AESTIMUM
Centro Studi di Estimo e di Economia Territoriale - Ce.S.E.T.

JUST ACCEPTED

Text-augmented house valuation: using embedding vectors and SHAP values

Lee Changro

Department of Real Estate, Kangwon National University, South Korea
Email: *spatialstat@naver.com*

Please cite this article as:
Changro, L. (2026). Text-augmented house valuation: using embedding vectors and SHAP values. **Aestimum**, *Just Accepted*.
DOI: 10.36253/aestim-19824

LEE CHANGRO

Department of Real Estate, Kangwon National University, South Korea.

E-mail: spatialstat@naver.com

Keywords: *Text data, Embedding vectors, Extreme gradient boosting, SHAP values, House valuation*

**Corresponding author*

ORCID*:*
LC: 0000-0002-7727-3168

Data Availability Statement*: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.*

Conflicts of Interest*: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.*

Text-augmented house valuation: using embedding vectors and SHAP values

With the recent advent of large language models, text data can now be readily processed and analyzed across various sectors; traditionally, this was not the case, and text data had proven challenging to analyze. The housing market uses and features a vast amount of textual data, such as property descriptions and neighborhood trend reports, but has yet to fully exploit this potential. This study has two objectives: first, to enhance house valuation accuracy by integrating text information into valuation models; and second, to interpret how text information contributes to improved performance. We convert the environmental descriptions of each house into embedding vectors, which are then incorporated into regression and extreme gradient boosting (XGB) models to estimate house prices. Additionally, we interpret the XGB model's results using Shapley Additive explanations (SHAP) values. The embedding vectors significantly improved the performances of both the regression and XGB models, and the embedding vector values aligned well with the semantic meaning of the environmental descriptions of each house. This study contributes to the literature by deepening our understanding of embedding vectors in the context of house valuations and housing markets.

1. Introduction

The rise of large language models (LLMs) has sparked significant increases in the use and application of text analysis across diverse industries (Linkon et al., 2024). At the core of this transformation are embedding vectors, which serve as numerical representations of raw text. By converting textual data, such as consumer reviews and legal contracts into numerical formats, embedding vectors enable the integration of textual information into quantitative models, like machine learning models. However, the housing market which is replete with textual data such as property descriptions, resident reviews, and market analysis reports, has yet to fully capitalize on the potential of these new tools.

In this study, we aim to estimate house prices in two contrasting regions of South Korea: a highly urbanized city and a typical rural county. Along with traditional variables such as site area and property age, we incorporate textual information about each house, transforming them into embedding vectors to enhance valuation accuracy. We begin by converting environmental descriptions into embedding vectors using three different embedding models, and then select the most relevant ones. These vectors are then integrated into regression and machine learning models to estimate house prices. Despite the interpretability challenges, we interpret the results obtained from the machine learning model and provide insights into how text information can be used in the valuation sector.

While the use of text data for property valuation has been explored in real estate literature (Baur et al., 2023; Zhang et al., 2024), these studies have been unable to explain how text data enhance valuation accuracy; this may partly be because of the black-box nature of machine learning models. This study addresses this gap by interpreting how textual information leads to improved performance. Thus, we contribute to the literature on house valuations and housing markets by deepening our understanding of embedding vectors.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature on embedding models and their applications. Section 3 describes the data and methodology used in this study. Section 4 presents the results and discusses their implications. Finally, Section 5 concludes with a summary of findings and suggestions for future research directions.

2. Literature review

2.1 Embedding models

Because computers operate using numerical data, it is essential to transform textual inputs into numerical representations that can be processed by computers. Embedding is a numerical representation designed to capture the meaning of the original data, and takes the form of a vector. For example, the sentence "The house is in a downtown area" might be represented by an embedding vector such as [0.8858, 0.2954, 0.4751]. Although this example illustrates a 3-dimensional vector, in practice, embedding vectors typically have more than 100 dimensions (Selva Birunda & Kanniga Devi, 2021; Patil et al., 2023).

To generate embedding vectors from text, embedding models such as GloVe, FastText, and BERT are used (Bojanowski et al., 2017). The process involves pre-processing the text, including tokenization, conversion to lowercase, and the removal of punctuation, followed by inputting the processed text into the selected embedding model. The model then maps words or sentences to dense numerical vectors in a high-dimensional space, and thus captures semantic relationships.

Many embedding models are currently available. Early models, such as GloVe and Word2Vec, assigned fixed vectors to words by using co-occurrence statistics or similar statistical methods for training (Mikolov et al., 2013; Pennington et al., 2014). Subsequently, more sophisticated models such as BERT and SBERT have emerged (Devlin et al., 2019; Reimers & Gurevych, 2019), that leverage transformer-based architectures and employ self-attention mechanisms to produce contextual embeddings (Vaswani et al., 2017). More recently, proprietary models such as text-embedding-3 series (Neelakantan et al., 2022) demonstrate substantial improvements in multilingual retrieval and performance compared to earlier embedding models. In addition, open-source efforts such as jina-embeddings-v3 (Sturua et al., 2024) leverage contrastive learning and parameter-efficient fine-tuning techniques to achieve state-of-the-art results. As of May 2025, more than 1.5 million natural language processing (NLP) models, including embedding models, are accessible on the Hugging-Face platform, where users can share models and datasets and witness the continuous development of new embedding models[1].

2.2 Applications of embedding vectors

Embedding vectors are versatile and can be applied to various domains to transform different types of data into numerical representations. They have been extensively employed in NLP, where they are used to train models that predict the sentiment of customer reviews or social media posts (AlSurayyi et al., 2019; Suryadi & Kim, 2018), classify text into predefined categories like spam detection in emails (Eswaraiah & Syed, 2024; Li & Gong, 2021), and facilitate machine translation by converting text from one language to another (Gheini et al., 2021; Mathur et al., 2019).

Beyond NLP, the number of avenues for the application of embedding vectors have continued to increase. They are used to represent products by encapsulating their features and relationships, thereby enabling personalized product recommendations (Baltescu et al., 2022; Xu et al., 2021). Additionally, embedding vectors can represent user behaviors and interactions, and help social platforms enhance the user experience (Zhao et al., 2022). At present, any data type can be converted into embedding vectors. For instance, images are now transformed into embedding vectors to recommend visually similar images or products to users (Girdhar et al., 2023; Ypsilantis et al., 2023).

Despite extensive research on embedding vectors in NLP and other areas, their application to the real estate sector remains relatively unexplored. The few existing studies are as follows: Lee (2022) predicted real estate prices by converting categorical variables such as zoning into embedding vectors. Some studies have used property descriptions to improve the accuracy of housing price estimates (Baur et al., 2023; Zhang et al., 2024). Property similarity search using image data has recently emerged as a promising extension of embedding vectors in automated valuation models. In this approach, deep learning models, such as vision transformers, are used to convert property images (e.g., interior, exterior, and street-view photos) into embedding vectors. These vectors capture visual characteristics, including design quality, layout, and environmental context. As a result, properties can be compared based on visual similarity, complementing traditional comparisons that rely solely on structured attributes (Deng, 2025). However, partly because of the black-box nature of machine learning models, these studies do not explain how textual data could contribute to performance improvements. In this study, we aim to use textual data, specifically environmental descriptions, to enhance housing valuation accuracy and attempt to interpret how such descriptions lead to improved performance.

[1] https://huggingface.co/models.

3. Data and methodology

3.1 Study area and dataset

Seongdong-gu and Gangjin-gun were selected for analysis. Seongdong-gu is one of Seoul's 25 districts and exemplifies a highly urbanized city. Gangjin-gun is one of 22 counties in Jeonnam Province, representing a typical rural and coastal area. These two areas are deemed relevant for analyzing houses in urban and rural settings[2]. Figure 1 illustrates the locations of the two study areas in the Korean Peninsula.

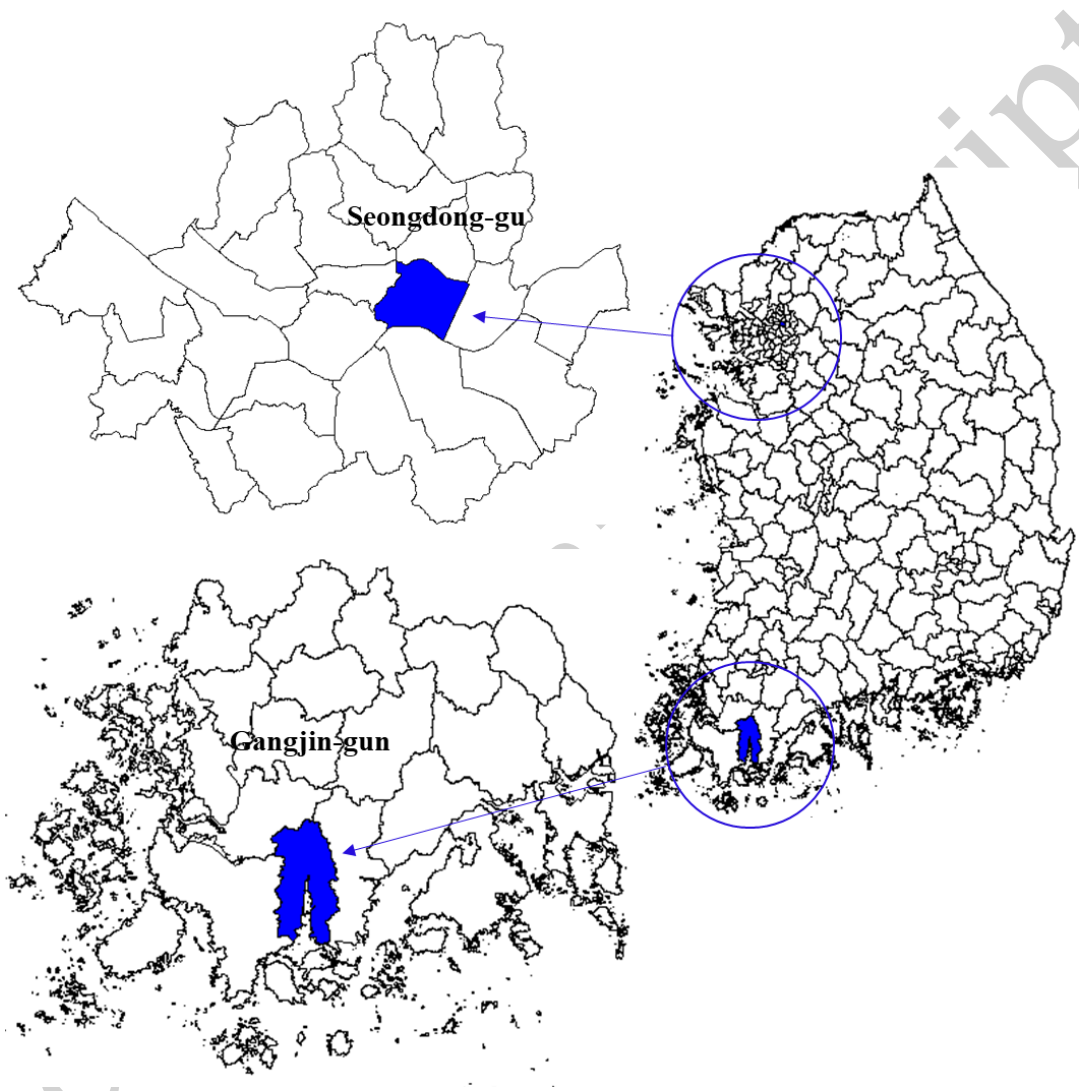

Figure 1. Study area.

The benchmark house data were obtained from the Ministry of Land, Infrastructure and Transport (MOLIT)[3]. MOLIT conducts annual surveys of these benchmark houses, making the results publicly available for stakeholders to use in transactions and investments, and for monitoring housing market trends. Price data were taken from 2024 surveys conducted by licensed property appraisers. The data included price information and various house characteristics, such as floor area and property age. Table 1 presents the descriptive statistics of the 1,670 benchmark houses (763 in Seongdong-gu and 907 in Gangjin-gun).

[2] As of August 2024, Seongdong-gu has a population of 276,000 with a density of 16,400 people per square kilometer, while Gangjin-gun has a population of 32,000 and a density of 65 people per square kilometer (KOSTAT, 2025). Both regions hold equal administrative status within South Korea's jurisdictional hierarchy.

[3] The data is publicly available at https://www.data.go.kr/data/15005267/fileData.do.

Table 1. Descriptive statistics of the data.

Seongdong-gu (n=763)	Min.	Mean	Median	Max.
Price (million KRW)	83.2	534.7	429.0	2,246.0
Site area (m^2)	28.1	129.2	116.0	413.0
Floor area (m^2)	30.0	233.4	186.1	1,028.1
Gangjin-gun (n=907)	Min.	Mean	Median	Max.
Price (million KRW)	2.6	28.9	14.2	362.0
Site area (m^2)	47.0	426.0	383.0	1,637.0
Floor area (m^2)	26.4	107.3	93.9	782.9

As the Table 1 shows, houses in Seongdong-gu are significantly more expensive than those in Gangjin-gun. Owing to high urbanization and the need for compactness, houses in Seongdong-gu typically try to make the most of the available land. This results in smaller lot sizes but larger floor areas compared to Gangjin-gun. For the purpose of evaluation, the data were randomly split; 80% was used for model training and the remaining 20% was reserved for model evaluation.

One piece of information in the benchmark house data includes environmental descriptions of each house. Table 2 presents examples of these descriptions in order of frequency. This study explicitly uses this information to estimate house prices.

Table 2. Environmental descriptions of houses.

No.	Description	Frequency
1	Remote and rural area	455
2	Middle-class residential area	216
3	Residential area in a downtown	205
…	… (Skipped for brevity)	…
35	High-density commercial district	1
36	Farming area around a local road	1
37	Emerging residential area	1

3.2 Creating embedding vectors

Figure 2 illustrates the research flow. First, textual information (environmental descriptions) is transformed into embedding vectors using three embedding models, from which the most relevant vectors are selected. Next, these embedding vectors are compressed into fewer dimensions to enhance the efficiency in subsequent valuation models. Two methods are employed for the house valuations: ordinary least squares (OLS) regression and extreme gradient boosting (XGB) models. Additionally, an OLS regression model without embedding vectors is used as the baseline. Finally, the XGB results are interpreted.

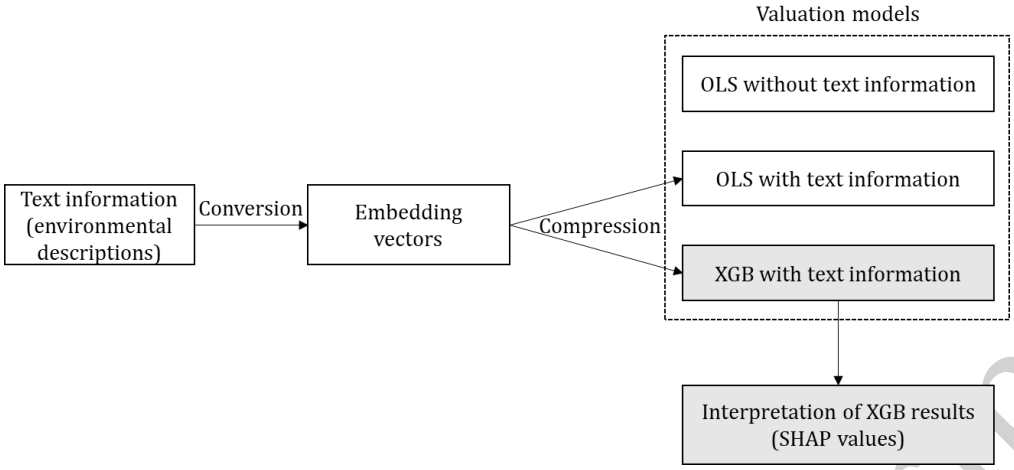

Figure 2. Flow of research.

To convert the environmental descriptions into embedding vectors, three embedding models are employed: a word-embedding model (glove-wiki-gigaword-50) and two sentence-embedding models (paraphrase-MiniLM-L12-v2 and paraphrase-mpnet-base-v2). The glove-wiki-gigaword-50 is pre-trained and represents words as 50-dimensional vectors using data from Wikipedia and the Gigaword corpus. Paraphrase-MiniLM-L12-v2 and paraphrase-mpnet-base-v2, are referred to as sentence embeddings A and B, respectively; both are pretrained models from the *Sentence Transformer* library. Sentence embedding A is a lightweight transformer model that produces 384-dimensional embedding vectors, whereas sentence embedding B, which is an enhanced version of BERT and RoBERTa, generates 768-dimensional embedding vectors[4].

Evaluating and selecting the relevant embedding models is challenging (Torregrossa et al., 2021). In this study, model selection was guided by two complementary criteria: semantic similarity and downstream valuation performance. To illustrate the relative representational quality of the three embedding models, the most common environmental description in the dataset, "Remote and rural area" (455 houses, 27% of the total), was used as a reference, and the five most similar descriptions were identified using cosine distance and presented in Table 3. While relying on a single reference sentence may not fully capture the diversity of environmental descriptions across the dataset, this comparison was intended as an interpretable demonstration rather than the sole basis for selection. Descriptions in word embedding showed less similarity to the reference, whereas those in sentence embeddings A and B were more similar. Ultimately, sentence embedding B was chosen for use in the subsequent valuation models as it outperformed sentence embedding A in a downstream model performance (i.e., XGB).

Table 3. Top five environmental descriptions similar to "Remote and rural area".

No.	Word embedding	Sentence embedding A	Sentence embedding B
1	Pure farming area (0.321)	Rural area near a village (0.074)	Rural area (0.062)
2	Existing industrial area (0.362)	Rural area (0.083)	Rural area on a mountainside (0.149)
3	High-density commercial district (0.415)	Rural area on a mountainside (0.167)	Pure farmland (0.212)
4	Existing residential area (0.428)	Rural area nearby the coast (0.186)	Rural area near the suburbs (0.268)
5	Emerging residential area (0.429)	Rural area near the suburbs (0.213)	Rural area nearby the coast (0.281)

Note: Numerical values indicate the cosine distance (= 1 – cosine similarity), where a value closer to zero indicates greater similarity.

[4] Details on these models are available at https://www.sbert.net/docs/sentence_transformer/pretrained_models.html. Prior to embedding, the environmental descriptions underwent text preprocessing steps. These included lowercasing all text, removing punctuation and special characters, and stripping leading and trailing whitespace. Tokenization was handled internally by each embedding model: the word embedding model tokenizes text at the word level, splitting descriptions into individual word tokens, while the two sentence embedding models employ sub-word tokenization via their respective built-in tokenizers from the *Sentence Transformers* library, which handle out-of-vocabulary terms more robustly.

Sentence embedding B consists of 768-dimensional vectors which can be inefficient if used in their entirety in subsequent valuation models. To enhance efficiency, sentence embedding B is reduced to two dimensions using principal component analysis (PCA). The explained variance ratio for these two dimensions (two components) is 37%, with the first and second components accounting for 22% and 15%, respectively: hereafter referred to as embedding components 1 and 2[5].

3.3 Valuation models

Three valuation models are used to estimate house prices: OLS without text information, OLS with text information, and XGB with text information. The first model acts as the baseline. XGB is a widely used machine learning model that has demonstrated excellent performance in various studies (Alshboul et al., 2022; Chen et al., 2015; Osman et al., 2021). An additional benefit of using XGB is the ease of computing Shapley Additive explanations (SHAP) values, which can be efficiently calculated for tree-based models, including gradient boosting.

The SHAP value begins by examining all possible subsets of explanatory variables, commonly known as coalitions. For each explanatory variable, its marginal contribution to the prediction is calculated when added to each possible coalition of the other explanatory variables. The SHAP value for an explanatory variable is the average of its marginal contributions across all possible coalitions (Lundberg and Lee, 2017). A positive SHAP value indicates that the explanatory variable positively contributes to the prediction or pushes it toward a higher value, whereas a negative SHAP value indicates the opposite effect. The absolute value of SHAP value indicates the strength of the contribution of the explanatory variable. A SHAP value of zero indicates that the explanatory variable has no impact on the prediction[6].

To estimate house prices, Equation (1) was used:

$$Price\ (log-transformed) = house\ type + site\ area + floor\ area + story + road + site\ shape + \\ + building\ frame + property\ age + house\ bearing + embedding\ component\ 1 + \\ + embedding\ component\ 2 \quad (1)$$

From the explanatory variables available in the benchmark house data, 11 were selected after reviewing their coefficients and significance. *House type* indicates the subcategory of a house, such as detached, attached, rowhouse, mixed type, and other variations. *Site area* refers to the area of the house site, whereas *floor area* is the total floor space within the house. *Story* denotes the number of floors; *road* categorizes the size of the road, including main, middle, narrow roads, and other variations; and *site shape* describes whether the house site is regular or irregular. *Building frame* includes types like reinforced concrete (RC), wooden frames, steel frames, brick masonry, and others. *Property age* is the number of years since a house was constructed and *house bearing* indicates the direction that the house faces (south or non-south). The final two variables were the principal components derived from the embedding vectors.

The XGB implementation details are as follows. The maximum depth of the decision tree was set to 3. Increasing this value can result in a more complex model and potentially cause overfitting. A learning rate of 1.0 was used. The number of boosting rounds was set to 20. In addition, the squared error was selected as the loss function[7].

4. Results and discussion

[5] The number of PCA components was selected based on an evaluation of a downstream model performance using XGB. Models incorporating 3, 5, and 7 components were compared against a two-component baseline using mean squared error (MSE). No meaningful improvement was observed beyond two components. Additionally, beyond the first two components, each additional component individually explained less than 10% of the variance, with diminishing marginal returns. Incorporating more components also increase the risk of overfitting, particularly given the relatively small sample size (763 in Seongdong-gu and 907 in Gangjin-gun). Thus, a two-component representation was retained as the reasonable specification.

[6] Details on the SHAP value implementation are provided in Lundberg and Lee (2017).

[7] These hyperparameters were determined through evaluation on a validation dataset, where key parameters (tree depth, learning rate, and number of boosting rounds) were varied and the resulting goodness of fit was assessed. The reported values reflect the configuration that yielded the best performance. We acknowledge that a more systematic tuning approach, such as grid search, was not employed, which remains a limitation of the current study.

4.1 Text-driven model improvement

Table 4 presents the results of the OLS analysis. In both Seongdong-gu and Gangjin-gun, embedding components 1 and 2 significantly enhanced the model performance. The adjusted R^2 values increased from 0.508 to 0.558 and from 0.633 to 0.751 in both regions, respectively. This indicates that text information, specifically environmental descriptions, contributes to improving valuation accuracy.

Table 4. Results of the OLS analysis.

Variables	Seongdong-gu		Gangjin-gun	
	Descriptions not included	Descriptions included	Descriptions not included	Descriptions included
Intercept	4.791*	4.581*	2.394*	2.224*
House type: rowhouse	-	-	0.381	0.455
House type: attached	-0.082	-0.010	-	-
House type: detached	0.010	0.018	0.306	0.264
House type: mixed type	-0.468*	-0.553*	-0.414	-0.623*
House type: other	-0.706*	-0.611*	0.060	0.208
Site area	0.009*	0.009*	0.000*	0.000*
(Skipped for brevity)	…	…	…	…
Embedding component 1	-	-0.265*	-	-0.393*
Embedding component 2	-	0.592*	-	-0.090*
Adjusted R^2	0.508	0.558	0.633	0.751

* Significant at the 5.0% level. The complete table is provided in the Appendix (Table A1).

Figure 3 shows the goodness-of-fit of the test dataset for the three models: OLS without text information, OLS with text information, and XGB with text information. In both regions, the OLS model incorporating environmental descriptions demonstrated a better fit than the OLS model without such information. Furthermore, the XGB model exhibited the best fit among the three models in both regions, making it a reliable choice for more detailed interpretations[8].

[8] In Seongdong-gu, the difference between OLS without text and with text may not be obvious, but the improvement from OLS with text to XGB with text is discernible: the latter model's predictions align more closely with observed values. In Gangjin-gun, the increasing alignment between predictions and observed values (from OLS without text, to OLS with text, and finally to XGB with text) is noticeable.

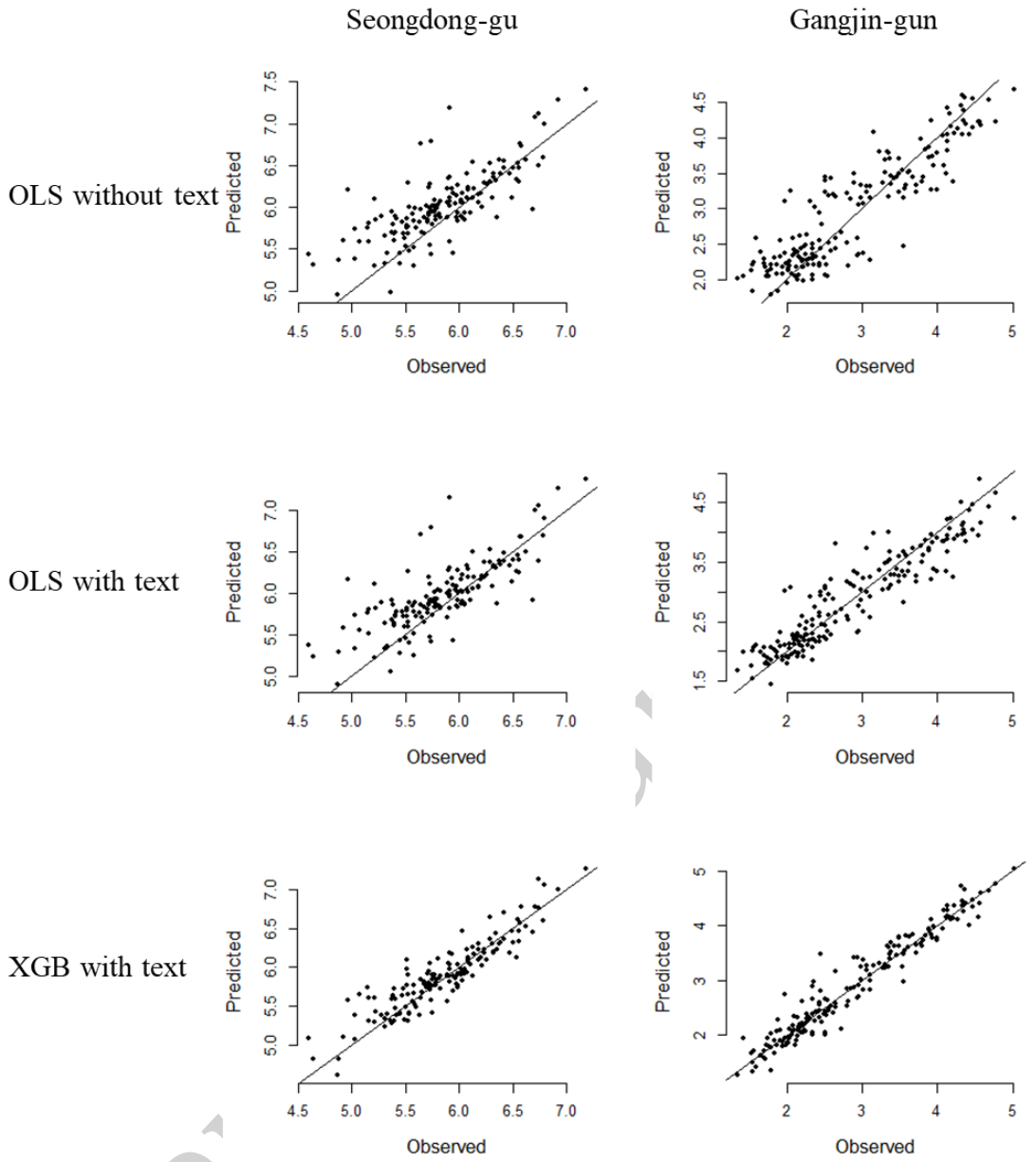

Figure 3. Goodness-of-fit.

4.2 SHAP value interpretation

The XGB model demonstrated the best fit, as shown in Figure 3. Figure 4 presents the importance of the explanatory variables, as determined by the SHAP values of the XGB model. Seongdong-gu, a highly urbanized area with scarce land, identifies site area as the most critical factor in estimating house prices. In contrast, Gangjin-gun, a rural area with abundant land, finds building-related factors such as property age to be more significant in house price estimations. In both regions, embedding components, especially the first one (embedding component 1), rank third in Seongdong-gu and second in Gangjin-gun[9].

[9] As shown in Figure 4, embedding component 2 ranked sixth in both Seongdong-gu and Gangjin-gun, indicating a non-negligible role in house valuation. As noted earlier, the two embedding components together capture 37% of the variance in the sentence embeddings, with components 1 and 2 accounting for 22% and 15%, respectively. By construction, PCA produces orthogonal components, meaning embedding component 2 captures a dimension of linguistic variation in environmental descriptions that is not represented by component 1. The semantic interpretation of component 2 is less straightforward than that of component 1, and therefore we focus on the interpretation of component 1 in this study.

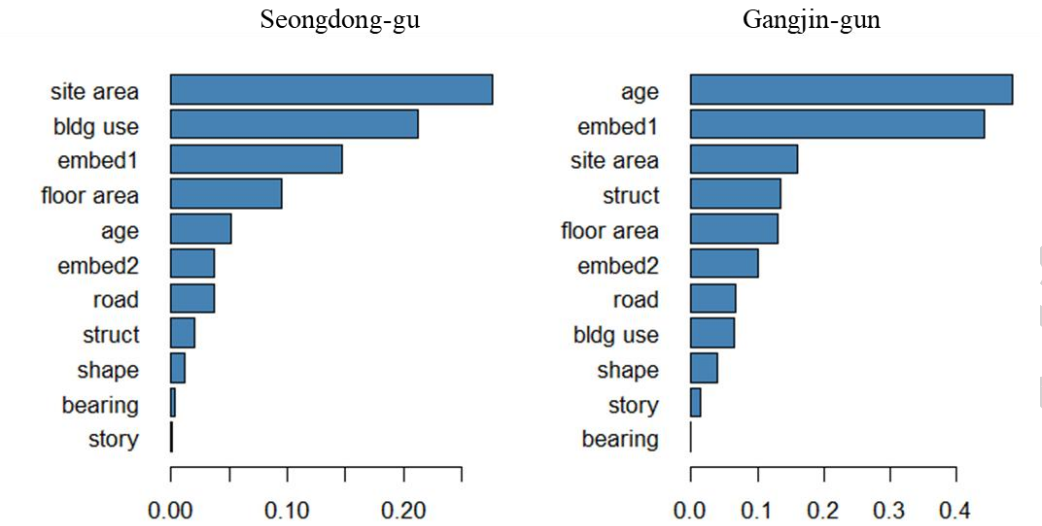

Figure 4. Importance of the explanatory variables.

Note: The scores on the horizontal axis represent the average absolute SHAP values for each variable.

Figure 5 illustrates the relationship between the SHAP values and embedding component 1. As confirmed in Table 4 (Results of the OLS analysis), the coefficients of embedding component 1 were negative (-0.265 and -0.393 for Seongdong-gu and Gangjin-gun, respectively), indicating that an increase in embedding component 1 reduces the outcome value (house price). Similarly, in both regions, as embedding component 1 increases, the SHAP value decreases in Figure 5.

In Seongdong-gu, a negative relationship between the SHAP values and embedding component 1 was observed, although there was some noise in the mid-values of embedding component 1 (a slight upward trend in this range). Notations ① and ② represent environmental descriptions that most significantly increased house prices: "a bustling market area" (-1.69) and "a general market area" (-1.44), respectively. Notations ③ and ④ represent environmental descriptions that most severely decreased house prices: "an established residential area" (-0.57) and "an area designated for development" (-0.26), respectively. Overall, the order of the embedding component values aligns well with the semantic meaning of the environmental descriptions. In large cities such as Seongdong-gu, houses in market areas tend to command higher prices, whereas those in residential areas or areas designated for development tend to have relatively lower prices.

In Gangjin-gun, ① and ② represent "a mix of residential and market areas" (-1.35) and "an existing market area" (-1.33), respectively, while ③ and ④ denote "a rural area nearby the coast" (1.63) and "a rural area on a mountainside" (1.82), respectively. This order of the embedding component values also aligns with the semantic meaning of environmental descriptions: in a rural county such as Gangjin-gun, houses in market areas generally have higher prices, whereas those nearby the coast or on a mountainside tend to be less expensive.

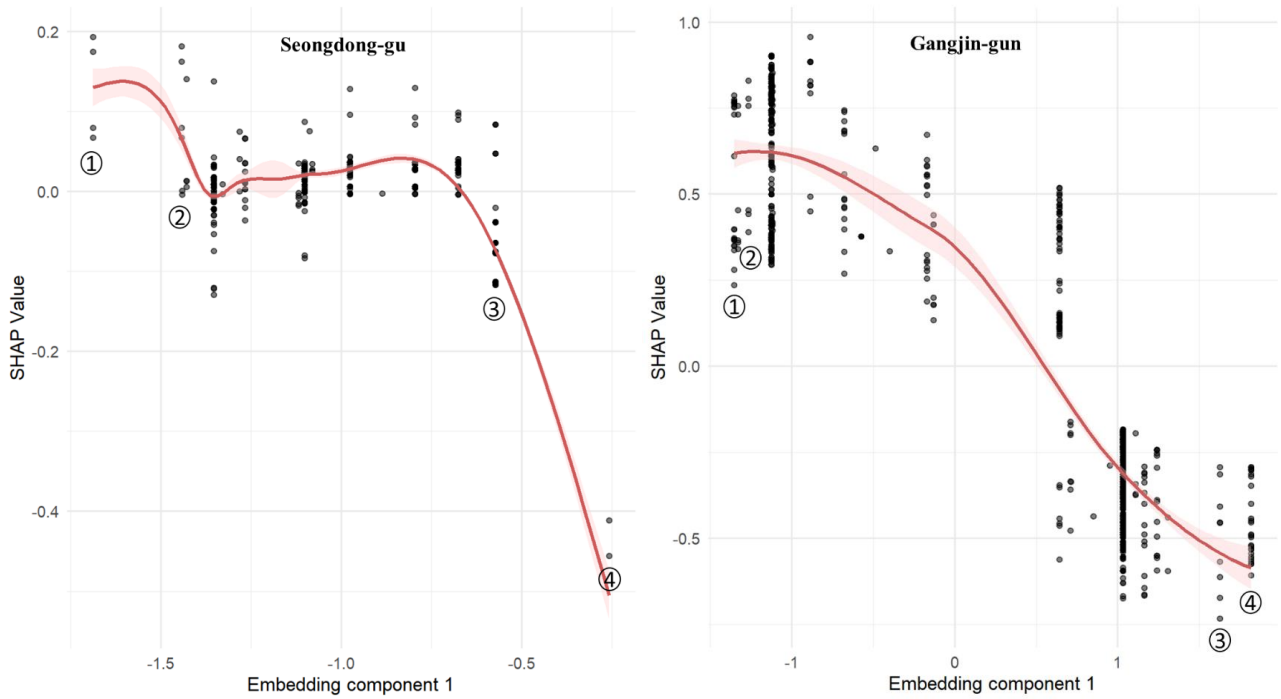

Figure 5. Relation between SHAP values and the embedding component 1.

Note: In Seongdong-gu, ① represents "a bustling market area" (-1.69), ② "a general market area" (-1.44), ③ "an established residential area" (-0.57), and ④ "an area designated for development" (-0.26). In Gangjin-gun, ① indicates "a mix of residential and market areas" (-1.35), ② "an existing market area" (-1.33), ③ "a rural area nearby the coast" (1.63), and ④ "a rural area on a mountainside" (1.82). The values in parentheses indicate the values of embedding component 1.

Figure 5 shows that embedding vectors that describe the environmental characteristics of a house can effectively capture the relationship between house prices and neighborhood features. The Figure 5 also demonstrates that the impact of the embedding vectors on house prices is intuitively comprehensible. This has several implications. First, embedding vectors can serve as a substitute for traditional location variables such as ZIP codes and geographical coordinates when such spatial data are unavailable. Second, when appropriately converted, embedding vectors can encode various latent housing features (e.g., neighborhood characteristics, architectural style, and interior finish quality) that traditional variables often fail to capture. Finally, numerical representations of text information can be expanded to capture the sentiment of the housing market, which is difficult to gauge using traditional metrics, such as sales volumes and construction permits. This enables more informed decision making regarding housing policies.

Housing markets are rich in textual data, including property descriptions, resident reviews, investor opinions, and market analysis reports. The ability to transform textual information into numerical representations is a valuable tool for crafting housing policies as it enhances policymakers' analytical capabilities. Aspects such as neighborhood livability and housing market sentiment, which are often challenging to measure and comprehend, can be better understood using this approach. By offering a more comprehensive view of the factors influencing market dynamics, embedding techniques are expected to provide housing policymakers with better insights into interventions such as zoning changes.

5. Conclusion

This study explored the integration of textual data, specifically environmental descriptions, into house valuation models using embedding vectors. By transforming the environmental descriptions into embedding vectors, we successfully integrated this textual information into both the OLS and XGB models. The inclusion of embedding components derived from the embedding vectors enhanced the performance of both the OLS and XGB models, as evidenced by the higher adjusted R^2 values and better goodness-of-fit. Furthermore, SHAP value analysis provided insights into the contribution of embedding components to price predictions, addressing the interpretability challenges often associated with machine learning models. These findings underscore the potential of leveraging textual data to refine house valuation methodologies.

Although this study successfully used environmental descriptions to improve house valuations, several avenues for future research remain unexplored. First, while this study highlights the value of environmental descriptions in improving house valuation models, the approach's generalizability may be limited. Expanding the scope to include diverse textual data, such as property descriptions or neighborhood reviews, could further enhance model performance. For example, owner-produced descriptions could introduce greater subjectivity, variability in structure, or idiosyncratic language, reducing prediction accuracy by introducing noise in the embedding space. Future applications should test robustness across different text sources and refine preprocessing techniques to improve generalizability beyond standardized texts. Second, exploring hybrid embedding models that integrate both text and visual data (e.g., property photographs or street views) could yield additional improvements. They are often referred to as multimodal embeddings and popular models such as CLIP (Contrastive Language-Image Pretraining) have demonstrated that combining textual and visual features captures complementary information. This study suggests that multimodal embeddings may address limitations in text-only models, paving the way for more comprehensive predictive analytics. Finally, extending this approach to other regions or countries presents a promising direction for future research. Regions with different languages or real estate traditions may require retraining or fine-tuning embedding model on locally relevant data. For instance, environmental descriptions in different cultural contexts might emphasize distinct attributes, such as proximity to religious sites or climate-related features, which may not be well-represented in a model trained on Korean real estate data.

In summary, these three areas, diverse text sources, multimodal embeddings, and cross-regional applications, warrant further exploration in future studies.

Declaration of Interests

The author declares no conflicts of interests.

References

Alshboul, O., Shehadeh, A., Almasabha, G., & Almuflih, A. S. (2022). Extreme gradient boosting-based machine learning approach for green building cost prediction. *Sustainability*, 14(11), 6651.

AlSurayyi, W. I., Alghamdi, N. S., & Abraham, A. (2019). Deep learning with word embedding modeling for a sentiment analysis of online reviews. *International Journal of Computer Information Systems and Industrial Management Applications*, 11, 227–241.

Baltescu, P., Chen, H., Pancha, N., Zhai, A., Leskovec, J., & Rosenberg, C. (2022). Itemsage: Learning product embeddings for shopping recommendations at pinterest. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA (pp. 2703-2711)*.

Baur, K., Rosenfelder, M., & Lutz, B. (2023). Automated real estate valuation with machine learning models using property descriptions. *Expert Systems with Applications*, 213, 119147.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Zhou, T. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1–4.

Deng, L. (2025). Real estate valuation with multi-source image fusion and enhanced machine learning pipeline. *PLoS One*, 20(5), e0321951.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171–4186).*

Eswaraiah, P., & Syed, H. (2024). A Hybrid Deep Learning GRU based Approach for Text Classification using Word Embedding. *EAI Endorsed Transactions on Internet of Things*, 10, 1–8.

Gheini, M., Ren, X., & May, J. (2021). Cross-attention is all you need: Adapting pretrained transformers for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic (pp. 1754–1765)*. Association for Computational Linguistics.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 15180–15190)*.

KOSTAT. (2025). Population and Household Survey. Statistics Korea. Available at: https://kostat.go.kr/ansk (Accessed 10 November 2025).

Lee, C. (2022). Enhancing the performance of a neural network with entity embeddings: an application to real estate valuation. *Journal of Housing and the Built Environment*, 37(2), 1057–1072.

Li, S., & Gong, B. (2021). Word embedding and text classification based on deep learning methods. *MATEC Web of Conferences*, 336, 06022.

Linkon, A. A., Shaima, M., Sarker, M. S. U., Nabi, N., Rana, M. N. U., Ghosh, S. K., & Chowdhury, F. R. (2024). Advancements and applications of generative artificial intelligence and large language models on business management: A comprehensive review. *Journal of Computer Science and Technology Studies*, 6(1), 225–232.

Lundberg, S., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), December 4-9, 2017, Long Beach, CA, USA)*.

Mathur, N., Baldwin, T., & Cohn, T. (2019, July). Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, July 28-August 2, 2019, Florence, Italy (pp. 2799–2808)*. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., & Weng, L. (2022). Text and code embeddings by contrastive pre-training. arXiv preprint arXiv:2201.10005.

Osman, A. I. A., Ahmed, A. N., Chow, M. F., Huang, Y. F., & El-Shafie, A. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal*, 12(2), 1545–1556.

Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A survey of text representation and embedding techniques in nlp. *IEEE Access*, 11, 36120–36146.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), October 25-29, 2014, Doha, Qatar (pp. 1532–1543)*.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China (pp. 3982–3992)*. Association for Computational Linguistics.

Selva Birunda, S., & Kanniga Devi, R. (2021). A review on word embedding techniques for text classification. In *Raj, J. S., Iliyasu, A. M., Bestak, R., Baig, Z. A. (Eds.). Innovative Data Communication Technologies and Application. Lecture Notes on Data Engineering and Communications Technologies, vol 59. Springer, Singapore (pp. 267–281)*.

Sturua, S., Mohr, I., Akram, M. K., Günther, M., Wang, B., Krimmel, M., & Xiao, H. (2024). jina-embeddings-v3: Multilingual embeddings with task lora. arXiv preprint arXiv:2409.10173.

Suryadi, D., & Kim, H. (2018). A systematic methodology based on word embedding for identifying the relation between online customer reviews and sales rank. *Journal of Mechanical Design*, 140(12), 121403.

Torregrossa, F., Allesiardo, R., Claveau, V., Kooli, N., & Gravier, G. (2021). A survey on training and evaluation of word embeddings. *International Journal of Data Science and Analytics*, 11(2), 85–103.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), December 4-9, 2017, Long Beach, CA, USA (pp. 6000–6010)*.

Xu, D., Ruan, C., Korpeoglu, E., Kumar, S., & Achan, K. (2021). Theoretical understandings of product embedding for e-commerce machine learning. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM 21) (pp. 256–264)*.

Ypsilantis, N. A., Chen, K., Cao, B., Lipovský, M., Dogan-Schönberger, P., Makosa, G., & Araujo, A. (2023). Towards universal image embeddings: A large-scale dataset and challenge for generic image representations. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023 (pp. 11290–11301).

Zhang, H., Li, Y., & Branco, P. (2024). Describe the house and I will tell you the price: House price prediction with textual description data. *Natural Language Engineering*, 30(4), 661–695.

Zhao, X., Jin, Z., Liu, Y., & Hu, Y. (2022). Heterogeneous information network embedding for user behavior analysis on social media. *Neural Computing and Applications*, 1–17.

Appendix

Table A1. Results of the OLS analysis.

Variables	Seongdong-gu		Gangjin-gun	
	Descriptions not included	Descriptions included	Descriptions not included	Descriptions included
Intercept	4.791*	4.581*	2.394*	2.224*
House type: rowhouse	-	-	0.381	0.455
House type: attached	-0.082	-0.010	-	-
House type: detached	0.010	0.018	0.306	0.264
House type: mixed type	-0.468*	-0.553*	-0.414	-0.623*
House type: other	-0.706*	-0.611*	0.060	0.208
Site area	0.009*	0.009*	0.000*	0.000*
Floor area	-0.001*	-0.002*	0.001*	0.001*
Story	0.084*	0.087*	0.319*	0.118
Road: middle	0.204*	0.204*	0.334	0.440
Road: narrow	0.184*	0.226*	-0.295	-0.022
Road: narrower	0.288*	0.314*	-0.696	-0.206
Road: narrowest	0.120	0.177*	-0.654	-0.195
Site shape: regular	0.144*	0.136*	0.204*	0.108*
Building frame: brick	0.246*	0.284*	0.792*	0.734*
Building frame: RC	0.375*	0.438*	1.475*	1.385*
Building frame: low quality wood	0.197	0.239	0.153*	0.237*
Building frame: high quality wood	-	-	1.730*	1.786*
Building frame: panel	-	-	0.740*	0.862*
Building frame: steel	-	-	1.142*	1.304*
Property age	-0.005*	-0.005*	-0.010*	-0.011*
House bearing: south	-0.063*	-0.038	0.021	0.051
Embedding component 1	-	-0.265*	-	-0.393*
Embedding component 2	-	0.592*	-	-0.090*
Adjusted R^2	0.508	0.558	0.633	0.751

*: Significant at the 5.0% level.