the results of Zamani and Schwartz (2017) and Tan and Guan (2021 show that *Twitter* language can be used to predict house price changes, at least for a small sample at the county and district levels, respectively. Except for this limited research, there is no general study that examines the relationship between social media and real estate markets. This lack of research is likely to be due to the complexity of social media data (Maynard et al., 2012), which is composed of the selection of relevant news and the linguistic peculiarities of social media messages. Nevertheless, social media sentiment promises to cover the perspectives and beliefs of various market participants and could therefore be even more informative than traditional sentiment measures. Thus, this study aims to create a standardized framework that enables investors of all kinds to create valuable social media sentiment indicators and better navigate the opaque real estate market.

Methodology and Data

Examining the relationship between *Twitter* messages and REIT market performance is a process that involves various techniques of natural language processing to handle the complexity of social media data. In the first step, tweets are selected based on firm identifiers and then filtered using signal words to ultimately obtain a corpus of relevant tweets. Hereby, the signal words are determined using a word embedding approach. Afterward, the remaining tweets are assigned to one of three sentiment categories through the application of three distinct classification techniques (a dictionary, a machine learning, and a deep learning approach). The purpose of comparing the three classifiers is to determine which one is best suited for analyzing the linguistic peculiarities of social media texts. In the last step, all classified messages are aggregated in monthly sentiment indices.

Five million posts from the past 10 years, up until December 2022, were collected from *Twitter,* as it has the highest message volume regarding stock markets (Xu and Cohen, 2018). Besides, we decided to collect tweets on US REITs, as posts regarding companies can be selected using "cashtags" and as the US market is the largest of its kind in the world. In the econometric equation, the dependent variable is the monthly market return, while the variable of interest is the sentiment variable created.

Results and their Contribution to Science and Practice

To the authors' knowledge, this is the first study to analyze the impact of social media sentiment on indirect real estate returns based on a comprehensive national dataset. The findings indeed suggest that there is a significant relationship between monthly sentiment and REIT returns, which occurs in two phases: a short-term speculative reaction and a greater longer-term reaction related to actual changes in the real estate market. While the conventional dictionary approach is capable of identifying this relationship, more sophisticated classifiers can achieve higher accuracy. For practitioners, this implies that the dictionary approach can serve as an initial step to comprehending the relationship, but they may opt for more advanced methods that offer precise results for investment decisions. Besides, this framework could be applied not only by investors but vice versa by REITs to understand and optimize their position in society and in the investor landscape.

5.2 Final Remarks

In the sense of Clayton's et al. (2009), early research on whether it is "[…] Fundamentals Versus Investor Sentiment" in real estate markets, many studies concluded that "fundamentals are the key driver of [...]" real estate markets; nonetheless, "sentiment also plays a […] role". Nowadays, it is evident that qualitative information, such as sentiment, serves as an important indicator when reliable information is scarce or unavailable and, therefore, acts as a substitute for investors when making decisions. To gain a better understanding of the market, it is crucial to include qualitative information in quantitative models. Since Clayton et al. (2009) have shown that sentiment is a significant driver of real estate markets, the real estate literature has focused on creating even more valuable sentiment indicators.

Since qualitative information is inherently difficult to measure, especially natural language processing has shown to be a promising opportunity to develop measures for such information. It enables the quantification of information from an ever-increasing number of textual sources. However, since the literature previously has focused on sentiment, which is only one type of qualitative information that can be extracted from textual data, there is still much unexplored valuable information within text. Therefore, this thesis extends the existing literature by extracting more and different types of information from previously analyzed sources – such as the proportion of new information, news coverage, and asset class-specific sentiment – and by investigating the impact of further sources, such as social media.

The findings of the papers demonstrate the importance of utilizing different types of qualitative information in real estate markets, and show that they can be just as valuable as sentiment itself. Each paper applies a different approach to improve the quality of market indicators and enhance our understanding of real estate markets. Hereby, they either create novel qualitative indicators or filter corpora by content to create more meaningful respectively tailormade sentiment indicators. In that sense, the first paper introduces a similarity measure, which determines new information within various documents. The paper concludes that companies providing valuable information in corporate disclosure to investors benefit from lower IPO underpricing. The second paper, based on newspaper articles, for the first time, applies a topic modelling approach to classify articles by their content. This, on the one hand, allows the creation of sentiment indicators that are specifically designed for an asset class, and on the other hand, quantifies the news coverage of each asset class. Both indicators are significant predictors of direct real estate market returns and increase our market understanding compared to general sentiment indicators. The third paper applies those methods to an entirely new corpus – social media – to identify, in a real estate context, relevant social media messages. Drawing from these many opinions, the results indicate that social media can serve as an earlier indicator of upcoming market developments compared to traditional news media.

Nonetheless, it is important to note that research in this domain is still at the very beginning. More sophisticated approaches are being developed, which will allow extracting even more information from textual data. This dissertation has been able to explore the content of the corpora on a high level, such as determining whether the content is similar to that of other documents or identifying which asset classes are being discussed in an article. However, it has not delved deeper into the content analysis. Hence, questions such as which specific information caused investors to act differently remain unanswered. To find answers, future research could focus on identifying specific factors, themes, or sentiment patterns within the textual data that directly influence investor behavior and decision-making processes. This would provide a more granular understanding of the relationship between textual content and real estate market reactions, ultimately contributing to more effective investment strategies and market forecasts.

One could argue that the ongoing professionalization and digitalization of the real estate industry, could create transparency, reduce information asymmetries, and ultimately diminish the impact of qualitative information. However, it will be a long time until real estate markets can be considered "transparent." Meanwhile, it can be observed that not only does professionalization progress, but the complexity of the real estate markets does as well, which ultimately is another driver for sentiment-induced trading behavior. In addition, increasing digitalization will provide new qualitative information, which can be a topic of textual analysis.

Consequently, qualitative data continues to hold significant value in the real estate sector. The insights provided by this thesis are of equal relevance to both practitioners and academics. Professionals can use the outlined methodologies to better anticipate future trends and optimize their investment strategies. Simultaneously, researchers are urged to delve deeper into the impact of qualitative aspects and enhance information extraction. In contrast to other realms of real estate investigation, the necessary data is readily available – it simply awaits exploration.

5.3 References

Beatty, R.P. and Ritter, J.R. (1986), “Investment banking, reputation, and the underpricing of initial public offerings”, Journal of Financial Economics, Vol. 15 1-2, pp. 213–232.

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

Beracha, E., Lang, M. and Hausler, J. (2019), “On the Relationship between Market Sentiment and Commercial Real Estate Performance—A Textual Analysis Examination”, Journal of Real Estate Research, Vol. 41 No. 4, pp. 605–638.

Bollen, J., Mao, H. and Zeng, X. (2011), “Twitter mood predicts the stock market”, Journal of Computational Science, Vol. 2 No. 1, pp. 1–8.

Broersma, M. and Graham, T. (2013), “Twitter as a News Source”, Journalism Practice, Vol. 7 No. 4, pp. 446–464.

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

Clayton, J., Ling, D.C. and Naranjo, A. (2009), “Commercial Real Estate Valuation: Fundamentals Versus Investor Sentiment”, The Journal of Real Estate Finance and Economics, Vol. 38 No. 1, pp. 5–37.

Dolvin, S.D. and Pyles, M.K. (2009), “REIT IPOs and the Cost of Going Public”, The Journal of Real Estate Finance and Economics, Vol. 39 No. 1, pp. 92–106.

Hanley, K.W. and Hoberg, G. (2010), “The Information Content of IPO Prospectuses”, Review of Financial Studies, Vol. 23 No. 7, pp. 2821–2864.

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

Li, T., van Dalen, J. and van Rees, P.J. (2018), “More than just Noise? Examining the Information Content of Stock Microblogs on Financial Markets”, Journal of Information Technology, Vol. 33 No. 1, pp. 50–69.

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

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

Marcato, G. and Nanda, A. (2016), "Information Content and Forecasting Ability of Sentiment Indicators: Case of Real Estate Market", Journal of Real Estate Research, Vol. 38 No. 2, pp. 165–204.

Maynard, D., Kalina, B. and Dominic, R. (2012), "Challenges in developing opinion mining tools for social media.", Proceedings of the@ NLP can u tag# usergeneratedcontent, pp. 15–22.

Paulussen, S. and Harder, R.A. (2014), "Social Media References in Newspapers", Journalism Practice, Vol. 8 No. 5, pp. 542–551.

Ruscheinsky, J., Kandlbinder, K., Alexander, M., Dietzel, K. and Schaefers, W. (2018a), "Forecasting the German Real Estate Market with Dictionary-based Textual Analysis", in 25th Annual European Real Estate Society Conference, Reading, United Kingdom, European Real Estate Society.

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

Sprenger, T.O., Tumasjan, A., Sandner, P.G. and Welpe, I.M. (2014), "Tweets and Trades: the Information Content of Stock Microblogs", European Financial Management, Vol. 20 No. 5, pp. 926–957.

Tan, M.J. and Guan, C. (2021), "Are people happier in locations of high property value? Spatial temporal analytics of activity frequency, public sentiment and housing price using twitter data", Applied Geography, Vol. 132, p. 102474.

Xu, Y. and Cohen, S.B. (2018), "Stock Movement Prediction from Tweets and Historical Prices", in Gurevych, I. and Miyao, Y. (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1970–1979.

Yadav, A. and Vishwakarma, D.K. (2020), "Sentiment analysis using deep learning architectures: a review", Artificial Intelligence Review, Vol. 53 No. 6, pp. 4335–4385.

Zamani, M. and Schwartz, H.A. (2017), "Using Twitter Language to Predict the Real Estate Market", in Lapata, M., Blunsom, P. and Koller, A. (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 28–33.