The Role of News Media Coverage and Sentiment in German Real Estate Markets

Dissertation zur Erlangung

des Grades eines

Doktors der Wirtschaftswissenschaften

Eingereicht an der Fakultät

für Wirtschaftswissenschaften

der Universität Regensburg

vorgelegt von

Franziska Plößl

Berichterstatter: Prof. Dr. Tobias Just

Prof. Dr. Wolfgang Schäfers

Tag der Disputation: 28.10.2022

The Role of News Media Coverage and Sentiment in German Real Estate Markets

Franziska Plößl

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 General Motivation

Especially in times of great uncertainty, private and institutional investors are increasingly looking for investment opportunities that promise both safety and stability. Oftentimes, investors strive to find stability in real estate markets. Real estate accounts for approximately 80 % of tangible assets in Germany (Just *et al.*, 2017), but even in the 21st century, real estate markets are still considered to be intransparent due to their heterogeneity, infrequent trading and information asymmetry. Thus, even with increasing digitalisation, the relevant information to make profound investment decisions is often unavailable in detail or with time-delays, especially regarding German real estate markets. As a result, prices in real estate markets do not fully reflect all information.

Decades earlier, Black (1986) referred to this as ‘noise’, describing the uncertainty concerning future market movements which may lead to market inefficiencies. In recent research, these factors have been linked to the concept of sentiment: the subjective judgements and behavioural characteristics of investors (Kearney and Liu, 2014). Baker and Wurgler (2007) state that there is no longer a question of whether investor sentiment affects prices, but how to measure investor sentiment and to quantify its effects.

Attempts to measure sentiment in real estate markets have been conducted by surveys or indicators derived from capital market data (e.g., Clayton *et al.*, 2008; Das *et al.*, 2014; Marcato and Nanda, 2016; Freybote and Seagraves, 2017). However, these traditional measures are mainly past-oriented, data collection is time- and cost-intensive, and surveys can suffer from selection bias. Yet the exponential growth in computational power and qualitative data contained in digitally available texts opens up new opportunities for performing text-based sentiment analysis. As of yet, new and continuously improving methods in computational linguistics are being developed and applied. Consequently, textual analysis enables researchers to measure influential factors that have historically been difficult or impossible to replicate (Loughran and McDonald, 2020). In real estate literature, text-based sentiment analysis often comprises the measurement of the tonality of real estate-related texts in order to classify positive or negative sentiment (e.g., Doran *et al.*, 2012; Walker, 2016; Nowak and Smith, 2017; Soo, 2018; Carstens and Freybote, 2021). However, reporting intensity on certain topics in the real estate sector may broaden the view of the role of sentiment. The concept of economic narratives, introduced by Shiller (2017), states that narratives can impact and shape economic outcomes. Economic narratives that reach out to a lot of market participants can affect their individual decision-making. In particular, the news media plays an important role in the dissemination of these narratives. Journalists have a key function: they view facts and provide filtered information complemented by their own assessments. In a sense, they act as gatekeepers for information (White, 1950). Hence, additional information from the news media could be of value to investors but it is not trivial to derive the relevant information from a daily flow of news.

Therefore, the aim of this dissertation is to extend the common measurement of text-based sentiment found in real estate literature not only by extracting the tonality, but also the reporting intensity of certain topics. Can the non-numeric information encoded in real estate-related news articles help to address intransparency and heterogeneity across both space and time in real estate markets?

Through the textual analysis of different German real estate news providers, two main indicators are obtained: news coverage and news sentiment. Hence, the first paper establishes the methodological framework with which to measure news coverage and news sentiment by investigating news articles with respect to whether real estate-related trends underlie cyclicity over a reporting period of 20 years. The second study examines the relationship between news coverage or news sentiment and total returns of the residential, office and retail asset class. Two trade and two daily newspapers are analysed to evaluate the quality of information that is provided for the targeted readership. Finally, the third paper focuses on the residential real estate market and investigates regional differences relating to how frequently German cities are reported upon, whether there are periods of media optimism or pessimism and whether these indicators are linked to house price movements.

1.2 Research Questions

An overview of the subject matter and the research questions addressed in the three papers is provided in this section. The first paper of the dissertation builds the foundation for the following sections by developing the methodological framework to quantify the news coverage, on the one hand, and the news sentiment, on the other hand, from real estate-related texts. Hence, the first paper primarily demonstrates the potential of the introduced approach, while this methodology is applied in paper 2 and 3, respectively.

Paper 1 — Cyclicity of Real Estate-related Trends: Topic Modelling and Sentiment Analysis on German Real Estate News

Paper 1 identifies and analyses the news coverage and sentiment of real estate-related trends in Germany to investigate whether these two indicators underlie cyclicity over a period of 20 years. Almost 170,000 newspaper articles provided by a major German real estate news provider are assigned to six trends through the integration of topic modelling and word embeddings into real estate analysis. Thereafter, a dictionary-based approach is applied using an industry-specific dictionary created by Ruscheinsky *et al.* (2018) to examine the level of optimistic or pessimistic language related to the trends. The research questions are as follows:

- Can the integration of topic modelling into real estate analysis contribute to the understanding of sentiment?
- How has news coverage of these six major trends developed over time in terms of quantitative media attention?
- Are there differences in tonality regarding the articles of the six trends?
- Does this tonality change over time?

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

Paper 2 examines the relationship between news coverage or news sentiment and total returns of the three main asset classes. Three sentiment indicators for the residential, office and retail market are generated from almost 137,000 articles originating from two trade and two daily newspapers by means of computational linguistic techniques including word embeddings, seeded topic modelling and dictionary-based sentiment analysis. Regarding this, the following research questions arise:

- Are different types of newspapers equivalent sources of information?
- Are there significant differences between news coverage and news sentiment obtained from either trade newspapers or daily newspapers?
- Is there a positive relationship between news coverage or news sentiment and the performance for each asset class?
- How do the different markets react in periods when both indicators are high or low?
- Do news coverage and news sentiment have predictive power over future performance for each asset class?

Paper 3 — News Coverage vs. Sentiment: Evaluating German Residential Real Estate Markets

Paper 3 investigates whether additional information quantified from the news flow, in particular reporting intensity, can help to increase transparency on housing markets. Hence, this paper examines the relationship between five different sentiment measures and residential real estate prices in Germany on a regional level. By means of natural language processing including word embeddings and a dictionary-based approach, almost 130,000 news articles are analysed regarding the seven largest cities in Germany. The research questions can be stated as follows:

- Is there a relationship between news coverage or news sentiment and German residential property prices?
- Are there differences on a regional level?
- Is there a causality flow from news coverage or news sentiment to changes in residential property prices or vice versa?

- Are there differences regarding the markets for houses and condominiums?

1.3 Submissions

This section provides details of authorships, submission to journals and publication status in addition to the previous sections for the three papers comprising this dissertation.

Paper 1 — Cyclicity of Real Estate-related Trends: Topic Modelling and Sentiment Analysis on German Real Estate News

Authors:

Franziska Plößl, Tobias Just, Lino Wehrheim

Submission:

Journal: Journal of European Real Estate Research

Submission Date: 12/28/2020

Current Status: Accepted (05/25/2021)

Paper 2 — Trade vs. Daily Press:

The Role of News Coverage and Sentiment in Real Estate Market Analysis

Authors:

Franziska Plößl, Nino Paulus, Tobias Just

Submission:

Journal: Journal of Property Research

Submission Date: 07/18/2022

Current Status: Under Review

Paper 3 — News Coverage vs. Sentiment:

Evaluating German Residential Real Estate Markets

Authors:

Franziska Plößl, Tobias Just

Submission:

Journal: International Journal of Housing Markets and Analysis

Submission Date: 07/20/2022

Current Status: Under Review

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.

Black, F. (1986), "Noise", *The Journal of Finance*, Vol. 44 No. 3, pp. 528–543.

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. (2008), "Commercial Real Estate Valuation: Fundamentals Versus Investor Sentiment", *SSRN Electronic Journal*.

Das, P.K., Freybote, J. and Marcato, G. (2014), "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.

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.

Just, T., Voigtländer, M., Eisfeld, R., Henger, R.M., Hesse, M. and Toschka, A. (2017), "Wirtschaftsfaktor Immobilien 2017", Gesellschaft für Immobilienwirtschaftliche Forschung e. V., Deutscher Verband für Wohnungswesen, Städtebau und Raumordnung e. V., Bundesarbeitsgemeinschaft Immobilienwirtschaft Deutschland und Haus & Grund Deutschland - Zentralverband der Deutschen Haus-, Wohnungs- und Grundeigentümer e. V., *IREBS International Real Estate Business School an der Universität Regensburg*.

Kearney, C. and Liu, S. (2014), "Textual Sentiment in Finance: A Survey of Methods and Models", *International Review of Financial Analysis*, Vol. 33, pp. 171–185.

Loughran, T. and McDonald, B. (2020), “Textual Analysis in Finance”, *Annual Review of Financial Economics*, Vol. 12, pp. 357–375.

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.

Nowak, A. and Smith, P. (2017), “Textual analysis in real estate”, *Journal of Applied Econometrics*, Vol. 32 No. 4, pp. 896–918.

Ruscheinsky, J.R., Kandlbinder, K., Schaefers, W., Dietzel, M.A. and Rochdi, K. (2018), “Predicting Real Estate Market Movements: the First Textual Analysis-Based-Sentiment Application in Germany”, *Working Paper*.

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

Soo, C.K. (2018), “Quantifying sentiment with news media across local housing markets”, *Review of Financial Studies*, Vol. 31 No. 10, pp. 3689–3719.

Walker, C.B. (2016), “The direction of media influence: Real-estate news and the stock market”, *Journal of Behavioral and Experimental Finance*, Vol. 10, pp. 20–31.

White, D.M. (1950), “The "gate keeper": A case study in the selection of news”, *Journalism Quarterly*, Vol. 27, pp. 383–390.

2. Cyclicity of Real Estate-related Trends: Topic Modelling and Sentiment Analysis on German Real Estate News

2.1 Abstract

Purpose — The purpose of this paper is to identify and analyse the news coverage and sentiment of real estate-related trends in Germany. Trends are considered as being stable and long-term. If the news coverage and sentiment of trends underlies cyclicity, this could impact investors' behaviour. For instance, in the case of increased reporting on sustainability issues, investors may be inclined to invest more in sustainable buildings, assuming that this is of growing importance to their clients. Hence, investors could expect higher returns when a trend topic goes viral.

Design/methodology/approach — With the help of topic modelling, incorporating seed words partially generated via word embeddings, almost 170,000 newspaper articles published between 1999 and 2019 by a major German real estate news provider are analysed

and assigned to real estate-related trends. This dataset is then analysed based on whether the tone of the news coverage of a specific trend is subject to change.

Findings — The articles concerning urbanisation and globalisation account for the largest shares of reporting. However, the shares are subject to change over time, both in terms of news coverage and sentiment. In particular, the topic of sustainability illustrates a clearly increasing trend with cyclical movements throughout the examined period. Overall, the digitalisation trend has a highly positive connotation within the analysed articles, while regulation displays the most negative sentiment.

Originality — To the best of the authors' knowledge, this is the first application to explore German real estate-newspaper articles regarding the methodologies of word representation and seeded topic modelling. The integration of topic modelling into real estate analysis provides a means through which to extract information in a standardised and replicable way. The methodology can be applied to several further fields like analysing market reports, company statements or social media comments on real estate topics. Finally, this is also the first study to measure the cyclicity of real estate-related trends by means of textual analysis.

Keywords — Trend Analysis, Textual Analysis, Word Embeddings, Topic Modelling, Sentiment Analysis, Machine Learning.

2.2 Introduction

According to Black, the effects of noise on our views of the world are profound. Noise implies uncertainty about future market movements which generates or amplifies market inefficiencies and thus enables trading in financial markets (Black, 1986). In the last few years, both a semantic and a conceptual shift have been attributed to the concept of noise. What Black laconically described as 'noise' more than 30 years ago is considered rather a sentiment today which can be described as being an investor's soft interpretation of different quantitative and qualitative information sets. Now there is ample evidence that sentiment has a profound impact on decisions and therefore on market developments (e.g., Tetlock, 2007; Ferguson *et al*., 2015; Liu, 2015). Understanding the shift factors for investor sentiment can be conducive to the development of more reliable analyses and forecasts of market indicators compared with simply considering quantitative indicators.

The concept of sentiment is related to Robert Shiller's idea of economic narratives (Shiller, 2017), as an economic narrative is the attempt to interpret a vast and vague set of rather qualitative information. Shiller argues that popular stories, with narratives reaching out to a large number of people, shape economic outcomes as they influence individual decision-making. These narratives do not need to be validated in a higher scientific sense; it suffices for them to possess mass appeal. As journalists are gatekeepers of information, both their selection and interpretation of news can lead to the rapid distribution of narratives. Hence, it is important to understand movements in these interpretations and representations by journalists. Or, as White describes it, "…that in his position as a 'gate keeper' the newspaper editor sees to it (…) that the community shall hear as a fact only those events which the newsman, as the representative of his culture, believes to be true.” (White, 1950, p. 390). With respect to real estate markets, Shiller illustrates his point by referring to historical cases, in which the stories of ever-increasing housing or land prices have enticed people towards imprudent investment decisions leading to speculative bubbles. These narratives typically comprise two different elements: First, it is important to understand the selection and frequency of a specific topic or narrative within the news flow. Second, the tonality of this topic is also important. The ability to quantify both the intensity and tonality of how a topic is reported can subsequently help the estimation of potential sentiment formation patterns on the news recipients’ part.

This duality is explicitly acknowledged in the methodological framework of this paper by combining both textual sentiment-analysis and topic modelling. The study therefore contributes to the growing literature which focusses on textual sentiment and, with methods developed in the field of natural language processing and computational linguistics, it is possible to broaden the view of the role of sentiment. This is included through automated topic analysis, which serves to identify topics that occur in a text corpus. In this vein, an in-depth analysis of several major narratives is performed — or real estate-related trends in the case of German real estate-news.

Identifying a relevant trend is not trivial. In scientific analysis, a megatrend is defined as being a long-term development that has started in the past and is expected to continue for several years to come. Megatrends have a significant impact on the economy, society or the environment, and cannot be influenced by individuals (Naisbitt, 1982; Groddeck and Schwarz, 2013). For real estate markets, many trends have been hitherto analysed (e.g., Zeitner and Peyinghaus, 2014; Saiz and Salazar, 2017). While any of these trends consists of several sub-trends (Vejlgaard, 2008), this article will focus on the broader concepts that are likely to be deemed relevant to real estate professionals. Therefore, the selection of real estate-related trends in this article follows Pfnür and Wagner (2020), who, by means of a comprehensive survey among 249 German real estate professionals, have isolated the six major trends most important to the German real estate industry. These six trends are globalisation, socio-demographic change, urbanisation, sustainability, digitalisation and changes in regulatory environment.

This selection is supported by the broad spectrum of literature on the impacts of any of these trends on real estate markets. There are many studies which focus on the impacts of globalisation (e.g., Currit and Easterling, 2009; Falkenbach and Toivonen, 2010), demographic change (e.g., Poterba, 2001; Brounen and Eichholtz, 2004; Just, 2013), urbanisation (e.g., Zhang, 2001; Bart, 2010), sustainability (e.g., Fuerst and McAllister, 2011; Toivonen and Viitanen, 2016) and digitalisation (e.g., Staub *et al*., 2016; Moring *et al*., 2018).

Every news article published between 1999 and 2019 in Immobilien Zeitung (*IZ*), a leading real estate newspaper for the German market, has been studied with a focus on these six trends, both in terms of news coverage and sentiment. For the given text corpus, the following questions are addressed: How has news coverage of these six major trends developed over time in terms of quantitative media attention? Is there a difference in tonality in articles between the six trends? Does this tonality change over time? Answers to these questions add to the understanding of how investor sentiment in real estate markets is emerging and changing over time, and how quickly and strongly a long-term narrative changes over time. The methodology to answer these questions presented in this paper extracts information from text in a standardised and replicable way. In future research, this methodology can be applied to a variety of research questions in the real estate industry using a broader set of text corpora.

The paper is organised as follows: Section 2.3 summarises the literature on topic modelling and sentiment analysis. The dataset is presented in section 2.4. In order to assign newspaper articles to the trends, the method of topic modelling with integrated seed words (section 2.5.2), partially generated via word embeddings (section 2.5.1), is presented. To examine trend-specific sentiment, a polarity score-based sentiment measure is provided (section 2.5.3). Section 2.6 shows the results with section 2.7 concluding and discussing the implications of the findings for the industry.

2.3 Literature Review

Topic modelling primarily concerns the derivation of underlying common topics — in this case real estate-related trends — from large bodies of unstructured textual data. This methodology for content analysis has been applied in different fields such as social media analytics (e.g., Zhao *et al*., 2011; Ghosh and Guha, 2013; Nordheim *et al*., 2018) or finance (e.g., Nguyen and Shirai, 2015; Cerchiello and Nicola, 2018; Aziz *et al*., 2019). The most common model, named Latent Dirichlet Allocation (*LDA*) (Blei *et al*., 2003), or extensions such as the Structural Topic Model (*STM*) (Roberts *et al*., 2013), are mainly used where *LDA* is a generative probabilistic model for collections of discrete data such as text corpora. However, there is only a limited amount of academic research which applies to topic models in the context of real estate market analysis. Winson-Geideman and Evangelopoulos (2013) used Latent Semantic Analyses (*LSA*) to structure the body of published research within four real estate-related journals over a period of almost 40 years. They utilised Singular Value Decomposition to extract socially constructed components of meaning. In addition, Evangelopoulos *et al*. (2015) presented methods and applications of *LSA* in order to investigate unstructured data in real estate research. Koelbl *et al*. (2020) applied *STM* to determine risk-factor topics discussed in the 10-K filings of US REITs and examined whether and how these topics affect the risk perceptions of investors.

Even now, there remains a lack of applications exploring real estate-related text corpora, such as news, especially for the German-speaking regions.

While topic modelling is still in its infancy in the field of real estate analysis, there is a rapidly growing body of literature on how to use textual analysis for measuring sentiment. Traditionally, investor sentiment has been measured by surveys or indicators derived from capital market data (e.g., Das *et al*., 2014; Freybote, 2016). Survey data usually suffers from selection bias and the problem that interviewees lack both any incentive or ability to reveal their true preferences. Sentiment derived from capital market indicators overstretches the concept of sentiment as it is difficult to disentangle the relative impact of soft and hard data. Thus, researchers have explored indirect measures of sentiment in real behaviour and text, i.e., in internet search volumes (e.g., Hohenstatt *et al*., 2011; Beracha and Wintoki, 2013) and through the examination of the potential tonality of texts. The latter is governed by two basic approaches. It can be performed either by using classifications through machine learning (e.g., Li, 2010; Hausler *et al*., 2018) or by using appropriate dictionaries. The dictionary-based approach refers to word lists which contain a sentiment category for each word, which tend to be either positive or negative in the majority of the studies (Feldman, 2013). A series of papers followed the seminal work by Tetlock (2007), who established this approach in the field of finance (for an overview, see Kearney and Liu, 2014). By building on a simple dictionary-based approach, Tetlock showed that media pessimism in articles of the Wall Street Journal column could help to predict downward pressure on stock market prices. In the real estate literature, this approach has been applied by Doran *et al*. (2012), who found that the tone of quarterly-earnings conference calls for publicly traded REITs has significant explanatory power for contemporaneous stock prices. Walker (2016) also investigated the relationship between the news media and shares of companies engaged in the housing market. Nowak and Smith (2017) ascertained that the comments section of the MLS listings provides information which improves the performance of hedonic pricing-estimates. Soo (2018) demonstrated that housing media sentiment has significant predictive power for future house prices which lead by nearly two years. Likewise, Ruscheinsky *et al*. (2018b) found a leading relationship between media sentiment and future REIT market movements. Moreover, Koelbl (2020) detected that higher levels of pessimistic (optimistic) language in the Management Discussion and Analysis of US REITs predict lower (higher) future firm performance. The first application of text-based sentiment analysis which focusses on the German real estate market was conducted by Ruscheinsky *et al*., (2018a) through the development of a domain-specific dictionary. This study also revealed a significant relationship between the extracted sentiment and the development of German residential real estate prices.

2.4 Data

The dataset for the textual analysis in this study contains all articles of the *IZ* published between January

1999 and December 2019. This news provider is chosen due to its ranking as one of the leading German real estate-related newspapers and the fact that its large digital archive covers a long time period. The *IZ* was founded in 1993 with a current circulation of 9,776 weekly issues (Q2-2020), a majority of 8,674 being sold to subscribers.[1] Moreover, the online visits averaged almost 850,000 per month in 2019.[2]

Pre-processing of the data involves the removal of punctuation marks, numbers, non-alphabetical and special characters and stop words. 'Stop words' relate to frequently occurring words which 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. More than 2,000 stop words are excluded from this analysis. Furthermore, illustrations, tables, English articles and editorial shortcuts are also excluded. The data is tokenised for the ensuing tasks. This process divides

[1] See https://cdn.iz.de/media/documents/iz-digital2020-de-v1.pdf.
[2] See https://www.immobilien-zeitung.de/mediadaten.

text into units (tokens) such as phrases, words and other meaningful entities. In this case, the text corpus is segmented by words.

The final dataset contains 168,012 articles with approximately 38 million words in total. The articles manifest a median length of 128 words, ranging from 35 to 4,253 words per article. Accordingly, due to its relatively short articles, the *IZ* provides compact industry-related news. Throughout the period covered in this paper, the *IZ* has increased in size: In 1999, the number of published articles amounted to 616 per quarter, which increased to 3,141 in 2019.

2.5 Methodology

2.5.1 Global Vectors Model

Apart from the numbers of topics, most topic model algorithms, such as *LDA*, extract topics without any prior specification by the researcher. This feature is a major advantage if the aim is to explore unknown texts. In this study though, the topics are known in advance, i.e., the six trends. As a consequence, an extension of *LDA* is applied, namely the seeded *LDA* model by Jagarlamudi *et al*. (2012), which allows the model to create topics around certain seed words provided by the researcher (see below).

Accordingly, the first step in the analysis consists of creating a list of appropriate seed words for each trend. Watanabe and Zhou (2020) argue that seed words should be both knowledge-based and frequency-based. However, King *et al*. (2017) demonstrate that even expert humans perform poorly and are unreliable regarding seed word-selection. Therefore, a word embeddings-approach, the Global Vector model (*GloVe*) introduced by Pennington *et al*. (2014) is applied as this method can be used to identify relationships between words such as synonyms. Given a text corpus of two sentences: ‘Company X bought a property in London.’ and ‘Corporation Y purchased a building in Berlin.’, the algorithm can detect for example the tokens ‘property’ and ‘building’ or ‘bought’ and ‘purchased’ etc. for word representation. Thus, for each trend proper seed words will be extracted from the newspaper articles by means of the *GloVe* model.

This unsupervised learning algorithm combines the advantages of global matrix factorisation and local context window methods to generate vectors for word representations. Matrix factorisation methods on the one hand decompose large matrices which capture statistical information of a text corpus and generate word representations of lower dimensional latent space in order to reduce computation time. While the concept of context window methods on the other hand is to predict linguistic patterns as linear relationships between the word vectors based on local context windows and perform better on word analogy tasks. Therefore, training of the *GloVe* algorithm is conducted on the basis of 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. Thus, the model generates a vector space of meaningful substructures. The objective function J of the weighted least squares regression model is specified as follows:

$$J = \sum_{i,j=1}^{V} f\big(X_{ij}\big)(w_i^T\, \widetilde{w}_j + b_i + \tilde{b}_j - log\, X_{ij})^2 \qquad (2.1)$$

where *V* is the size of vocabulary; X_{ij} 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. As word pairs exhibit different occurrence frequencies in the text corpus, a weighting function readjusts the cost for each word pair to prevent learning simple common word pairs. Given a co-occurrence count higher 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 detailed description). Lastly, the learning algorithm draws word vectors that are similar to the word vectors of the trends to be examined within the given text corpus.

2.5.2 Seeded LDA Model

In a second step, the seeded *LDA* model by Jagarlamudi *et al*. (2012) is used for topic classification of the newspaper articles as mentioned above. This seeding approach, combined with interactive topic modelling, fosters the examination of the text corpus while simultaneously guiding the topic discovery by prior selection of seed words. News coverage per trend can be calculated subsequently. In contrast to the traditional *LDA* model, the seeded *LDA* model improves topic-word and document-topic distribution through the use of seed words s, which means that the definition of each topic k as a Dirichlet-multinomial distribution is extended to a mixture of two distributions, where the parameter π controls the probability of drawing a word from the regular topic distribution ϕ_k^r versus the seed topic distribution ϕ_k^s. The generative story of the seeded *LDA* model is given by Jagarlamudi *et al*. (2012) as follows:

1. For each $k = 1 \cdots T$,
   (a) Choose regular topic $\phi_k^r \sim$ Dirichlet(β_r)
   (b) Choose seed topic $\phi_k^s \sim$ Dirichlet(β_s)
   (c) Choose $\pi_k \sim$ Beta(1, 1)
2. For each seed set $s = 1 \cdots S$,
   (a) Choose group-topic distribution $\psi_s \sim$ Dirichlet(α)
3. For each document d,
   (a) Choose a binary vector $\vec{b}$ of length S
   (b) Choose a document-group distribution $\zeta^d \sim$ Dirichlet($\tau\vec{b}$)
   (c) Choose a group variable g ∼ Multinomial(ζ^d)
   (d) Choose $\theta_d \sim$ Dirichlet(ψ_g) // of length T
   (e) For each token $i = 1 \cdots N_d$:
      i. Select a topic $z_i \sim$ Multinomial(θ_d)
      ii. Select an indicator $x_i \sim$ Bernoulli(π_{z_i})
      iii. if x_i is 0
         • Select a word $w_i \sim$ Multinomial($\phi_{z_i}^r$)
      iv. if x_i is 1
         • Select a word $w_i \sim$ Multinomial($\phi_{z_i}^s$)

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; θ (ζ) describes the document-topic (document-group) distribution and α, β, τ are hyperparameters that are used to control the learning process (see *Seeded LDA Model* in the appendix for detailed description).

By conducting the described model, the suitability of the seeded topic modelling-approach is in terms of extracting the news coverage of trends from real estate-related news as well as the variability of news coverage over time will be tested.

2.5.3 Unbounded Polarity Score

In order to measure the news sentiment in a third step, the dictionary-based approach is chosen for three reasons. Firstly, it reduces the subjectivity of researcher decisions, secondly the methodology is scalable and lastly it can be easily replicated (Loughran and McDonald, 2016). The calculation of a polarity score at document level requires two dictionaries: one dictionary which classifies words as being either positive or negative (polarity) and another dictionary which contains valence shifters (negators, amplifiers, de-amplifiers and adversative conjunctions). These valence shifters can strengthen, weaken or even flip a word’s prior polarity. To cite an example, the word ‘good’ exhibits a prior polarity of +1, which would be changed to -1 if preceded by ‘not’ or increased if preceded by ‘very’. Therefore, each document is broken into an ordered set of words, and the tonality of an overall document is denoted by the weighted polarity of the words contained. The equation used by the algorithm is given by Rinker (2019) as follows:

$$\delta = \frac{c'_{i,j}}{\sqrt{w_{i,jn}}} \tag{2.2}$$

where δ represents the unbounded polarity score; $c'_{i,j}$ indicates the summed weighted context clusters around the polarised words and $w_{i,jn}$ is the word count (see *Unbounded Polarity Score* in the appendix for detailed description).

By assigning a polarity value to each newspaper article, the suitability of the sentiment analysis for classifying the given trends as optimistic or pessimistic based on real estate-related news as well as the variability of the news sentiment over time will be investigated.

2.6 Results

2.6.1 Global Vectors Model

For the seed word selection, the *GloVe* model generates word representations of the six trends on the basis of the given text corpus, presented in table 2.1. 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.

Table 2.1 Word Representations of Trends

Trend	***Word Vectors***
Globalisation	internationalisation [0.343]; urbanisation [0.263]; global [0.251]; digitalisation [0.247]; professionalisation [0.246]; uncertainty [0.237]; capital markets [0.232]; harmonisation [0.232]; diversification [0.223]; recession [0.219]
Demography	demographic [0.365]; depopulation [0.323]; immigration [0.309]; housing demand [0.299]; housing conditions [0.253]; population ageing [0.249]; societal [0.246]; change [0.252]; population [0.226]; provision for old age [0.213]
Urbanisation	population growth [0.304]; urban sprawl [0.275]; rent growth [0.274]; rural depopulation [0.248]; rent increase [0.245]; shortage [0.245]; influx [0.234]; upheaval [0.232]; urbanity [0.231]; housing shortage [0.226]
Sustainability	sustainable [0.581]; energy efficiency [0.482]; ecological [0.473]; green [0.460]; certification [0.438]; cost-effectiveness [0.437]; criteria [0.425]; dgnb [0.387]; standards [0.382]; climate protection [0.358]
Digitalisation	digital [0.560]; innovations [0.452]; process [0.450]; ecommerce [0.445]; solutions [0.436]; business models [0.397]; technologies [0.384]; challenges [0.380]; transformation [0.374]; proptechs [0.339]
Regulation	capital investment code [0.410]; regulations [0.409]; introduction by law [0.392]; bafin (German Federal Financial Services Regulatory Authority) [0.367]; transparency [0.353]; policies [0.337]; legislator [0.303]; tightening of the law [0.440]; organisations [0.295]

Notes: Table 2.1 displays results of the *GloVe* model with 10 word vectors representing the one for each trend. The degree of similarity measured as a percentage is shown in brackets.
Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH.

For instance, the word vector ‘internationalisation’ provides the highest correspondence level of 34.3 % for ‘globalisation’. For the vector ‘sustainability’, for example, ‘sustainable’ (58.1 %) and ‘energy efficiency’ (48.2 %) possess high levels of similarity. Word representations for ‘digitalisation’ are given by ‘digital’ (56.0 %), ‘innovations’ (45.2 %) and ‘process’ (45.0 %). Selecting reliable seed words in order to perform the following task of seeded topic modelling results in more than 50 seed words for each trend (see table 2.6 in the appendix).

2.6.2 Seeded LDA Model

By using the created seed word dictionary, each article is assigned to the most likely topic/trend by the algorithm. Table 2.2 provides the most probable words which appear in the text corpus for each trend. One additional topic is created during this process containing all articles (15.2 %) that could not be assigned to any of these trends. Thus, the process implies that articles cannot be assigned to more than one trend.

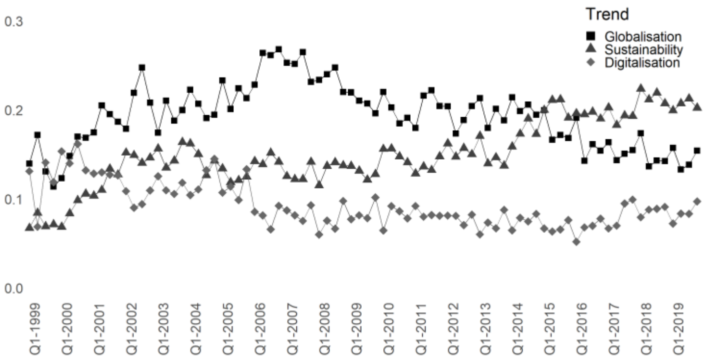
Table 2.2 Description of Seeded Topics

Trend	***Most probable Words***
Globalisation	funds [17,800]; asset [10,051]; real estate-funds [7,952]; portfolio [8,516]; investor [5,859]; europe [7,431]; equity [5,280]; volume [5,536]; global [3,212]; portfolios [3,791]
Demography	cities [6,685]; region [5,663]; stock [7,946]; decrease [3,616]; trend [3,596]; growth [4,521]; returns [2,645]; regions [2,591]; investments [5,677]; inhabitants [2,909]
Urbanisation	city [22,339]; areal [9,757]; district [7,441]; area [6,457]; district development [5,168]; old [5,518]; city centre [7,301]; plan [5,032]; centre [4,486]; city (the word *city* is also used in German) [8,071]
Sustainability	facility management [2,877]; facility [3,926]; service provider [4,306]; international [6,680]; headquarters [5,296]; services [3,557]; energy [2,433]; asset [10,051]; industry [2,930]; technical [3,023]
Digitalisation	online [3,559]; data [3,175]; internet [2,940]; secure [4,207]; information [6,042]; success [3,776]; system [1,645]; quality [4,093]; platform [1,783]; solutions [1,924]
Regulation	federal court [1,892]; policies [3,690]; municipality [3,387]; judgement [3,186]; spd (German party) [3,184]; contract [5,185]; cooperative building company [3,548]; principal [2,076]; federal government [2,431]; organization [3,012]

Notes: Table 2.2 displays results of the seeded *LDA* model with 10 most probable words per trend. The general frequency of the words in the text corpus is shown in brackets.
Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH.

From 1999 to 2019, the trend urbanisation is the most frequently reported upon, accounting for 19.22 % of the text corpus. This topic is dominated by words relating to urban development such as ‘city’, ‘areal’ or ‘district development’. The trend globalisation also comprises almost a fifth of all articles (19.04 %) and is characterised by ‘funds’, ‘asset’ or ‘portfolio’. With 16.17 % of the articles, sustainability is the third most reported upon trend. Articles relating to demographic change account for 11.29 % of the dataset while the model attributes 10.32 % of the articles to regulation. With only 8.60 % of the articles, the trend digitalisation is the least reported.

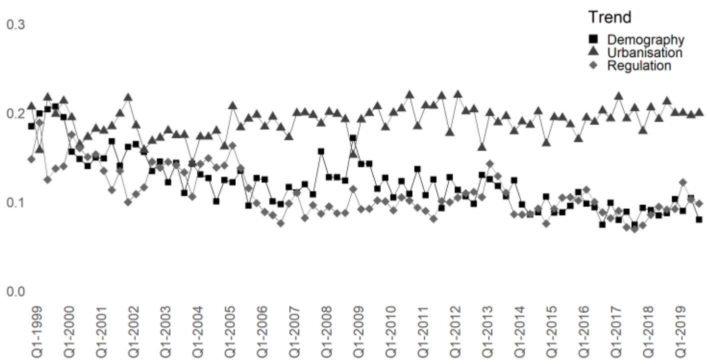
Figure 2.1 Article Shares over Time (1)



Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH.

Figures 2.1 and 2.2 show the development of the article shares on a quarterly basis. Hence, in relative terms, a strong upward trend can be seen over time, especially for articles on sustainability issues. The news coverage of urbanisation accounts for a high proportion of the text corpus throughout the entire period, whereas the shares for demography and regulation are declining. Reports on globalisation demonstrated an upward trend until the financial crisis; since then, reporting has been significantly less frequent. At the time of the dotcom bubble, articles relating to digitalisation were frequently published, but reporting on this topic has remained at a relatively low level in subsequent years. Remarkably, this also remains valid for the very recent past.

Figure 2.2 Article Shares over Time (2)



Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH.

Nevertheless, the results show (short-term) fluctuations in the news coverage, varying in intensity for the individual trends. Therefore, the key descriptive statistics of the article shares on a quarterly basis are provided in table 2.3. The standard deviation (SD), as well as the amplitude (ampl), measured as being the difference between the highest and lowest share for each trend, are strongest for globalisation (SD: 0.0357; ampl: 15.42) and sustainability (SD: 0.0303; ampl: 15.36), whereas this difference is weakest for urbanisation (SD: 0.0165; ampl: 7.64). For comparison, the topic containing all other articles is characterised by a high standard deviation as well. The statistics imply that sustainable topics are discussed more frequently during some periods and significantly less in others, while the news flow on urbanisation is significantly more stable throughout the entire period covered. In the case of globalisation, for example, the sharpest decline within a year can be seen from 2008 to 2009 after the financial crisis, while the strongest increase in sustainable topics can be found from 2002 to 2003 following the introduction of the German Energy Saving Regulation.

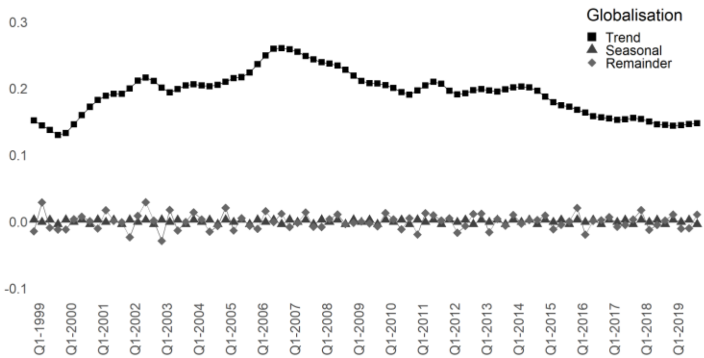
Table 2.3 Descriptive Statistics of Article Shares over Time

Trend	*SD*	*Amplitude*	*Min*	*Max*
Globalisation	0.0357	15.42	11.38 %	26.80 %
Demography	0.0303	13.29	7.43 %	20.71 %
Urbanisation	0.0165	7.64	14.36 %	22.00 %
Sustainability	0.0374	15.63	6.72 %	22.36 %
Digitalisation	0.0244	10.99	5.18 %	16.17 %
Regulation	0.0259	11.96	6.94 %	18.90 %
Other	0.0357	14.19	7.41 %	21.60 %

Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH.

In a next step, the generated time series has been decomposed into trend, seasonal and remainder components using the Seasonal-Trend Decomposition Procedure based on Loess in order to consider seasonal effects (Cleveland *et al*., 1990). This method is applied as any type of seasonality can be handled and the seasonal component is allowed to change over time. After excluding the seasonal and remainder component, the results are confirmed as the highest standard deviations are given by sustainability (SD: 0.0361), globalisation (SD: 0.0326) and demography (SD: 0.0270), followed by regulation, digitalisation and urbanisation. Thus, there are longer-term shifts in reporting on some of the underlying trends and not only seasonal effects. Figure 2.3 illustrates the decomposed time series for globalisation.

Figure 2.3 Article Share over Time — Seasonal-Trend Decomposition



Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH.

The results of the seeded *LDA* model show that the news coverage of trends by the *IZ* varies over time. This could be of relevance to real estate investors who focus on properties that largely depend on one specific trend. For instance, the investment in nursing home-facilities in times when the demographic topic is intensively covered by the media, urban residential properties in phases with a vast amount of reporting on urbanisation or properties with a good energy standard in periods with media focus on sustainability. However, ultimately, investor sentiment is not only shaped by a topic's frequency, but also by the tonality of the respective articles upon it.

2.6.3 Unbounded Polarity Score

Several dictionaries in different languages have been developed for the purpose of measuring the tonality of specific words, some even specifically for the economic domain. However, each industry has its own idiosyncratic word meanings, which cannot be directly transferred and applied to other industries. Consequently, industry-specific dictionaries can reduce measurement errors (Loughran and McDonald, 2011). A real estate-specific German dictionary was created by Ruscheinsky *et al*. (2018a), which contains 8,144 negative words (-1) and 5,745 positive words (+1). It was validated by a representative survey of real estate professionals. Given this input, the tonality of the articles in the dataset can be approximated by counting the polarised words and by taking into account the valence shifters. The shifters are manually documented in another dictionary with 76 entries forming a second input.

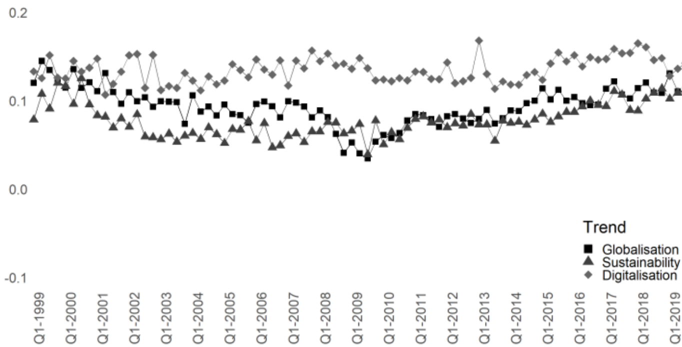
Table 2.4 Descriptive Statistics of Polarity Scores

Trend	***Mean***	***Median***	***SD***	***Min***	***Max***
Globalisation	0.0923	0.0923	0.1328	-0.5802	0.7811
Demography	0.1134	0.1235	0.1378	-0.5714	0.8247
Urbanisation	0.1009	0.1049	0.1165	-0.4851	0.7020
Sustainability	0.0815	0.0645	0.1181	-0.6495	0.8201
Digitalisation	0.1360	0.1457	0.1306	-0.7183	0.7131
Regulation	-0.0072	0.0000	0.1617	-0.7765	0.6406
Other	0.0860	0.0868	0.1247	-0.7671	0.7414
Total	0.0868	0.0919	0.1348	-0.7765	0.8247

Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH.

Table 2.4 summarises the descriptive statistics of the resulting polarity scores per trend. By and large, articles on digitalisation achieve the highest polarity score and average at 0.1360, followed by demography (0.1134), urbanisation (0.1009), globalisation (0.0909) and sustainability (0.0815). With a score of -0.0072, regulation is the only trend with a weakly negative connotation.

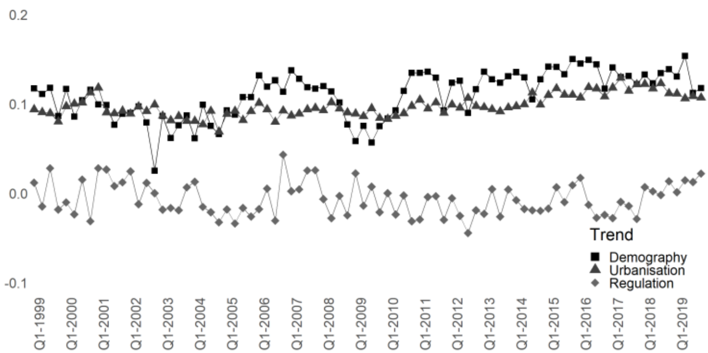
Figure 2.4 Polarity Scores over Time (1)



Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH.

With an average value of 0.0868, the overall reporting of the *IZ* can be considered as being neutral. This result is not surprising, for at least two reasons. First, polarity-bearing parts such as editorials, guest features and other opinion pieces comprise only a section of the newspaper. Additionally, the dictionary does not measure opinion per se, but general tonality instead. Even a strong opinion can be coated in a comparatively neutral style. Still, it seems plausible that opinion pieces show higher sentiment scores than news. Secondly, outlets such as the *IZ* or other professional newspapers can be expected to be fact-oriented, especially compared with mass media. However, table 2.4 as well as figures 2.4 and 2.5, exhibit considerable differences between the six trends which contain valuable information for investors.

Figure 2.5 Polarity Scores over Time (2)



Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH.

Over the course of 21 years, articles associated with regulation have clearly been written in a more negative tone. What is more, figures 2.4 and 2.5 show that the development of the polarity scores for the trends are slightly cyclical, too. After the financial crisis, most trends, particularly those of globalisation and demography, experienced a temporary drop in polarity values, while others, especially those of urbanisation and digitalisation, have fared much better during the crisis. In the last few years, the overall tonality of articles on trends has again improved, likely reflecting the strong market-environment for German real estate investments.

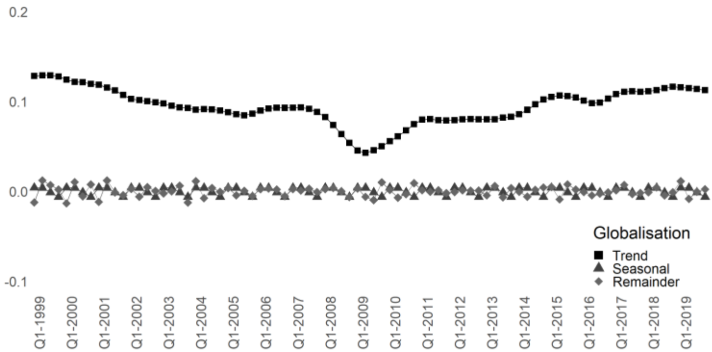
Table 2.5 Descriptive Statistics of Polarity Scores over Time

Trend	***SD***	***Amplitude***	***Min***	***Max***
Globalisation	0.0217	11.03	3.45 %	14.48 %
Demography	0.0260	12.85	2.50 %	15.34 %
Urbanisation	0.0123	6.00	6.88 %	12.88 %
Sustainability	0.0189	8.59	3.91 %	12.50 %
Digitalisation	0.0139	6.11	10.67 %	16.78 %
Regulation	0.0187	8.76	-4.46 %	4.31 %
Other	0.0181	8.25	4.19 %	12.44 %

Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH.

Table 2.5 provides an overview of the descriptive statistics of the aggregated polarity scores per trend on a quarterly basis. The values for the trend demography (SD: 0.0260; ampl: 12.83) and globalisation (SD: 0.0217; ampl: 11.03) fluctuate the strongest, while the amplitude in tonality is smallest for urbanisation (SD: 0.0123; ampl: 6.01) and digitalisation (SD: 0.0139; ampl: 6.11). After adjusting for the seasonal and remainder component, the results are confirmed as they are primarily driven by a change in the trend component, see figure 2.6 for globalisation. Demography (SD: 0.0226) and globalisation (SD: 0.0200) show the highest standard deviations, followed by sustainability, regulation, digitalisation and urbanisation.

Figure 2.6 Polarity Scores over Time — Seasonal-Trend Decomposition



Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH.

The results suggest that the sentiment analysis captures the specific tone for each trend and that there are differences in time. In this sense, investors who bet on a globalisation or demographic trend are for more at risk of imputed sentiment in newspaper articles changing and thus also impacting overall market sentiment.

2.7 Conclusion

This study examined how frequently real estate-related trends are discussed and how optimistic or pessimistic they are reflected in real estate news over a period of 21 years. Therefore, the introduced methodological framework is a combination of both topic modelling and textual sentiment analysis. The newspaper articles are assigned to six real estate-related trends by applying the seeded *LDA* model. Urbanisation, globalisation and sustainability account for the largest shares of reporting while digitalisation is the least frequently reported upon, even in the last few years. News coverage of these respective trends is subject to measurable change over time. Particularly regarding sustainability, a clear upward trend can be identified, while the reporting on globalisation of real estate has eroded since the financial crisis.

While overall, reporting in the *IZ* can be classified as being predominantly neutral with regards to tonality there are significant differences between articles on the specific topics. Reporting on digitalisation achieves the highest polarity scores while regulatory issues yield the lowest. What is more, the sentiment measures for the respective trends vary over time, and this is especially valid for demography and globalisation, while reporting on urbanisation and digitalisation appears to follow a form of tonality which is more stable.

As this study examined a large text corpus over a long period of time, the results provide new insights: the news coverage of trends is following short-term and strong movements, although these trends are typically described as being long and stable processes, since the seeded *LDA* algorithm is capable of quantifying the narratives in such a way as to enable statistical analysis, and these findings remain valid after adjusting for seasonal and remainder components. This can impact investor sentiment and thus investment decisions, as investors also build their expectations and assessments on the foundation of media interpretation. In this sense, machine-learning techniques of natural language-processing allow the evaluation of a vast amount of qualitative information which could be beneficial for investors.

The findings of this study reach far beyond the analysis of trends, as the methods of topic modelling and polarity calculus can be applied to many other research questions. First, these results are limited by using a sole data source and comparatively broad and generic topics. The analysis is able to yield additional information when applied to various text corpora. Does the news coverage and sentiment of real estate topics in the mass media or corporate publications follow similar cycles to those in this analysis? Second, the classification of articles depends on the seed words chosen, which could vary using other data. Third, the sentiment measure is based on weightings. Therefore, sentiment analysis could be performed at sentence level or aspect-based. Fourth, the quality and suitability of dictionaries is also crucial for the process of examining the text corpus.

In this vein, this study addresses a plethora of new related research questions in order to evaluate the analysts' views and quantifying qualitative data, i.e., reducing noise in real estate markets. Compared with stock or bond markets, real estate markets are considered to be less transparent and heterogeneous while the presented methods can improve transparency and efficiency in the real estate industry. Furthermore, the relationship between textual measures and proxies for the trends could be examined in future research by investigating whether newspapers provide leading (predicting) or lagging (reporting) information. Then it would prove reasonable to dismantle the trend bundles into individually underlying trends: the demographic trend consists of secular developments in fertility, mortality and migration; urbanisation trends can affect different asset classes very differently. Thus, further research could investigate whether the news coverage or sentiment is related to market movements of specific asset classes.

To the best knowledge of the authors, this is the first study to apply the seeded *LDA* algorithm to the field of real estate. Quantifying both the news coverage and sentiment could be of value to investors and influence both opinion-building and decision-making as it helps to better understand the overall market narrative dominating during a specific market phase. Thus, a wide range of research projects are expected to build on and improve the approach undertaken in this study.

2.8 References

Aziz, S., Dowling, M.M., Hammami, H. and Piepenbrink, A. (2019), "Machine Learning in Finance: A Topic Modeling Approach", available at SSRN: https://ssrn.com/abstract=3327277 or http://dx.doi.org/10.2139/ssrn.3327277.

Bart, I.L. (2010), "Urban sprawl and climate change: A statistical exploration of cause and effect, with policy options for the EU", *Land Use Policy,* Vol. 27 No. 2, pp. 283–292.

Beracha, E. and Wintoki, M.B. (2013), "Forecasting Residential Real Estate Price Changes from Online Search Activity", *Journal of Real Estate Research*, Vol. 35 No. 3, pp. 283–312.

Black, F. (1986), "Noise", *The Journal of Finance,* Vol. 44 No. 3, pp. 528–543.

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

Brounen, D. and Eichholtz, P. (2004), "Demographics and the global office market—Consequences for property portfolios", *Journal of Real Estate Portfolio Management,* Vol. 10 No. 3, pp. 231–242.

Cerchiello, P. and Nicola, G. (2018), "Assessing News Contagion in Finance", *Econometrics*, Vol. 6 No. 1, pp. 1–19.

Cleveland, R.B., Cleveland, W.S., McRae, J.E. and Terpenning, I. (1990), "STL: A Seasonal-Trend Decomposition Procedure Based on Loess", *Journal of Official Statistics,* Vol. 6 No. 1, pp. 3–73.

Currit, N. and Easterling, W.E. (2009), "Globalization and population drivers of rural-urban land-use change in Chihuahua, Mexico", *Land Use Policy*, Vol. 26 No. 3, pp. 535–544.

Das, P.K., Freybote, J. and Marcato, G. (2014), "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.

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. 355–380.

Falkenbach, H. and Toivonen, S. (2010), "Effects of international investments on the Finnish commercial real estate market", *Journal of Real Estate Literature*, Vol. 18 No. 2, pp. 313–328.

Feldman, R. (2013), "Techniques and Applications for Sentiment Analysis", *Communications of the ACM*, Vol. 56 No. 4, pp. 82–89.

Ferguson, N.J., Philip, D., Lam, H.Y.T. and Guo, J. (2015), "Media Content and Stock Returns: The Predictive Power of Press", *Multinational Finance Journal*, Vol. 19 No. 1/2, pp. 1–31.

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

Fuerst, F. and McAllister, P. (2011), "Green Noise or Green Value? Measuring the Effects of Environmental Certification on Office Values", *Real Estate Economics,* Vol. 39 No. 1, pp. 45–69.

Ghosh, D.D. and Guha, R. (2013), "What are we 'tweeting' about obesity? Mapping tweets with Topic Modeling and Geographic Information System", *Cartography and Geographic Information Science*, Vol. 40 No. 2, pp. 90–102.

Groddeck, V. von and Schwarz, J.O. (2013), "Perceiving megatrends as empty signifiers: A discourse-theoretical interpretation of trend management", *Futures,* Vol. 47, pp. 28–37.

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.

Hohenstatt, R., Käsbauer, M. and Schäfers, W. (2011), “'Geco' and its Potential for Real Estate Research - Evidence from the US Housing Market”, *Journal of Real Estate Research*, Vol. 33 No. 4, pp. 471–506.

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*, USA, pp. 204–213.

Just, T. (2013), *Demografie und Immobilien*, 2nd ed., Oldenbourg, München.

Kearney, C. and Liu, S. (2014), “Textual Sentiment in Finance: A Survey of Methods and Models”, *International Review of Financial Analysis*, Vol. 33, pp. 171–185.

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

Koelbl, M., Schuierer, R., Steininger, B.I. and Schaefers, W. (2020), “Is Mandatory Risk Reporting Informative? Evidence from US REITS Using Machine Learning for Text Analysis”, available at SSRN: https://ssrn.com/abstract=3686492.

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.

Liu, S. (2015), “Investor Sentiment and Stock Market Liquidity”, *Journal of Behavioral Finance*, Vol. 16 No. 1, pp. 51–67.

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.

Moring, A., Maiwald, L. and Kewitz, T. (2018), *Bits and Bricks: Digitalisierung von Geschäftsmodellen in der Immobilienbranche*, Springer Gabler, Wiesbaden.

Naisbitt, J. (1982), *Megatrends: Ten new directions transforming our lives*, Warner Books, New York.

Nguyen, T.H. and Shirai, K. (2015), "Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, Beijing, pp. 1354–1364.

Nordheim, G. von, Boczek, K. and Koppers, L. (2018), "Sourcing the Sources: An analysis of the use of Twitter and Facebook as a journalistic source over 10 years in The New York Times, The Guardian, and Süddeutsche Zeitung", *Digital Journalism*, Vol. 6 No. 7, pp. 807–828.

Nowak, A. and Smith, P. (2017), “Textual analysis in real estate”, *Journal of Applied Econometrics*, Vol. 32 No. 4, pp. 896–918.

Pennington, J., Socher, R. and Manning, C. (2014), “Glove: Global Vectors for Word Representation”, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Qatar, pp. 1532–1543.

Pfnür, A. and Wagner, B. (2020), “Transformation of the real estate and construction industry: empirical findings from Germany”, *Journal of Business Economics*, Vol. 2 No. 1, p. 79.

Poterba, J.M. (2001), “Demographic Structure and Asset Returns”, *The Review of Economics and Statistics*, Vol. 83 No. 4, pp. 565–584.

Rinker, T. (2019), Package ‘sentimentr’: Calculate Text Polarity Sentiment, R 3.4.0, available at: https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf.

Roberts, M.E., Stewart, B.M., Tingley, D. and Airoldi, E.M. (2013), “The Structural Topic Model and Applied Social Science”, *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation.*

Ruscheinsky, J.R., Kandlbinder, K., Schaefers, W., Dietzel, M.A. and Rochdi, K. (2018a), “Predicting Real Estate Market Movements: the First Textual Analysis-Based-Sentiment Application in Germany”, *Working Paper*.

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

Saiz, A. and Salazar, A. (2017), “Real Trends: The Future of Real Estate in the United States”, Capital One Banking/Massachusetts Institute of Technology, available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3100934.

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

Soo, C.K. (2018), “Quantifying sentiment with news media across local housing markets”, *Review of Financial Studies*, Vol. 31 No. 10, pp. 3689–3719.

Staub, P., Stucki, M. and Wettstein, A. (2016), *Digital Real Estate: Bedeutung und Potenziale der Digitalisierung für die Akteure der Immobilienwirtschaft*, SVIT Verlag AG, Zürich.

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.

Toivonen, S. and Viitanen, K. (2016), “Environmental scanning and futures wheels as tools to analyze the possible future themes of the commercial real estate market”, *Land Use Policy*, Vol. 52, pp. 51–61.

Vejlgaard, H. (2008), *Anatomy of a trend*, McGraw-Hill, New York.

Walker, C.B. (2016), “The direction of media influence: Real-estate news and the stock market”, *Journal of Behavioral and Experimental Finance*, Vol. 10, pp. 20–31.

Watanabe, K. and Zhou, Y. (2020), “Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches”, *Social Science Computer Review*, available at: https://journals.sagepub.com/doi/10.1177/0894439320907027.

White, D.M. (1950), “The "gate keeper": A case study in the selection of news”, *Journalism Quarterly*, Vol. 27, pp. 383–390.

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.

Zeitner, R. and Peyinghaus, M. (2014), *Megatrends und ihr Einfluss auf die Immobilienwirtschaft, Process Management Real Estate Monitor*, Hochschule für Technik und Wirtschaft Berlin & pom + international GmbH, Berlin/Frankfurt am Main.

Zhang, T. (2001), “Community features and urban sprawl: the case of the Chicago metropolitan region”, *Land Use Policy*, Vol. 18 No. 3, pp. 221–323.

Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. and Li, X. (2011), “Comparing Twitter and Traditional Media Using Topic Models”, *Advances in Information*, pp. 338–349.

2.9 Appendix

Seeded LDA Model

The seeded *LDA* model by Jagarlamudi *et al*. (2012) improves both topic-word and document-topic probability distributions of the *LDA* model by Blei *et al*. (2003).

Therefore, the traditional definition of each topic *k* as a Dirichlet-multinomial distribution ϕ_k over words is extended to a regular topic ϕ_k^r and a seed topic ϕ_k^s distribution. The seed topic distribution can only generate seed words from the given seed set while the regular topic distribution can generate any word. The following generative process updates the topic assignment of documents.

First, the model generates multinomial distributions for seed and regular topics. The parameter π_k controls the probability of drawing a word from the seed topic distribution or the regular one.

Second, each seed set *s* is associated with a group-topic distribution ψ_s which is a distribution over the regular topics. The hyperparameters α and β of the first and second step refer to the regular *LDA* model.

Third, a list of seed sets for each document d is generated and represented by the binary vector $\vec{b}$. Given this vector, which defines a mean of a Dirichlet distribution, a document-group distribution ζ^d can be sampled. The concentration of this Dirichlet to a hyperparameter τ is set by the researcher. A group variable g is drawn from the resulting multinomial which groups the documents that show the same seed set. Then, the document-topic distribution θ_d from a Dirichlet distribution with the group's-topic distribution can be chosen in order to ensure a relation of the topic distributions of documents within each group. Afterwards, a topic z_i for each token is generated. After a topic is chosen, either the seed or the regular topic distribution is selected based on the indicator x_i which is drawn from a Bernoulli distribution. Given the selected distribution a word can be generated.

Unbounded Polarity Score

The algorithm to calculate the unbounded polarity score δ introduced by Rinker (2019) first breaks each sentence into an ordered bag of words, for instance $w_{1,2,3}$ represents the third word of the second sentence of the first paragraph. Each word is then compared to a dictionary of polarised words *pw* (tagged with +1 and -1). Next, a polarised context cluster $c_{i,j,l}$ is defined with *n* words before and after the *pw* by the researcher. Given a set of valence shifters the words of a polarised context cluster are either tagged as neutral $w_{i,j,k}^{0}$, negator (e.g., not) $w_{i,j,k}^{n}$, amplifier (e.g., very) $w_{i,j,k}^{a}$, de-amplifier (e.g., less) $w_{i,j,k}^{d}$ or adversative conjunction (e.g., but) $w_{i,j,k}^{\mathrm{ac}}$. Amplifiers (de-amplifiers) increase (decrease) the polarity, negators flip the polarity and an adversative conjunction before (after) the *pw* up-weights (down-weights) the cluster. The researcher provides the weight *z*. Given the respective polarity each *pw* is furthermore weighted by the function and the amount of valence shifters surrounding it. The equation is not affected by neutral words, but the word count *n*.

To calculate δ the weighted context clusters $c_{i,j,l}$ are summed $c'_{i,j}$ and divided by the square root of the word count $\sqrt{w_{i,\ jn}}$.

$$\delta = \frac{c'_{i,j}}{\sqrt{w_{i,\ jn}}} \tag{2.2}$$

where:

$$c'_{i,j} = \sum((1 + w_{amp} + w_{deamp}) \cdot w^{p}_{i,j,k}(-1)^{2+w_{neg}}) \tag{2.2.1}$$

$$w_{amp} = (w_b > 1) + \sum(w_{neg} \cdot (z \cdot w^{a}_{i,j,k})) \tag{2.2.2}$$

$$w_{deamp} = max(w_{deamp'}, 1) \tag{2.2.3}$$

$$w_{deamp'} = (w_b < 1) + \sum(z(-w_{neg} \cdot w^{a}_{i,j,k} + w^{d}_{i,j,k})) \tag{2.2.4}$$

$$w_b = 1 + z_2 * w_{b'} \tag{2.2.5}$$

$$w_{b'} = \sum(|w_{adversative\ conjunction}|, \ \ldots, \ w^{p}_{i,j,k}, \ w^{p}_{i,j,k}, \ \ldots, \ |w_{adversative\ conjunction}| \cdot -1) \tag{2.2.6}$$

$$w_{neg} = \left(\sum w^{n}_{i,j,k}\right) mod\ 2 \tag{2.2.7}$$

Table 2.6 Selected Seed Words

Trend	***Seed Words***
Globalisation	capital markets; competition; competitiveness; complexity; consolidated; consolidation; cooperation; countries; distribution; diversification; diversify; economic crisis; economic growth; economic situation; economy; europe; european; expanding; financial crisis; financial market crisis; financial markets; fund; fund manager; global; globalisation; harmonisation; institutional; interest rate environment; international; internationalisation; liberalisation; low interest rate environment; market leader; market participants; outsourcing; professionalisation; real estate cycles; real estate funds; realignment; recession; restructuring; sign; standardisation; strategy; strategy change; structural change; subprime crisis; uncertainty; unification; union; upswing; world; world economy; world trade; worldwide
Demography	age distribution; age appropriate; aging; ambulant; apartments for senior citizens; barrier free; beds; care; care facility; care of senior citizens; care places; care properties; care requirements; caregivers; change; clinics; day care; dementia; demographic; demography; development; facilities; health; health care; healthy; household figures; human population; in need of care; local supply; maintenance; medical; migration movements; mobile; mobility; nursing; nursing beds; nursing home; obsolescence; old; population; population decrease; population development; population forecast; provision for old age; residences; residents; senior centre; senior citizen real estate; senior citizen residences; senior citizens; services; shortage of supply; shrinking; staff; stationary; support; types of living
Urbanisation	agglomeration; air pollution; appreciation; area; area potentials; centre; centrality; city; city centre; city migration; city planner; connection; crowding out; cultural; densification; density; development plan; district development; downtown; gentrification; green areas;

	hospitality; housing shortage; immigration; individualisation; influx; metropolis; metropolitan; metropolitan areas; metropolitan city; metropolitan region; migration; noise exposure; quality of life; regional research; rent increase; rental growth; rental price; rental price growth; residency; residential districts; rural; rural depopulation; shortage of land; small town; town; university; urban; urban centre; urban development; urban growth; urban sprawl; urban studies; urbanistic; urbanity; urbanisation
Sustainability	added value; additional costs; building technology; carbon dioxide; certification; certified; climate; climate neutral; climate protection; climate summit; cold; consumption; continuously; cost saving; dgnb (German Sustainable Building Council); district heating; ecological; economic efficiency; efficiency; efficient; emission; energetic; energy; energy consumption; energy costs; energy efficiency; energy management; energy saving; energy transition; environment; environmentally aware; environmentally friendly; facility management; flexibility; green; greenhouse gas; heat recovery; leed; life cycle; neutral; operating costs; optimization; power; primary energy; quality; raw material; recycling; renewable; resources; saving; standards; sustainability; sustainability; sustainability fund; sustainable; warmth
Digitalisation	ai; analogue; app; artificially; automation; blockchain; building information modelling; business models; challenge; cloud; communication; computing; data; data collection; data protection; digital; digitalisation; ecommerce; electronically; engineering; information; innovation; innovation centre; innovative; integrated; intelligence; interface; internet; internet access; it; modelling; networked; networking; online; operations; platform; processes; progress; proptechs; real time; server; shared; smart; smartphone; software; solutions; startup; system; tech; technologies; technology; the age of; transformation; virtual; web

Regulation	adjustments; associations; authorities; bafin (German Federal Financial Services Regulatory Authority); bestellerprinzip (German law); cap; citizens; coalition; come into force; community; cover; decisions; discussion; district level; draft law; draft law; election; federal council; federal government; financial services supervision; fines; free amount; funding; german investment code; government; housing markets; implementation; incentive; intensification; intervention; investor protection; law; legislation; legislator; limitation; local authority; measures; mietpreisbremse (German law); national level; national level; new regulation; policy; politician; public; reform; regulation; regulations; regulatory; rental cap; revision; rule; rules; specifications; state; subsidise; support; transparency

Notes: Table 2.6 displays the results of the *GloVe* model and the knowledge-based seed words.
Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH.

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 coverage of real estate topics or news sentiment and total returns of the asset classes of residential, office and retail. Using methods of natural language processing, including word embeddings, 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.

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. Hence, a substantial body of literature, quantifying the stock market-related effects of investor sentiment, has since evolved. Contributing to this body of literature, this paper uses textual analysis to examine the impact of newspapers on investors and thus on real estate markets. We separate the newspaper articles by asset class, recognising a varying level of responsiveness of certain markets to shifts and shocks in sentiment, and find that news coverage and sentiment are Granger-causal for future asset class returns.

Increasing computing power and the resulting possibilities in the field of textual analysis have established the derivation of sentiment indicators from corporate publications, newspapers or social media in the last decade and supplemented or even partly replaced the formerly widespread survey-based indicators as well as capital market indicators for measuring changes in sentiment. Through the application of textual analysis in their pioneering works, Tetlock (2007), Das and Chen (2007), and Loughran and McDonald (2011) have found significant correlations between textual sentiment indicators and stock returns, return volatility, and trading volume. However, similar research investigating the role of text-based sentiment in real estate markets is still limited. There is a variety of studies analysing the importance of further factors beyond the conventional economic fundamentals on real estate markets. Lin *et al.* (2009), for example, explained REIT returns through use of investor sentiment obtained from changes in discounts of closed funds, and Ling *et al.* (2014) applied a survey-based sentiment as well as a sentiment derived from eight sentiment proxies to commercial real estate markets. Given this importance of investors’ sentiment in real estate markets, Beracha *et al.* (2019), Hausler *et al.* (2018), and Ruscheinsky *et al.* (2018b) extended sentiment analysis by means of a textual sentiment analysis in the commercial and securitised real estate sector respectively. When deriving sentiment from newspaper headlines, these studies 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 also news-based sentiment indicators should be asset class-specific.

Despite the necessity for asset class-specific sentiment indicators, 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 this reason, a comprehensive approach was chosen for this study, in order to facilitate the examination of the different publishing priorities of the media. For investors, such differences can play an important role as they base their decisions to some extent on only part of the available information. Hence, we ascertain asset class-specific sentiment indicators by applying a classification algorithm to the real estate articles of leading trade newspapers and daily newspapers in Germany. The dual approach in this study on asset class-specific news coverage and news sentiment is not only consistent with the textual sentiment already explored by prior research, but also addresses the concept of 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. 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 assigned to the real estate asset classes of residential, office, and retail. An 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, with which to assign articles to an asset class. This step enabled 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 in all time periods. Ultimately, the intertemporal links between reporting intensity and sentiment, on the one hand, and the performance of 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.

Indeed, we found that trade newspapers and daily newspapers differ in their reporting, which can be at least partially explained by considering their contrasting target groups. Furthermore, newspapers are particularly suitable for predicting returns in one to two quarters, with trade newspapers appearing to be a more reliable source.

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. Both the textual analysis procedure and methodology regarding the analysis of our hypotheses are described in section 3.4, while the dataset is presented 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

Topic Modelling

Textual analysis means the conversion of qualitative and highly unstructured information within a text via standardised algorithms into comparable quantitative data. This results in major challenges in natural language processing, as language can be ambivalent and vague, and this can be disparate to the information contained within it as well as the intentions of authors (Loughran and McDonald, 2020). To overcome these issues, researchers commenced the development of classification algorithms in the early 2000s, which analyse sets of documents (the corpus) and obtain joint topics describing each document in the corpus. This classification of topic modelling brings unstructured text into a form which enables intelligent searches as well as a statistical characterisation of documents and hence has been applied in multiple fields 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 the field of finance to classify news, e.g., from newspapers or social media, as either buy, sell or hold recommendations. 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 published research in four real estate related journals over a period

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

of 37 years to identify the main topics in real estate research while 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 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 usual *LDA* algorithm, especially when the documents are pre-selected by shared themes. Since the algorithm is designed to extract commonalities from a corpus of unstructured text, it fails to identify subtopics of a set of documents pre-selected by topics. Furthermore, if specific subtopics are targeted ex ante, they do not necessarily have to be included in the topics suggested by the algorithm.

Jagarlamudi *et al.* (2012) addressed this problem and developed a semi-supervised *LDA* algorithm. The so-called seeded Latent Dirichlet Allocation 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 seeded *LDA*, Mai (2021), for example, extracts narratives from the New York Times from the past 150 years and ascertains that these extracted narratives can serve as strong market predictors. Likewise, Antweiler and Frank (2006) cluster Wall Street Journal corporate news stories from 1973 to 2001 into business incidents and identify an overreaction to news on the American stock markets. The only application of seeded *LDA* in the context of real estate is in the form of a recent paper of Ploessl *et al.* (2021), in which the authors use seeded *LDA* to assign almost 170,000 German 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 21-year period.

Sentiment Analysis

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.*, 2008; Freybote and Seagraves, 2017). While the former usually suffer from selection bias and the problem 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) is the study, we consider most likely to be the pioneer in addressing these problems, in which sentiment is extracted from a textual corpus. Tetlock 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 since (e.g., Feldman *et al.*, 2010; Loughran and McDonald, 2011; García, 2013). One of the biggest criticisms of the dictionary-based approach is its dependence on pre-defined dictionaries, which, in addition, should be domain-specifically adjusted (Henry and Leone, 2016), since the meaning of a word is sometimes inextricably tied to its context. 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 identify significant relationships between market sentiment and real estate market developments. Clayton *et al.* (2008) and Ling *et al.* (2014) estimate that their survey-based investor-sentiment indicators are closely linked to market returns in subsequent periods. Further, Hausler *et al.* (2018) show that a ML-based sentiment indicator obtained from professional financial news leads the US securitised and direct commercial real estate markets. Soo (2018) is one of the few studies to analyse residential markets, also finding that changes in housing media sentiment to have significant predictive power for future changes in house prices that lead by nearly two years. By focusing on the German residential market, Ruscheinsky *et al.* (2018a) created a German domain-specific dictionary and revealed a statistically significant relationship between the extracted sentiment of the most important German real estate newspaper and the development of residential real estate prices. Marcato and Nanda (2016) are the only ones to analyse both residential and commercial real estate markets. According to them, survey-based sentiment indices, however, varying in importance regarding each market, which may reflect a greater level of responsiveness of certain markets to shifts and shocks in sentiment.

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:

H1. *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 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:

H2. *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.*, 2008). Consequently, the sentiment indices should have a predictive power over the future performance of an asset class, leading to the third hypothesis:

H3. *News Coverage and sentiment have predictive power over future performance for each asset class.*

3.4 Methodology

3.4.1 News Coverage — Seeded LDA and Global Vectors 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[4] to an asset class, is required as the arti-

[4] Pre-processing of the articles involves the removal of punctuation marks, numbers, non-alphabetical and special characters, and stop words. 'Stop words' relate to frequently occurring words that have

cles are not tagged by keywords that allowed for inference to the asset class discussed in an article. Therefore, a seeded *LDA* 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 seeded *LDA* 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 seeded *LDA* 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

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.

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. Hence, a word embeddings-approach, the Global Vector model (*GloVe*), introduced by Pennington *et al.* (2014), is applied in order to select suitable seed words from the corpus. *GloVe* can be used to identify relationships between words, also ensuring that the selected seed words occur in the text corpus. For instance, the unsupervised learning algorithm can reveal, considering the two sentences 'Company X bought a property in London.' and 'Corporation Y purchased a building in Berlin.', the tokens 'property' and 'building' or 'bought' and 'purchased' etc. for word representation. To generate these vectors for word representation, the algorithm combines the advantages of global matrix factorisation and local context window methods.[5] For this reason, the

[5] 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

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:

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

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

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.

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).

Given the 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, in semi-supervised seeded *LDA* each topic k is defined by two distributions, the regular topic distribution ϕ_k^r, and an additional seed topic distribution ϕ_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 $k = 1 \cdots T$,
   (a) Choose regular topic $\phi_k^r \sim$ Dirichlet(β_r)
   (b) Choose seed topic $\phi_k^s \sim$ Dirichlet(β_s)
   (c) Choose $\pi_k \sim$ Beta(1, 1)
2. For each seed set $s = 1 \cdots S$,
   (a) Choose group-topic distribution $\psi_s \sim$ Dirichlet(α)
3. For each document d,
   (a) Choose a binary vector $\vec{b}$ of length S
   (b) Choose a document-group distribution $\zeta^d \sim$ Dirichlet($\tau\vec{b}$)
   (c) Choose a group variable g ~ Multinomial(ζ^d)
   (d) Choose $\theta_d \sim$ Dirichlet(ψ_g) // of length T
   (e) For each token $i = 1 \cdots N_d$:
      i. Select a topic $z_i \sim$ Multinomial(θ_d)
      ii. Select an indicator $x_i \sim$ Bernoulli(π_{z_i})
      iii. if x_i is 0
         - Select a word $w_i \sim$ Multinomial($\phi_{z_i}^r$)
      iv. if x_i is 1
         - Select a word $w_i \sim$ Multinomial($\phi_{z_i}^s$)

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 (*NC*) 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 $NC \in [0; 1]$, with a higher value meaning that an asset class is discussed more frequently.

3.4.2 News Sentiment — Unbounded Polarity Score

In the second step of the natural language processing procedure, articles' polarities are determined using a dictionary-based 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[6] is compared to the polarised

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

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 *Unbounded Polarity Score* in the appendix 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 sentiment of a specific newspaper towards an asset class. Hence, news sentiment (*NS*) 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.[7]

[7] 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.

3.4.3 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:

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

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

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, 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:

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

The components of the *VAR* are tested for stationarity through the use of an Augmented Dickey-Fuller Test. 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 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 conducting Granger causality tests (Granger, 1969). 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.

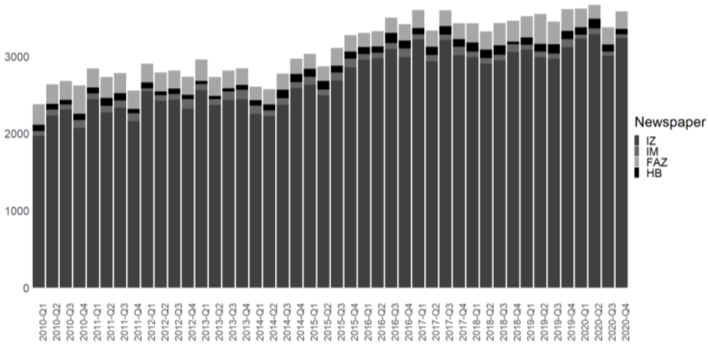
3.5 Data

3.5.1 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.[8] 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*.

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

Figure 3.1 Number of Articles over Time by Newspaper



Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH, Immobilien Manager Verlag IMV GmbH & Co. KG, Frankfurter Allgemeine Zeitung GmbH, Handelsblatt GmbH.

The articles of the *IZ* have an average length of 229 words and 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)[9] and offer detailed information to the reader. In order to reduce the

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

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.[10]

3.5.2 INREV Returns

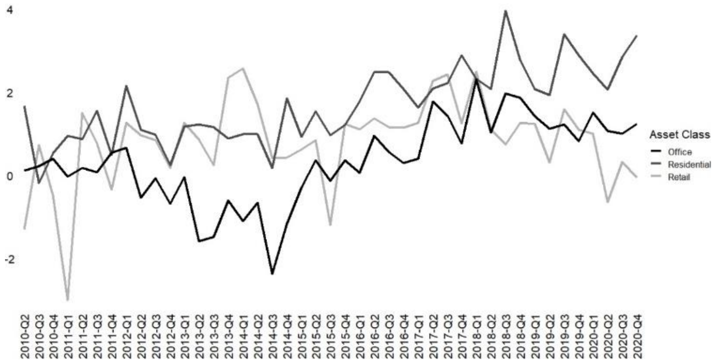
The dependent performance variable is based on the German Vehicles Index extracted from the European Association for Investors in Non-Listed Real Estate Vehicles (*INREV*) database. *INREV*'s German Vehicles Index comprises the data of 49 German vehicles which have invested in Germany, having provided quarterly total return data for each asset class since 2003. During the observation period, the vehicles hold, on average, properties totalling a gross asset value of 14.4 billion EUR whereby the allocation amongst residential, office and retail is almost equal. However, this index is merely an approximation regarding the actual performance of

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

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 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.

Figure 3.2 Percentage Total Return over Time by Asset Class



Data Source: INREV.

3.5.3 Macroeconomic and Real Estate 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.* (2008), Ling *et al.* (2014), Walker (2014), Das *et al.* (2014), 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 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.[11] Descriptive statistics for

[11] 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

INREV returns and control variables are presented in table 3.1.

Table 3.1 Descriptive Statistics

	Mean	***Median***	***SD***	***Min***	***Max***
Return Residential	0.0171	0.0167	0.0092	-0.0018	0.0396
Return Office	0.0036	0.0038	0.0100	-0.0237	0.0231
Return Retail	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 Office	0.0271	0.0410	0.0739	-0.1662	0.1612
TV Retail	0.0186	0.0179	0.0952	-0.1803	0.3015
INT Residential	-0.0003	-0.0002	0.0007	-0.0018	0.0013
INT Office	0.0106	0.0060	0.0131	-0.0110	0.0370
INT Retail	-0.0014	-0.0010	0.0022	-0.0060	0.0030

Data Source: INREV, Federal Statistical Office of Germany, boerse.de Finanzportal, JLL, BNP Paribas Real Estate, NAI Global, Deutsche Bundesbank.

lead to the improvement of estimation results, or data was not available on asset class level (e.g., construction turnover).

3.6 Results

3.6.1 News Coverage

Following the approach of Watanabe and Zhou (2020) our selection of reliable knowledge-based 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 *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.

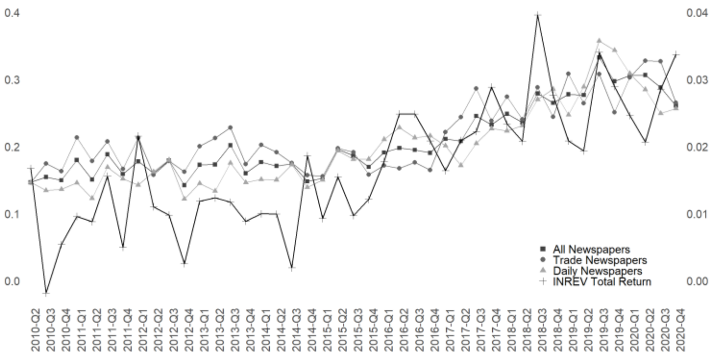
After assigning the articles to the corresponding asset class through the seeded *LDA* algorithm, the *NC* indices, as shown in Figures 3.3 to 3.5, can be computed.[12] Hereby, retail is the most frequently reported market, with a mean share of 27.63 % of all articles, followed by office (23.48 %), and then residential (20.74 %).[13] 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

[12] 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 *IZ*. Furthermore, indices for trade and daily newspapers are based on the mean of the underlying newspapers.

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

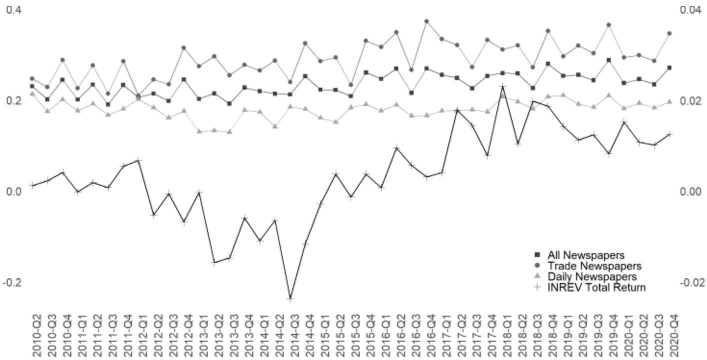
2015. Retail is initially more present in daily newspapers, which may be traced back to its presence in everyday life. Moreover, it is the asset class, throughout which the *NC* indices of both types of newspapers move the least synchronously. Even though both indices are positively correlated, as for all asset classes, the reporting intensity of the daily newspapers has decreased significantly since 2015, while it has remained stable in the trade papers.

Figure 3.3 Residential — News Coverage over Time



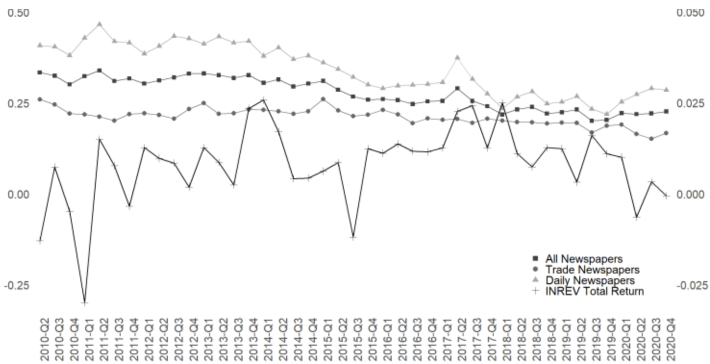
Notes: News Coverage (left), INREV Index Return (right).
Data Source: INREV, IZ Immobilien Zeitung Verlagsgesellschaft mbH, Immobilien Manager Verlag IMV GmbH & Co. KG, Frankfurter Allgemeine Zeitung GmbH, Handelsblatt GmbH.

Figure 3.4 Office — News Coverage over Time



Notes: News Coverage (left), INREV Index Return (right).
Data Source: INREV, IZ Immobilien Zeitung Verlagsgesellschaft mbH, Immobilien Manager Verlag IMV GmbH & Co. KG, Frankfurter Allgemeine Zeitung GmbH, Handelsblatt GmbH.

Figure 3.5 Retail — News Coverage over Time



Notes: News Coverage (left), INREV Index Return (right).
Data Source: INREV, IZ Immobilien Zeitung Verlagsgesellschaft mbH, Immobilien Manager Verlag IMV GmbH & Co. KG, Frankfurter Allgemeine Zeitung GmbH, Handelsblatt GmbH.

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. The same applies to the office market, in which correlations are highest using one to four period lagged *NC* (between 0.49 and 0.56). 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 *NC* of retail and its total return is even negative (-0.18) and highest (-0.32) when total return leads *NC* by two quarters. These observations are consistent when disaggregating *NC* indices to both types of newspapers and they are validated through a set of bivariate *VAR*s: 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 news coverage between trade and daily newspapers,[14] which seem to be driven by 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 *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

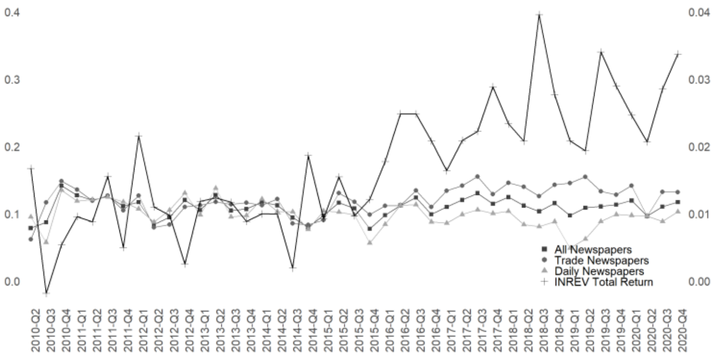
[14] See table 3.8 in the appendix for the descriptive statistics of the generated indices.

daily newspapers, as revealed by Figures 3.6 to 3.8.[15] 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 *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 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

[15] See table 3.8 in the appendix for the descriptive statistics of the generated indices.

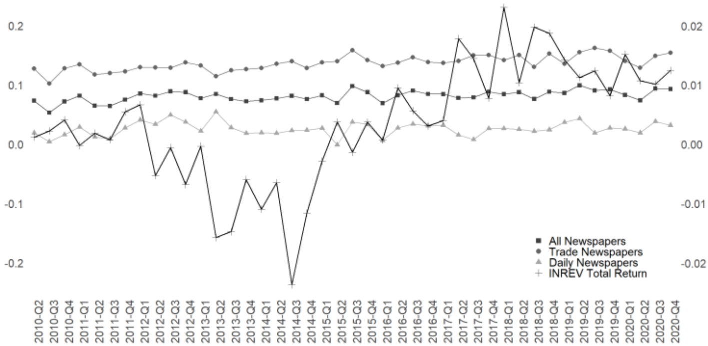
newspapers have been significantly more optimistically written in their articles on the residential market than the daily newspapers. This divergence thus begins simultaneously with the previously described increase in residential reporting intensity, which can possibly be seen as an indicator of changing market conditions.

Figure 3.6 Residential — News Sentiment over Time



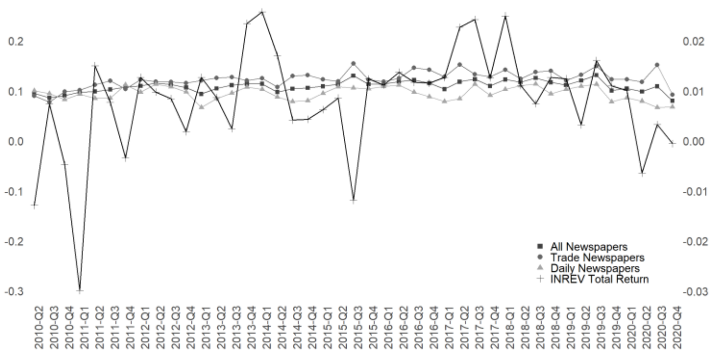
Notes: News Sentiment (left), INREV Index Return (right).
Data Source: INREV, IZ Immobilien Zeitung Verlagsgesellschaft mbH, Immobilien Manager Verlag IMV GmbH & Co. KG, Frankfurter Allgemeine Zeitung GmbH, Handelsblatt GmbH.

Figure 3.7 Office — News Sentiment over Time



Notes: News Sentiment (left), INREV Index Return (right).
Data Source: INREV, IZ Immobilien Zeitung Verlagsgesellschaft mbH, Immobilien Manager Verlag IMV GmbH & Co. KG, Frankfurter Allgemeine Zeitung GmbH, Handelsblatt GmbH.

Figure 3.8 Retail — News Sentiment over Time



Notes: News Sentiment (left), INREV Index Return (right).
Data Source: INREV, IZ Immobilien Zeitung Verlagsgesellschaft mbH, Immobilien Manager Verlag IMV GmbH & Co. KG, Frankfurter Allgemeine Zeitung GmbH, Handelsblatt GmbH.

To see whether this applies, we compare *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 *NS* indices. Additionally, it seems that, as expected, more optimistic reporting is related to higher returns. The use of a set of bivariate *VAR*s 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 *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 regarding 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 and Granger Causality

The objective of this analysis is to examine the ability of news coverage and sentiment measures to predict 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, as specified in equation 3.8, for each asset class and newspaper type.

Table 3.2 Residential — VAR 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 displays 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 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 displays 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 3.4 Retail — VAR 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 displays 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.

Focusing on news coverage, 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 newspapers 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.

In terms of news sentiment, one can ascertain a statistically significant relationship between *NS* 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 property returns but not on commercial property returns. Nonetheless, for the 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, 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 news-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-based 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 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 news-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.

Table 3.5 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 displays 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 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 displays 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.

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 coverage or sentiment 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 indicators are 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 news-based 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-based 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 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?", available at SSRN: https://ssrn.com/abstract=878091.

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. (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. 44 No. 3, pp. 528–543.

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

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

Clayton, J., Ling, D.C. and Naranjo, A. (2008), "Commercial Real Estate Valuation: Fundamentals Versus Investor Sentiment", *SSRN Electronic Journal*.

Das, P.K., Freybote, J. and Marcato, G. (2014), "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.

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. 355–380.

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.

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

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., 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. 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", *Pro-*

ceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, USA, pp. 204–213.

Jianfeng S., Arjun M., Bing L., Qing L., Huayi L. and Xiaotie D. (2013), "Exploiting Topic based Twitter Sentiment for Stock Prediction", *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Vol. 2, 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*, Cham: Palgrave Macmillan.

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 Schaefers, W. (2020), "Is Mandatory Risk Reporting

Informative? Evidence from US REITS Using Machine Learning for Text Analysis", available at SSRN: https://ssrn.com/abstract=3686492.

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, pp. 357–375.

Mai, D. (2021), "Economic Narratives and Market Outcomes: A Semi-supervised Topic Modeling Approach", available at SSRN: https://ssrn.com/abstract=3990324.

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", *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, Beijing, pp. 1354–1364.

Nordheim, G. von, Boczek, K. and Koppers, L. (2018), "Sourcing the Sources: An analysis of the use of Twitter and Facebook as a journalistic source over 10 years in The New York Times, The Guardian, and Süddeutsche Zeitung", *Digital Journalism*, Vol. 6 No. 7, pp. 807–828.

Pennington, J., Socher, R. and Manning, C. (2014), "Glove: Global Vectors for Word Representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Qatar, 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*, Vol. 14 No. 3, pp. 381–400.

Rinker, T. (2019), *Package 'sentimentr': Calculate Text Polarity Sentiment*, R 3.4.0, available at:

https://cran.r-project.org/web/packages/157entiment/157entiment.pdf.

Ruscheinsky, J.R., Kandlbinder, K., Schaefers, W., Dietzel, M.A. and Rochdi, K. (2018a), "Predicting Real Estate Market Movements: the First Textual Analysis-Based-Sentiment Application in Germany", *Working Paper*.

Ruscheinsky, J.R., Lang, M. and Schaefers, W. (2018b), "Real estate media sentiment through textual analysis", *Journal of Property Investment and 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*, Princeton University Press.

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*, 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*, available at: https://journals.sagepub.com/doi/10.1177/0894439320907027.

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.

3.9 Appendix

Unbounded Polarity Score

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

where:

$$c'_{i,j} = \sum((1 + w_{amp} + w_{deamp}) \cdot w^{p}_{i,j,k}(-1)^{2+w_{neg}}) \tag{3.3.1}$$

$$w_{amp} = (w_b > 1) + \sum(w_{neg} \cdot (z \cdot w^{a}_{i,j,k})) \tag{3.3.2}$$

$$w_{deamp} = max\left(w_{deamp'}\,, 1\right) \tag{3.3.3}$$

$$w_{deamp'} = (w_b < 1) + \sum(z(-w_{neg} \cdot w^{a}_{i,j,k} + w^{d}_{i,j,k})) \tag{3.3.4}$$

$$w_b = 1 + z_2 * w_{b'} \tag{3.3.5}$$

$$w_{b'} = \sum(|w_{adversative\ conjunction}|,\ \ldots,\ w^{p}_{i,j,k},\ w^{p}_{i,j,k},\ \ldots,\ |w_{adversative\ conjunction}| \cdot -1) \tag{3.3.6}$$

$$w_{neg} = \left(\sum w^{n}_{i,j,k}\right)\ mod\ 2 \tag{3.3.7}$$

Table 3.7 Selected Seed Words

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

***Notes**:* Table 3.7 displays the selection of seed words for the three asset classes of residential, office and retail; all results in this paper were estimated in German and have been translated, with expressions consisting of two or more words in English appearing as single words in German.; for example, the results of the *GloVe* model show for the word vector 'residential property' the highest correspondence level of 27.93 % for the vector 'condominium' or 'housing' (26.97 %) in the largest data set of *IZ.*

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

Notes: Table 3.8 displays descriptive statistics of quarterly sentiment indicators between 2010 and 2020 for the three asset classes of residential, office and retail subdivided by trade newspapers and daily newspapers. *NC* is news coverage, *NS* is news sentiment.

4. News Coverage vs. Sentiment: Evaluating German Residential Real Estate Markets

4.1 Abstract

Purpose — To investigate whether additional information of the permanent news flow, especially reporting intensity, can help to increase transparency in housing markets, this study examines the relationship between news coverage or news sentiment and residential real estate prices in Germany at a regional level.

Design/methodology/approach — Using methods in the field of natural language processing, in particular word embeddings and dictionary-based sentiment analyses, we derive five different sentiment measures from almost 320,000 news articles of two professional German real estate news providers. These sentiment indicators are used as covariates in a first difference fixed effects regression in order to investigate the relationship between news coverage or news sentiment and residential real estate prices.

Findings — The empirical results suggest that the ascertained news-based indicators have a significant positive relationship with residential real estate prices. It appears that the combination of news coverage and sentiment proves to be a reliable indicator. Furthermore, the extracted sentiment measures lead residential real estate prices up to two quarters. Finally, the explanatory power increases when regressing on prices for condominiums compared with houses, implying that the indicators may rather reflect investor sentiment.

Originality — To the best of the authors' knowledge, this is the first paper to extract both the news coverage and news sentiment from real estate-related news for regional German housing markets. The approach presented in this study to quantify additional qualitative data from texts is replicable and can be applied to many further research areas on real estate topics.

Keywords — Textual Analysis, News Sentiment, News Coverage, Word Embeddings, Dictionary-based Approach, Machine Learning, Residential Real Estate Markets.

4.2 Introduction

In an increasingly digital economy, newspapers occasionally appear like a relict from the past, at least in paper form. Today, information flows rapidly on digital platforms, on social media, TV and radio. And yet, newspapers still exist in the 21st century, accounting for a media market share of 14.5 % in Germany (Statista, 2022), they have adapted to the new media environment by establishing an online presence and by offering tailor-made formats, special-interest papers and magazines.

Evidently, there is still demand for the specific combination of information and entertainment offered by newspapers. Newspapers are not able to surpass the internet and TV-stations with regards to speed, current relevance and scale and scope of information. However, there are situations in which time-constrained readers value a pre-selection of information. What is more, when information is uncertain or when data is missing, readers appreciate an assessment and evaluation of opaque information sets. In such a scenario, a balanced commentary can be more valuable to the readers than pure speed and volume of information. Thus, reading newspapers can be especially interesting for real estate market participants, as the heterogeneity of this asset class prevents full transparency, and the complexity of the regulatory and market environment make it — particularly for small-scale actors — virtually impossible to follow and evaluate the permanent news flow.

Primarily, readers look for information when reading the news. But they may be impelled by the sentiment the journalist consciously or subconsciously has conveyed. This sentiment could bias both the journalists' choice of topics and tonality of the article. Thus, reading the news does not only change the level of information on an asset class, but can also reflect or even affect overall market sentiment: the prevailing attitude of investors to the expected price development in a market. These new pieces of information and changes in sentiment can be considered early indicators for changes in hard market indicators like prices, rents or returns. Then, even better-informed market professionals will find value in reading the news, as they may be able to gain an understanding of future market perceptions.

In this article we analyse the relationship between news coverage or news sentiment of two German real estate news providers (*Immobilien Zeitung* and *Thomas Daily*) and residential property prices. In this paper we provide answers to four research questions: Is there a relationship between news coverage or news sentiment and German residential property prices? Are there differences at regional levels? Is there a causality flow from news coverage or news sentiment to changes in residential property prices or vice versa? Are there differences regarding the markets for houses and condominiums?

The remainder of this paper is organised as follows: Section 4.3 discusses the literature on sentiment analysis in the context of real estate and deduces the hypothesis. The dataset is presented in section 4.4, while the textual analysis procedure and methodology are described in section 4.5. The results are displayed in section 4.6, leading to the conclusion in section 4.7, which discusses the implications for the real estate industry.

4.3 Literature Review and Hypothesis Development

Textual Analysis in Economics and Finance

Sentiment analysis has become a relevant area of research in economics and finance, applying methods evolving from survey-based analysis or indicators derived from capital market data (e.g., Clayton *et al.*, 2008; Das *et al.*, 2014; Marcato and Nanda, 2016; Freybote and Seagraves, 2017) to computational linguistic techniques. Recent developments in data processing and analysis have enabled sentiment to be extracted from large bodies of digitally available texts. The work of Tetlock (2007) constitutes one of the first applications of the dictionary-based approach quantifying the sentiment of news articles from The Wall Street Journal. Through the assignment of words to dictionaries which contain categorised words of positive or negative sentiment, Tetlock has shown that media pessimism helps to predict downward pressure on stock market prices. A number of research papers followed on the basis of Tetlock's work by adopting the dictionary-based approach (e.g., Feldman *et al.*, 2010; Loughran and McDonald, 2011; García, 2013; Kearney and Liu, 2014). However, this method is challenged by two main aspects: first, the requirement of pre-defined word lists and second, the need for discipline-specific or even market-specific dictionaries (Loughran and McDonald, 2011; Henry and Leone, 2016; Nowak *et al.*, 2021). To address these issues, various other studies have employed a machine learning-based approach such as support vector machines and Naïve Bayes (e.g., Pang *et al.*, 2002; Schumaker and Chen, 2009; Li, 2010; Hausler *et al.*, 2018). In the debate about the ability of either humans or machines to create sentiment dictionaries, Loughran and McDonald (2020) argue that humans may arbitrate the nuance of words better. Moreover, since the dictionary-based approach is easy to replicate and traceable, it is still the most commonly used method.

Textual Analysis in Real Estate

Several studies have applied textual analysis in the context of real estate research, following the dictionary-based approach. Regarding the field of REITs, the paper of Doran *et al*. (2012), for example, was successful in showing that the tone of quarterly-earnings conference calls for publicly traded REITs has significant explanatory power for stock price movements. A leading relationship between media sentiment and future REIT market movements was also found by Ruscheinsky *et al.* (2018b). By investigating the language in the Management Discussion and Analysis of REITs, Koelbl (2020) concluded that higher levels of pessimistic (optimistic) language predicts lower (higher) future firm performance. Likewise, Carstens and Freybote (2021) found that the tone in REIT financial statements can predict total returns in the subsequent quarter, given periods of poorly performing real estate markets. Only a few studies have investigated the dictionary-based approach concerning residential real estate markets. For instance, Nowak and Smith (2017) were able to improve the performance of hedonic pricing-estimates with the textual information contained in the comments section of the MLS listings. The work of Soo (2018) provides evidence of the fact that local housing media sentiment has predictive power for future house prices leading by nearly two years. By the development of an industry-specific dictionary for the German-speaking area, Ruscheinsky *et al.* (2018a) have so far been the first to apply the dictionary-based approach to the German housing market, finding a significant relationship between the measured sentiment and the development of residential real estate-prices. Moreover, Nowak *et al.* (2021) have shown that the information provided in real estate agents' remarks concerning properties can be used to address heterogeneity in housing markets and have again demonstrated that in-sample prediction errors in the pricing models can be reduced.

Besides of the extraction of tonality from real estate-related texts, the reporting intensity of specific topics becomes of interest to sentiment analysis, which is based on the concept of economic narratives introduced by Shiller (2017). These economic narratives can have an impact on economic outcomes as they influence individual decision-making if they reach a large audience. A fast distribution of these narratives is attained in particular by journalists through their selection and interpretation of information. In order to understand the shifts in media attention, a few studies have attempted to measure the news coverage of certain topics in addition to their tonality. To explore unstructured bodies of texts, topic model algorithms such as Latent Dirichlet Allocation (*LDA*) (Blei *et al.*, 2003) are often applied (e.g., Zhao *et al.*, 2011; Nguyen and Shirai, 2015; Cerchiello and Nicola, 2018; Nordheim *et al.*, 2018; Aziz *et al.*, 2019). These generative probabilistic models primarily concern the derivation of underlying common topics. However, there is limited research that applies topic modelling in the context of real estate market analysis: Ploessl *et al.* (2021) used seeded *LDA* for the assignment of German news articles to six real estate-related trends and revealed that both the news coverage and sentiment of these expected stable trends show cyclical elements within a 21-year period. Another algorithm, the structural topic model, was employed by Koelbl *et al.* (2021) for the detection of risk-factors in the 10-K filings of REITs in order to investigate how these could affect the risk perceptions of investors. Ploessl *et al.* (2022) examined the relationship between news coverage or sentiment of the main asset classes and total returns, suggesting that trade newspapers outperform daily newspapers in the prediction of future total returns. Still, there remains a lack of research concerning the analyses of real estate-related texts: first, the intensity of reporting and second, the precise measurement of tonality, and third, regional differences in residential real estate markets, especially for the German-speaking area.

Hypothesis Development

Based on the studies of Soo (2018) and Ruscheinsky *et al.* (2018a), this paper focusses on residential real estate but assigns the news articles of the text corpus to the seven major regional housing markets in Germany by means of computational linguistic techniques. In addition, the effect of not only the tonality but also the frequency of news media is examined. Following Ploessl *et al.* (2022), a positive relationship between news coverage or sentiment and residential property prices is expected. The hypothesis is derived from the idea that trade newspapers typically report more frequently and positively on prospering markets and therefore predominantly reflect investor sentiment (Tetlock, 2007; Walker, 2016; Beracha *et al.*, 2019; Ploessl *et al.*, 2022):

H1. *News coverage and news sentiment are positively related to residential property prices.*

As real estate markets are characterised by intransparency and are only slowly adjusting to new information

(Baum *et al.*, 1996; Clayton *et al.*, 2008), sentiment indicators are generally assumed to lead real estate market movements. According to Shiller's (2017) idea, being that economic narratives can shape individual and collective investor decision-making and that newspapers are an important source of information for professionals and investors, the extracted information from texts could have predictive power over future residential property prices, resulting in the second hypothesis:

H2. *News coverage and news sentiment Granger-cause residential property prices.*

Lastly, it can be expected that the explanatory power is better for condominium prices than for house prices because trade newspapers predominantly target professional real estate investors, who most frequently invest in condominiums:

H3. *News coverage and news sentiment correlate more strongly with condo prices than house prices.*

4.4 Data

Two types of datasets are used: the text corpus given by two major German real estate news providers on the one hand and residential property prices and economic variables for each city (Berlin, Duesseldorf, Frankfurt, Hamburg, Cologne, Munich and Stuttgart) on the other hand. The sample period is from 2008 to 2018 with a quarterly data frequency.

4.4.1 Textual Corpus

The text corpus for this study consists of a large dataset of 319,956 news articles, containing 52.3 million words, published between Q1-2008 and Q4-2018 by two professional German real estate news providers: Immobilien Zeitung (*IZ*) and Thomas Daily (*TD*). *IZ* is the leading newspaper in the real estate industry with a weekly circulation of 11,000 copies while *TD* is part of the US-based CoStar Group, and provides the Thomas Daily Morning News daily for 33,000 subscribers. The news of *IZ* consist of an average length of 227 words

(SD: 246; Min: 36; Max: 4,253) and those of *TD* of 130 words (SD: 92; Min: 24; Max: 3,322).

The pre-processing of the news articles involves the removal of punctuation marks, numbers, non-alphabetical 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 context of real estate, after which the data is tokenised meaning that the text is divided into units (tokens) such as phrases, words and other meaningful entities. Here, the text corpus is segmented by words.

4.4.2 Real Estate and Macroeconomic Data

For the replication of the German residential real estate markets, the average price per m² of houses (*House Price*) and condominiums (*Condo Price*) is used for each city provided by empirica regio for the period from Q1-2008 to Q4-2018. The data basis of empirica regio is an extensive collection of property market data from more than 100 sources (listing data) and is conducted on a quarterly basis at city level.[16]

This study also includes macroeconomic and real estate control variables at city level that have demonstrated an impact on real estate markets, based on previous research on housing markets (Walker, 2016; Freybote and Seagraves, 2017; Ruscheinsky *et al.*, 2018a) and which had been provided by the statistical offices of the German federal states. Typically, the value of an asset refers to its present discounted value of future cash flow, which is why the average rent per m² (*Rent*) provided by empirica regio is used as one explanatory variable. As several housing studies emphasise the impact of labour market variables on housing demand

[16] Transaction data for property prices are not publicly available in Germany and are only rarely made available for research purposes by the valuation committees for property values in the respective cities. When comparing the transaction prices of the valuation committee and the listing prices of empirica regio for Berlin, similar mean values and standard deviations result, as well as a correlation of 0.98 between the two time series. Thus, it is assumed that the prepared data basis of empirica regio, which is based on hedonic pricing, serves as a very good approximation of the price development.

(Nakajima, 2011; Soo, 2018), the number of unemployed (*Unemp*) is included. To account for the housing supply, the number of building permits (*Buildper*) and the construction turnover (*Consturn*) are used. Lastly, the home loan interest rate (*Int*), which has an impact on housing demand and prices (Mayer and Sinai, 2009; Taylor, 2013), is taken into account and is available at the Bundesbank.[17] Table 4.1 presents descriptive statistics of residential property prices and the control variables transformed into their first differences for the entire seven cities (see table 4.10 in the appendix for descriptive statistics at city level, before data transformation).

[17] To replicate the economic situation of the respective cities, the Gross Domestic Product has also been included to the model, but did not improve the regression results. Furthermore, the average wages and salaries as well as population have also been tested as macroeconomic control variables.

Table 4.1 Descriptive Statistics

	Mean	***Median***	***SD***	***Min***	***Max***
House Price	0.5353	0.4300	1.4423	-11.4300	7.8900
House Price new	0.5678	0.4300	1.4939	-12.0700	8.2300
Condo Price	0.5959	0.4800	0.7054	-1.7600	3.0300
Condo Price new	0.7307	0.5700	0.7901	-2.3700	3.6300
Rent	0.0878	0.0800	0.1537	-0.4100	0.6800
Rent new	0.0969	0.0900	0.1666	-0.4100	0.7300
Unemp	-0.0836	-0.1667	0.3890	-0.9911	1.4112
Buildper	0.0039	-0.0400	0.9045	-2.6100	3.0700
Consturn	0.1664	0.0209	2.1670	-11.3446	14.6293
Int	-0.0693	-0.0567	0.1941	-0.4633	0.4833

Notes: Table 4.1 displays descriptive statistics of quarterly variables transformed into their first differences between 2008 and 2018 for all seven cities. *House (Condo) Price* is the average price per m² for residential properties. *Rent* is the average rental price per m² for residential properties. *New* indicates the average price per m² for new construction of residential properties. *Unemp* is the number of unemployed people. *Buildper* is the number of building permits. *Consturn* is the construction turnover, and *Int* is the home loan interest rate.

4.5 Methodology

4.5.1 Sentiment Measures

In order to extract information from the news articles, a two-step natural language processing procedure is applied, enabling the derivation of news coverage (*NC*) indicators in addition to the news sentiment (*NS*) indicators. In the first step, the method of word embeddings is used to assign the news articles referring to residential properties in the seven respective cities. A dictionary-based approach is applied in the second step to measure the tonality of these news articles.

News Coverage — Global Vectors Model

Topic model algorithms are often utilised to classify unstructured texts or large bodies of textual data. Since topic models, including the extensions of the traditional *LDA* algorithm, discover more general topics, this method is less applicable to this study at regional level: the topics of interest are very specific due to the subdivision of the residential market into seven cities. Hence, a word-embeddings approach, the Global Vector model

(*GloVe*), introduced by Pennington *et al.* (2014), is applied. *GloVe* can be used to identify relationships between words, including, e.g., synonyms. For instance, the unsupervised learning algorithm can detect the tokens ‘house’ and ‘building’ for word representation and therefore combines the advantages of global matrix factorisation and local context window methods.[18] In this vein, the algorithm generates a vector space of meaningful substructures because the training of the algorithm is only performed on non-zero elements in a global word-word co-occurrence statistic X, as opposed to the entire sparse matrix or the local context windows within a large text corpus. The objective function J of the weighted least squares regression model is specified by Pennington *et al.* (2014):

[18] 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 predicts linguistic patterns as linear relationships between the word vectors based on local context windows in order to perform better on word analogy tasks.

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

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. A weighting function readjusts the cost for each word pair to prevent the learning of simple common word pairs, as word pairs exhibit different occurrence frequencies in a text corpus. Given a co-occurrence count higher 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). Given the generated word representations of the text corpus, word vectors with high levels of similarity can be drawn for the residential market and for the seven cities. Then, the news articles can be compared with these word vectors and each article that contains at least one word representation for the residential market and one for a city is tagged for this regional housing market. With this process, a news article can be tagged, e.g., for the housing market in Berlin and the housing market in Hamburg, if the news article reports upon both cities. On the basis of this assignment, the news coverage indicator (*NC*) is calculated as follows:

$$NC_{i,t} = \frac{r_{i,t}}{n_t} \tag{4.2}$$

where $r_{i,t}$ describes the number of news articles regarding the residential market in a city *i* in period *t* and n_t is the total number of articles in period *t*. With $NC \in [0; 1]$, a higher value indicates that a housing market is discussed more frequently.

News Sentiment — Unbounded Polarity Score

A dictionary-based approach is applied for the second step of the textual analysis in order to extract sentiment from the news articles. For this approach, a sentiment dictionary classifying words into pre-defined categories of positive or negative sentiment is required

(Tetlock, 2007). According to Loughran and McDonald (2016), this classification has several advantages; as the subjectivity of researcher decisions is avoided, it is easily scalable with publicly available dictionaries and also easily replicable. However, the use of industry-specific dictionaries is recommended, as there are industry-specific word meanings, which has been shown by the studies of Doran *et al.* (2012) and Henry and Leone (2016). The dictionary used for this study, namely the German Real Estate Sentiment Dictionary (*GRESD*), was developed by Ruscheinsky *et al.* (2018a). The dictionary has been validated by a representative survey of German real estate professionals and contains 8,144 negative and 5,745 positive words.

Following the approach of Rinker (2019), each word w of a news article is compared with the polarised words of the sentiment dictionary *GRESD*, which are tagged either with +1 or -1. Thereafter, a polarised context cluster $c_{i,j,l}$ is defined, given n words before and after each polarised word by the researcher. In addition to the traditional ‘bag-of-words’ approach, a second dictionary containing 76 valence shifters is incorporated in order to achieve a more precise measurement of tonality. Valence shifters are negators (e.g., not) $w^{n}_{i,j,k}$, amplifiers (e.g., very) $w^{a}_{i,j,k}$, de-amplifiers (e.g., less) $w^{d}_{i,j,k}$ or adversative conjunctions (e.g., but) $w^{\mathrm{ac}}_{i,j,k}$, which can strengthen, weaken, or reverse the polarity of a word, and where the researcher provides the weight z. Subsequently, a valence shifter affects the score of the polarised context cluster. If no valence shifter is detected, the cluster is tagged as neutral $w^{0}_{i,j,k}$. Finally, 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 $\sqrt{w_{i,\,jn}}$:

$$\delta = \frac{c'_{i,j}}{\sqrt{w_{i,jn}}} \tag{4.3}$$

where:

$$c'_{i,j} = \sum((1 + w_{amp} + w_{deamp}) \cdot w^{p}_{i,j,k}(-1)^{2+w_{neg}}) \qquad (4.3.1)$$

$$w_{amp} = (w_b > 1) + \sum(w_{neg} \cdot (z \cdot w^{a}_{i,j,k})) \qquad (4.3.2)$$

$$w_{deamp} = max\left(w_{deamp'}, 1\right) \qquad (4.3.3)$$

$$w_{deamp'} = (w_b < 1) + \sum(z(-w_{neg} \cdot w^{a}_{i,j,k} + w^{d}_{i,j,k})) \qquad (3.3.4)$$

$$w_b = 1 + z_2 * w_{b'} \qquad (3.3.5)$$

$$w_{b'} = \sum(|w_{adversative\ conjunction}|, \ \dots, \ w^{p}_{i,j,k}, \ w^{p}_{i,j,k}, \ \dots, \ |w_{adversative\ conjunction}| \cdot -1) \qquad (4.3.6)$$

$$w_{neg} = \left(\sum w^{n}_{i,j,k}\right) mod\ 2 \qquad (4.3.7)$$

The resulting polarities of the news articles can then be aggregated on a quarterly basis and the news sentiment indicator (*NS*) can be determined as:

$$NS_{i,t} = \frac{\sum_{1}^{n} \delta_{i,t}}{n_t} \qquad (4.4)$$

where $\delta_{i,t}$ denotes the polarity score of a news article referring to the housing market in city i in period t and n_t is the total number of articles in period t. The sentiment scores are determined using article length weighted polarity scores, with a higher (lower) score than 0 indicating a positive (negative) sentiment, respectively. In order to investigate how residential real estate markets react in times when, e.g., both *NC* and *NS* achieve high values, the indicators are combined by addition:[19]

$$NCS_{i,t} = NC_{i,t} + NS_{i,t} \tag{4.5}$$

Hence, a high *NCS* score, for example, would indicate that a city's housing market is in the media spotlight and is being positively reported upon. Additionally, the

[19] Despite the fact that the standard deviations of *NC* and *NS* differ, the differences are minor and do not require 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.

ratio of positive (*POS*) and negative (*NEG*) news articles is calculated as several studies have revealed that media pessimism appears to be a more reliable indicator than media optimism (e.g., Tetlock, 2007; Ruscheinsky *et al.*, 2018a; Koelbl, 2020):

$$POS_{i,t} = \frac{\delta_{i,t}^{+}}{n_t} \tag{4.6}$$

$$NEG_{i,t} = \frac{\delta_{i,t}^{-}}{n_t} \tag{4.7}$$

where $\delta_{i,t}^{+}$ ($\delta_{i,t}^{-}$) is the number of news article referring to the housing market in a city *i* with a positive (negative) polarity score in period *t* and n_t comprises the total number of articles in quarter *t*. Both *POS* and *NEG* range from 0 to 1, whereby a higher (lower) value indicates a greater level of optimism (pessimism).

4.5.2 Model Specification

In order to test the ability of the generated sentiment measures to enable a better explanation of the residential property prices, the following first difference fixed effects models are estimated for all five sentiment indicators (*NC*, *NS*, *NCS*, *POS*, *NEG*):

$$\Delta\, house\ price_{i,t} = \beta_0 + \beta_1\, \Delta\, controls_{i,t} + \beta_2\, \Delta\, sentiment\ indicator_{i,t} + \Delta\, u_{i,t} \quad (4.8.1)$$

$$\Delta\, condo\ price_{i,t} = \beta_0 + \beta_1\, \Delta\, controls_{i,t} + \beta_2\, \Delta\, sentiment\ indicator_{i,t} + \Delta\, u_{i,t} \quad (4.8.2)$$

where i denotes the city, and t the period. Besides the sentiment indicator, the regression equations each include a vector of control variables containing economic and real estate variables, while u_{it} represents the error term. In order to account for heteroscedasticity and serial correlation, the models are estimated using robust standard errors.

The relationship between residential property prices and the sentiment measures is further investigated by performing Granger causality tests (Granger, 1969).

Panel Granger causality can aid the improvement of the understanding of the lead-lag relationships between sentiment and prices and to eventually determine whether news' sentiment has predictive power and vice versa.

4.6 Results

4.6.1 Sentiment Measures

News Coverage — Global Vectors Model

In the first step of the textual analysis, the *GloVe* model generates the word vectors on the basis of the given text corpus, enabling the word representations for residential properties and the cities' names to be retrieved. Extracts of the word representations are presented in table 4.2: All results in this paper were estimated in German and have been translated, and expressions consisting of two or more words in English appeared as single words in German. For example, the word vector 'dwelling house' provides the highest correspondence level of 61.2 % for 'residential building', followed by 'multi-family house' of 53.1 % or 'condominium' of 46.4 %.

Word representations for the cities, for instance Berlin, are given by ‘berliner’ (75.0 %), city districts such as ‘berlinmitte’ (55.2 %) and ‘capital’ (41.5 %). In general, the assignment of the news articles for the cities results in word vectors of the cities’ respective districts, street names and city characteristics such as ‘harbor city’ (41.8 %) for Hamburg or ‘mainmetropolis’ (31.2 %) for Frankfurt.

Table 4.2 Results Global Vectors Model

Word Representations	
Residential	dwelling house [0.612]; multifamily house [0.531]; living area [0.560]; condominium [0.464]; rental apartment [0.448]; town house [0.358]; new built apartment [0.326]; housing supply [0.263]; household [0.215]; social housing [0.211]
Berlin	berliner [0.750]; berlinmitte [0.552]; capital [0.415]; charlottenburg [0.414]; friedrichshain [0.391]; kreuzberg [0.347]; neukoelln [0.365]; schoeneberg [0.312]; marzahn [0.286]; steglitz [0.224]
Duesseldorf	duesseldorfer [0.731]; carlstadt [0.504]; derendorf [0.39 4]; golzheim [0.366]; koenigsallee [0.374]; flingernsued [0.319]; duesseltal [0 .268]; friedrichsstadt [0.245]; unterbilk [0.241]; bilk [0. 237]
Frankfurt	frankfurter [0.750]; altfrankfurt [0.513]; frankfurtinnens tadt [0.429]; nordend [0.423]; westend [0.418]; bockenh

	eim [0.377]; mainmetropolis [0.312]; sachsenhausen [0.307]; frankfurt airport [0.225]; gallus [0.218]
Hamburg	hamburger [0.761]; hamburgmitte [0.573]; hansestadt [0.469]; harbor city [0.418]; altona [0.393]; hamburgnord [0.350]; wandsbek [0.339]; eimsbuettel [0.321]; bergedorf [0.319]; harburg [0.276]
Cologne	koelner [0.754]; koelninnenstadt [0.664]; rodenkirchen [0.446]; lindenthal [0.410]; nippes [0.388]; koelnchorweiler [0.347]; porz [0.256]; koelnkalk [0.252]; ehrenfeld [0.231]; muehlheim [0.226]
Munich	muenchner [0.660]; muenchenaltstadt [0.559]; muenchensolln [0.552]; maxvorstadt [0.479]; bogenhausen [0.447]; isarvorstadt [0.394]; schwabing [0.377]; ludwigsvorstadt [0.259]; lehel [0.253]; bergamlaim [0.213]
Stuttgart	stuttgarter [0.742]; stuttgartwest [0.658]; stuttgartost [0.644]; stuttgartsued [0.642]; cannstadt [0.472]; vaihingen [0.412]; moehringen [0.360]; muehlhausen [0.291]; birkach [0.285]

Notes: Table 4.2 displays results of the *GloVe* model with 10 word vectors representing the housing market and the cities. The degree of similarity measured as a percentage is shown in brackets.

After classifying the news articles based on the word representations, the news coverage referring to the cities' housing markets can be calculated. In total, 74,201 news articles have been tagged, equalling 23.2 % of the given dataset. The market most reported upon is that of Berlin, resulting in 5.5 % of all news articles in the text corpus. It is not surprising that the capital of Germany, being the most populous, is the most frequently topic occurring. Still, compared with the other cities, Berlin is underrepresented by 12.4 % in the given dataset in relation to its population. Berlin is followed by Frankfurt (4.2 %), which is denoted by one of the smallest populations but a high price level and is thereby overrepresented by 10.6 %. Hamburg, the second largest city in terms of population, accounts for 3.9 %. The most expensive city, Munich, comprises merely 3.4 % of the dataset. The two cities in North Rhine-Westphalia, Duesseldorf and Cologne, amount to the same reporting shares of 2.3 %, while Stuttgart (1.6 %) is the least frequently reported upon but is also marked by high prices. In summary, the article shares of the other five cities approximately represent the size of the city in terms of population.

News Sentiment — Unbounded Polarity Score

In the second step of the textual analysis, the polarity score is calculated for each news article in order to evaluate whether the news article is written in more positive or negative language. Overall, a relatively neutral, slightly positive reporting style of the two professional news providers can be concluded through a mean polarity score for all news articles of 0.0716 (SD: 0.1300, Min: -0.8487, Max: 1.0169). The most 'positively' reflected residential markets are Stuttgart (0.1115) and Munich (0.1105). The largest cities, in terms of population, Hamburg and Berlin also achieve high scores of 0.1097 and 0.1029, respectively. These are followed by news articles associated with the housing markets of Duesseldorf (0.0998), Cologne (0.0979) and Frankfurt (0.0958).

Lastly, the polarity scores are calculated on a quarterly basis to generate the news sentiment indicator. With *NC* and *NS*, the further sentiment measures *NCS*, *POS* and *NEG* can be computed. Table 4.3 presents descriptive statistics of all generated news-based sentiment indicators for the text corpus; individual descriptive statistics for *IZ* and *TD* and at city level can be seen in table 4.11 and 4.12 in the appendix, respectively. The sentiment indicators are also transformed into their first differences for the regression analysis.

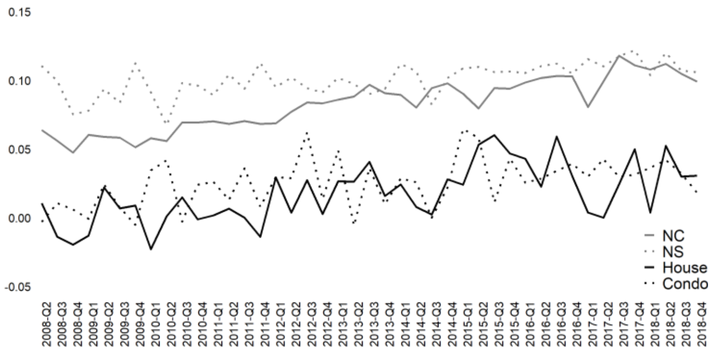
Table 4.3 Descriptive Statistics of Sentiment Measures

	Mean	***Median***	***SD***	***Min***	***Max***
NC	0.0500	0.0475	0.0221	0.0125	0.1180
NS	0.1035	0.1033	0.0129	0.0656	0.1393
NCS	0.1535	0.1516	0.0256	0.1032	0.2355
POS	0.6908	0.6918	0.0438	0.5267	0.7978
NEG	0.3092	0.3082	0.0438	0.2022	0.4733

Notes: Table 4.3 displays descriptive statistics of quarterly variables between 2008 and 2018 for all seven cities. *NC* is news coverage, *NS* is news sentiment, *NCS* is the combination of both indices, *POS* (*NEG*) is the positive (negative) ratio of news sentiment.

To illustrate the development of the indicators over time and at regional level, figure 4.1 shows the news article shares and the polarity scores for Berlin.

Figure 4.1 Berlin — News Coverage and News Sentiment over Time



Data Source: IZ Immobilien Zeitung Verlagsgesellschaft mbH, THOMAS DAILY GmbH, empirica regio GmbH.

The reporting intensity has increased throughout the period of our analysis: In 2008 the news coverage was ranging about 5 % and increased up to 12 % in 2017, while the measured sentiment was subject to greater fluctuations during the financial crisis and has then been slowly trending up. The growth rate of the average price per m² for houses in Berlin turns negative during the financial crisis and then also shows an increase up to 6 % in 2015; condo prices were less affected by the financial crisis and have also been rising at oscillating growth rates until 2015. Two peaks in growth rates are evident in both house and condo price data: 2012 and 2015. After both peaks, growth rates fluctuations fell. Overall, the sentiment indicator *NC* is highly correlated with condo prices of 0.88 and only slightly less with house prices (0.81). Positive correlations also result between *NS* and condo (house) prices given by 0.60 (0.58). This first observation already points in the direction of hypothesis 1, being that the sentiment measures are positively associated with residential property prices. This appears to be the case for news coverage; with regards to news sentiment, the picture is less clear for the smaller cities (Cologne, Munich and Stuttgart; see correlations for all cities in table 4.13 in the appendix).

4.6.2 Estimation Results

After extracting the information from the news articles and generating the corresponding time series for the different sentiment measures, their ability to better explain residential property prices will be tested using a first difference fixed effects regression model according to equation 4.8.1 and 4.8.2, and further investigated by performing bidirectional Granger causality tests. The results are displayed in table 4.4 (*House Price*) and 4.5 (*Condo Price*), and are run for each of the five measures of language.

Concerning the estimated results on house prices in table 4.4, the coefficient on *NC* shows a positive sign and is statistically significant at the 10 % level, suggesting that frequent reporting is associated with a positive price development. Moreover, this result is supported by the associated Granger Causality test, which demonstrates that *NC* runs ahead of the price movements by up to two quarters. The coefficient on *NS* is also positive and statistically significant at the 10 % level, but the Granger causality running from *NS* to house prices remains unsubstantiated by our tests, while the combined indicator *NCS* also shows a positive and statistically significant relationship with house prices, and Granger-causes prices up to two quarters. The models testing the ratios of positive or negative language also display the expected sign for the corresponding sentiment measure and are statistically significant, but do not indicate Granger causality. When regressing on average condo prices (table 4.5), the entire five sentiment measures also show the expected sign and are statistically significant at the 1 % level, with the exception of *NC*, which is significant at the 5 % level. Additionally, Granger causality runs from the sentiment indicators to condo prices, which also holds for two lags when referring to *NC* and *NCS*. Furthermore, higher values for adjusted R^2 can be observed in comparison with table 4.4.

Table 4.4 Estimation Results on House Price

House Price	***NC***	***NS***	***NCS***	***POS***	***NEG***
Rent	0.270	0.291	0.202	0.230	0.230
Unemp	0.389 *	0.499 *	0.508 *	0.495 *	0.495 *
Buildper	0.179	0.184	0.182	0.177	0.177
Consturn	0.004	0.008	0.012	0.010	0.010
Int	0.457 **	0.352 **	0.405 *	0.413 *	0.413 *
Sentiment Indicator	0.104 *	0.154 *	0.141 **	0.029 *	-0.029 *
Intercept	0.017 **	0.017 ***	0.019 ***	0.013 **	0.013 **
N	301	301	301	301	301
Adj. R^2	0.133	0.160	0.161	0.144	0.144
Panel Granger Causality					
Sentiment Indicator (-1)	1.91 *	1.23	3.05 ***	1.31	1.31
Sentiment Indicator (-2)	2.35 **	0.02	2.39 ***	0.34	0.34
House Price (-1)	0.55	0.51	0.18	1.02	1.02
House Price (-2)	0.99	0.34	0.19	0.35	0.35

Notes: Table 4.4 displays the results of the estimated regression models on house prices with quarterly sentiment indicators transformed into their first differences: news coverage (*NC*), news sentiment (*NS*), the combination of both indices (*NCS*), the positive (*POS*) and negative (*NEG*) ratio of *NS*. The set of the control variables includes the first difference of the average rental price per m² (*Rent*), the number of unemployed (*Unemp*), the number of building permits (*Buildper*), the construction turnover (*Consturn*) and the home loan interest rate (*Int*). Robust standard errors to heteroscedasticity and autocorrelation are used. Granger Causality values are shown for both directions. * denotes significance at the 10 % level, ** at the 5 % level and *** at the 1 % level. The regression is based on 301 observations according to the seven cities from Q2-2008 to Q4-2018.

Overall, these findings support the first hypothesis, being that news coverage and sentiment are positively associated with residential property prices. The predictive power of *NC* is given for both types of residential real estate, while the predictive power of *NS* can be observed for condo prices only. Accordingly, the results partially support the second hypothesis. As regressing on condo prices improves explanatory power and relevance of the news-based indicators, the third hypothesis, being that *NC* and *NS* more strongly correlate with condo than house prices cannot be rejected. These observations lead to the conclusion that the extracted textual information of professional news providers actually capture professional investor sentiment (Tetlock, 2007; Walker, 2016; Beracha *et al.*, 2019; Ploessl *et al.*, 2022) because if, for example, the interests of tenants were reflected more in the reporting of the newspapers, then rising prices would likely be perceived more negatively. Hence, this indicates that institutional investors may benefit from an advantage in terms of information. However, it appears to be worth evaluating both *NC* and *NS* as the combined indicator *NCS* gives the impression of being the most reliable indicator. Finally, the results highlight the fact that real estate markets are particularly subject to sentiment due to their characteristics such as low transparency, illiquidity and time lags in construction or transaction, and performing textual analysis can therefore be used to address this heterogeneity in housing markets.

Table 4.5 Estimation Results on Condo Price

Condo Price	*NC*	*NS*	*NCS*	*POS*	*NEG*
Rent	0.794 ***	0.733 ***	0.705 ***	0.705 ***	0.705 ***
Unemp	0.230 **	0.299 **	0.289 ***	0.295 **	0.295 **
Buildper	0.023 *	0.024 *	0.023 *	0.021 *	0.021 *
Consturn	-0.006	-0.004	-0.002	-0.003	-0.003
Int	-0.177 **	-0.191 **	-0.164 **	-0.161 **	-0.161 **
Sentiment Indicator	0.058 **	0.075 ***	0.053 ***	0.014 ***	-0.014 ***
Intercept	0.020 ***	0.022 ***	0.004 ***	0.020 ***	0.020 ***
N	301	301	301	301	301
Adj. R^2	0.174	0.210	0.194	0.187	0.187
Panel Granger Causality					
Sentiment Indicator (-1)	2.19 **	2.84 **	2.24 ***	1.68 *	1.68 *
Sentiment Indicator (-2)	1.74 *	0.79	1.87 **	0.41	0.41
Condo Price (-1)	1.26	0.59	0.51	1.50	1.50
Condo Price (-2)	0.74	0.70	0.76	1.21	1.21

Notes: Table 4.5 displays the results of the estimated regression models on condo prices with quarterly sentiment indicators transformed into their first differences: news coverage (*NC*), news sentiment (*NS*), the combination of both indices (*NCS*), the positive (*POS*) and negative (*NEG*) ratio of *NS*. The set of the control variables includes the first difference of the average rental price per m² (*Rent*), the number of unemployed (*Unemp*), the number of building permits (*Buildper*), the construction turnover (*Consturn*) and the home loan interest rate (*Int*). Robust standard errors to heteroscedasticity and autocorrelation are used. Granger Causality values are shown for both directions. * denotes significance at the 10 % level, ** at the 5 % level and *** at the 1 % level. The regression is based on 301 observations according to the seven cities from Q2-2008 to Q4-2018.

4.6.3 Robustness Tests

In order to check the robustness of the presented results, further models were fitted concerning the measuring of the residential real estate prices and the measuring of sentiment. Firstly, the average price per m² for new construction of residential properties is employed as the independent variable, on the assumption that the extracted sentiment affects new property prices the most quickly. Secondly, the five measures of language are divided by the two news providers *IZ* and *TD*. This also enables an initial indication of the composition of the targeted readership of the media.

The empirical results on new house prices, as shown in table 4.6, confirm the findings that *NC* and *NS* are positively associated with residential property prices. Contrary to expectations, adjusted R^2 decreases slightly for the models. Still, the Granger causality tests also show that *NC* and *NCS* lead new house prices by one quarter. Referring to the results on new condo prices in table 4.7, the main findings also continue to prevail, but the results deviate slightly in terms of size and significance of the five sentiment measures. Concerning the Granger causality tests, the null hypothesis is again rejected for all models at the first lag for the sentiment indicators, hence suggesting that there is a statistically significant Granger-causality running from the sentiment measures to new condo prices. However, the results are also indicative of the fact that new condo prices Granger-cause *NC*, i.e., there may be a feedback loop between *NC* and new condo prices. Given that this is the only identified statistically significant coefficient running in the opposite direction, we do not attribute a lot of emphasis to this result.

Table 4.6 Estimation Results on House Price, new Construction

House Price	*NC*	*NS*	*NCS*	*POS*	*NEG*
Rent	0.336	0.410	0.342	0.352	0.352
Unemp	0.480 *	0.544 *	0.559 *	0.557 *	0.557 *
Buildper	0.183	0.188	0.187	0.183	0.183
Consturn	0.016	0.017	0.020	0.019	0.019
Int	0.301 **	0.206 **	0.241 **	0.247 **	0.247 **
Sentiment Indicator	0.139 **	0.103 *	0.105 *	0.023 *	-0.023 *
Intercept	0.011	0.010	0.012	0.007	0.007
N	301	301	301	301	301
Adj. R^2	0.139	0.149	0.131	0.143	0.143
Panel Granger Causality					
Sentiment Indicator (-1)	2.42 **	0.85	2.90 ***	1.27	1.27
Sentiment Indicator (-2)	1.17	0.21	2.39 **	0.51	0.51
House Price (-1)	0.28	0.08	0.22	0.79	0.79
House Price (-2)	1.06	0.63	0.69	1.51	1.51

Notes: Table 4.6 displays the results of the estimated regression models on house prices of new construction with quarterly sentiment indicators transformed into their first differences: news coverage (*NC*), news sentiment (*NS*), the combination of both indices (*NCS*), the positive (*POS*) and negative (*NEG*) ratio of *NS*. The set of the control variables includes the first difference of the average rental price per m² (*Rent*), the number of unemployed (*Unemp*), the number of building permits (*Buildper*), the construction turnover (*Consturn*) and the home loan interest rate (*Int*). Robust standard errors to heteroscedasticity and autocorrelation are used. Granger Causality values are shown for both directions. * denotes significance at the 10 % level, ** at the 5 % level and *** at the 1 % level. The regression is based on 301 observations according to the seven cities from Q2-2008 to Q4-2018.

Table 4.7 Estimation Results on Condo Price, new Construction

Condo Price	***NC***	***NS***	***NCS***	***POS***	***NEG***
Rent	0.660 ***	0.607 ***	0.581 **	0.587 **	0.587 **
Unemp	0.213 **	0.277 **	0.266 **	0.262 **	0.262 **
Buildper	0.045 *	0.046 *	0.044 *	0.043 *	0.043 *
Consturn	-0.006	-0.001	-0.001	-0.002	-0.002
Int	-0.073 **	-0.082 **	-0.056 **	-0.054 **	-0.054 **
Sentiment Indicator	0.065 **	0.068 ***	0.046 **	0.010 *	-0.010 *
Intercept	0.025 ***	0.027 ***	0.028 ***	0.026 ***	0.026 ***
N	301	301	301	301	301
Adj. R^2	0.160	0.186	0.173	0.163	0.163
Panel Granger Causality					
Sentiment Indicator (-1)	1.82 *	3.27 ***	2.78 **	1.73 *	1.73 *
Sentiment Indicator (-2)	2.43 **	1.71 *	1.65 *	0.81	0.81
Condo Price (-1)	1.47	0.70	0.71	1.22	1.22
Condo Price (-2)	1.99 *	0.26	0.90	0.60	0.60

Notes: Table 4.7 displays the results of the estimated regression models on condo prices of new construction with quarterly sentiment indicators transformed into their first differences: news coverage (*NC*), news sentiment (*NS*), the combination of both indices (*NCS*), the positive (*POS*) and negative (*NEG*) ratio of *NS*. The set of the control variables includes the first difference of the number of the average rental price per m² (*Rent*), the number of unemployed (*Unemp*), the number of building permits (*Buildper*), the construction turnover (*Consturn*) and the home loan interest rate (*Int*). Robust standard errors to heteroscedasticity and autocorrelation are used. Granger Causality values are shown for both directions. * denotes significance at the 10 % level, ** at the 5 % level and *** at the 1 % level. The regression is based on 301 observations according to the seven cities from Q2-2008 to Q4-2018.

The information contained in the two text sources of *IZ* and *TD* is also compared on condo prices, as presented in table 4.8 and 4.9. Each of the five sentiment indicators shows the expected sign and all are statistically significant for both news providers. More precisely, *IZ* only indicates Granger causality running from *NC* and *NCS* to condo prices, while all sentiment measures of *TD* lead condo prices. Again, the combined indicator appears to be a more reliable indicator than the single indicators, as it is significant for both information providers and Granger-causes condo prices. However, the explanatory power for the divided datasets is reduced, when compared with table 4.5, which is why the analysis of several professional real estate-related news providers in addition to a high number of news articles may be worth considering. Overall, the findings imply that valuable additional information can be extracted from news articles, especially for investors and professional market participants.

Table 4.8 Estimation Results on Condo Price, using IZ

Condo Price	*NC*	*NS*	*NCS*	*POS*	*NEG*
Rent	0.837 ***	0.668 ***	0.656 ***	0.704 ***	0.704 ***
Unemp	0.198 ***	0.237 **	0.251 **	0.264 **	0.264 **
Buildper	0.026 *	0.020 *	0.019 *	0.016 *	0.016 *
Consturn	-0.002	-0.004	-0.005	-0.002	-0.002
Int	-0.199	-0.166	-0.150	-0.186 *	-0.186 *
Sentiment Indicator	0.091 *	0.038 ***	0.031 **	0.007 *	-0.007 *
Intercept	0.019 ***	0.023 ***	0.023 ***	0.021 ***	0.021 ***
N	301	301	301	301	301
Adj. R^2	0.176	0.193	0.186	0.178	0.178
Panel Granger Causality					
Sentiment Indicator (-1)	1.75 *	0.33	1.93 *	0.42	0.42
Sentiment Indicator (-2)	2.22 **	0.42	1.89 *	0.10	0.10
Condo Price (-1)	0.35	0.29	0.32	0.03	0.03
Condo Price (-2)	1.43	0.62	0.02	0.06	0.06

Notes: Table 4.8 displays the results of the estimated regression models on condo prices with quarterly sentiment indicators by *IZ* transformed into their first differences: news coverage (*NC*), news sentiment (*NS*), the combination of both indices (*NCS*), the positive (*POS*) and negative (*NEG*) ratio of *NS*. The set of the control variables includes the first difference of the average rental price per m² (*Rent*), the number of unemployed (*Unemp*), the number of building permits (*Buildper*), the construction turnover (*Consturn*) and the home loan interest rate (*Int*). Robust standard errors to heteroscedasticity and autocorrelation are used. Granger Causality values are shown for both directions. * denotes significance at the 10 % level, ** at the 5 % level and *** at the 1 % level. The regression is based on 301 observations according to the seven cities from Q2-2008 to Q4-2018.

Table 4.9 Estimation Results on Condo Price, using TD

Condo Price	***NC***	***NS***	***NCS***	***POS***	***NEG***
Rent	0.745 ***	0.809 ***	0.813 ***	0.759 ***	0.759 ***
Unemp	0.231 **	0.297 ***	0.298 **	0.265 **	0.265 **
Buildper	0.021 *	0.026 *	0.027 *	0.025 *	0.025 *
Consturn	-0.005	-0.003	-0.003	-0.002	-0.002
Int	-0.157 *	-0.185 *	-0.184 *	-0.148 *	-0.148 *
Sentiment Indicator	0.150 *	0.140 ***	0.137 ***	0.106 ***	-0.106 ***
Intercept	0.021 ***	0.020 ***	0.020 ***	0.020 ***	0.020 ***
N	301	301	301	301	301
Adj. R^2	0.171	0.189	0.186	0.175	0.175
Panel Granger Causality					
Sentiment Indicator (-1)	3.52 ***	1.86 *	4.14 ***	2.08 **	2.08 **
Sentiment Indicator (-2)	1.25	0.66	1.64 *	0.60	0.60
Condo Price (-1)	0.93	1.16	0.98	0.01	0.01
Condo Price (-2)	1.07	0.52	0.02	0.10	0.10

Notes: Table 4.9 displays the results of the estimated regression models on condo prices with quarterly sentiment indicators by *TD* transformed into their first differences: news coverage (*NC*), news sentiment (*NS*), the combination of both indices (*NCS*), the positive (*POS*) and negative (*NEG*) ratio of *NS*. The set of the control variables includes the first difference of the average rental price per m² (*Rent*), the number of unemployed (*Unemp*), the number of building permits (*Buildper*), the construction turnover (*Consturn*) and the home loan interest rate (*Int*). Robust standard errors to heteroscedasticity and autocorrelation are used. Granger Causality values are shown for both directions. * denotes significance at the 10 % level, ** at the 5 % level and *** at the 1 % level. The regression is based on 301 observations according to the seven cities from Q2-2008 to Q4-2018.

4.7 Conclusion

Real estate markets are characterised by heterogeneity and a lack of transparency, which can imply significant information asymmetry among market participants. Hence, the existing literature states that sentiment plays an important role in the pricing of real estate. In this study, we employ five different news-based sentiment measures while additionally considering the reporting intensity in order to address information asymmetry. For this analysis, a large text corpus of real estate-related news articles, spanning eleven years of reporting, was analysed regarding the German regional housing markets to determine to which extent the house price movements are anticipated in real estate news coverage and tonality of the articles. We investigate whether there is a relationship between news coverage or news sentiment and residential real estate prices at regional level, and whether there is a causality flow from news coverage or news sentiment to changes in residential property prices or vice versa, in addition to whether there are differences in the markets for houses and condominiums.

The findings of our analysis show that the obtained news-based indicators have a significant impact on residential real estate prices beyond all control variables. What is more, an increase in the intensity of reporting upon the housing market of a city is related to an increase in average purchase prices. The same applies to the news sentiment: more optimistic language correlates with increasing prices. The combination of these two measures proves to be a reliable indicator in all tested models. We also find that the extracted sentiment measures lead residential real estate prices up to two quarters. Finally, the explanatory power of the sentiment indicators is higher for condo prices than for house prices. Since condos are often of interest of institutional investors, this result indicates that information contained in the analysed news articles tends to reflect investor sentiment, which is consistent with prior studies.

In this vein, newspapers contain relevant information for market actors to better understand the market movements, narratives and sentiment. To balance the information asymmetries on real estate markets, it may be worth investing time and money in the evaluation of the flow of information in the media, even for small-scale market participants or information platforms. Hence, the information derived from news articles could thus be valuable to investors by influencing both opinion-formation and decision-making.

This additional qualitative information in textual form is available digitally and in real time and can therefore be extracted in a standardised and reproducible way without time delays, using our presented approach. Additionally, the textual analysis approach can be used to answer further research questions, such as e.g., whether news coverage and sentiment could be indicative for the likelihood of a specific policy intervention. The analytical setting could also be used to extract more information from further textual sources, such as social media posts, corporate publications, publications by special interest groups, or by political actors.

Future research could be carried out with the aim of overcoming the limitations of this study. Although the overall sample size for this study comprises 320,000 articles, it still spans merely one decade and two news providers. Furthermore, the chosen dictionary-based approach could be challenged and compared with other dictionaries. Beyond that, the comparison of less supervised algorithms and the abilities of humans to classify sentiment should be conducted, as language is a complex communication tool, implying that irony, sarcasm, exaggerations or metaphors may not be interpreted correctly. Therefore, one can finally conclude that the application of textual analysis and computational linguistics is still in its beginnings. Due to this, we expect that future studies to build on and improve the presented methods and data sets.

4.8 References

Aziz, S., Dowling, M.M., Hammami, H. and Piepenbrink, A. (2019), "Machine Learning in Finance: A Topic Modeling Approach", available at SSRN: https://ssrn.com/abstract=3327277.

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, pp. 36–49.

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.

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

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.

Cerchiello, P. and Nicola, G. (2018), "Assessing News Contagion in Finance", *Econometrics*, Vol. 6 No. 1, pp. 1–19.

Clayton, J., Ling, D.C. and Naranjo, A. (2008), "Commercial Real Estate Valuation: Fundamentals Versus Investor Sentiment", *SSRN Electronic Journal*.

Das, P.K., Freybote, J. and Marcato, G. (2014), "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.

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, pp. 915–953.

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.

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., 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. 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.

Kearney, C. and Liu, S. (2014), “Textual Sentiment in Finance: A Survey of Methods and Models”, *International Review of Financial Analysis,* Vol. 33, pp. 171–185.

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

Koelbl, M., Schuierer, R., Steininger, B.I. and Schaefers, W. (2021), “Can Risks be Good News? Revealing Risk Perception of Real Estate Investors using Machine Learning”, *Working Paper*.

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

Loughran, T. and McDonald, B. (2011), “When is a Liability Not a Liability? Textual analysis, Dictionaries, and 10-Ks”, *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, pp. 357–375.

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.

Mayer, C.J. and Sinai, T. (2009), "U.S. House Price Dynamics and Behavioral Economics", Policy Making insights from Behavioral Economics, *Federal Reserve Bank of Boston*.

Nakajima, M. (2011), "Understanding House-Price Dynamics", *Business Review*, No. Q2, pp. 20–28.

Nguyen, T.H. and Shirai, K. (2015), "Topic Modeling based Sentiment Analysis on Social Media for Stock Market Prediction", *Proceedings of the 53rd Annual*

Meeting of the Association for Computational Linguistics, Beijing, pp. 1354–1364.

Nordheim, G. von, Boczek, K. and Koppers, L. (2018), "Sourcing the Sources: An analysis of the use of Twitter and Facebook as a journalistic source over 10 years in The New York Times, The Guardian, and Süddeutsche Zeitung", *Digital Journalism*, Vol. 6 No. 7, pp. 807–828.

Nowak, A. and Smith, P. (2017), "Textual analysis in real estate", *Journal of Applied Econometrics*, Vol. 32 No. 4, pp. 896–918.

Nowak, A.D., Price, B.S. and Smith, P.S. (2021), "Real Estate Dictionaries Across Space and Time", *The Journal of Real Estate Finance and Economics*, Vol. 62 No. 1, pp. 139–163.

Pang, B., Lee, L. and Vaithyanathan, S. (2002), "Thumbs up? Sentiment Classification using Machine Learning Techniques", *Proceedings of the ALC-02 conference on Empirical methods in natural language processing, Association for Computional Linguistics, Morristown*, pp. 79–86.

Pennington, J., Socher, R. and Manning, C. (2014), “Glove: Global Vectors for Word Representation”, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Qatar*, 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*, Vol. 14 No. 3, pp. 381–400.

Ploessl, F., Paulus, N. and Just, T. (2022), “Trade vs. Daily Press: The Role of News Coverage and Sentiment in Real Estate Market Analysis”, *Working Paper*.

Rinker, T. (2019), *Package ‘sentimentr’: Calculate Text Polarity Sentiment*, R 3.4.0, available at: https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf.

Ruscheinsky, J.R., Kandlbinder, K., Schaefers, W., Dietzel, M.A. and Rochdi, K. (2018a), “Predicting

Real Estate Market Movements: the First Textual Analysis-Based-Sentiment Application in Germany", *Working Paper*.

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

Schumaker, R.P. and Chen, H. (2009), "Textual analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System", *ACM Transactions on Information Systems*, Vol. 27 No. 2, pp. 1–19.

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

Soo, C.K. (2018), "Quantifying sentiment with news media across local housing markets", *Review of Financial Studies*, Vol. 31 No. 10, pp. 3689–3719.

Statista (2022), "Zeitschriften in Deutschland", available at: https://de.statista.com/statistik/studie/id/6598/dokument/zeitschriften-statista-dossier/.

Taylor, J.B. (2013), *Getting off track: How government actions and interventions caused, and worsened the financial crisis*, Hoover Press.

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. (2016), "The direction of media influence: Real-estate news and the stock market", *Journal of Behavioral and Experimental Finance*, Vol. 10, pp. 20–31.

Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. and Li, X. (2011), "Comparing Twitter and Traditional Media Using Topic Models", *Advances in Information*, pp. 338–349.

4.9 Appendix

Table 4.10 Descriptive Statistics per City

	Mean	***Median***	***SD***	***Min***	***Max***
House Price_B	2,311.11	2,029.50	638.63	1,733.00	3,774.00
House Price_D	3,162.84	3,224.00	676.21	2,220.00	4,637.00
House Price_F	3,352.68	3,233.50	641.69	2,616.00	4,698.00
House Price_H	2,989.57	2,988.50	506.61	2,303.00	4,066.00
House Price_C	2,721.59	2,656.00	425.54	2,182.00	3,824.00
House Price_M	5,752.18	5,751.00	1,343.65	3,622.00	8,452.00
House Price_S	3,841.89	3,770.50	722.17	2,758.00	5,086.00
House Price new_B	2,523.41	2,331.00	569.14	2,018.00	4,005.00
House Price new_D	3,579.00	3,606.50	734.76	2,480.00	5,136.00
House Price new_F	3,636.57	3,538.50	570.75	2,823.00	4,774.00
House Price new_H	3,167.43	3,141.50	535.23	2,433.00	4,273.00
House Price new_C	2,860.57	2,730.50	646.85	2,094.00	4,599.00
House Price new_M	5,840.75	5,820.50	1,186.42	3,858.00	8,611.00
House Price new_S	4,272.86	4,310.00	813.31	2,963.00	5,959.00
Condo Price_B	2,152.07	1,971.50	762.77	1,302.00	3,803.00
Condo Price_D	2,256.66	2,222.00	504.66	1,624.00	3,290.00
Condo Price_F	2,814.59	2,492.00	883.73	1,905.00	4,910.00
Condo Price_H	2,725.43	2,709.50	676.50	1,756.00	3,963.00
Condo Price_C	2,180.05	2,068.00	534.92	1,599.00	3,242.00
Condo Price_M	4,295.91	4,137.00	1,511.59	2,433.00	7,133.00
Condo Price_S	2,595.91	2,333.00	788.22	1,751.00	4,096.00
Condo Price new_B	3,163.09	2,951.00	934.93	2,141.00	5,394.00
Condo Price new_D	3,429.66	3,184.50	870.60	2,433.00	5,167.00
Condo Price new_F	3,616.20	3,299.00	965.52	2,576.00	6,154.00
Condo Price new_H	3,559.02	3,551.00	753.32	2,498.00	4,977.00
Condo Price new_C	3,182.05	3,104.00	569.06	2,517.00	4,276.00
Condo Price new_M	5,253.25	5,115.50	1,579.99	3,304.00	8,324.00
Condo Price new_S	3,801.68	3,416.50	1,096.31	2,664.00	6,022.00
Rent_B	7.47	7.57	1.48	5.30	10.26
Rent_D	8.61	8.58	0.71	7.28	10.06
Rent_F	10.63	10.79	1.18	8.93	12.69
Rent_H	9.43	9.65	0.79	7.82	10.72
Rent_C	8.90	8.85	0.86	7.64	10.55
Rent_M	12.66	12.69	1.85	10.32	16.07
Rent_S	9.55	9.60	1.29	7.90	11.90
Rent new_B	10.51	10.93	1.74	7.56	13.27
Rent new_D	10.88	10.96	0.89	8.97	12.40

Rent new_F	12.26	12.22	1.17	10.58	14.41
Rent new_H	11.40	11.76	0.86	9.52	12.62
Rent new_C	10.17	10.22	0.88	8.80	11.81
Rent new_M	14.27	14.11	2.21	11.45	18.09
Rent new_S	11.44	11.54	1.39	9.59	13.88
Unemp_B	11.59	11.72	2.04	7.63	14.90
Unemp_D	8.60	8.80	0.94	6.50	9.96
Unemp_F	7.16	7.42	0.97	4.97	9.00
Unemp_H	7.53	7.48	0.69	6.00	9.06
Unemp_C	9.47	9.55	0.91	7.40	11.34
Unemp_M	4.99	5.07	0.65	3.43	6.20
Unemp_S	5.50	5.53	0.67	3.97	6.93
Buildper_B	544.95	531.50	118.74	345.00	804.00
Buildper_D	75.23	69.00	36.60	6.00	186.00
Buildper_F	123.77	121.00	45.78	41.00	236.00
Buildper_H	431.39	422.00	111.45	244.00	696.00
Buildper_C	193.57	197.00	69.31	72.00	348.00
Buildper_M	338.70	334.00	76.77	130.00	513.00
Buildper_S	89.16	68.00	64.64	23.00	273.00
Consturn_B	164,729.18	137,564.00	95,894.93	48,295.00	402,354.00
Consturn_D	11,209.43	10,383.00	6,834.50	2,592.00	30,755.00
Consturn_F	28,902.41	26,279.50	10,176.48	13,266.00	53,335.00
Consturn_H	94,961.00	91,387.00	32,029.82	34,593.00	215,154.00
Consturn_C	52,058.14	52,638.50	17,457.52	26,208.00	99,781.00
Consturn_M	35,447.64	34,847.50	10,581.73	12,354.00	56,210.00
Consturn_S	3,332.56	3,419.40	686.60	2,128.45	4,326.20
Int_B-S	3.11	3.00	1.12	1.69	5.26

Notes: Table 4.10 displays descriptive statistics of quarterly variables between 2008 and 2018 for each city. *House* (*Condo*) *Price* is the average price per m² in EUR for residential properties. *Rent* is the average rental price per m² in EUR for residential properties. *New* indicates the average price per m² for new construction of residential properties. *Unemp* is the of the number of unemployed people in %. *Buildper* is the number of building permits. *Consturn* is the construction turnover in 1,000 EUR and *Int* is the home loan interest rate in %. *B* is Berlin; *D* is Duesseldorf; *F* is Frankfurt; *H* is Hamburg; *C* is Cologne; *M* is Munich and *S* is Stuttgart.

Table 4.11 Descriptive Statistics of Sentiment Measures of IZ and TD

	Mean	***Median***	***SD***	***Min***	***Max***
NC_iz	0.0217	0.0182	0.0118	0.0048	0.0641
NS_iz	0.1198	0.1207	0.0184	0.0641	0.1695
NCS_iz	0.1415	0.1417	0.0221	0.0776	0.2001
POS_iz	0.7632	0.7701	0.0584	0.5690	0.9057
NEG_iz	0.2368	0.2299	0.0584	0.0943	0.4310
NC_td	0.0285	0.0275	0.0119	0.0060	0.0626
NS_td	0.0918	0.0926	0.0144	0.0541	0.1322
NCS_td	0.1203	0.1204	0.0179	0.0614	0.1645
POS_td	0.6382	0.6410	0.0557	0.4000	0.8261
NEG_td	0.3618	0.3590	0.0557	0.1739	0.6000

Notes: Table 4.11 displays descriptive statistics of quarterly variables between 2008 and 2018 for all seven cities generated from the text corpus subdivided into *IZ* and *TD*. *NC* is news coverage, *NS* is news sentiment, *NCS* is the combination of both indices, *POS* (*NEG*) is the positive (negative) ratio of news sentiment.

Table 4.12 Descriptive Statistics of Sentiment Measures per City

	Mean	*Median*	*SD*	*Min*	*Max*
NC_B	0.0834	0.0850	0.0189	0.0475	0.1180
NC_D	0.0343	0.0358	0.0070	0.0159	0.0476
NC_F	0.0644	0.0634	0.0145	0.0348	0.0891
NC_H	0.0595	0.0596	0.0109	0.0386	0.0838
NC_C	0.0345	0.0359	0.0082	0.0212	0.0521
NC_M	0.0513	0.0503	0.0118	0.0255	0.0748
NC_S	0.0248	0.0254	0.0053	0.0125	0.0349
NS_B	0.1012	0.1042	0.0120	0.0673	0.1221
NS_D	0.0994	0.0995	0.0102	0.0789	0.1202
NS_F	0.0948	0.0930	0.0105	0.0656	0.1136
NS_H	0.1090	0.1086	0.0091	0.0915	0.1272
NS_C	0.0983	0.0976	0.0121	0.0736	0.1289
NS_M	0.1110	0.1111	0.0100	0.0905	0.1360
NS_S	0.1116	0.1120	0.0149	0.0848	0.1393
NCS_B	0.1846	0.1864	0.0275	0.1228	0.2355
NCS_D	0.1337	0.1335	0.0135	0.1032	0.1563
NCS_F	0.1592	0.1573	0.0210	0.1127	0.1964
NCS_H	0.1685	0.1693	0.0168	0.1347	0.2071
NCS_C	0.1328	0.1309	0.0136	0.1039	0.1582
NCS_M	0.1623	0.1642	0.0134	0.1260	0.1914
NCS_S	0.1365	0.1335	0.0161	0.1038	0.1710
POS_B	0.6915	0.6997	0.0354	0.6031	0.7529
POS_D	0.6722	0.6767	0.0420	0.5267	0.7571
POS_F	0.6584	0.6591	0.0349	0.5862	0.7258
POS_H	0.7182	0.7232	0.0365	0.6390	0.7857
POS_C	0.6736	0.6793	0.0406	0.5761	0.7500
POS_M	0.7067	0.7133	0.0319	0.6475	0.7763
POS_S	0.7161	0.7106	0.0440	0.6377	0.7978
NEG_B	0.3085	0.3003	0.0354	0.2471	0.3969
NEG_D	0.3278	0.3233	0.0420	0.2429	0.4733

NEG_F	0.3416	0.3410	0.0349	0.2742	0.4138
NEG_H	0.2818	0.2768	0.0365	0.2143	0.3610
NEG_C	0.3264	0.3208	0.0406	0.2500	0.4239
NEG_M	0.2933	0.2868	0.0319	0.2237	0.3525
NEG_S	0.2839	0.2894	0.0440	0.2022	0.3623

Notes: Table 4.12 displays descriptive statistics of quarterly variables between 2008 and 2018 for each city. *NC* is news coverage, *NS* is news sentiment, *NCS* is the combination of both indices, *POS* (*NEG*) is the positive (negative) ratio of news sentiment. *B* is Berlin; *D* is Duesseldorf; *F* is Frankfurt; *H* is Hamburg; *C* is Cologne; *M* is Munich and *S* is Stuttgart.

Table 4.13 Correlations of Prices and Sentiment Measures

House Price	***NC***	***NS***	***NCS***	***POS***	***NEG***
Berlin	0.8066	0.5788	0.8061	0.5269	-0.5269
Duesseldorf	0.6743	0.2670	0.5491	0.2837	-0.2837
Frankfurt	0.7828	0.6171	0.8484	0.5370	-0.5370
Hamburg	0.8840	0.4575	0.8260	0.3566	-0.3566
Cologne	0.7170	-0.0056	0.4257	0.2821	-0.2821
Munich	0.8449	-0.3544	0.4812	-0.0761	0.0761
Stuttgart	0.7630	-0.0532	0.2008	-0.0806	0.0807

Condo Price	***NC***	***NS***	***NCS***	***POS***	***NEG***
Berlin	0.8753	0.5962	0.8608	0.5532	-0.5532
Duesseldorf	0.7045	0.2012	0.5152	0.2345	-0.2345
Frankfurt	0.7997	0.5863	0.8446	0.4740	-0.4740
Hamburg	0.8846	0.4485	0.8216	0.3757	-0.3757
Cologne	0.7631	-0.0434	0.4194	0.2527	-0.2527
Munich	0.8701	-0.3151	0.5327	-0.0511	0.0510
Stuttgart	0.6981	0.0520	0.2768	-0.0222	0.0223

Notes: Table 4.13 displays the correlations between house prices or condo prices and the sentiment measures (no transformation of data). *NC* is news coverage, *NS* is news sentiment, *NCS* is the combination of both indices, *POS* (*NEG*) is the positive (negative) ratio of *NS*.

5. Conclusion

This dissertation investigates whether the non-numeric information revealed in real estate-related news media can contribute towards increased transparency in real estate markets and can support market participants in making profound investment decisions. This is achieved by extending the common measurement of text-based sentiment in the real estate literature by not only quantifying the tonality but also the reporting intensity of specific topics or asset classes. The following section summarises the motivation, research design and main findings of the three contributing papers and concludes with limitations and further research opportunities.

5.1 Executive Summary

Paper 1 — Cyclicity of Real Estate-related Trends: Topic Modelling and Sentiment Analysis on German Real Estate News

Megatrends are described as being stable and long-term developments and have a significant impact on the economy, society and the environment, but cannot be influenced by individuals (Naisbitt, 1982; Groddeck and Schwarz, 2013). Through which trends is the real estate industry affected, and are these subject to an expected trend pattern or even cyclicity? Pfnür and Wagner (2020) have identified six trends of the most importance to the real estate industry by means of a survey among 249 German real estate professionals. When these trends underlie cyclicity, this may have an impact on investor behaviour: if news coverage of sustainable real estate increases, investors might assume that this is of growing importance to their clients and invest more in sustainable buildings and therefore expect higher returns.

Hence, this paper analyses the news coverage and news sentiment of these six trends from almost 170,000 news articles published over a period of 21 years by a major German news provider, the *Immobilien Zeitung*. Methods of natural language processing are used to determine the news coverage of each trend: topic modelling, incorporating seed words partially identified via word embeddings. Thereafter, by applying a dictionary-based approach, the news sentiment, the level of optimistic or pessimistic language, is quantified. In this respect, an industry-specific German dictionary for the real estate industry created by Ruscheinsky *et al.* (2018), which was validated by a representative survey among real estate professionals, is employed.

The results show that the trends of urbanisation, globalisation and sustainability account for the largest shares of reporting, while digitalisation is — contrary to expectations — the least frequently discussed. An upward trend can only be identified for the topic of sustainability, while the article shares dedicated to globalisation has declined since the financial crisis. Overall, the tonality of the analysed trade newspaper can be classified as being relatively neutral, but there are significant differences between the trend topics: the highest polarity scores are achieved on digitalisation, while regulatory issues display the most negative sentiment. However, the shares are subject to change over time, both in terms of news coverage and sentiment, even after adjusting for seasonal and remainder components.

As the news coverage and sentiment of the trends follow short-term and strong movements, in contrast to their characteristic of being long and stable processes, the results provide new insights: the investment decisions of market participants and investors may be affected, as they also build up their expectations and assessments of the information provided by the news media. In this vein, this is the first paper to integrate seeded topic modelling and word embeddings into real estate analysis to provide a means through which to extract information in a standardised and replicable way.

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

Even in the age of digitalisation, real estate markets are still comparatively intransparent. Information is not fully available, leading to market inefficiencies (Black, 1986). Thus, investors are also looking for market-relevant information and do so by reading news articles, but is this investment of time and money viable? And are different types of newspapers an equivalent source of information for different markets? Several studies have highlighted the importance of sentiment in real estate markets due to this heterogeneity (e.g., Clayton *et al.*, 2008; Hausler *et al.*, 2018; Beracha *et al.*, 2019), while attempting to measure general market sentiment. Marcato and Nanda (2016) have attributed varying degrees of importance of survey-based and market fundamental-based indicators to residential and commercial real estate. Hence, news-based sentiment indicators should also be derived for the individual asset classes.

In addition, it is assumed that news media meets different informational needs depending on the targeted readership. The understanding of such differences may be valuable to investors as their decision-making is also based on the conveyed narratives and sentiment by the news media.

In this respect, this study examines the relationship between news coverage or sentiment and total returns of the asset main classes (residential, office and retail). Computational linguistics, namely seeded topic modelling, word embeddings and dictionary-based sentiment analysis enables us to derive three sentiment indicators for each of the three asset classes from almost 137,000 articles published by two trade (*Immobilien Zeitung* and *Immobilien Manager*) and two daily (*Frankfurter Allgemeine Zeitung* and *Handelsblatt*) newspapers in the period between 2010 and 2020. Subsequently, to investigate whether news coverage and sentiment either lead or lag in terms of the performance of assets classes, a vector autoregression framework is applied.

The regression results indicate divergent reporting positions of the two types of newspapers: Daily newspapers report in a more pessimistic tone in times of increasing returns in the residential market than the trade newspapers. Moreover, we find a significant relationship between the changes in the news-based sentiment indicators and the development of total returns within the respective asset classes. This is the case in particular for the capital growth component, indicating that the information revealed in the news reflects investor sentiment. Thus, trade newspapers outperform daily newspapers in the prediction of future total returns. Overall, the sentiment indicators Granger-cause total returns, but regarding retail markets, the generated indicators also lag prior returns. Finally, the study suggests that the news-based sentiment indicators are more suitable for explaining future market developments in comparison with survey-based indicators.

To the best knowledge of the authors, this is the first study to extract both the news coverage and sentiment for the different real estate asset classes and different types of newspapers, respectively. Extracting qualitative data from real estate-related news media provides valuable insights, facilitating a better understanding of the future market movements that are not based solely on fundamental changes.

Paper 3 — News Coverage vs. Sentiment: Evaluating German Residential Real Estate Markets

Market participants are looking for new information when following the daily flow of information in the news media. Indirectly, however, they may be influenced by the sentiment conveyed by the journalists, as this profession functions as a gatekeeper for information (White, 1950). In a sense, the news does not simply change the level of information concerning a certain asset class, but can shape the overall market sentiment, i.e., the behaviour of investors over the expected price developments in a certain market. Accordingly, changes in sentiment can indicate changes in hard market data, such as prices or return. In order to gain an insight into the future market perceptions, it can be beneficial for investors to better understand the mechanisms of the news media.

Based on the studies of Soo (2018) and Ruscheinsky *et al.* (2018), this paper examines the relationship between news coverage or sentiment and residential real estate prices in Germany on a regional level. By means of textual analysis, using word embeddings and a dictionary-based approach, five different sentiment measures based on almost 320,000 news articles published by two German professional news providers (*Immobilien Zeitung* and *Thomas Daily*) in the context of real estate between 2008 and 2018 are obtained. This is followed by a first difference fixed effects regression to investigate the markets for houses and condominiums in the seven biggest cities in Germany.

Most importantly, the results cannot reject the hypothesis that the news-based indicators have a significant positive relationship with residential real estate prices and also the second hypothesis as they lead the prices up to two quarters. More precisely, the combined indicator based on news coverage and sentiment proves to be a reliable indicator and highlights the importance of analysing reporting intensity in addition. Lastly, regressing on prices for condominiums leads to an increased explanatory power when compared with houses, which indicates that the indicators tend to reflect investor sentiment and therefore supports the third hypothesis.

In summary, the study shows that the application of algorithms in the field of computational linguistics for real estate-related text corpora can be used to gain an insight into future market perceptions as residential real estate markets are particularly subject to sentiment. Even traditional newspapers still hold market-relevant information for investors in order to enable a better understanding of the market narratives and sentiment.

5.2 Final Remarks and Further Research

The literature in the context of real estate has shown that real estate markets in particular are subject to sentiment due to their characteristics, such as low transparency, illiquidity and time lags in construction or transaction. Still, actors in the real estate industry predominately rely on quantitative data sets that are oftentimes of limited availability for their decision-making. At present, the potential of qualitative data is becoming acknowledged by academia and in practice, whilst the narratives in the news media have so far been omitted and not taken into account. Through the extension of the common measures of text-based sentiment through the implementation of an indicator that captures the news coverage, this dissertation contributes to the existing literature by investigating whether news coverage and sentiment extracted from real estate-related texts lead to the provision of new information and help to explain future real estate market movements.

In this vein, the first paper of this dissertation illustrates the potential of the introduced approach, while the application is undertaken in paper 2 and 3. The analysis of various textual sources in the news media for different markets and asset classes in Germany shows that the qualitative information of real estate-related news providers is truly informative and relates to returns and price developments. The consideration of news coverage promises to improve the understanding of how markets respond to narratives. Furthermore, the results indicate that text-based sentiment is less volatile than survey-based sentiment and might therefore be a more solid indicator. Moreover, trade newspapers provide relatively accurate data, which is why multiple newspapers should be processed in terms of news coverage and sentiment, enabling faster and better-informed decisions.

Nevertheless, this dissertation is constrained by certain limitations: First, the present analysis focusses on German real estate markets. There are already several studies concerning text-based sentiment in English-speaking countries, but the approach undertaken in this study should be tested in different areas, i.e., at a national and regional level. Future research could also investigate specific asset classes such as healthcare facilities or logistic properties to further explore the importance of textual analysis in even more intransparent markets. Second, qualitative information, mainly from newspapers, has been quantified in this case, but the potential applications are not limited to the news media. For instance, the investigation of market reports, corporate publications, professional research publications or social media content could be further sources for qualitative data. Beyond this, the first methods for evaluating audio and video in a standardised way are currently in their infancy but have the potential to open up a wide range of possible applications. Third, the algorithms employed in this dissertation have emerged in a new field of natural language processing and are being criticized for several reasons. The quality and suitability of dictionaries is crucial for the texts to be analysed.

Hence, industry-specific wordlists provided in multiple languages would be beneficial for academia. Also, unsupervised learning algorithms should be further compared with the abilities of humans to classify sentiment. Furthermore, content analysis of text documents is not a trivial task as topic modelling algorithms detect rather general topics, but extended algorithms already enable more differentiated applications. However, the given limitations of the recent algorithms will encourage researchers to address these issues and improve the techniques to extract further and precise information from texts.

In conclusion, the three contributing papers of this dissertation illustrate that the digitally available news media sources facilitate the improvement of the predictions of future market movements by means of newly developed techniques in computational linguistics. As information is a valuable resource, an understanding of the narratives and sentiments set in the real estate industry by the news media can support market participants in the making of profound decisions.

5.3 References

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. 44 No. 3, pp. 528–543.

Clayton, J., Ling, D.C. and Naranjo, A. (2008), "Commercial Real Estate Valuation: Fundamentals Versus Investor Sentiment", *SSRN Electronic Journal*.

Groddeck, V. von and Schwarz, J.O. (2013), "Perceiving megatrends as empty signifiers: A discourse-theoretical interpretation of trend management", *Futures*, Vol. 47, pp. 28–37.

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.

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.

Naisbitt, J. (1982), *Megatrends: Ten new directions transforming our lives,* Warner Books, New York.

Pfnür, A. and Wagner, B. (2020), “Transformation of the real estate and construction industry: empirical findings from Germany”, *Journal of Business Economics,* Vol. 2 No. 1, p. 79.

Ruscheinsky, J.R., Kandlbinder, K., Schaefers, W., Dietzel, M.A. and Rochdi, K. (2018), “Predicting Real Estate Market Movements: the First Textual Analysis-Based-Sentiment Application in Germany”, *Working Paper*.

Soo, C.K. (2018), “Quantifying sentiment with news media across local housing markets”, *Review of Financial Studies,* Vol. 31 No. 10, pp. 3689–3719.

White, D.M. (1950), “The "gate keeper": A case study in the selection of news”, *Journalism Quarterly,* Vol. 27, pp. 383–390.