ISSN: 1648-715X/eISSN: 1648-9179

VILNIUS TECH
Vilnius Gediminas Technical University

INTERNATIONAL JOURNAL OF STRATEGIC PROPERTY MANAGEMENT

2025
Volume 29
Issue 2
Pages 93–101
https://doi.org/10.3846/ijspm.2025.23637

EVALUATING EMBEDDING MODELS FOR TEXT CLASSIFICATION IN APARTMENT MANAGEMENT

Changro LEE*

Department of Real Estate, Kangwon National University, 1 Kangwondaehak-gil, 24341 Chuncheon, Gangwon-do, Republic of Korea

Article History:
- received 26 November 2024
- accepted 21 February 2025

Abstract. The recent proliferation of embedding models has enhanced the accessibility of textual data classification. However, the crucial challenge is evaluating and selecting the most effective embedding model for a specific domain from a vast number of options. In this study, we address this challenge by assessing the performance of embedding models based on their effectiveness in downstream tasks. We analyze consultation records maintained by an apartment management body in South Korea, and convert this textual data into numerical representations using various embedding models. The vectorized text is then categorized using a *k*-means clustering algorithm. The downstream task, specifically, the classification of consultation records, is evaluated using a quantitative metric (Silhouette score) and qualitative approaches (domain-specific knowledge and visual inspection). The qualitative approaches yield more reliable results than the quantitative approach. These findings are expected to be valuable for the various stakeholders in property management.

Keywords: embedding model, text data, clustering, domain-specific knowledge, apartment management.

*Corresponding author. E-mail: *spatialstat@naver.com*

1. Introduction

The rise of generative AI is transforming the industrial landscape, with large language models (LLMs) leading the way. In LLMs, embedding models are important components because they provide numerical representations of text that capture semantic relationships and contextual information. Many organizations and researchers have developed numerous embedding models. Although the abundance of embedding models has simplified text data analysis, the evaluation and selection of the most effective model for a specific task from the vast number of available options remains challenging.

Embedding models can be evaluated in several ways. They can be assessed independently using standard benchmark datasets, such as WordSim-353, SimLex-999, or STS Benchmark, to gauge the quality of the embeddings (García-Ferrero et al., 2021; Abe et al., 2022; Liu et al., 2024). However, a more effective approach involves evaluating their performance based on how well they support downstream tasks. This method helps identify the embedding model that is optimal for a domain-specific downstream task because no single embedding model is universally suitable for all tasks.

This study analyzes the consultation records pertaining to apartment management in South Korea. These records primarily comprise textual data, including enquiries and responses from residents, property managers, and management professionals. We convert these textual data into numerical data using various embedding models. The vectorized text is then fed into a *k*-means clustering algorithm to categorize the records. We conduct a comprehensive evaluation of the embedding models by assessing the performance of the downstream task, specifically, the classification of consultation records in property management. Thereafter, the clustering results are evaluated using both quantitative and qualitative approaches. Based on these evaluations, we determine the effectiveness of the embedding models used in the upstream stage.

This study contributes to the literature in two ways. First, although textual data analysis is commonly conducted in many fields, it is relatively rare in real estate, particularly in property management. This study applies various embedding models to property management text data. Second, the performance of the embedding models is evaluated using both quantitative and qualitative approaches. Such qualitative approaches have rarely been employed before; however, in this study, they yielded more reliable results than the quantitative approach. The findings are expected to provide valuable insights for the various stakeholders engaged in property management.

The remainder of this paper is structured as follows: Section 2 reviews the embedding models and their evaluation methods. Section 3 describes the dataset and methodology used in this study. Section 4 presents the clustering results, evaluates the embedding models, and discusses their implications for property management. Finally, Section 5 concludes the study and suggests directions for future research.

2. Literature review

2.1. Embedding models

An embedding model transforms high-dimensional text data into low-dimensional vectors, thereby capturing semantic meanings and relationships (Li et al., 2020). In natural language processing, words or sentences are converted into dense vectors, which dramatically facilitates mathematical operations by mapping discrete data (such as words) into continuous vector spaces (Yang et al., 2014). These vectors preserve the context and meaning of the original text, thereby enabling tasks such as document classification and sentiment analysis. Table 1 presents the example of an original sentence and its corresponding embedding vector.

Table 1. Example of an embedding vector

Original sentence	Embedding vector
"Hello, property manager!"	[0.25, –0.13, 0.45, 0.67, –0.34, 0.12, 0.89, –0.56, ...]

Note: The actual embedding vector would be much longer.

Numerous pre-trained embedding models have recently been developed by various organizations and researchers. The number of such models is constantly rising as new models are regularly added.[1] Term frequency–inverse document frequency (TF-IDF) is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents. It is one of the earliest models for text representation that focuses on the frequency of terms, while reducing the weight of commonly used words (Aizawa, 2003; Ramos, 2003). Following TF-IDF, word embedding models such as Word2Vec have emerged, which represent words in a continuous vector space, and capture the semantic similarities between words by placing similar words closer together in the vector space. Word2Vec uses neural networks to learn word associations, significantly improving the quality of text representations (Mikolov et al., 2013). More recently, transformers, such as bidirectional encoder representations from transformers (BERT) and generative pre-trained transformers (GPT), have revolutionized natural language processing by using self-attention mechanisms to capture contextual relationships in text more effectively (Vaswani et al., 2017; Rodrawangpai & Daungjaiboon, 2022). Transformers can process entire sentences simultaneously, allowing for a better understanding and generation of the human language.

2.2. Evaluation of embedding models

Textual data analysis has become readily implementable owing to the current abundance of embedding models. However, the challenge lies in evaluating the numerous embedding models. Although embedding models can be evaluated independently, a more effective approach is to assess their performance by assessing how well a downstream task performs when using each embedding model. This approach helps identify the optimal embedding model for a specific downstream task, as no single embedding model is universally relevant for all downstream tasks. In this study, the downstream task involves the clustering of consultation records for property management. Thus, our focus is on evaluating the quality of the clustering results and ultimately determining the effectiveness of the upstream embedding models.

In clustering, the evaluation approach depends on data availability. Several convenient metrics can be used when a labelled dataset is available. Classification accuracy is a common metric. It measures the proportion of correctly classified cases relative to the total number of cases. It is straightforward to understand and widely used in the literature (Janecek et al., 2008; Kilimci & Akyokuş, 2019). Precision, recall, and the F1-score are often used for imbalanced datasets or when the costs of false positives and false negatives differ (Yacouby & Axman, 2020). The area under the receiver operating characteristics curve (AUC-ROC) is a probability curve, and particularly useful for binary classification (Jaskowiak et al., 2022).

When a labeled dataset is unavailable, as in this study, the following techniques can be used. First, the Silhouette score can be used to measure how similar a case is to its own cluster compared with other clusters (Shahapure & Nicholas, 2020). It is frequently used for measuring clustering quality, and ranges from –1 to +1, where a higher value indicates better clustering.[2] Second, visualization of the clustering results can also be useful. The clustering results can be represented graphically using t-distributed stochastic neighbor embedding (t-SNE) or principal component analysis, and researchers can observe whether they form distinct clusters (Bajal et al., 2022). Finally, domain-specific knowledge can guide the evaluation, enabling researchers to assess whether the clustering results capture findings revealed in previous studies (Dash et al., 2022). In this study, we assess clustering quality using the Silhouette score, visual inspection, and domain-specific knowledge of property management. Based on these evaluations, we determine the effectiveness of the embedding models.

[1] As of November 2024, the number of embedding models available for text classification exceeds 70,000 on the Hugging Face Model Hub (https://huggingface.co/models?pipeline_tag=text-classification&sort=trending).

[2] This technique was first proposed by Rousseeuw (1987), where the numerical formula was also provided.

This study contributes to the literature in two ways: by applying embedding models to property management and combining quantitative and qualitative evaluations.

- Although textual data analysis is common in many fields, it is rarely used in real estate, particularly property management. This study applies various embedding models to property management text data.
- The performance of the embedding models is evaluated by assessing a downstream clustering task using a quantitative metric (Silhouette score) and qualitative approaches (domain-specific knowledge and visual inspection). Qualitative approaches have rarely been employed in the literature; however, they yielded more reliable results than the quantitative approach.

3. Method

3.1. Dataset

In South Korea, apartments are predominant among residential housing units, with apartments comprising approximately 14.7 million or 80% of the country's 18.8 million housing units (Korean Statistical Information Service, 2021). This high preference for apartments is driven by factors such as high population density and urbanization pressure, unlike many other countries where single-family houses are more common. To support the management of apartments, many local governments operate support centers that offer various services, including administrative guidance, dispute resolution, training and education, and community development initiatives (Eun et al., 2015).

Among these services provided, administrative guidance is considered most essential. Support centers have consultation offices where residents and property managers can seek advice and information through phone calls or online platforms. To ensure the delivery of quality services, these offices are staffed with qualified experts, such as accountants, legal advisors, and housing policy professionals. All consultations are documented and recorded, with the option to share the data with third parties if the caller or writer consents. This study analyzed a dataset from the Seoul Support Center for Apartment Management, specifically examining 9,023 telephone consultations from March 2017.[3] Examples of consultation records are presented in Table 2.

Table 2. Examples of consultation records (total number of records is 9,023)

Date	Enquiry	Response
2017-03-02	In the event of a water leak from the pipe in the laundry room, who is responsible for the repairs and how is this determined?	The responsibility for repairing the water leak depends on whether the laundry room is designated for public or private use. If the laundry room is intended for public use, the management office is responsible for arranging the repair work and covering the cost. On the other hand, if the laundry room is designated for private use by the residents, the owner of the apartment unit directly above the laundry room is responsible for paying for the repairs
2017-03-03	How should I handle the splitting of deductions from management expenses and recording expenses when processing retained earnings?	Please consult the Accounting Standards Explanation Guide on the Disposal of Retained Earnings

3.2. Embedding models and the clustering algorithm

To convert consultation records (text data) into numerical representations, four embedding models are used: TF-IDF, Word2Vec, Masked and Permuted Network (MPNet), and General Text Embeddings (GTE). As one of the earliest frequency-based models for text vectorization, TF-IDF generated an output embedding dimension of 83,111. This implies that each record was converted into an 83,111-dimensional vector, where the dimensions correspond to the number of tokens used in the TF-IDF. However, such high dimensionality is impractical for subsequent tasks such as clustering analysis. To improve computational efficiency, the dimensions were reduced to 500 using principal component analysis, resulting in a 9,023 × 500 matrix that was fed into the clustering analysis.

Word2Vec, proposed by Mikolov et al. (2013), was utilized in this study through its Gensim version, an open-source library for natural language processing. Word2Vec's output embedding dimension needs to be specified by a researcher, and was set to 300 for this study. Therefore, each record was converted into a 300-dimensional vector.

MPNet is a pre-trained model that combines the architectures of BERT and the eXtreme Language model NETwork (XLNet).[4] Its output embedding dimension was 768, converting each record into a 768-dimensional vector.

[3] The consultation records used for analysis include both enquiries and responses. The original records were in Korean and translated into English before analysis. Using English text simplifies natural language processing (NLP) tasks, both in data preprocessing and in finding relevant pre-trained embedding models. Since Korean is less commonly used in NLP, the availability of relevant pre-trained embedding models is extremely limited. Translating the input to English is a practical approach for conducting NLP tasks on less common languages (Guja et al., 2024). This method leverages the extensive availability and sophistication of models for English, enhancing the potential for accurate analysis. In this study, translation was performed using Google Translation APIs.

[4] In this study, the pre-trained model "sentence-transformers/all-mpnet-base-v2" from the Sentence Transformers library was employed. Further details on MPNet can be found in Song et al. (2020).

GTE is a pre-trained embedding model based on transformer architecture. In this study, a small variant of this model was used to balance performance and computational load.[5] Its output embedding dimension was 384. Each record was converted into a 384-dimensional vector.

The *k*-means clustering algorithm was employed to categorize the consultation records. The algorithm partitions a dataset into *k* distinct groups and aims to uncover the underlying patterns in the data (Morissette & Chartier, 2013). Owing to its ease of implementation and wide accessibility, the *k*-means clustering algorithm was chosen to classify the consultation records.

4. Results

4.1. Clustering results

The *k*-means clustering algorithm requires a researcher to specify the number of clusters in advance. The elbow method is used to determine this number. This method illustrates the relationship between the number of clusters and within-cluster sum of squared distances (WCSS). The optimal number of clusters is identified at the point where the rate of decrease in WCSS sharply declines, forming an "elbow" shape. Figure 1 shows the elbow plots for clustering based on the four embedding models. While TF-IDF- and GTE-based clustering exhibit a clear elbow at six clusters, the optimal number is less distinct for Word2Vec- and MPNet-based clustering. Based on TF-IDF- and GTE-based clustering, this study sets the number of clusters at six (k = 6).

Table 3 presents the *k*-means clustering results derived from the four embedding models. The figures in parentheses below each embedding model's name denote the size of the input features. For TF-IDF, each of the 9,023 consultation records was transformed into a 500-dimensional vector and subsequently fed into the *k*-means clustering algorithm for classification.

While Word2Vec, MPNet, and GTE yielded clusters with relatively balanced case numbers, TF-IDF produced skewed clusters. Cluster 3 contained 4,992 cases, constituting 55% of the total, whereas clusters 4 and 5 contained 361 and 394 cases, respectively. Several methods for evaluating the quality of clustering results have been proposed. This study employed Silhouette scores, domain-specific knowledge, and visual inspection.

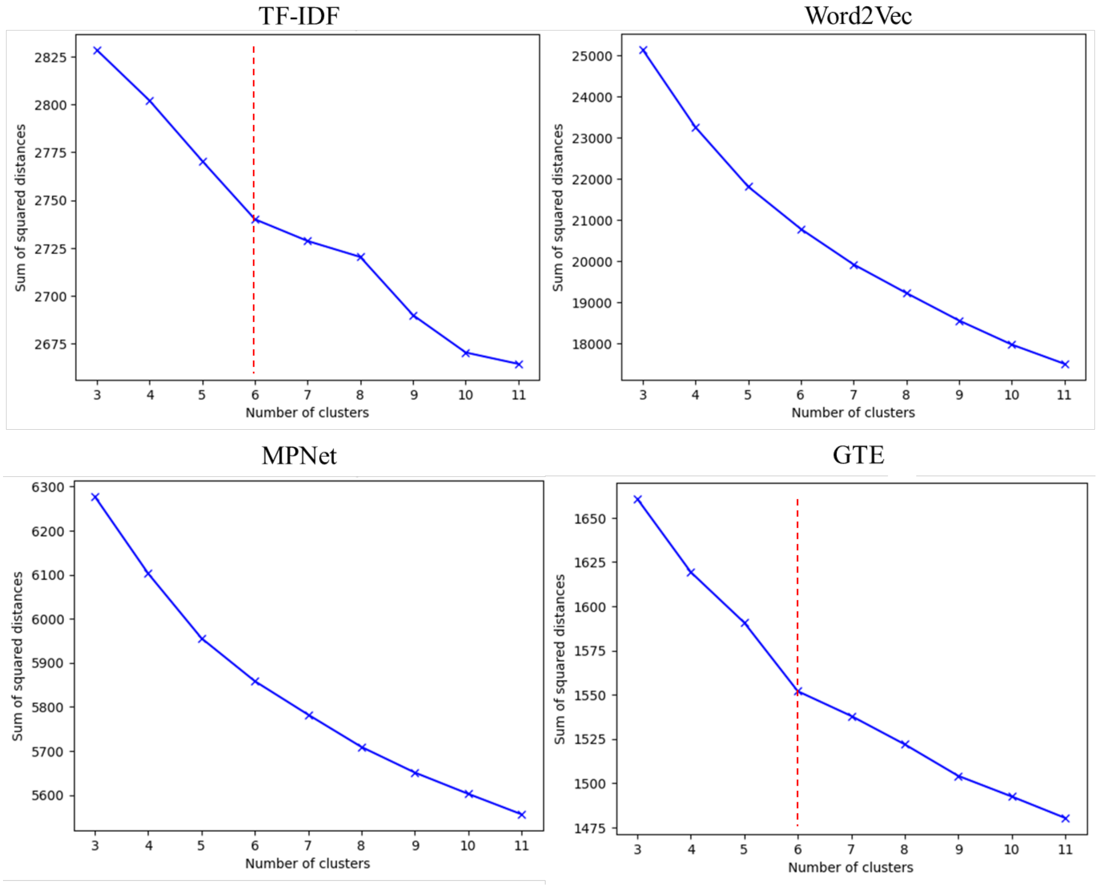


Figure 1. Elbow plots

[5] The pre-trained model "thenlper/gte-small" was used in this study, which is available from the Hugging Face Transformers library. More information on GTE is provided by Li et al. (2023).

Table 3. Clustering results derived from the four embedding models[6]

Cluster	TF-IDF (9,023×500)	Word2Vec (9,023×300)	MPNet (9,023×768)	GTE (9,023×384)
1	1,171	1,224	1,527	1,701
2	1,573	1,483	961	1,230
3	4,992	2,599	2,284	1,537
4	361	1,205	1,645	839
5	394	1,367	954	1,713
6	532	1,145	1,652	2,003
Sum	9,023	9,023	9,023	9,023

4.2. Evaluation

4.2.1. Silhouette scores

According to Kaufman and Rousseeuw (2009), Silhouette scores typically indicate the following: values above 0.50 suggest that a reasonable structure has been found and the clusters are well separated and defined, while values below 0.25 imply that no substantial structure has been found and the clusters may not be meaningful. The Silhouette scores for the clustering results derived from each embedding model are listed in Table 4. Based on the average scores in the final column, MPNet-based clustering proved to be the most effective. Word2Vec- and GTE-based clustering demonstrated moderate effectiveness, and TF-IDF-based clustering was the least effective.

Table 4. Silhouette scores for clustering results derived from the four embedding models

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Average
TF-IDF	0.42	0.42	0.35	0.35	0.27	0.64	0.41
Word2Vec	0.28	0.29	0.42	0.78	0.63	0.75	0.53
MPNet	0.69	0.80	0.81	0.77	0.67	0.42	0.69
GTE	0.43	0.34	0.67	0.77	0.56	0.33	0.52

4.2.2. Domain-specific knowledge

Table 5 presents an overview of the major topics revealed in previous studies on housing management in South Korea. Because these issues were identified and investigated intensively by domain experts, they can be considered reference topics (RT) in apartment management.

To leverage domain-specific knowledge, we employ a keyword-extraction algorithm to extract keywords from each cluster in the four clustering results. A smaller and faster variant of BERT is used for this task[7], and then these extracted keywords are compared with previous studies' findings (Table 5).

Table 5. Reference topics revealed in previous studies

	RT	Description	Source
1	Effective budgeting	Budgeting is essential for apartment management, covering maintenance, repair, administrative expenses, and compliance, ensuring efficient fund allocation for routine maintenance and repair	Eun et al. (2015), Hyun and Lee (2021)
2	Vender management	Vendor management is a common practice in apartment management, involving the selection and contracting of companies that offer various services, such as garbage collection and elevator safety inspections	Eun et al. (2015), Byun (2016)
3	Residents' Representative Council	Residents' Representative Council is the main decision-making body in apartment complexes. Elections for these positions can be competitive. The law requires an election commission to ensure a transparent election process	Byun (2016), Hyun and Lee (2021)
4	Noise-between-floors conflicts	These refer to disputes arising from excessive noise transmission between neighboring units, particularly from upper floors	Kim (2024)
5	Long-term repair plans	Property managers must create long-term repair plans and secure funds from owners for major repairs, following government guidelines on minimum reserves	Shin and Lee (2022)
6	Interpreting regulations and their application to specific cases	Support centers frequently receive enquiries about how to understand and apply relevant regulations to practical situations	Eun et al. (2015)

Table 6 presents the results of the keyword extraction process. The 'Keywords' column lists the three primary keywords extracted for each row. For instance, the first row contains "accounting, account reconciliation, and billing," which aligns with RT 1 in Table 5. The cluster with these keywords was consistently identified across all the four embedding models. Similarly, the second and third rows correspond to RT 2 and RT 3 respectively, and were detected by all four embedding models. However, the cluster associated with RT 4 was only found in two of the four models, TF-IDF and GTE. Similarly, clusters linked to RT 5 and RT 6 were observed in two models each: TF-IDF and GTE for RT 5, and MPNet and GTE for RT 6. Keywords in the remaining rows did not match any RTs in Table 5 and were thus labeled as NA or not-applicable.

[6] As shown in Table 3, the dimensionality of the embedding vectors from the four models varies from 300 to 768. While embedding vector size could impact the performance of clustering analyses, with larger vectors often performing better, this was not observed in the subsequent results of this study.

[7] The pre-trained model "paraphrase-MiniLM-L6-v2" is employed in keyword extraction. This model is based on a smaller version of the BERT model, with "L6" indicating that it has 6 transformer layers. It computes the cosine similarity between consultation records and potential keywords, and selects the phrases with the highest similarity scores as keywords. Further details can be found in Senel et al. (2022).

Table 6. Extracted keywords in the four clustering results

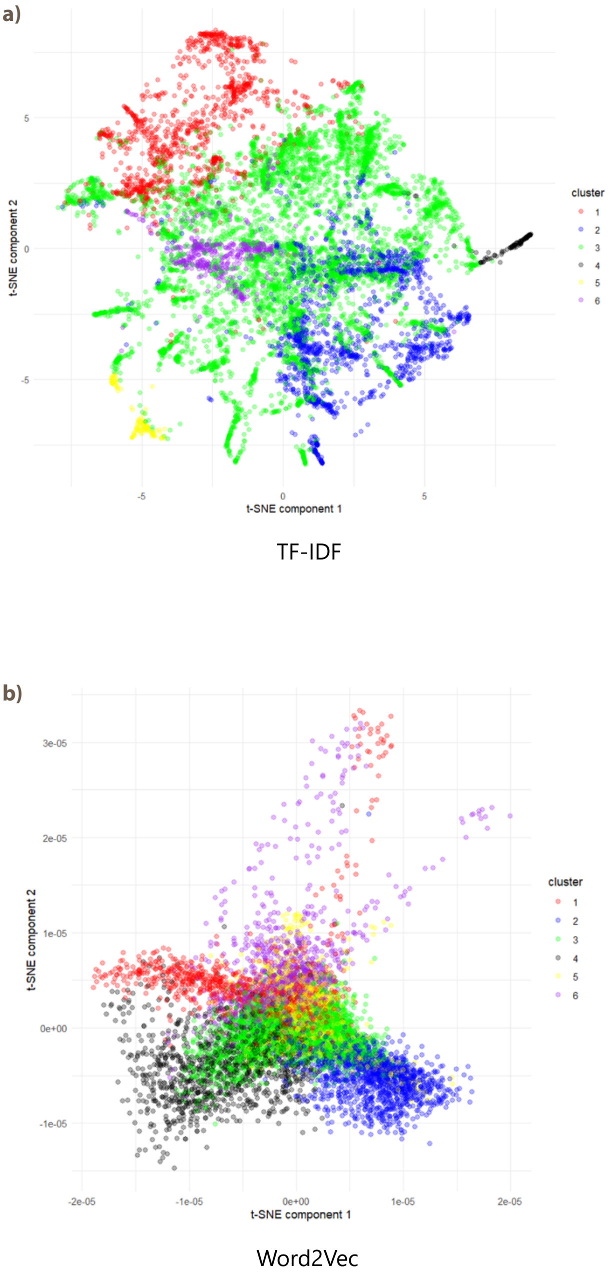
Keywords	RT	TF-IDF	Word2Vec	MPNet	GTE
Accounting, account reconciliation, billing	1	○	○	○	○
Bidding, selection guideline, vendor	2	○	○	○	○
Candidate, representative, residents	3	○	○	○	○
Noise, dispute, floors	4	○	×	×	○
Repair, reserve, liability	5	○	×	×	○
Regulation, enforcement decree, law	6	×	×	○	○
Owner, owner-occupied, local government	NA	×	○	○	×
Budget, year, amount	NA	○	×	×	×
Complaint, facilities, confirmation	NA	×	○	×	×
Housing, tenant, building	NA	×	○	×	×
Complaints, enquiry, information	NA	×	×	○	×

From the TF-IDF to the GTE columns, each column is marked with six circles representing the six clusters (k = 6) determined by each method. All six clusters identified by the GTE-based clustering matched the RTs listed in Table 5. Five, four, and three clusters were matched for TF-IDF-, MPNet-, and Word2Vec-based clusterings, respectively. Based on Table 6, GTE ranks as the best-performing method, followed by TF-IDF, MPNet, and Word2Vec.

4.2.3. Visual inspection

The evaluation of the four embedding models varied when the Silhouette scores were used and domain-specific knowledge was leveraged. In this context, visualizing the clustering results can aid evaluation, as researchers can examine whether distinct clusters are formed in each embedding model. The clustering outcomes were visualized using t-SNE in the 2D space,[8] as shown in Figure 2.

Each panel shows the distribution of 9,023 observations (consultation records) with six clusters overlaid. The GTE plot demonstrates well-defined and separated clusters, with each color (representing a cluster) occupying a relatively distinct area with minimal overlap. The points within each cluster are also more compact compared to other clustering results. Furthermore, the clusters in the GTE plot do not blend together as much as in other results. In the TF-IDF and MPNet-based clusterings, the separation between clusters is not as clean as with GTE. There is more overlap between adjacent clusters, especially in the central regions, and the clusters tend to be more spread out compared to GTE. The Word2Vec plot shows the least distinct cluster formations, with clusters being difficult to differentiate due to considerable overlap. The "clusters" in the Word2Vec plot appear more as a smeared mass with color variations rather than distinct groups. In summary, GTE emerges as the top-performing method, followed by TF-IDF and MPNet, with Word2Vec ranking fourth. Therefore, visual inspection supports the domain knowledge-based evaluation.[9]

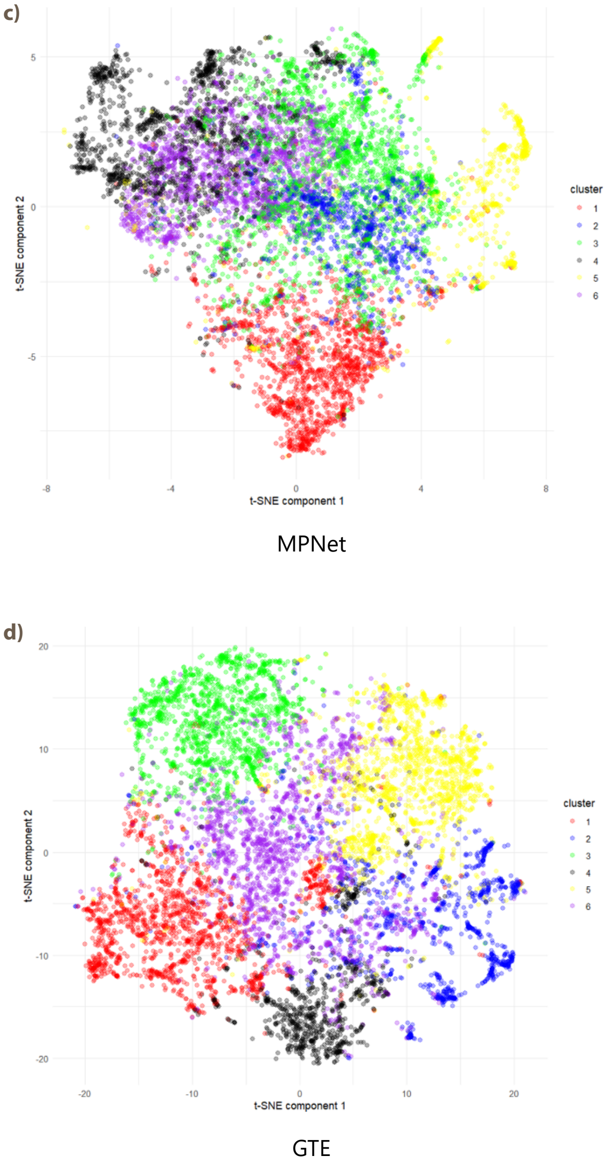


TF-IDF

Word2Vec

[8] t-SNE is an algorithm frequently used for dimensionality reduction. Implementation details are concisely explained in Platzer (2013), and Pareek and Jacob (2021).

[9] Silhouette scores: MPNet > Word2Vec > GTE > TF-IDF. Domain knowledge and visual inspection: GTE > TF-IDF, MPNet > Word2Vec.



MPNet

GTE

Figure 2. Visualization of clustering results in t-SNE 2D space[10]

4.3. Implications for property management

Table 6 reveals that each RT is present in more than one embedding model, whereas each non-RT is found in only one embedding model, with the exception of the "owner, owner-occupied, local government" topic, which appears in two models (Word2Vec and MPNet). This indicates the reliability of the domain knowledge-based evaluation, which is further validated through a visual inspection of the clustering results.

When selecting a single embedding model for consultation records in the context of property management, the preferred choice would be the GTE. This involves converting consultation records into embedding vectors (resulting in a 9,023 × 384 matrix) and categorizing them into six groups. These groups were: effective budgeting, vendor management, residents' representative council, noise-between-floors conflicts, long-term repair plans, and interpreting regulations and their application to specific cases.[11]

This study examined 9,023 consultation records from the Seoul Support Center in March 2017. However, these services are offered throughout the year by all 17 provincial governments in South Korea, making their classification and documentation by humans impractical. The approach proposed in this study offers a promising solution to this challenge. By selecting a domain-relevant embedding model and clustering algorithm, this study demonstrated that residents' enquiries and complaints can be systematically and efficiently classified. The categorized results are expected to aid the 17 support centers for apartment management in identifying enquiry trends and enhancing operational strategies.

5. Conclusions

This study analyzed 9,023 consultation records compiled by the Seoul Support Center for Apartment Management in March 2017. These records are textual data; thus, we converted them into numeric data using the four embedding models: TF-IDF, Word2Vec, MPNet, and GTE. The vectorized text from these models was fed into a *k*-means clustering algorithm, and six clusters were identified. This study conducted a comprehensive evaluation of the four embedding models by assessing how well the downstream task performs. In this study, the downstream task was to classify consultation records for property management. The clustering results were assessed using a quantitative metric (Silhouette score) and qualitative approaches (domain-specific knowledge and visual inspection). Qualitative approaches, which have rarely been employed in this field, have proven to be more accurate and reliable than the quantitative approach. Through these evaluations, we

[10] The range of t-SNE component values does not have a specific meaning. The relative distances between data points in the t-SNE plot are what matters, as they indicate the similarity or dissimilarity between data points.

[11] TF-IDF and Word2Vec generate static embeddings, where a word has the same vector representation regardless of its surrounding context. In contrast, GTE produces dynamic, contextual embeddings, where the representation of a word varies based on the sentence in which it appears. In our case, with consultation records, the same word used in different contexts (e.g., "The property manager fixed the broken window" and "The property manager apologized for the broken promise") may have significantly different meanings. This contextual adaptability allows GTE to perform better. Additionally, GTE is built based on nodes and edges in a graph, enabling a broad comprehension of linguistic features. GTE's underlying graph structure allows it to capture longer-range dependencies between phrases, compared to the sequential processing models like MPNet. This might explain the higher performance of GTE, especially when the consultation records have longer descriptions of the enquiry at hand. While these interpretations are plausible, we acknowledge that further analysis is needed to confirm these hypotheses. Visualizing the embeddings themselves might reveal additional insights.

ultimately determined the effectiveness of the four embedding models. In the context of consultation records, the best-performing embedding model was the GTE.

With the abundance of available embedding models, document classification has become readily implementable. The challenge is to select an effective embedding model for a specific task. This study underscores the importance of domain-specific knowledge and qualitative evaluations in the selection and application of embedding models for domain-specific tasks. The approach adopted in this study is expected to provide insights into similar applications in other domains.

This study was conducted using a dataset from a single month and a single support center, which may not fully represent the diversity of enquiries received throughout the year and across different provinces. We expect that embedding models pre-trained on text data more similar to our domain, such as other types of customer service records or text focused on housing management, would likely perform better than those trained on vastly different data, like social media text. A larger dataset would generally lead to better clustering results, as it provides a more representative sample of the different issues. The quality of the textual data could also impact results; noisy, inconsistent, or poorly formatted data, such as typos and abbreviations, could negatively affect the representation power of embedding models. In summary, domain similarity, data volume, and data quality would impact the results. Therefore, future studies should apply the approach used in this study to different types of documents to confirm the generalizability of the findings.

Regarding the evolving nature of embedding models, we anticipate the following trends. Future embedding models will likely become increasingly complex and incorporate more parameters, leading to a better capture of semantic nuances and improved performance. Additionally, we expect the development of more sophisticated fine-tuning techniques for domain-specific applications. These fine-tuning techniques could help adapt embedding models to new datasets more effectively. In short, the continued development of embedding models means that the optimal model for a certain task may evolve over time.

References

Abe, K., Yokoi, S., Kajiwara, T., & Inui, K. (2022, November). Why is sentence similarity benchmark not predictive of application-oriented task performance? In *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems* (pp. 70–87). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.eval4nlp-1.8

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, *39*(1), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

Bajal, E., Katara, V., Bhatia, M., & Hooda, M. (2022). A review of clustering algorithms: Comparison of DBSCAN and K-mean with oversampling and t-SNE. *Recent Patents on Engineering*, *16*(2), 17–31. https://doi.org/10.2174/1872212115666210208222231

Byun, W. J. (2016). Improving transparency in apartment management. *Practice & Theory of Civil Law*, *19*(2), 79–107. https://doi.org/10.21132/minsa.2016.19.2.03

Dash, T., Chitlangia, S., Ahuja, A., & Srinivasan, A. (2022). A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports*, *12*(1), Article 1040. https://doi.org/10.1038/s41598-021-04590-0

Eun, N., Kwak, D., Chae, H., & Jee, E. (2015). Roles of housing management support center and development plan. *Journal of the Korean Housing Association*, *26*(6), 169–180. https://doi.org/10.6107/JKHA.2015.26.6.169

García-Ferrero, I., Agerri, R., & Rigau, G. (2021, November). Benchmarking meta-embeddings: What works and what does not. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3957–3972). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.333

Guja, A., Siwiak, M., & Siwiak, M. (2024). *Staring data analytics with generative AI and Python*. Manning Publications.

Hyun, S. H., & Lee, E. K. (2021). Determining factors of multi-family housing management policy. *Letter of Korean Policy Sciences*, *25*(3), 35–62. https://doi.org/10.31553/kpsr.2021.9.25.3.35

Janecek, A., Gansterer, W., Demel, M., & Ecker, G. (2008, September). On the relationship between feature selection and classification accuracy. In *New challenges for feature selection in data mining and knowledge discovery* (pp. 90–105). PMLR.

Jaskowiak, P. A., Costa, I. G., & Campello, R. J. (2022). The area under the ROC curve as a measure of clustering quality. *Data Mining and Knowledge Discovery*, *36*(3), 1219–1245. https://doi.org/10.1007/s10618-022-00829-0

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.

Kilimci, Z. H., & Akyokuş, S. (2019, September). The evaluation of word embedding models and deep learning algorithms for Turkish text classification. In *2019 4th International Conference on Computer Science and Engineering (UBMK)* (pp. 548–553). IEEE. https://doi.org/10.1109/UBMK.2019.8907027

Kim, S. H. (2024). Investigating noise-between-floors crimes and their characteristics. *Public Security Research*, *38*(3), 95–128.

Korean Statistical Information Service. (2021). *Population and housing census*. Daejeon City.

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., & Li, L. (2020). *On the sentence embeddings from pre-trained language models*. arXiv.

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., & Zhang, M. (2023). *Towards general text embeddings with multi-stage contrastive learning*. arXiv.

Liu, Y., Chen, W., Liu, H., Zhang, Y., Zhang, M., & Qu, H. (2024). Biologically plausible sparse temporal word representations. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(11), 16952–16959. https://doi.org/10.1109/TNNLS.2023.3290004

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

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, *9*(1), 15–24. https://doi.org/10.20982/tqmp.09.1.p015

Pareek, J., & Jacob, J. (2021). Data compression and visualization using PCA and T-SNE. In *Advances in Information Communication Technology and Computing: Proceedings of AICTC 2019* (pp. 327–337). Springer Singapore. https://doi.org/10.1007/978-981-15-5421-6_34

Platzer, A. (2013). Visualization of SNPs with t-SNE. *PloS ONE*, *8*(2), Article e56883. https://doi.org/10.1371/journal.pone.0056883

Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. *Proceedings of the First Instructional Conference on Machine Learning*, *242*(1), 29–48.

Rodrawangpai, B., & Daungjaiboon, W. (2022). Improving text classification with transformers and layer normalization. *Machine Learning with Applications*, *10*, Article 100403. https://doi.org/10.1016/j.mlwa.2022.100403

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Senel, L. K., Schick, T., & Schütze, H. (2022). *CoDa21: Evaluating language understanding capabilities of NLP models with context-definition alignment*. arXiv.

Shahapure, K. R., & Nicholas, C. (2020, October). Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 747–748). IEEE. https://doi.org/10.1109/DSAA49011.2020.00096

Shin, Y. J., & Lee, C. H. (2022). Factors influencing management expenses and long-term repair plans: Evidence from apartments in Busan. *Tax Accounting Research*, *72*, 31–50.

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and permuted pre-training for language understanding. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 16857–16867). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/c3a690be93aa602ee2dc0ccab5b7b67e-Paper.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems* (pp. 79–91). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.eval4nlp-1.9

Yang, B., Yih, W. T., He, X., Gao, J., & Deng, L. (2014). *Embedding entities and relations for learning and inference in knowledge bases*. arXiv.