REMV
REAL ESTATE MANAGEMENT AND VALUATION
Journal homepage: www.remv-journal.com

MINING REAL ESTATE DATA: A SYSTEMATIC REVIEW OF TEXT-BASED APPROACHES

Joanna Węgrzyn*[1]

[1] *Department of Real Estate and Investment Economics, Krakow University of Economics, ul. Rakowicka 27, 31-510 Kraków, Poland, email: wegrzynj@uek.krakow.pl, ORCID: 0000-0003-3327-6313*

** Corresponding author*

ARTICLE INFO

Keywords:

systematic literature review, science mapping, text mining, topic modelling, real estate

JEL Classification:

C45, C81, R30

ABSTRACT

The rise of Big Data has transformed information processing, with text mining techniques gaining prominence across various disciplines, including real estate market analysis. The research aim is to provide a comprehensive overview of this field, highlighting emerging trends and potential areas for future research. To explore the application of text mining approaches, the study presents a systematic literature review (SLR) combining performance analysis, science mapping techniques and topic modeling. The analysis reveals a noticeable increase in research activity post-2015. Key application areas of text mining techniques include housing and real estate markets, urban planning and social media, smart cities and public spaces, online reviews and consumer feedback, and sustainable development. Additionally, the research identifies challenges related to data quality, standardization, and computational efficiency. The study concludes by identifying future research directions and potential applications for real estate practice.

Citation: Węgrzyn, J. (2026). Mining real estate data: A systematic review of text-based approaches. *Real Estate Management and Valuation*, *34*(1), 1-15. https://doi.org/10.2478/remav-2026-0001

1. Introduction

The era of Big Data is undoubtedly transforming the way we think about, collect, and process information. Large volumes of data are often gathered from distributed sources and generated dynamically. Nevertheless, this aspect alone does not completely capture the essence of the Big Data concept. The uniqueness of Big Data arises from its four main characteristics (Stephens-Davidowitz, 2017). First, Big Data introduces new types of data. Second, data extracted from new sources are often anonymized, which is why Big Data appears to provide a more reliable representation of human behavior. Third, it enables a more granular focus on small populations. Finally, Big Data facilitates numerous causal experiments. Stephens-Davidowitz (2017) refers to these four characteristics as the four powers of Big Data.

Researchers across various disciplines are increasingly exploring ways to integrate Big Data into empirical studies. The potential of large-scale textual data, in particular, has sparked interest in fields such as social sciences, economics, and urban studies. Advanced computational tools now allow for the extraction of meaningful insights from vast amounts of unstructured textual information, opening new avenues for interdisciplinary research.

As data becomes more accessible, the demand for efficient processing and interpretation continues to rise. Advanced algorithms make it possible to uncover hidden patterns, identify trends, and gain deeper insights into studied phenomena. In this context, text mining techniques are playing an increasingly significant role. These methods are also gaining popularity in real estate market analysis.

The objective of this paper is to identify significant trends, essential topics, and research gaps regarding the application of text mining techniques within the real estate market. For that purpose, a systematic review of the literature will be performed.

The paper is structured as follows: The next parts of the introductory section explain the idea of text mining. Section 2 outlines the research methodology, while Section 3 deals with empirical data and introduces data visualization. Section 4 provides a discussion, and finally, Section 5 concludes with key takeaways and final remarks.

1.1. Origins of text mining

It is essential to begin an exploration of the advancement of text mining techniques by first defining a concept that seems to be intrinsically linked to it—textual analysis. Textual analysis is a methodology that involves understanding language, symbols, and/or pictures present in texts to gain information regarding how people make sense of and communicate life and life experiences (Allen, 2017), whereas text mining (sometimes called text data mining) can be perceived as a process of exploratory data analysis that leads to the discovery of unknown information, or to answers to questions for which the answer is not currently known (Hearst, 1999). While both fields involve the examination of textual data, textual analysis focuses on interpreting and understanding meaning, often within linguistic, literary, or social contexts, whereas text mining mainly focuses on computational methods to extract structured information and uncover hidden patterns.

Establishing a clear distinction between these concepts provides a foundation for understanding how advancements in machine learning, neural networks, and NLP have contributed to the evolution of text mining methodologies. An interesting insight into the development of modern text mining can be drawn from the trends depicted in the chart (Figure 1). The graph highlights the remarkable rise in the frequency of terms such as "machine learning", "neural networks" and "natural language processing" (NLP) over time, underscoring their growing importance in this domain. These technological advancements have played a pivotal role in the evolution of text mining methodologies (Lula, 2005). The surge in "machine learning" and "neural networks", especially after 2010, illustrates the increasing reliance on sophisticated algorithms for text analysis.

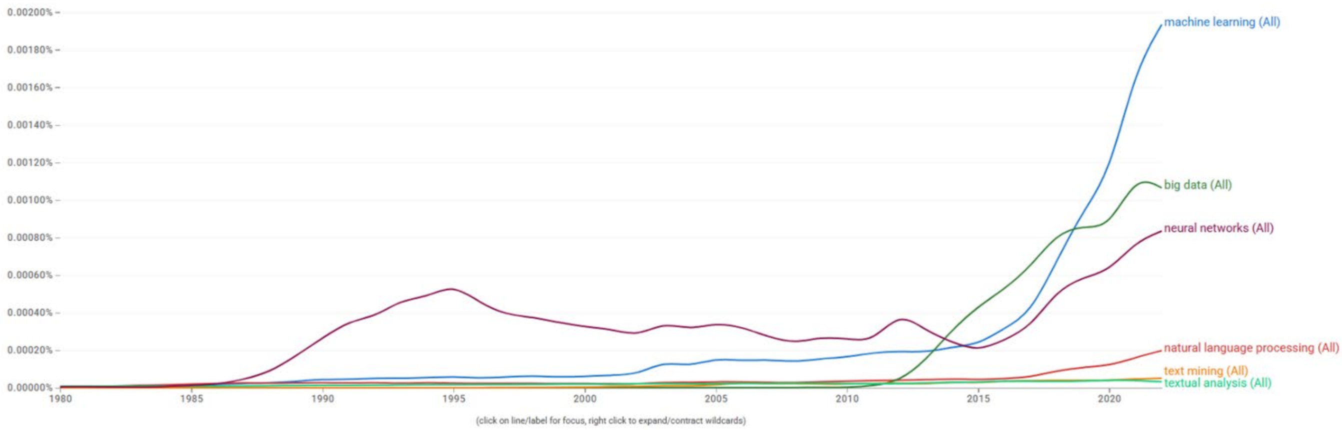


Fig. 1. Trend in the frequency of the terms connected to text mining over the period 1980–2022. *Source*: https://books.google.com/ngrams/.

Neural networks emerged from research conducted in the field of artificial intelligence, especially works concerning the construction of models of basic structures occurring in the brain. Neural networks are particularly effective for solving problems where there are complex, non-linear relationships between input data and output predictions. In the context of text mining, they are powerful tools for data classification. Neural networks, mainly deep learning architectures, have revolutionized text processing by achieving significant breakthroughs in understanding language semantics and context.

Machine learning (ML) emerged in the mid-20th century as a method to enable computers to learn patterns from data (Hotho et al., 2005). Early algorithms such as decision trees, Naïve Bayes, and support vector machines (SVMs) were instrumental in text classification and sentiment analysis. Machine learning has enabled automated systems to identify patterns and extract meaningful insights from unstructured textual data.

NLP evolved from rule-based methods to statistical models, and eventually to deep learning-based approaches. The introduction of word embeddings in the early 2010s enabled better representation of textual data, significantly improving text mining capabilities.

Across NLP, ML, and neural networks, the basic principle remains the same: the model learns from historical data to predict or classify new, unseen data. The quality and size of the training data are critical, as the model's ability to generalize to new examples depends on how diverse and representative the training data are. In other words, the training data sets determine how machine learning systems “draw conclusions”. Therefore, apart from the undoubted advantages and huge development potential, the use of text mining techniques poses certain risks. Crawford put it aptly when she wrote that if all the apples in a data set were red and none were green, the machine learning system would not recognize that a green apple was an apple at all (Crawford, 2021).

Despite the power of text mining as a technological tool, human involvement remains critical. As machine learning systems are only as good as the data they are trained on, there is always a risk of biases in the data influencing the outcomes. Therefore, while text mining has revolutionized data analysis and provided powerful tools for uncovering patterns, the human factor is indispensable for critical analysis, interpretation, and addressing potential biases in the data. Both fields complement each other, with textual analysis providing depth and context, and text mining offering scalability and the ability to analyze vast amounts of data.

1.2. Text mining procedure and techniques

A subject of text mining can be any symbol stored in digital form. Therefore, images and video files are also defined as objects of text mining (Grimmer & Stewart, 2013). However, this research only focuses on text as data expressed in characters.

The main tasks in which text mining techniques can be used are: information extraction, text classification (e.g. sentiment analysis), discovering hidden topics in large corpora (topic modeling, e.g. LDA), clustering and document organization, etc. (Aggarwal & Zhai, 2012). A typical text mining study comprises the following stages (Antons et al., 2020):

1) data gathering,
2) data preprocessing,
3) content analysis and
4) integration of the text mining findings and results into the study.

Gathering raw text data in a text mining context means collecting data from various sources such as websites, databases, reports, and social media. Data preprocessing transforms raw text into a structured format by removing noise and inconsistencies. It employs such techniques as tokenization, stemming, lemmatization, etc. Text preprocessing aims to turn unstructured text data into structured data in the form of numbers so that mathematical and statistical algorithms can be applied (Elder et al., 2012). The content analysis applies different algorithms such as clustering, classification, sentiment analysis, and topic modeling algorithms. Once the data has been operationalized as variables, it can be integrated into econometric modeling.

Various academic works took up the challenge to provide an overview of recent research trends in text mining. The latest of them suggest that research on big data, social media analysis, and emotion analysis (sentiment analysis) is emerging as the newest research trend (Jung & Lee, 2020). The systematic literature review presented in the following part of the paper is certainly not comprehensive. However, it can be regarded as an attempt to shed light on these trends in the area of real estate by systematizing the use of text mining in research papers.

2. Material and methods

In this research, a systematic literature review (SLR) methodology was employed. This method consists of three stages: planning, implementation, and reporting (Tranfield et al., 2003).

In the planning phase, the research objectives are defined, and the criteria for selecting relevant studies are established. In general, the research questions that can be addressed by SLR oscillate around three main topics, which are (Booth et al., 2016)effectiveness questions (e.g. What effect does intervention X, compared with intervention Y?), methodology questions (e.g. considering research method used in previous studies), and conceptual questions (e.g. a theory that can be used to describe the examined phenomenon). The study takes a methodological approach. The main purpose of the study is to identify the main trends, key topics, and research gaps in the area of real estate, considering text mining techniques. For the purpose of the study, a set of inclusion criteria was established:

- a list of 44 top real estate journals published by the American Real Estate Society was chosen as a referring database (Table A.1 in the Appendix);
- keywords set that was used for advanced search considering the title, abstract and keywords of papers; the keywords list consists of the following elements: “natural language processing”, “NLP”, “text mining”, “textual data mining”, “computer assisted text analysis”, “topic model*”, “sentiment analysis”, “bag-of-words”, “text preprocessing”, “Latent Dirichlet Allocation”, “textual analysis”.

In the execution stage, the actual data collection process is carried out. It was assumed that the papers would be retrieved from the Web of Science and Scopus databases. The search process resulted in the identification of 234 papers that met the established criteria (as of January 21, 2025). The selected studies were then screened to remove irrelevant entries. Since the selected set of articles comes from chosen English-language journals with no time restrictions applied, the primary criterion for excluding a paper from further analysis was its irrelevance to the research topic. Abstract screening resulted in a final set of 74 papers (the string for Web of Science used to obtain the final dataset, which comprises 74 papers, is enclosed in Table A.2 in the Appendix).

The data were subsequently exported and processed using the following programs: *CitNetExplorer*, *VOSviewer, R (bibliometrix)*, and *Python (litstudy)*. These tools enable the creation of maps that visualize keyword co-occurrence, citations, and author collaborations. The mapped clusters facilitate the identification of networks of connections between publications and their authors, highlighting potential research gaps that may serve as inspiration for future scientific investigations.

The reporting stage involves documenting and presenting the findings. This stage also includes a discussion of the review process and any limitations encountered during the study. Conclusions are drawn based on the findings, and, where applicable, suggestions for future research are provided. Using bibliometric analysis - as a tool for the analysis of the selected set of papers - the study aims to address the posed research questions. The techniques for bibliometric analysis comprise two main categories(Donthu et al., 2021):

1) performance analysis and
2) science mapping (e.g. citation analysis, keyword occurrence, and topic modeling).

Performance analysis is a bibliometric technique used to evaluate the productivity and impact of research entities such as authors, institutions, journals, and countries. It focuses on measuring quantitative aspects of scientific output, including publication counts, citation metrics, and journal impact. Citation analysis is a basic technique for science mapping, assuming that citations reflect intellectual linkages between publications when one publication cites the other. Based on this assumption, they can be combined into one cluster (Tomczyk et al., 2024). Mapping keyword co-occurrence is another technique that can help uncover the central themes within a vast collection of publications. It may also highlight emerging trends and gaps in the field. Keywords serve as a representation of a document's content and facilitate the connection between different publications. When multiple publications share one or more keywords, it typically suggests a relationship between them, indicating that they explore similar research topics. Topic modeling, particularly techniques like Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF), is a computational method for uncovering thematic structures in large text corpora. That is why topic modeling aligns closely with science mapping.

The potential of bibliometric analysis is revealed when performance analysis and science mapping can be combined (Solorzano & Plevris, 2022). That is why both techniques were used in the research. Together with the following SLR procedure, the study ensures a systematic and transparent approach, which is essential for conducting a rigorous bibliometric analysis.

3. Results

3.1. Performance analysis

This figure (Figure 2) illustrates the trend in the number of academic publications related to text mining in real estate over a specified period.

The data show a noticeable increase in research activity in recent years, particularly after 2015, suggesting a growing interest in applying text mining techniques to real estate analysis. Peaks in publication output may correspond to technological advancements in machine learning and natural language processing (NLP) or increased availability of large-scale real estate data.

The prevailing number of publications originated from the journals "*Cities*" and "*Urban Studies*". The accompanying diagram ranks the ten most influential journals according to their publication volume (Figure 3).

The distribution of citations across these journals suggests that a few key outlets dominate the field, potentially shaping the discourse and directing future research. Journals specializing in urban economics appear prominently, highlighting the focus on those issues. The emergence of newer journals in recent years indicates that the research domain is broadening beyond its traditional core themes and topics.

The following figure lists the ten most cited papers in the analyzed dataset (Figure 4).

eISSN: 2300-5289

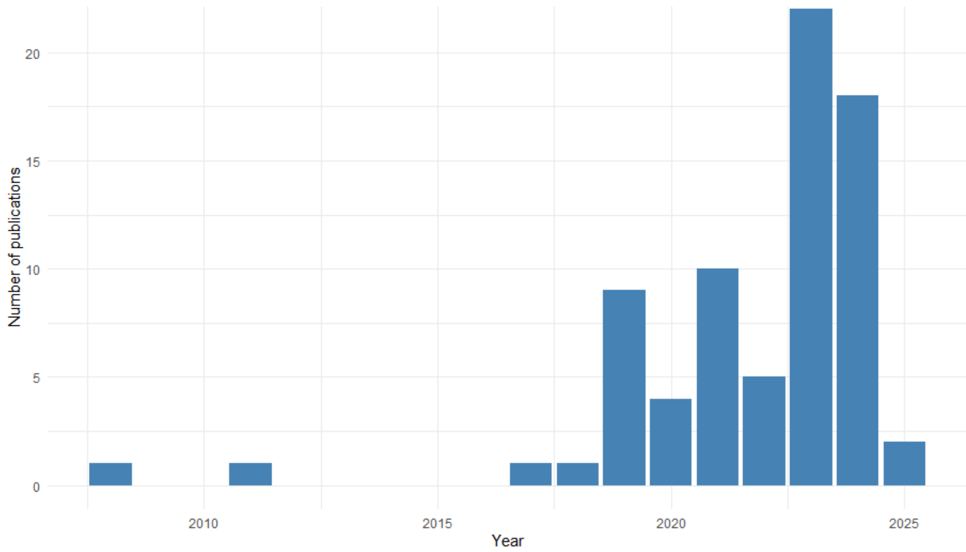


Fig. 2. Number of publications by year. *Source*: own study.

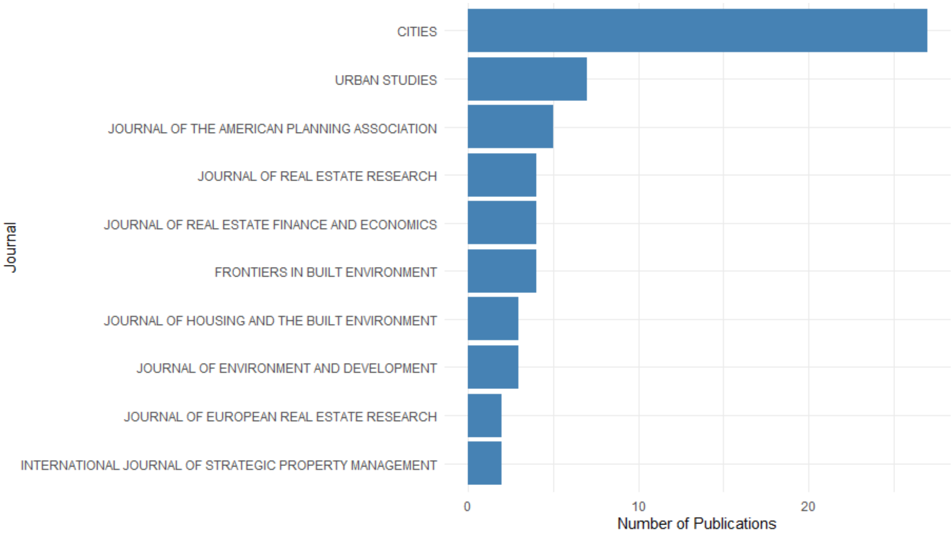


Fig. 3. Ten most influential journals. *Source*: own study.

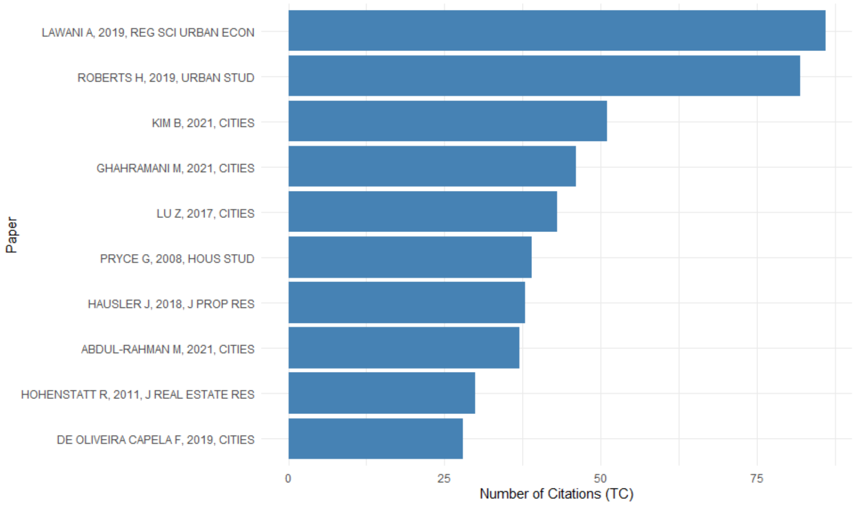


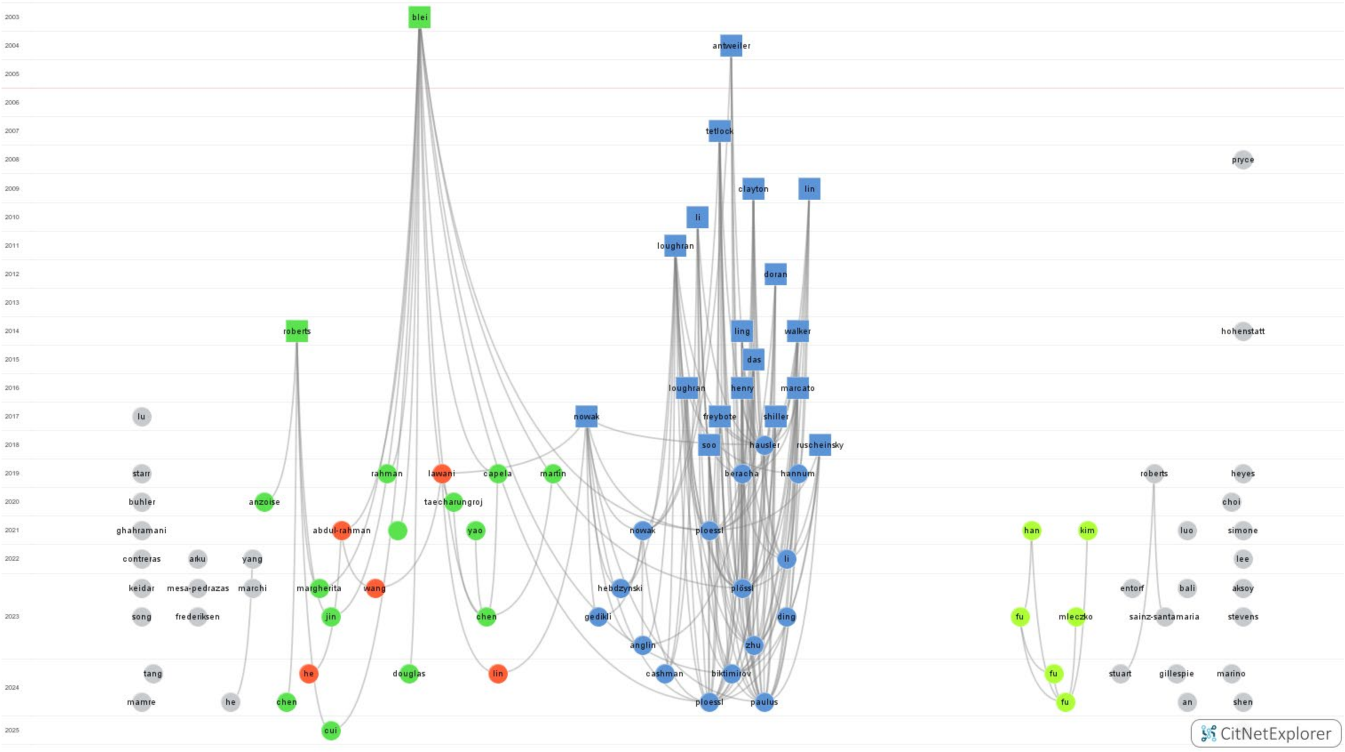
Fig. 4. Ten most influential papers. *Source*: own study.

eISSN: 2300-5289 | Received 2025-02-13 | Revised 2025-06-17 | Accepted 2025-07-04

For example, Lawani et al. (2019) studied the relationship between guest reviews and the price set by hosts on the Airbnb platform in Boston. They incorporated sentiment analysis into classical hedonic price modeling. Roberts et al. (2019) used data obtained from Twitter to investigate citizens' emotions while using urban green spaces. Also Kim et al. (2021) focused on urban issues to build a foundation for automatic monitoring of citizen complaints. These studies share a common research thread. They explore the role of text mining in real estate by focusing on citizen and consumer opinions. These works exemplify the effective utilization of Big Data, providing a new perspective on the analysis of researched populations, which may include citizens or consumers of particular services or products.

3.2. Citations network

For citation-based information retrieval, *CitNetExplorer* was utilized (see Figure 5). This tool proves to be particularly helpful when a comprehensive overview of the literature on a certain research topic is required. It also helps to trace the origins of a scientific idea or research concept (van Eck & Waltman, 2014).



**names in squares come from outside the analyzed set (20 publications), and names in circles come from the analyzed set (74 articles)*

Fig. 5. The citation network. *Source*: own study.

This network visualizes maps of citation relationships between publications. Each circle represents a paper, with citation links flowing upwards over time. The vertical axis corresponds to time, with a publication's position along this axis determined by its year of publication. The horizontal positioning of publications reflects their proximity within the citation network, meaning that closely related studies are positioned near each other (van Eck & Waltman, 2014).

Four clusters of research were identified, indicating distinct thematic areas within the literature. The presence of distinct clusters suggests that text mining applications in real estate can be categorized into multiple thematic areas. Papers that are centrally positioned within the blue cluster exhibit a high degree of connectivity through citation references. Seminal studies within this cluster include the works of Antweiler and Frank (2004) as well as Tetlock (2007). Antweiler and Frank (2004) have studied the effect of online messages on Wall Street stock market volatility. Also, Tetlock (2007) measured the interactions between the media and the stock market. He found that negative media sentiment precedes lower returns the next day. Additionally, he found that unusually high or low pessimism predicts high market trading volume. Both works relate to the study of tone and emotions expressed in publications on stock market investments. They form the foundation on which a line of publications using sentiment analysis to research the real estate market was built. In this vein, Ploessl and Just

eISSN: 2300-5289

(2024) investigated news coverage and its sentiment in the German residential real estate markets. Their analysis revealed that news-based indicators significantly influence residential real estate prices, with both reporting intensity and optimistic sentiment correlating with price increases.

The green cluster has its origins in works by (Blei et al., 2003) and (Roberts et al., 2014). Blei et al. (2003) pioneered *Latent Dirichlet Allocation* (LDA), a generative probabilistic model for collections of text data. Roberts et al. (2014) advanced structural topic modeling (STM) on recent developments in machine learning. LDA revolutionized topic modeling by providing a systematic way to identify latent themes in large text datasets. STM further refined this by integrating contextual information, making topic discovery more flexible and interpretable for diverse research applications, see Roberts et al. (2014).

The citation visualization revealed two primary research themes within the analyzed dataset, which are instrumental in advancing methodologies for text mining applications in real estate research. The first thread, represented by the blue cluster, focuses on the use of sentiment analysis in real estate markets. The second thread, represented by the green cluster, revolves around topic modeling. The network structure of citations also reveals isolated subfields. That may indicate areas where research is emerging but has not yet been widely integrated with the broader literature.

3.3. Keyword occurrence

The most popular keywords in the examined dataset include the following: *textual analysis* (12 occurrences), *Twitter* (12), *sentiment analysis* (12), *investor sentiment* (10), and *social media* (9). The timeline graph (Figure 6) displays the evolution of frequently used keywords across the analyzed publications.

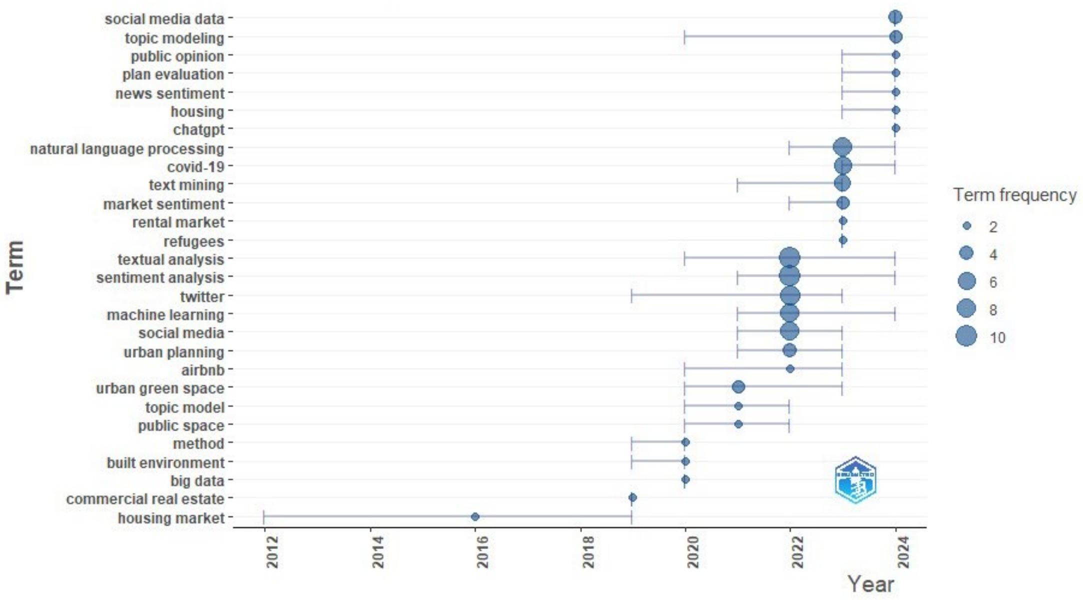


Fig. 6. Timeline of most frequent author keywords. *Source*: own study.

Figure 6 provides insight into how specific terms have gained prominence over time. The initial studies concentrated on the housing market. However, the introduction of new keywords indicates the development of innovative research directions.

The keyword clustering visualization (Figure 7) offers insights from another angle by illustrating the grouping of terms based on their co-occurrence patterns. Typically, when co-occurrence data is analyzed, an initial transformation is conducted on the dataset. The objective of this transformation is to extract similarities from the data or, more precisely, to standardize the dataset (Van Eck & Waltman, 2009). In this instance, the distance between keywords was standardized through association strength, which is the default configuration in the VOSviewer software.

Analyzing these clusters allows for the identification of prevailing themes and emerging areas of interest within real estate text-mining research. Nodes with strong internal connections suggest well-established research themes, such as *sentiment analysis*, *Twitter* and *social media*. Loosely connected keywords may indicate emerging areas with potential for further exploration, e.g. *crisis* and *immigration, race and housing*, or *ChatGpt* and *plan evaluation*.

The analysis reveals four thematic clusters; blue and green on the left side of the chart are more related to real estate markets. The blue cluster represents investors' perspective, and the green cluster is focused on socioeconomic aspects of housing. On the right side of the map, there are located two other clusters, which combine a wide range of urban topics. The yellow cluster is more tightly connected to urban management, facing challenges such as climate change. Under the red one cluster, there are grouped such topics as local government, urban planning, policy and public space.

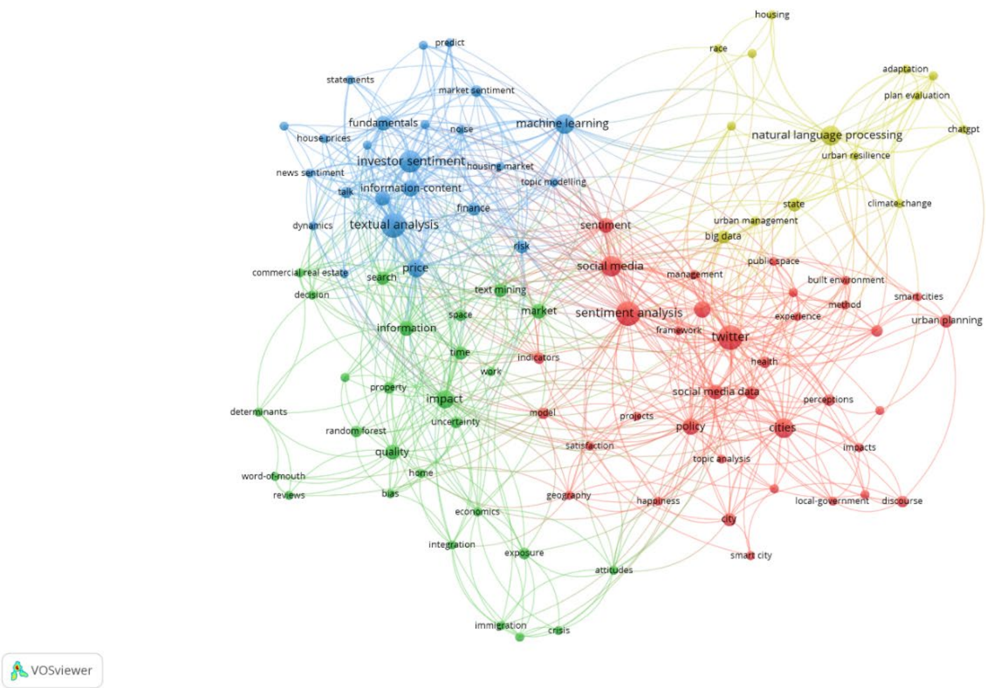


Fig. 7. Clusters of all keywords (no of keywords: 101). *Source*: own study.

3.4. Topic modeling

To train a Non-Negative Matrix Factorization (NMF) topic model and visualize the topics as word clouds, the Python package litstudy was used (Heldens et al., 2022). Topic modeling techniques systematically identify a collection of themes, with each theme characterized by a concise group of terms. Then, specific weights corresponding to these themes are allocated to each document. Considering the specificity measures of topic modeling, a model with fewer, more specific topics is generally preferred. This leads to better-defined and more coherent topics, while a higher number of less specific topics may result in greater overlap and ambiguity (see Lula, 2018). Based on this assumption, variants with four, five, and six topics were investigated. Keeping in mind the need for the best interpretability of the obtained results, five topics were extracted based on the words used in document abstracts (Figure 8).

The word clouds provide an intuitive way to understand dominant themes and keywords in text mining research related to real estate. A brief overview of each of the five topics is essential:

- topic 1 (housing and real estate market, 2): the share of the topic is 26%, this topic focuses on housing, real estate, and market trends; keywords such as housing market, estate, and prices suggest discussions about property values, market dynamics,
- topic 2 (urban planning and social media): the share of the topic is 27%, this topic relates to city planning, policy-making, and the influence of social media on urban development; the presence of planning, cities, big data, and social media indicates an interest in how planning data stored in digital forms and data from social media can help to understand urban environments,
- topic 3 (smart cities and public spaces): the share of the topic is 22%, this topic is centered around urban policies, sustainability, and smart city initiatives; words like public, urban, spaces, cities,

eISSN: 2300-5289

and green suggest discussions on how technology can be used to zoom into urban life,
- topic 4 (online reviews and consumer feedback): the share of the topic is 10%, this topic is about consumer behavior, online platforms, and product or service evaluations; keywords like quality, price, listings, and reviews suggest an interest in how users perceive and rate different products or services on the real estate market,
- topic 5 (sustainable development): the share of the topic is 15%, this topic focuses on sustainability, adaptation, and urban resilience in the face of environmental and economic challenges; words like sustainable, development, adaptation, and communication indicate discussions on long-term planning for sustainable cities.

The results of topic modeling can be enhanced by incorporating the data shown in the table below, which highlights the three most representative papers for each of the five topics.

Finally, the diagram below visualizes the distribution of five topics across various academic journals (Figure 9).

Fig. 8. Visualization of topics using word clouds. *Source*: own study .

Table 1

The most representative works for 5 topics

Authors and year	Paper Title	Journal Title	T1	T2	T3	T4	T5
		Topic 1 (housing and real estate market)					
Pryce & Oates (2008)	Rhetoric in the language of real estate marketing	Housing Studies	**1.00**	0.00	0.00	0.00	0.00
Hohenstatt & Kaesbauer (2014)	GECO's Weather Forecast for the UK Housing Market: To What Extent Can We Rely on Google Econometrics?	Journal of Real Estate Research	**0.99**	0.00	0.00	0.00	0.01
Beracha et al. (2019)	On the Relationship between Market Sentiment and Commercial Real Estate Performance-A Textual Analysis Examination	Journal of Real Estate Research	**0.96**	0.00	0.00	0.00	0.04
		Topic 2 (urban planning and social media)					
Fu et al. (2023)	Using Natural Language Processing to Read Plans A Study of 78 Resilience Plans From the 100 Resilient Cities Network	J. of the American Plan. Assoc.	0.00	**0.95**	0.00	0.00	0.05
Mleczko & Desmond (2023)	Using natural language processing to construct a National Zoning and Land Use Database	Urban Studies	0.14	**0.86**	0.00	0.00	0.00
Tang et al. (2024)	Land revenue and government myopia: Evidence from Chinese cities	Cities	0.07	**0.86**	0.07	0.00	0.00
		Topic 3 (smart cities and public spaces)					
Mesa-Pedrazas et al. (2023)	The new town square: Twitter discourses about balconies during the 2020 lockdown in Spain	Cities	0.03	0.00	**0.97**	0.00	0.00
Yang et al. (2022)	Quantifying the relationship between public sentiment and urban environment in Barcelona	Cities	0.00	0.00	**0.94**	0.00	0.05
Arku et al. (2022)	Highlighting smart city mirages in public perceptions: A Twitter sentiment analysis of four African smart city projects	Cities	0.00	0.05	**0.87**	0.00	0.09

Topic 4 (online reviews and consumer feedback)							
Lawani et al. (2019)	Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston	Regional Science and Urban Ec.	0.03	0.00	0.00	**0.97**	0.00
Lin & Yang (n.d.)	The price of short-term housing: A study of Airbnb on 26 regions in the United States	Journal of Housing Economics	0.16	0.02	0.00	**0.83**	0.00
Hebdzyński (2023)	Quality information gaps in housing listings: Do words mean the same as pictures?	Journal of Housing and the Built Env.	0.21	0.05	0.00	**0.73**	0.00
Topic 5 (sustainable development)							
Marchi et al. (2023)	Sustainability communication of tourism cities: A text mining approach	Cities	0.00	0.00	0.00	0.00	**1.00**
Frederiksen & Banks (2023)	Can Mining Help Deliver the SDGs: Discourses, Risks and Prospects	Journal of Environment & Dev.	0.00	0.00	0.00	0.00	**1.00**
Jin et al. (2023)	Empirical evidence of urban climate adaptation alignment with sustainable development: Application of LDA	Cities	0.00	0.09	0.12	0.00	**0.79**

Source: *own* study.

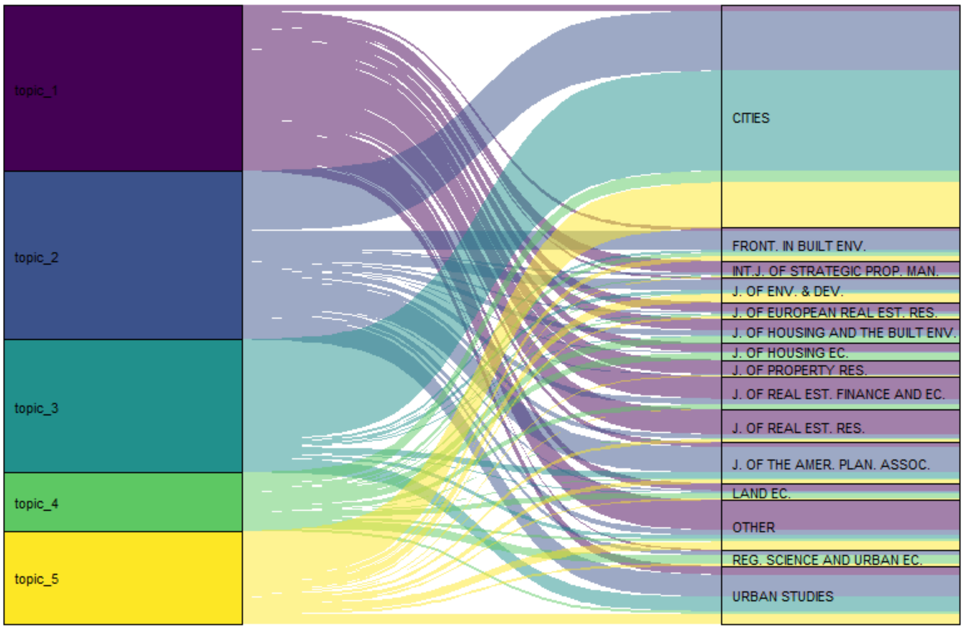


Fig. 9. Flow of topics by source**.** *Source*: own study.

The left side represents the topics, while the right side lists the journals in which related research is published. The width of the flows indicates the strength of the connection between a topic and a journal The journals "Cities" and "Urban Studies" are distinguished as leading publication sources, covering a diverse array of subjects, with particular emphasis on topics 2, 3, and 5. The distribution of topics among the journals is generally balanced; however, certain publications, like the "Journal of Real Estate Finance and Economics" and the "Journal of Real Estate Research," exhibit a more specialized focus on the housing and real estate markets.

In summary, performance analysis and science mapping collectively provide a comprehensive overview of the research landscape, trends, and key contributions in the application of text mining techniques in real estate market analysis.

4. Discussion

The research aim was to identify significant trends, essential topics, and research gaps regarding the application of text mining techniques within the real estate market.

Conducted systemic literature review showed that the application of text mining in real estate analytics has advanced significantly during the last decade, driven by developments in machine learning and natural language processing. Nearly two-thirds of the scholarly articles discuss topics related to urban issues, while the remaining papers tackle the problem of real estate markets. The analysis of citations indicated that the methodologies employed in both research streams predominantly rely on sentiment analysis and topic modeling.

One of the most promising areas of research is the use of sentiment analysis for a more immediate and dynamic analysis of price fluctuations and market trends (Winson-Geideman, 2018). Unlike in the past, when researchers had to rely on quarterly reports or other delayed sources to assess market sentiment, the era of Big Data offers a new approach. Nowadays, it provides real-time insights. Information is now generated continuously and is relatively unrestricted. As citation analysis revealed sentiment analysis was incorporated into real estate research from studies on the financial markers. Comparing the property market to the equity market in terms of liquidity can be challenging; however, some research areas offer a valuable outlet for sentiment analysis, e.g. (REITs). Through the analysis of online reviews, property descriptions, and insights from investors, researchers can also identify key patterns that influence market dynamics across other segments of the real estate sector.

Furthermore, topic modeling enhances the classification of real estate-related content, facilitating more structured and insightful data analysis. Topic modeling, particularly through the use of Structural Topic Models (STM), also has the potential to explore the spatial-temporal trajectories of topics, see Cui et al. (2025).

Analyzing keyword clusters also helps identify additional areas worth exploring. Keywords located at the edges of the co-occurrence network may indicate emerging areas with potential for further investigation. Based on this analysis, two emerging trends were identified: (1) discrimination in the housing market and (2) the use of large language models to evaluate planning documents and trace policies in urban planning.

Paraphrasing Goodwin (2021), it is difficult to predict the direction of real estate research in the coming years. However, further technological advancements are undoubtedly expected. Despite these advancements, several challenges persist.

First, text mining techniques have their drawbacks. For example, overfitting in machine learning models can result in models that perform well on training data but poorly on real-world data, limiting their predictive power. Careful validation and awareness of data limitations are crucial for responsible implementation (Renigier-Biłozor et al., 2022).

Secondly, although it is reasonable to assume that facts, opinions, and other information should be communicated as clearly as possible, it is essential to recognize that language can also be used to obscure meaning. Winson-Geideman (2018) indicates that a comprehensive understanding of content requires awareness not only of what is being conveyed but also of the intent behind it. What is more, the concept of Big Data is closely linked to influence and authority, raising concerns regarding data privacy and algorithmic bias. The increasing reliance on large-scale data processing introduces ethical dilemmas, including the potential misuse of personal information, a lack of informed consent, and the reinforcement of societal inequalities through biased models. To uphold ethical standards and mitigate these risks, it is essential to ensure transparency, accountability, and fairness in data-driven research. In his book Science Fiction, Stuart Ritchie (2021) advocates for developing tools and algorithms that can detect and correct biases in data, ensuring that research outcomes are more reliable and less prone to manipulation. Such advancements could play a crucial role in enhancing the integrity of scientific practices by addressing the ethical concerns of privacy, consent, and fairness.

One of the key strengths of this study is its comprehensive systematic literature review, which provides a structured and in-depth analysis of text mining applications in real estate research. The use of bibliometric techniques, such as citation and keyword co-occurrence analysis, enhances the study's credibility by identifying prevailing research trends and gaps. Additionally, by integrating topic modeling, the study highlights significant methodological advancements and their practical implications for market analysis.

However, some limitations should be acknowledged. First, the study primarily relies on data from a selected set of academic journals, which may exclude relevant papers published outside the examined data set. Second, while the study emphasizes sentiment analysis and topic modeling, other emerging techniques, such as transformer-based NLP models (e.g., BERT or GPT), could further enhance predictive accuracy and contextual understanding. Future research should explore the integration of these advanced AI models to refine real estate text mining methodologies. Additionally, addressing data standardization issues and ethical concerns, particularly regarding bias and transparency in automated decision-making, remains an essential area for further investigation.

5. Conclusions

This study demonstrates that text-based data mining is a valuable tool for extracting insights from vast real estate datasets. The integration of NLP and machine learning techniques has shown promise in improving market predictions, understanding consumer sentiment, and refining property valuation models. The analysis reveals that sentiment analysis, named entity recognition, and topic modeling are among the most effective techniques for structuring unstructured real estate data.

However, the field still faces several challenges. Balancing the benefits of Big Data with ethical considerations and regulatory oversight remains a critical challenge in maintaining a transparent research process. Data standardization issues, ethical concerns, and computational limitations must be addressed to maximize the potential of text-based approaches. The study also highlights the need for improved methodologies to integrate textual and numerical data, ensuring more comprehensive analyses.

The findings of the enclosed paper have practical implications for real estate practitioners, including property investors and urban planners. First, the integration of text mining techniques, such as sentiment analysis and topic modeling, can provide investors with real-time insights into market trends and consumer preferences, enabling more informed decision-making. For instance, analyzing online reviews or social media sentiment could help investors identify emerging hotspots or shifts in demand. Urban planners can leverage these tools to monitor public sentiment about urban projects or policies, allowing for more responsive and citizen-centric planning. Additionally, the ability to uncover hidden patterns in large datasets can aid in identifying risks or opportunities that might otherwise go unnoticed, such as neighborhoods dynamics. Finally, the systematic mapping of research gaps highlighted in the paper could guide practitioners toward innovative strategies, such as using AI-driven tools for evaluating urban plans or addressing issues like housing affordability and racial equity in urban development. These applications bridge the gap between academic research and actionable strategies in real estate practice.

Looking ahead, advances in AI-driven text mining, coupled with enhanced interpretability methods, will likely expand the practical applications of these techniques in real estate research (Renigier-Biłozor & Janowski, 2024).Researchers should also consider the impact of regulatory frameworks on data accessibility and usage. Ultimately, bridging the gap between academic research and industry applications will be crucial for the continued evolution of text mining in real estate analytics.

Funding Sources

This work was supported by the Krakow University of Economics, Poland [PRW/WPOT/2024/0038].

References

Aggarwal, C. C. ., & Zhai, ChengXiang. (2012). *Mining text data*. Springer.

Allen, M. (2017). *Textual analysis*. The SAGE Encyclopedia of Communication Research Methods, 1754–1756. https://doi.org/10.4135/9781483381411

Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. *R & D Management, 50*(3), 329–351. https://doi.org/10.1111/radm.12408

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance, 59*(3), 1259–1294. https://doi.org/10.1111/j.1540-6261.2004.00662.x

Arku, R. N., Buttazzoni, A., Agyapon-Ntra, K., & Bandauko, E. (2022). Highlighting smart city mirages in public perceptions: A Twitter sentiment analysis of four African smart city projects. *Cities (London, England), 130*, 103857. Advance online publication.

Beracha, E., Lang, M., & Hausler, J. (2019). On the relationship between market sentiment and commercial real estate performance-A textual analysis examination. *Journal of Real Estate Research, 41*(4), 605–638. https://www.jstor.org/stable/10.2307/26857143 https://doi.org/10.22300/0896-5803.41.4.605

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. *Journal of Machine Learning Research,* (3), 993–1022.

Booth, A., Sutton, A., & Papaioannou, D. (2016). *Systematic Approaches to a Successful Literature Review (Second Editon)*. Sage (Atlanta, Ga.).

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. In the atlas of AI*. Yale University Press., https://doi.org/10.2307/j.ctv1ghv45t https://doi.org/10.12987/9780300252392.

Cui, N., Malleson, N., Houlden, V., Yan, Y., & Comber, A. (2025). Using Twitter to understand spatial-temporal changes in urban green space topics based on structural topic modelling. *Cities (London, England), 157*, 105601. Advance online publication. https://doi.org/10.1016/j.cities.2024.105601

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research, 133*(March), 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Elder, J., Miner, G., & Nisbet, B. (2012). *Practical text mining and statistical analysis for non-structured text data aplications*. Academic Press.

Frederiksen, T., & Banks, G. (2023). Can mining help deliver the SDGs: Discourses, risks and prospects. *Journal of Environment & Development, 32*(1), 83–106. https://doi.org/10.1177/10704965221139759

Fu, X., Li, C., & Zhai, W. (2023). Using natural language processing to read plans: A study of 78 resilience plans from the 100 resilient cities network. *Journal of the American Planning Association, 89*(1), 107–119. https://doi.org/10.1080/01944363.2022.2038659

Goodwin, K. R. (2021). Thirty years of housing research. *Journal of Housing Research, 30*(2), 107–110. https://doi.org/10.1080/10527001.2021.1985907

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Hearst, M. A. (1999). *Untangling text data mining.* Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 3–10. www.aaai.org/ https://doi.org/10.3115/1034678.1034679

Hebdzyński, M. (2023). Quality information gaps in housing listings: Do words mean the same as pictures? *Journal of Housing and the Built Environment, 38*(4), 2399–2425. https://doi.org/10.1007/s10901-023-10043-z

Heldens, S., Sclocco, A., Dreuning, H., van Werkhoven, B., Hijma, P., Maassen, J., van Nieuwpoort, R. V. (2022). litstudy: A Python package for literature reviews. SoftwareX, 20, 101207. https://doi.org/10.1016/j.softx.2022.101207

Hohenstatt, R., & Kaesbauer, M. (2014). GECO's weather forecast for the UK housing market: To what extent can we rely on Google econometrics? *Journal of Real Estate Research, 36*(2), 253–282. https://doi.org/10.1080/10835547.2014.12091387

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *Journal for Language Technology and Computational Linguistics, 20*(1), 19–62. https://doi.org/10.21248/jlcl.20.2005.68

Jin, S., Stokes, G., & Hamilton, C. (2023). Empirical evidence of urban climate adaptation alignment with sustainable development: Application of LDA. *Cities (London, England), 136*, 104254. Advance online publication. https://doi.org/10.1016/j.cities.2023.104254

Jung, H., & Lee, B. G. (2020). Research trends in text mining: Semantic network and main path analysis of selected journals. *Expert Systems with Applications, 162*, 113851. Advance online publication. https://doi.org/10.1016/j.eswa.2020.113851

Kim, B., Yoo, M., Park, K. C., Lee, K. R., & Kim, J. H. (2021). A value of civic voices for smart city: A big data analysis of civic queries posed by Seoul citizens. *Cities (London, England), 108*, 102941. Advance online publication. https://doi.org/10.1016/j.cities.2020.102941

Lawani, A., Reed, M. R., Mark, T., & Zheng, Y. (2019). Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston. *Regional Science and Urban Economics, 75*, 22–34. https://doi.org/10.1016/j.regsciurbeco.2018.11.003

Lin, W., & Yang, F. (2023). *The price of short-term housing: A study of Airbnb on 26 regions in the United States.* Available at SSRN: https://ssrn.com/abstract=4583093 or http://dx.doi.org/10.2139/ssrn.4583093

Lula, P. (2005). *Text mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych. [Text mining as a tool for obtaining information from text documents].* www.statsoft.pl/czytelnia.html67

Lula, P. (2018). *Statystyczne modelowanie zawartości dokumentów tekstowych [Statistical modeling of the content of text documents].* Wydawnictwo Uniwersytetu Ekonomicznego.

Marchi, V., Marasco, A., & Apicerni, V. (2023). Sustainability communication of tourism cities: A text mining approach. *Cities (London, England), 143*, 104590. Advance online publication. https://doi.org/10.1016/j.cities.2023.104590

Mesa-Pedrazas, A., Nogueras-Zondag, R., & Duque-Calvache, R. (2023). The new town square: Twitter discourses about balconies during the 2020 lockdown in Spain. *Cities (London, England), 143*, 104595. Advance online publication. https://doi.org/10.1016/j.cities.2023.104595

Mleczko, M., & Desmond, M. (2023). Using natural language processing to construct a national zoning and land use database. *Urban Studies (Edinburgh, Scotland), 60*(13), 2564–2584. https://doi.org/10.1177/00420980231156352 PMID:39822385

Ploessl, F., & Just, T. (2024). News coverage vs sentiment: Evaluating German residential real estate markets. *International Journal of Housing Markets and Analysis, 17*(2), 395–417. https://doi.org/10.1108/IJHMA-07-2022-0102

Pryce, G., & Oates, S. (2008). Rhetoric in the language of real estate marketing. *Housing Studies, 23*(2), 319–348. https://doi.org/10.1080/02673030701875105

Renigier -Biłozor, M., & Janowski, A. (2024). Human-machine synergy in real estate similarity concept. *Real Estate Management and Valuation, 32*(2), 13-30. DOI: https://doi.org/10.2478/remav-2024-0010

Renigier-Biłozor, M., Janowski, A., Walacik, M., & Chmielewska, A. (2022). Human emotion recognition in the significance assessment of property attributes. *Journal of Housing and the Built Environment, 37*, 23–56. https://doi.org/10.1007/s10901-021-09833-0

Ritchie, S. (2020). *Science fictions: Exposing fraud, bias, negligence and hype in science*. Random House.

Roberts, H., Sadler, J., & Chapman, L. (2019). The value of Twitter data for determining the emotional responses of people to urban green spaces: A case study and critical evaluation. *Urban Studies (Edinburgh, Scotland), 56*(4), 818–835. https://doi.org/10.1177/0042098017748544

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Solorzano, G., & Plevris, V. (2022). Computational intelligence methods in simulation and modeling of structures: A state-of-the-art review using bibliometric maps. *Frontiers in Built Environment, 8*(8), 1049616. Advance online publication. https://doi.org/10.3389/fbuil.2022.1049616

Stephens-Davidowitz, S. (2017). *Everybody lies: What the internet can tell us about who we really are.* Bloomsbury Publishing.

Tang, X., Wang, W. J., & Liu, W. X. (2024). Land revenue and government myopia: Evidence from Chinese cities. *Cities (London, England), 154*, 105393. Advance online publication. https://doi.org/10.1016/j.cities.2024.105393

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Tomczyk, P., Brüggemann, P., & Doligalski, T. (2024). *The automation of science? Possibilities and boundaries of ai applications for conducting systematic literature reviews.* Https://Doi.Org/10.1142/S0218213024500234

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14*, 207–222. https://doi.org/10.1111/1467-8551.00375

van Eck, N. J., & Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology, 60*(8), 1635–1651. https://doi.org/10.1002/asi.21075

van Eck, N. J., & Waltman, L. (2014a). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics, 8*(4), 802–823. https://doi.org/10.1016/j.joi.2014.07.006

van Eck, N. J., & Waltman, L. (2014b). *Systematic retrieval of scientific literature based on citation relations: Introducing the CitNetExplorer tool* (pp. 13–20). BIR@ ECIR. www.citnetexplorer.nl

Winson-Geideman, K. (2018). Sentiments and Semantics. *Source: Journal of Real Estate Literature, 26*(1), 3–12. https://doi.org/10.1080/10835547.2018.12090471

Yang, L., Marmolejo Duarte, C., & Marti Ciriquian, P. (2022). Quantifying the relationship between public sentiment and urban environment in Barcelona. *Cities (London, England), 130*, 103977. Advance online publication. https://doi.org/10.1016/j.cities.2022.103977

Appendices

Table A.1

List of journals selected for the analysis

Journal Title	Real estate finance rank	Real estate & urban economics rank	Built environment rank
Building Research & Information			3
Cities			4
City			2
Critical Housing Analysis		1	1
Environment & Planning A		4	4
Environment & Planning B		3	4
Frontiers in Built Environment			2
Housing and Society		2	2
Housing Policy Debate		3	3
Housing Studies	3	3	
Housing, Theory, and Society		2	
International Journal of Housing Markets and Analysis		2	2
International Journal of Strategic Property Management	2		
International Real Estate Review	2	2	
International Journal of Urban and Regional Research		2	
Journal of African Real Estate Research	1	1	1
Journal of Corporate Real Estate	2		2
Journal of Environment & Development			2
Journal of European Real Estate Research	2	2	2
Journal of Financial Management of Property and Construction	1		
Journal of Housing and the Built Environment		2	3
Journal of Housing Economics	3	3	
Journal of Housing Research	3	3	
Journal of Property Investment and Finance	3		
Journal of Property Research	3		3
Journal of Property, Planning, and Environmental Law			2
Journal of Real Estate Finance and Economics	4	4	
Journal of Real Estate Literature	2	2	2
Journal of Real Estate Portfolio Management	3		
Journal of Real Estate Practice and Education	1	1	1
Journal of Real Estate Research	4	4	
Journal of Sustainable Real Estate	2		2
Journal of the American Planning Association		4	4
Journal of Urban Economics		4	
Land Economics		3	
Pacific Rim Property Journal			1
Pacific Rim Property Research Journal	2	2	
Property Management	2		2
Real Estate Economics	4	4	
Real Estate Management and Valuation	1		1

Real Estate Taxation	1		
Regional Science and Urban Economics		4	
The Appraisal Journal	2		2
Urban Studies		4	4

Source: Own study based on ARES Real Estate Journal List and the list of scientific journals.

Table A.2

The query string to the set of 74 papers in the Web of Science Database

DO=(10.1016/j.cities.2020.102741 10.1080/02673030701875105 10.1016/j.cities.2017.01.004 10.1080/09599916.2018.1551923 10.1016/j.regsciurbeco.2018.11.003 10.1108/JERER-08-2018-0036 10.1016/j.cities.2019.04.009 10.3368/le.95.2.157 10.1177/0042098017748544 10.1016/j.cities.2018.12.014 10.1108/JCRE-11-2018-0043 10.3389/fbuil.2019.00083 10.1177/0042098019896711 10.1177/0042098019873824 10.1177/0042098019828997 10.1108/JERER-12-2020-0059 10.3389/fbuil.2021.690071 10.3389/fbuil.2021.619283 10.1016/j.cities.2021.103395 10.1080/01944363.2020.1831401 10.1007/s11146-019-09740-w 10.1016/j.cities.2020.103070 10.1016/j.cities.2020.102986 10.1016/j.cities.2021.103273 10.1016/j.cities.2020.102941 10.3389/fbuil.2022.839770 10.1016/j.cities.2022.103857 10.1016/j.cities.2022.103977 10.3846/ijspm.2022.16255 10.1007/s10901-021-09885-2 10.1080/09599916.2023.2199764 10.1080/01944363.2022.2038659 10.1007/s11146-022-09900-5 10.1177/00420980231156352 10.1016/j.cities.2023.104590 10.1016/j.cities.2023.104399 10.1016/j.cities.2023.104254 10.1016/j.cities.2023.104595 10.1016/j.cities.2022.104054 10.1111/1540-6229.12399 10.1177/10704965221139759 10.1007/s10901-022-09985-7 10.1016/j.jue.2023.103588 10.1016/j.cities.2023.104475 10.1177/10704965231157318 10.1016/j.regsciurbeco.2023.103913 10.1007/s11146-021-09865-x 10.1016/j.cities.2023.104365 10.1016/j.cities.2022.104094 10.1177/10704965231205020 10.1007/s10901-023-10043-z 10.3846/ijspm.2023.19092 10.1108/IJHMA-07-2022-0102 10.3368/le.100.2.071522-0056R 10.1080/01944363.2024.2309259 10.1016/j.cities.2024.104857 10.1080/08965803.2024.2340915 10.1016/j.jhe.2024.102025 10.1016/j.cities.2024.105383 10.1016/j.cities.2024.104900 10.1016/j.jhe.2024.102005 10.1016/j.cities.2024.104933 10.1080/01944363.2023.2292674 10.1080/08965803.2024.2357869 10.1016/j.cities.2024.105257 10.1016/j.cities.2024.105016 10.1016/j.cities.2024.105393 10.1177/00420980241297088 10.1007/s11146-022-09917-w 10.1080/01944363.2023.2271893 10.1016/j.cities.2024.105601 10.1177/00420980241246913) OR TI=((Weather Forecast for the UK Housing Market: To What Extent Can We Rely on Google Econometrics) OR (On the Relationship between Market Sentiment and Commercial Real Estate Performance-A Textual Analysis Examination))

Source: own study.