# Geo-NLP Insights: Unveiling Residential Real Estate Trends Through Textual Analysis

Marta Bottero, Federico Dell'Anna(✉), Sara Monaci, and Simone Persico

Interuniversity Department of Regional and Urban Studies and Planning (DIST), Politecnico di Torino, Viale Mattioli 39, 10125 Turin, Italy
{marta.bottero,federico.dellanna,sara.monaci, simone.persico}@polito.it

**Abstract.** The physical and design aspects of buildings, the availability of urban infrastructures and services, and the presence of landscape amenities significantly influence consumer preferences for residential living spaces. Capturing these preferences can provide insights on spatial changes in consumer behavior. To address this challenge, this paper introduces a novel methodological framework that combines Natural Language Processing (NLP) with geospatial data analysis, focusing particularly on insights derived from real estate listings. This innovative approach enables an in-depth examination of urban market dynamics by allowing for a comprehensive analysis that includes temporal shifts and spatial differences. By employing advanced analytical techniques, such as lexical analysis of listings and geospatial exploration, the framework is designed to decode complex market narratives and uncover detailed consumer preferences.

The methodological framework of this paper aims to construct a comprehensive model by integrating textual, spatial, and temporal data from listings, thereby offering a complete perspective of the urban real estate market. The integration of these diverse data sources from real estate listings provides critical insights into consumer preferences and market trends, equipping stakeholders with essential information to make informed decisions. Ultimately, this approach not only broadens the understanding of the real estate market but also aids stakeholders in crafting strategic initiatives that align with the sector's evolving demands. This method facilitates informed decision-making and strategic planning in the real estate domain by leveraging the rich data found in property listings.

**Keywords:** Text Mining · Online housing listing · Advertising keywords

## 1 Introduction

Real estate is a complex field of investigation which is essential in the study of urban transformations. As it is well known, the analysis of real estate market plays an important role in the understanding of urban dynamics, highlighting the most important factors affecting housing demand and price mechanisms.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024 Workshops, LNCS 14821, pp. 163–175, 2024.
https://doi.org/10.1007/978-3-031-65308-7_12

According to the literature, there are different building's characteristics which influence market value and they can be grouped according to specific classes [1]: i) Location (or extrinsic) characteristics, which refer to the whole building in which the building unit is located; ii) Position (or intrinsic) characteristics, which consider the specific building unit; iii) Technological characteristics, which describe the typological properties of the building unit; iv) Productive characteristics, which take into account the legal and economic condition of the property.

Traditionally, real estate market analysis has been developed constructing conventional databases that display data belonging to the aforementioned characteristics, like location, surface, floor, number of bathrooms, building age, distance from parks and green areas, price and others [2].

More recently, the analysis of data has significantly improved also in the context of real estate market, thanks to the new technologies and tools belonging to data science. In this domain, a very important role is played by Machine Learning algorithms which are used to derive the estimation of a set of essential indicators [3].

In this paper, we focus on big data analytics on real estate, with specific reference to online real estate marketing; in particular, we consider advertising information to reflect on consumers preferences and how they are exhibited in the actual market trends. Indeed, real estate advertisements can offer numerous details and important information about the properties in a certain urban area. The automatic extraction of keywords from the text of the real estate advertisements can help in better understanding the area under investigation.

Our paper aims at investigating the use of the Natural Language Processing (NLP) in the real estate market analysis. Starting from text data coming from property listing portals, we applied different algorithms to clarify the possible outcomes as well as the advantages of this approach in the study of real estate.

The basic idea is that, differently from conventional databases, the use of NPL on online advertisements can capture more evidently the actual trends going on in the market and provide qualitative understanding that could be very useful for delivering a more complete picture. In this sense, the use of NPL on online listing could be more effective than transaction-based data in highlighting relevant keywords that could be employed for forecasting future trends in the market.

## 2 Background and State of the Art

The application of NLP in real estate offers a nuanced approach to understanding market dynamics through the lens of textual data from listings and reviews. The application of NLP in real estate provides a sophisticated means of deciphering market dynamics through the analysis of textual data found in listings and reviews.

Cadorel et al. (2023) delve into the intricacies of extracting valuable insights from the complex, domain-specific language of French real estate listings [4]. Their exploration into various text representation models reveals that bespoke, custom-tailored approaches significantly outshine generic pre-trained models in capturing the nuanced linguistic features specific to real estate ads.

Further enriching the understanding of urban real estate markets, Blanchi et al. (2022) introduce a novel methodology that marries NLP techniques with Geographic Information Systems (GIS) [5]. Their work on extracting geographic knowledge from real estate ads and identifying named entities and relationships offers a deep dive into the social and spatial structures that underpin urban spaces. This integration not only adds a spatial dimension to the analysis but also paves the way for an exploration of how location impacts market dynamics and consumer preferences, underscoring the significance of spatial analysis in comprehending the complexities of real estate market dynamics.

The study by Su et al. (2021) presents a hierarchical hedonic modeling approach based on semantic and sentimental analysis of online housing advertisements across five Chinese megacities, enriching our understanding of the dynamics between landscape amenities and the real estate market, specifically rental prices [6]. This study advances the discourse by not only affirming the value of landscape amenities in urban planning and real estate valuation but also demonstrating the variability of this value across different urban contexts. It highlights that the perceived value of landscape amenities, as reflected through online housing listings, significantly influences rental pricing strategies, thereby impacting the broader rental market dynamics.

While existing literature excels in textual and spatial analyses, the temporal dynamics of the market - especially in reaction to external events like the COVID-19 pandemic - are vividly captured in the study by Lee and Lee (2023) [7]. Their analysis demonstrates how online listings can reflect temporal shifts in consumer preferences, highlighting a growing demand for properties with enhanced views and more space, alongside a declining emphasis on proximity to transportation and business districts.

A common theme across these studies is the strategic use of NLP to derive insights from textual data, with Su et al. (2021) and Blanchi et al. (2022) providing a pivotal contribution by integrating spatial context through georeferenced data. Not least, the literature reveals the presence of the study by Lee and Lee (2023) that incorporates temporal data to provide insight into the property market over time. Despite the advancements, the literature presents a gap: the lack of an integrated model that simultaneously considers spatial, temporal, and textual data to provide a holistic view of the real estate market. Such a model could offer unprecedented insights into the dynamic interplay between market trends, consumer preferences, and external influences.

Upcoming research should aim to develop integrated models that encompass spatial, temporal, and textual dimensions to fully understand the complexities of the real estate market. Acknowledging the advancements in the application of NLP in real estate and the insights gained from spatial and temporal analyses, the overarching goal of this research is to bridge the identified gap in the literature. We propose to develop an integrated model that synergistically incorporates spatial, temporal, and textual dimensions to offer a comprehensive understanding of the real estate market. This proposed model aims to deliver a robust framework capable of unveiling the intricate relationships between market trends, consumer preferences, and external factors, providing stakeholders with a powerful tool to navigate and strategize in the complex landscape of real estate, particularly in the face of unforeseen events and market shifts.

## 3 Methodology

The proposed methodological framework aims to provide a detailed picture of market trends and consumer behaviors starting from real estate listings considering space and time. It incorporates spatial analysis to understand how geographic factors influence property values and buyer preferences. By temporal analysis it is possible monitoring market trends over time to understand changes in preferences and identify possible cyclical pattern.

The model is designed to be a robust framework that not only reveals the intricate relationships between market trends, consumer preferences and external factors, but also provides stakeholders with a powerful tool to navigate and strategize within the complex industry landscape real estate. This is especially important in the face of unexpected events and market changes, such as economic downturns or global pandemics, which require rapid and informed responses from market participants.

By providing stakeholders - from developers and investors to urban planners and policy makers - a comprehensive tool, this framework aims to improve decision-making processes, enabling more targeted marketing strategies, informed investment decisions and effective policy formulations. Using this model, developers can gain insights into the most desirable features and emerging trends in residential and commercial real estate. This allows developers to design and build properties that align more closely with current market demands, ensuring greater marketability and return on investment. The tool can highlight areas with rising property values and predict future hotspots, thus allowing urban planners to offer detailed insights into how different areas of a city are evolving in terms of demographic changes, housing demand and infrastructure needs. This information is crucial for planning public services, zoning and transportation networks that efficiently support urban growth and sustainability. Policymakers can use the model to understand the impact of regulations and economic policies on the real estate market. This model allows stakeholders to anticipate market changes more accurately and respond quickly. In a competitive market landscape, the ability to quickly adapt to new developments and consumer preferences provides a crucial competitive advantage, enabling stakeholders to take proactive rather than reactive measures.

The proposed methodological framework provides notable advantages over traditional real estate valuation models. While traditional models tend to rely on static comparisons and isolated numeric attributes to estimate property value, this framework integrates spatial and temporal data, starting from textual information [8, 9]. By utilizing advanced analytics such as NLP and network analysis, the framework provides a dynamic, predictive view of the market, enabling stakeholders to identify emerging trends and make proactive decisions. This holistic approach allows for a more comprehensive understanding of how geographic shifts, evolving consumer behaviors, and economic disruptions impact real estate values, offering a nuanced alternative to conventional models.

This section outlines the overall workflow that encompasses the stages from data collection to the development of local insights useful for making informed decisions. In the subsequent sub-sections, we will delve into the complexities of our research methodology, which is succinctly encapsulated in the flowchart in Fig. 1. This methodological diagram traces the path from the initial data collection to the nuanced analysis and visualization of the results. We begin our journey with a systematic collection of real estate

listings, endowed with textual and geospatial data. Subsequently, we have meticulously preprocessed the text to distill the data into a corpus ready for an in-depth NLP analysis. Computational techniques based on statistical ranking are employed to sort and prune the data, ensuring the refinement of the most salient terms. A review by architectural and civil engineering experts helps verify the consistency of the extracted features and ensures the accuracy of the results. The culmination of this process is a sophisticated network analysis, paired with a data visualization that brings the abstract data to life, transforming it into a vibrant illustration of interconnected terms and their geospatial significance.

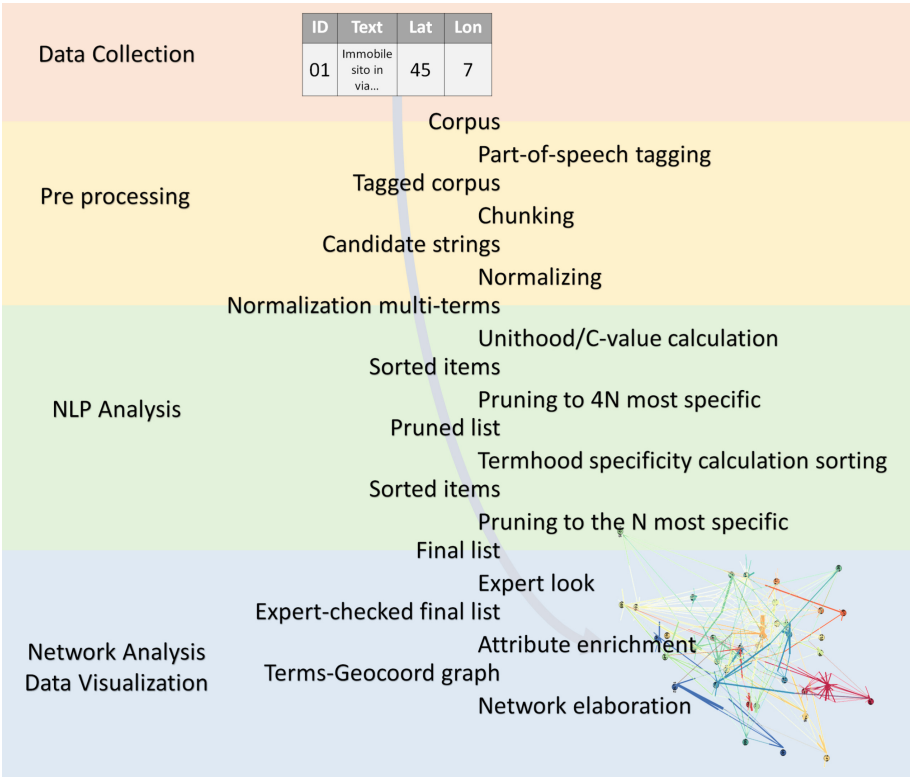


**Fig. 1.** Methodological workflow.

### 3.1 Data Collection

The framework unfolds with a data collection plan centered on acquiring a dataset enriched with detailed textual descriptions for each real estate listing, which is indispensable for the NLP segment of the study. The textual content, detailing property features, amenities, and other qualitative aspects, is essential for an NLP analysis that seeks to

capture the nuances and specific terminologies that resonate with both the inherent and the perceived market value of properties.

To conduct a nuanced analysis, it is vital to integrate these detailed textual descriptions with exact geographic data, whether that be coordinates or address-based information amenable to geolocation methods. This crucial preliminary stage ensures that the resulting dataset comprises exhaustive narrative descriptions, setting the stage for a comprehensive, text-based analytical process that will later be augmented by spatial analysis employing Geographic Information Systems (GIS) technology.

Moreover, a targeted focus on particular types of buildings is advisable, allowing for an exploration into the preferences of demographic segments that possess the financial means for certain property standards. This focused approach aims to uncover the preference trends and standards desired by specific income brackets within the population.

### 3.2 Data Preprocessing

Before the actual analysis both textual description and the geographic coordinates field needs a pre-processing phase. The linguistic preprocessing consists of three basic steps: part of speech tagging, chunking, and stemming.

The Part Of Speech tagging tool categorizes terms based on their grammatical category, such as noun, adjective, verb, or adverb. To process the ads we identified CorTexT Manager [15] which employs Natural Language Toolkit [10], commonly known as NLTK, for English and TreeTagger for other languages [11, 12].

During the next Chunking phase, tags are used to identify noun phrases in the corpus, that is, a sequence of nouns and adjectives. This phase creates the set of possible multi-terms. Eventually, with the normalization phase, minor spelling inconsistencies between multi-terms, including the presence or absence of hyphens, are refined, assuming, for example, that the phrases "windowed bathroom" and "windowed bathroom" belong to the same class.

The stemming phase is dedicated to manage multiple terms that can be grouped together if they have the same stem. For example, singular and plurals are automatically placed into the same class (for example, "windows" and "window" are two possible forms of the stem "window"). Porter's stemmer is used for English [13], whereas Snowball stemmers are used for French, German, and Italian [10].

Geographic coordinates are the other essential aspect to be addressed. Since the goal is to cluster the extracted terms based on geographic location, a simple and intuitive path to follow is to group ads by proximity by defining a "Microzone" attribute. In detail, a microzone, smaller in area compared to a neighborhood, represents a portion of the municipal territory that exhibits homogeneity in terms of its location, urban planning, historical-environmental aspects, socioeconomic characteristics, as well as the provision of urban services and infrastructure. These microzones therefore represent uniform areas that delineate submarkets within the real estate market. In each microzone, the real estate units are uniform in terms of typological characteristics, construction period, and predominant use. Often, real estate listing sites utilize this subdivision to facilitate users' search for properties. The microzone will be assigned a pair of latitude and longitude

values, calculated through the average or median value of ads located in the specific zone, and will be used to map the information retrieved.

### 3.3 NLP Analysis

Cortext function Terms Extraction automatically identifies terms pertaining to a given corpus by using Natural Language Processing tools that allow us to identify not only simple terms but also multi-terms (called n-grams). For large datasets, terms will only be extracted from sample sub-corpuses with a fixed number of documents (selected at random from the original sequential datasets). Regardless of the sampling approach used, identified phrases will be indexed throughout the whole corpus.

Frantzi et al. (2000) proposed the C-value approach to determine the unithood of multi-term stems [14]. Unithood is the degree to which a sequence of words is able to form a stable lexical unit (Wong, 2009) [15]. The frequency of terms can be calculated at the document level (i.e., the number of distinct documents in which they appear) or at the sentence level (which includes repetitions of a term across sentences within a document). The most relevant terms are chosen based on a balance of specificity and frequency. By default, specificity is calculated as a chi2 score (sum of all terms). It is also conceivable to utilize a simpler G2 score (GF-IDF for Global Frequency/Inverse Document Frequency derived from the well-known Term Frequency/Inverse Document Frequency TF-IDF). The G2 score is based on the notion that intriguing terms are likely to be repeated in the same document. One can optionally disable the role of specificity in the final ranking of extracted phrases, such that only the top N most frequent terms are retrieved (frequency). Our option has been to use the chi2 specificity, since the G2 score could "reward" repetitive ads beyond measure.

The ranking phase provide an entire list of possible multi-terms organized into stemmed classes, from which the N most relevant will be chosen. Linguists often make two assumptions when attempting to identify the most significant multi-terms in a corpus: pertinent terms occur more frequently, and longer phrases are more likely to be relevant. To sort the list of candidate phrases, a basic statistical criterion is used. To get the total number of occurrences (frequency), we count all multi-terms in the corpus that correspond to a certain stemmed class. In this phase, if two candidate multi-terms are nested, the larger chain's frequency is increased. For example, if "windowed bathroom" appears in a document, we only increment the multi-stem "window bathroom," not the smaller stem "bathroom". Items are then sorted by unithood [10], and the list is reduced to 4N multi-stems, starting with the greatest C-value [11]. This stage eliminates less frequent multi-stems, but more importantly, it enables second-order analysis of the remaining list.

Finally, Cortext use a similar method to Van Eck et al. (2011) to eliminate irrelevant multi-terms that may still be prevalent [16]. The logic is that irrelevant words should have an unbiased distribution in relation to other terms on the list, implying that neutral terms can appear in any text in the corpus independent of the precise thematic they address. We generate a co-occurrence matrix M for each item in the list. A multi-stem's termhood ($\theta$) is calculated by adding its chi-square values to those of the other classes in the list. We rank the list depending on $\theta$, and only the N most specific multi-stems are conserved.

### 3.4 Data Visualization

The extracted terms are now indexed in the original dataset, which also contains geocoded data. Specifically, each apartment in our dataset contains the longitude and latitude coordinates, which have been previously used to calculate the geographic coordinate of specific microzones. The microzone is an urban area of variable size defined in our original dataset that we use for clustering the textual information retrieved. Using the Network Mapping feature we will be able to map the extracted terms with the microzone, and export the bipartite graph that will be shaped with Gephi [17], a dedicated software for Network Analysis. The bipartite adjective refers to the heterogenous nature of the network that is composed by two different types of nodes: the Microzones (containing spatial information) and the Terms (the descriptive labels extracted with NLP).

In detail, we will use the Geo Layout plugin that will exploit the longitude and latitude coordinates to position each microzone. The position of the microzone node will then be fixed to apply a second clustering algorithm that will group terms around and in between the microzone, without losing spatial information, but actually exploiting the previously positioned nodes and the weight of the relation with each microzone to properly position the extracted Terms. Force Atlas is the force-directed algorithm used for clustering because of its readability and the peculiarity of using edge weights to position each node, which in our case is critical for identifying the most consistent position for terms that straddle two or more microzones. The final step is to resize labels based on Terms weights, with the goal of highlighting the most relevant features.

The term nodes are enriched with features that categorize them with parameters drawn from the literature, such as intrinsic or extrinsic characteristics, green infrastructures, mobility/accessibility, different facilities such as medical, educational, commercial or leisure ones, type of investment, and finally toponyms.

One aspect that requires attention in the data visualization phase are high-degree nodes that may map terms to incongruous positions due to the force-directed dynamics of the algorithm. One solution is to filter low-weight edges so as to favour terms that move close to the microzone with the highest correlation with the specific term. The label adjustment phase, the purpose of which is to increase the readability of the word map, must also be approached carefully to minimize the displacement of terms that must remain associated with the microzone.

### 3.5 Guidelines Definition

In the final phase of the methodological framework, a structured procedure is established to translate analysis results into actionable guidelines. This process involves synthesizing insights gained from identifying impactful keywords, mapping real estate preferences, and observing temporal variations into a coherent set of recommendations. Rather than focusing on future predictions, the framework analyses changes in the past to identify evolving market trends and shifts in consumer behaviour over time.

The utilization of these guidelines by real estate operators, urban planners, and other stakeholders is multifaceted. For real estate agents and marketers, the guidelines can offer data-driven strategies for optimizing property listings to align with buyer preferences, thereby improving engagement and potentially accelerating sales [18]. Urban

planners and policymakers can leverage these insights to inform decisions on zoning, infrastructure development, and the provision of public services, ensuring that urban interventions are closely aligned with the evolving needs and desires of the residential market [19]. Furthermore, these guidelines can serve as a valuable resource for investors and developers by highlighting emerging trends and preferences, aiding in the strategic planning of future projects. Through this approach, the guidelines aim to bridge the gap between market analysis and practical application, fostering informed decision-making that benefits a broad spectrum of stakeholders in the real estate sector.

## 4 Discussion

The application of this framework is anticipated to yield several key outcomes that will enhance the understanding of real estate market dynamics.

The insights derived from this procedure are expected to provide substantial benefits to various stakeholders. Real estate agents, for instance, will gain a deeper understanding of effective listing marketing strategies, enabling them to craft more appealing property descriptions. In a domain where habits and established rhetoric cover the micro-narratives of real estate marketing, the methodology used can lead to a more effective and resonant communication with respect to the specific characteristics of different markets. In fact, the data-driven approach used - able to integrate NLP and SNA for the visualisation of textual, spatial, and temporal data - can be a useful support for the definition of a "real-estate semantic of the city". The semantic could be a powerful tool not only for the elaboration of real-estate listings, but also of a more articulated real-estate marketing strategy.

For policymakers and urban planners, the findings will offer valuable guidance on targeting urban interventions and services to meet the community's housing preferences and needs, thereby enhancing the quality of urban living environments. In the context of the data-driven practice applied in fact, a further step could be taken that specifically concerns the reactions of users to the listings already included in the dataset. Each listing in fact also shows the number of online views, how many times the real-estate agency was reached by message, etc. By integrating the reaction analysis with the textual, spatial, and temporal analysis of the listings described above, it would also be possible to understand better which listings are actually most resonant for a certain audience in a selected geographical area. This further methodological step could provide significant indications for policy makers and urban planners addressing urban and social needs, as well as the real estate market priorities.

One of the primary expected results is the identification of the most impactful keywords, pinpointing property features and services frequently mentioned in listings that significantly attract buyer interest. This analysis will shed light on what aspects of properties resonate most with potential buyers, informing agents and sellers about the elements that should be highlighted in marketing efforts. Additionally, the framework aims to map real estate preferences across different geographic regions. By visualizing the spatial distribution of these preferences, it will become apparent which property characteristics are particularly valued in specific locales. This spatial analysis will reveal regional variations in market demands, offering a nuanced view of how location influences property desirability.

A further outcome includes the investigation of trends and temporal variations within the real estate market. This part of the analysis will track how preferences have evolved over time, identifying any emerging trends that could signify shifts in the residential market landscape.

As other scholars have pointed out [7], disruptive events - e.g. the COVID19 pandemic - have changed housing demands by favouring different dimensions than in the past. It would therefore be interesting to apply the described framework longitudinally by comparing the semantics of advertisements before and after COVID19 and how this has evolved in specific geographical areas. In addition to this and in a predictive way, events of a different nature but impacting on housing needs, could also be taken into account: for example, the Olympic Games and the redevelopment of entire urban areas. Understanding these temporal patterns is crucial for stakeholders aiming to anticipate market movements and adapt their strategies accordingly.

Despite the anticipated contributions of this framework, it is important to acknowledge potential limitations. These may include the availability and quality of data, the complexity of integrating and analysing diverse datasets, and the generalizability of the findings across different markets and contexts. Recognizing these limitations will be essential for accurately interpreting the results and applying the insights in practice.

## 5 Conclusions

### 5.1 Key Findings and Implications

The main outcome of this study is the development of a methodological framework capable of analysing the textual data of property advertisements with lexical, spatial, and temporal analyses. This integration would potentially allow for a detailed examination capable of capturing the complexity of real estate markets and revealing the evolution of consumer preferences and market trends over time and across geographic locations.

Using advanced lexical analysis techniques, this research could reveal important insights into consumer preferences and market trends from the analysis of property listings. It could indicate that specific attributes of properties, such as their location, technological features, and proximity to amenities, could have a significant impact on consumer decisions and property values. Furthermore, the analysis could highlight the critical role of landscape amenities in influencing consumer preferences and their effect on market prices, particularly in urban settings. This point could emphasize the importance of incorporating environmental and urban design considerations into real estate development and marketing strategies.

Furthermore, the study could illuminate how changes over time, including the effects of external events such as the COVID-19 pandemic, may influence consumer preferences and market dynamics. Information gained from the analysis of property listings over time would be essential for industry players who wish to anticipate changes in the market and adapt their strategies accordingly.

The methodology and prospective results of this research could have profound implications for future real estate market analyses. By showing the importance of combining spatial, temporal, and textual data from property advertisements, this study could pave

the way for more in-depth and accurate market analyses. Real estate professionals, policy makers and urban planners could use this information to make informed decisions, adapt marketing strategies and design urban developments that meet changing consumer demands and preferences. Through the examination of property advertisements, stakeholders could gain a clearer understanding of what determines the value of the property market and how to effectively communicate the attractiveness of properties to potential buyers or renters.

### 5.2 Future Possible Implementation of the Proposed Framework

The methodological framework outlined could be validated thanks to his application on specific geographic and/or urban areas. The collection of real-estate listings in the urban area of Turin (Italy) for example, could represent a significant dataset to be analyzed through the NLP approach described in the above sections. The city of Turin presents many interesting features: it has a densely populated historical center with high value properties, as long as multiple neighborhoods close to university campuses located in different areas of the city. Those areas may be also important fields to observe the transformations of the real estate marketing strategies facing a demand driven by students' housing needs. Turin also presents different amenities such as an extensive river area running through the city, multiple parks and gardens, a significant proximity to the Alps which impacts deeply on the city landscape and may also affect the demand of a real estate property with a scenic view. Moreover Turin, as a medium-size urban agglomerate in Northwest Italy, and in consideration of all the other features, could provide an interesting test-site for further applications of the methodology in different urban areas with similar characteristics at national or European level. Those analyzed features, also considered in a comparative perspective involving different urban areas, could highlight the critical role of landscape amenities in influencing consumer preferences and their effect on market prices, particularly in urban settings. That could emphasize the importance of incorporating environmental and urban design considerations into real estate development and marketing strategies. Turin area could finally represent an ideal location to validate the space dimension of the methodological approach proposed in the paper.

In relation to the real estate changes over time, the methodological approach could be validated thanks to the textual analysis of Turin's property listings in two different periods of time: in a pre-COVID19 phase (2018) and in a post-COVID19 time frame (2023). Through a longitudinal and comparative approach based on the pandemic disruptive event, many changes in the real estate marketing could be observed. As hypotheses, we may anticipate for example, an increased offer of properties located close to amenities such as the river area, or parks and gardens, or an enhanced offer of properties provided with balconies and easily accessible courtyards. Moreover, we may also notice a change in the real estate listings referred the University campuses areas: maybe we could observe an overall increase of those properties sales in consideration of the difficulties encountered by the owners during the pandemic, and the prolonged absence of the students attending classes online.

Future directions for this study could involve refining the integration of NLP with spatial statistics by focusing on detailed real estate listings for a clearer understanding of market trends [20]. Enhancing geographic tagging of property descriptions and features

could help identify emerging consumer preferences, spatial patterns in property demand, and neighborhood development trends. More sophisticated machine learning models for predictive analytics could yield accurate forecasts of market hotspots and shifts in property values. This comprehensive approach will empower stakeholders to make data-driven decisions in urban planning, investment strategy, and real estate development.

**Disclosure of Interests.** The authors declare that they have no competing interests.

## References


1. Forte, C., De' Rossi, B.: Principi di economia ed estimo. Etas libri, Milan (1979)
2. d'Amato, M., Coskun, Y. (eds.): Property Valuation and Market Cycle. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-09450-7
3. Valier, A., Micelli, E.: Automated models for value prediction: a critical review of the debate. Valori e Valutazioni. **24**, 151–161 (2020)
4. Cadorel, L., Tettamanzi, A.G.B.: A Comparative Study of Text Representations for French Real-Estate Classified Advertisements Information Extraction. In: ENIGMA-23, CEUR Workshop Proceedings, vol. 3495, pp. 55–63 (2023)
5. Blanchi, A., Fusco, G., Emsellem, K., Cadorel, L.: Studying Urban Space from Textual Data: Toward a Methodological Protocol to Extract Geographic Knowledge from Real Estate Ads. In: Gervasi, O., Murgante, B., Misra, S., Rocha, A.M.A.C., Garau, C. (eds.) Computational Science and Its Applications – ICCSA 2022 Workshops, pp. 520–537. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-10562-3_37
6. Su, S., He, S., Sun, C., Zhang, H., Hu, L., Kang, M.: Do landscape amenities impact private housing rental prices? A hierarchical hedonic modeling approach based on semantic and sentimental analysis of online housing advertisements across five Chinese megacities. Urban For. Urban Greening **58**, 126968 (2021). https://doi.org/10.1016/j.ufug.2020.126968
7. Lee, J., Lee, K.O.: Online listing data and their interaction with market dynamics: evidence from Singapore during COVID-19. J. Big Data **10**, 99 (2023). https://doi.org/10.1186/s40537-023-00786-5
8. Casini, L., Marone, E., Scozzafava, G.: La scuola estimativa italiana, gli International Valuation Standard (IVS) e il Codice delle valutazioni immobiliari: i problemi di natura metodologica e applicativa. Aestimum **83**, 69–81 (2024). https://doi.org/10.36253/aestim-15327
9. Acciani, C., Fucilli, V., Bozzo, F., Petrontino, A., Sardaro, R.: The market comparision approach (MCA) for the valuation of woodlands (2016)
10. Bird, S., Klein, E., Loper., E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly, Beijing (2009)
11. Schmid, H.: Improvements in Part-of-Speech Tagging with an Application to German. Presented at the (1999)
12. Schmidt, H.: Probabilistic part-of-speech tagging using decision trees. Presented at the (1994)
13. Porter, M.F.: An algorithm for suffix stripping. Program **14**, 130–137 (1980). https://doi.org/10.1108/eb046814
14. Frantzi, K., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the C-value/NC-value method. Int. J. Digit. Libr. **3**, 115–130 (2000). https://doi.org/10.1007/s007999000023
15. Wong, W.: Determination of Unithood and Termhood for Term Recognition. In: Handbook of Research on Text and Web Mining Technologies (2011)

16. van Eck, N.J., Waltman, L.: Text mining and visualization using VOSviewer. ArXiv. abs/1109.2, (2011)
17. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. Proc. Int. AAAI Conf. Web Soc. Media **3**, 361–362 (2009). https://doi.org/10.1609/icwsm.v3i1.13937
18. Guo, J.Y., Bhat, C.R.: Operationalizing the concept of neighborhood: application to residential location choice analysis. J. Transp. Geogr. **15**, 31–45 (2007). https://doi.org/10.1016/j.jtrangeo.2005.11.001
19. Waddell, P.: UrbanSim: modeling urban development for land use, transportation, and environmental planning. J. Am. Plann. Assoc. **68**, 297–314 (2002). https://doi.org/10.1080/01944360208976274
20. Miranker, M., Giordano, A.: Text mining and semantic triples: spatial analyses of text in applied humanitarian forensic research. Digital Geogr. Soc. **1**, 100005 (2020). https://doi.org/10.1016/j.diggeo.2020.100005