

Forecasting Patent Trends and Emerging Technologies in the Construction Industry Using BERTopic

Sirikan Samranrom*, Watchareewan Jitsakul

Department of Information Technology Management, Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand

Received 27 December 2025; Received in revised form 11 June 2026

Accepted 19 June 2026; Available online 30 June 2026

ABSTRACT

This study aims to forecast patent trends and exploring technologies in the construction industry using BERTopic, with patent data as the primary database. In addition, the study forecasts trends in patent registrations related to BIM technology using the Moving Average, Single Exponential Smoothing, and ARIMA methods. The results indicate that ARIMA(0,2,1) yields the lowest mean squared error (MSE) of 1.12. Furthermore, in exploring technology trends of BIM technology through keyword-based BERTopic analysis, the findings show that patents related to BIM technology can be classified into four main topics: (1) Topic 0, related to BIM-based methods and systems, such as the application of BIM in construction systems, management, and methodologies; (2) Topic 1, related to Augmented Reality (AR) and Virtual Reality (VR); (3) Topic 2, related to Digital Twin technology, including Digital Twinning methods integrated with BIM; and (4) Topic 3, related to interior and decoration design using BIM. The findings of this study are intended to support construction firms, research and development departments, and technology policymakers in making informed strategic decisions regarding BIM adoption, innovation investment, and long-term technology planning in the construction sector.

Keywords: BIM; BERTopic; Forecast; Patent

1. Introduction

The construction industry is currently undergoing rapid technological transformation, particularly driven by the adop-

tion of Building Information Modeling (BIM). BIM is a comprehensive information modeling system used for the digital design, planning, and management of

construction projects (scheduling, analyses, cost evaluation, etc.) [1, 2]. This technology facilitates efficient data exchange and collaboration, substantially reduces errors in the construction process, and enhances the accuracy of cost and schedule management [3]. However, despite the widespread implementation of BIM technology, the developmental trajectory and its integration with exploring technologies remain in a state of flux, necessitating in-depth analysis to forecast future directions. Consequently, a thorough understanding of these trends is crucial for informing innovation strategies for organizations and practitioners within the construction sector.

One of the information sources that concretely reflects the overall landscape of technological development is patents, which represent innovations that have been officially registered and legally protected. Therefore, patent data analysis can be used to track exploring technological directions and global trends in research and development. Patent analysis helps enhance understanding of innovation and technological development over time. Related studies include analyses of BIM technology development paths that provide in-depth insights into knowledge diffusion and industry knowledge networks. These studies aim to identify key technologies and future directions of BIM through patent citation network analysis, employing methods such as Main Path Analysis and Cluster Analysis [4,6]. Accordingly, this study seeks to conduct a quantitative analysis of patent data to forecast technological trends in BIM and exploring technologies in the construction industry, providing a comprehensive perspective on the evolution and future directions of BIM-related technologies.

2. Literature Review

Patent analysis is a strategic process that involves organizing, processing, and extracting knowledge from patent documents in order to transform technical and legal information into actionable insights that can support decision-making. Patents constitute the largest and continuously updated repository of technological information, containing detailed technical disclosures that may not be available in other types of academic literature. Consequently, patent data have been widely utilized by researchers for a variety of purposes, ranging from the analysis of technological trends to the development of product innovations.

Research utilizing Social Network Analysis (SNA) includes the work of Pan et al. [5], who examined patents related to Building Information Modeling (BIM) in the construction industry. They developed a text mining framework integrating several techniques: (1) SNA; (2) Latent Dirichlet Allocation (LDA) for topic clustering to extract key thematic content from patent documents; and (3) Natural Language Processing (NLP). Their findings indicate that BIM is increasingly being integrated with exploring technologies—such as the Internet of Things (IoT), blockchain, and Geographic Information Systems (GIS)—to facilitate smart city development and enhance data management efficiency. Furthermore, Mazlumi [9] employed quantitative patent analysis focusing on International Patent Classification (IPC) co-occurrence networks. By applying SNA, the study constructed and analyzed the co-occurrence structures of IPC codes in IoT patents, utilizing various metrics including centrality measures (node importance), edge measures (inter-node relationships), and community detection to identify network clusters.

Park et al. [6] examined the structure and knowledge flow within BIM technology by utilizing patent citation networks, integrated with patent data mining and Social Network Analysis (SNA). The findings demonstrate that BIM applications have evolved beyond design, expanding significantly into energy management, smart city integration (GIS), and big data analytics.

Zhang et al. [7] investigated the construction of innovation pathways by leveraging expired patents through a multi-methodological approach. This framework integrates Functional-Oriented Search (FOS) for cross-domain technology identification, the BERTopic model for advanced semantic analysis, TRIZ theory for systematic technology migration, and the Best-Worst Method (BWM) for prioritizing innovative solutions.

Furthermore, text mining has been extensively adopted as a primary instrument for patent analysis to extract insights and forecast technological trajectories beyond the scope of BIM. For instance, Bamakan [8] utilized text mining to analyze unstructured data and develop patent maps for strategic decision support. Similarly, Charmanas [10] explored deep learning-based NLP techniques to streamline the identification of complex technical topics, thereby reducing the cognitive load on domain experts, particularly within the 6G domain.

Furthermore, several studies have adopted BERTopic, a next-generation topic modeling approach that leverages word embeddings for advanced data analysis. For instance, Mun [12] utilized BERTopic to extract field-level topic patterns from U.S. patent abstracts, examining the correlation between technological evolution and the reactivation of previously overlooked patents. Kim [13] applied this model to identify ten core topics within the customer service in-

dustry, demonstrating its efficacy in forecasting annual technological trends. To address complex and rapidly shifting domains, Jiang [14] proposed an enhanced dynamic topic modeling framework based on BERTopic to capture the temporal evolution of 6G technologies, thereby improving both analytical accuracy and interpretability. Similarly, Lu [15] employed BERTopic to delineate knowledge pathways in leading agricultural universities, emphasizing the linkage between scientific innovation and technical application. Finally, Xian [16] implemented the model to cluster academic literature and track shifts in research priorities over time.

Existing BIM patent studies primarily rely on citation analysis, Social Network Analysis (SNA), Latent Dirichlet Allocation (LDA), and keyword-frequency approaches. While these methods provide useful insights, they often suffer from limited semantic representation and difficulty capturing contextual relationships among patent documents. In contrast, BERTopic leverages transformer-based embeddings that preserve semantic meaning and contextual information, enabling more accurate topic discovery from complex technical texts. Furthermore, previous BIM patent studies rarely integrate topic modeling with time-series forecasting to examine both technological themes and future innovation trajectories simultaneously. This study addresses these gaps by combining BERTopic-based semantic topic modeling with patent trend forecasting methods. Consequently, this study employs a quantitative framework to analyze BIM technological trends and explore potential future development directions in construction technologies. The selected forecasting models (Moving Average, Single Exponential Smoothing, and ARIMA) were cho-

sen because they represent widely adopted time-series forecasting approaches with different assumptions and forecasting characteristics. Moving Average (MA) is effective for smoothing short-term fluctuations and identifying underlying trends, while Single Exponential Smoothing (SES) is suitable for datasets where recent observations may provide stronger predictive information than older observations. ARIMA was included because of its capability to model temporal dependencies and autocorrelation patterns commonly observed in technology and innovation indicators. Since patent registration data may exhibit technology-driven growth, fluctuations, and evolving temporal patterns, comparing these forecasting approaches provides a comprehensive assessment of future BIM innovation trends and enables the selection of the most appropriate model for the dataset. This integrated methodology provides robust insights into the evolutionary trends and outlook of the BIM landscape.

3. Research Methodology

This section presents the methodological foundations underlying the proposed transparency assessment framework. The study integrates three complementary analytical techniques: the Analytic Hierarchy Process (AHP), the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), and Factor Analysis. Factor Analysis is employed to identify latent dimensions associated with corruption and transparency indicators, AHP is used to determine the relative importance of these dimensions through expert judgment, and TOPSIS provides a systematic mechanism for aggregating the weighted criteria and ranking foundations according to their transparency performance. Together, these methods establish a coherent quanti-

tative framework for supporting evidence-based transparency assessment and governance evaluation.

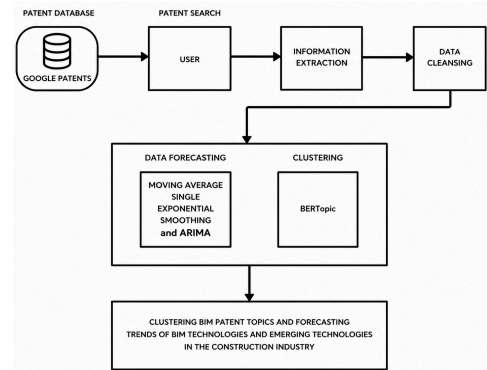


Fig. 1. Research framework.

The Framework consists of four main steps:

- (i) **Data Collection:** Patent data were collected from Google Patents [17] on November 10, 2025 using the search query: (BIM) (Building Information Modeling) (G06Q) (G06F) (G06T) before:priority:20250101 after:priority: 20100101. The IPC classes G06F, G06T, and G06Q were selected for inclusion in the search query based on evidence from previous studies. These classifications cover key technology domains related to digital data processing, image processing and visualization, and business or management information systems, which are commonly associated with BIM technologies. Therefore, the use of these IPC classes helps improve the relevance and precision of the retrieved patent dataset. The search initially retrieved 1,374 patent documents from multiple patent offices including USPTO, CNIPA, KIPO, WIPO, and EPO.

- (ii) **Data Preparation and Processing:** The initial phase involved a comprehensive refinement of the selected patent dataset. This was done to filter out any noise, such as duplicate entries, incomplete records, or irrelevant information. Regarding the textual components were applied a series of standard preprocessing techniques: all text was normalized to lowercase, and we stripped away punctuation numbers, and special symbols that did not contribute to the core meaning. Furthermore, were removed... common stopwords such as "the," "and," and "of" to focus specifically on high-value keywords. By extracting key attributes—such as the title, abstract, registration year, and IPC codes—we converted the raw data were converted into a structured format optimized for analysis. For example, a patent titled 'BIM (building information modeling) forward automatic design method of building structure based on Revit' was processed into: 'bim building information modeling forward automatic design method building structure based on revit'. Integrating this cleaned text with its corresponding metadata, including the 2022 registration year and IPC code G06F, provided a finalized dataset ready for the next stages of the study.
- (iii) **Patent trend forecasting** was conducted using time-series forecasting methods, including Moving Average (MA), Single Exponential Smoothing (SES), and Autoregressive Integrated Moving Average (ARIMA) These methods were selected because they are well-established time-series

forecasting methods that are suitable for relatively small datasets and provide different levels of modeling complexity. The patent dataset used in this study consists of annual patent counts over a limited observation period, making these methods appropriate and reliable for trend forecasting.

- (iv) The identification of key technology topics was conducted through keyword clustering using the BERTopic model to detect exploring technology trends in the future. Patent abstract data were analyzed using the BERTopic model, which serves as a keyword clustering technique to identify the key technology topics represented in the patent corpus.

All data analyses were conducted using Python 3.12. Topic modeling was performed using the BERTopic framework in combination with the SentenceTransformer model (all-MiniLM-L6-v2) for document embedding generation. Dimensionality reduction was carried out using UMAP (Uniform Manifold Approximation and Projection) with cosine distance, while topic clustering was performed using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). Data preprocessing and management were performed using pandas and numpy, while data visualization was generated using matplotlib. Model performance was evaluated using the Mean Squared Error (MSE) metric from the scikit-learn library.

3.1 Forecasting methods and measurement of forecasting accuracy

3.1.1 Moving Average (MA)

MA is a statistical smoothing technique used in time-series analysis to reduce short-term fluctuations and highlight under-

lying trends by calculating the average of observations within a specified time window. As new observations become available, the oldest observations are removed from the calculation, resulting in a continuously updated average. The method is widely used for trend analysis and forecasting due to its simplicity and effectiveness in filtering random variations in data, a process formally articulated by the expression in Eq. (3.1).

$$MA = \frac{A_1 + A_2 + \dots + A_n}{n}, \quad (3.1)$$

where n is the number of observations in the dataset. A is each of the data points.

3.1.2 Single Exponential Smoothing (SES)

SES is a time series forecasting method for data without clear trends or seasonal patterns, a process formally articulated by the expression in Eq. (3.2).

$$F_{t+m} = \alpha y_t + (1 - \alpha)F_t, \quad (3.2)$$

where F_{t+m} is single exponential smoothed value in period $t+m$ for $m = 1, 2, 3, 4, \dots$, y_t is the actual value in time period t . α is the unknown smoothing constant to be determined with value lying between 0 and 1. F_t is the forecast or smoothed value of period t .

3.1.3 Autoregressive Integrated Moving Average (ARIMA)

ARIMA is a statistical forecasting method that uses past values, trends, and errors in time-series data to indicate potential future points. It is denoted as ARIMA(p, d, q), where p is the lag order, d is the degree of differencing, and q is the order of the moving average. The general formula expresses the differenced variable y'_t as a linear combination of its past values and past error terms:

$$y'_t = c + \phi_1 y'_{t-1} + \dots + \phi_p y'_{t-p}$$

$$+ \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \quad (3.3)$$

where y'_t is the time series data that has been differenced (integrated) to make it stationary. c is a constant. ε_t is the white noise error (or residual) at time t . ϕ is Autoregressive (AR). θ is Moving Average (MA).

3.1.4 Mean Squared Error (MSE)

MSE serves as a pivotal metric in model evaluation, quantifying the average of the squared disparities between the observed (actual) values and the corresponding values estimated (predicted) by the model. Its derivation is formally represented by the formula delineated in Eq. (3.4).

$$MSE = \frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2, \quad (3.4)$$

where n is the number of observations in the dataset. y_t is the actual value at time t . $\hat{y}_t$ is the predicted value at time t .

3.1.5 The Root Mean Square Error (RMSE)

RMSE stands as a widely employed metric for quantitatively assessing the discrepancy between a model's projected values and the empirically observed data points. The computation of this measure involves taking the square root of the average of all squared error terms, a process formally articulated by the expression in Eq. (3.5).

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}. \quad (3.5)$$

3.1.6 Mean Absolute Percentage Error (MAPE)

MAPE is a forecasting metric derived using the explicit mathematical formula presented in Eq. (3.6). This measure

effectively expresses accuracy as a percentage of the error, relative to the actual observation. Fundamentally, a reduced MAPE value is directly indicative of superior forecasting precision and robustness, signifying that the model's projections are tightly aligned with the realized values.

$$MAPE = 100\% \times \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|, \quad (3.6)$$

where A_t is the actual value at a given time t . F_t is the forecasted value at time t .

3.2 BERTopic

BERTopic is a topic modeling technique used to extract topics from large collections of text by combining natural language processing (NLP) with transformer-based language models such as BERT (Bidirectional Encoder Representations from Transformers). Unlike Latent Dirichlet Allocation (LDA), which relies on a bag-of-words representation and requires the number of topics to be predefined, BERTopic leverages contextual embeddings from transformer-based language models to capture semantic relationships between documents. This enables BERTopic to automatically determine the optimal number of topics without prior assumptions, and to produce more coherent, interpretable topic representations — particularly when applied to short technical texts such as patent abstracts. The method consists of three main steps:

- (i) Dimensionality Reduction via UMAP: The initial phase involves applying Uniform Manifold Approximation and Projection (UMAP) to perform dimensionality reduction of the vector representations. This procedure is critical for streamlining the processing speed and computational

efficiency of the data, while simultaneously ensuring the preservation of the essential underlying topological and relational structures among the individual data points.

- (ii) Clustering via HDBSCAN: Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is applied to aggregate proximate vectors into distinct clusters, or topics. This algorithm is specifically utilized for its efficacy in identifying and segregating data points that do not exhibit sufficient density to be assigned to any cluster, classifying them as noise.
- (iii) Keyword Extraction using c-TF-IDF: Following clustering, class-based Term Frequency–Inverse Document Frequency (c-TF-IDF) is employed. This technique facilitates the identification of the most representative and salient words or phrases within each derived cluster (topic), thereby providing a clear and interpretable semantic summary for every identified thematic group.

4. Results and Discussion

After screening and cleaning, 724 BIM-related patents remained for topic modeling analysis. Among these, 153 patents with granted status were used for patent trend forecasting. The remaining 571 records represent patent applications that were published but had not yet been granted. The country with the highest number of patent registrations is China (59.48%), followed by South Korea (32%) and the United States (5.28%), as shown in Fig. 2. The dominance of BIM-related patents in China and South Korea may reflect strong government support for digital

construction, smart cities, and Industry 4.0 initiatives. These policies encourage R&D investment, patent protection, and technology commercialization. As a result, BIM innovation appears to be concentrated in a few leading markets, which should be considered when interpreting global technology trends and future developments.

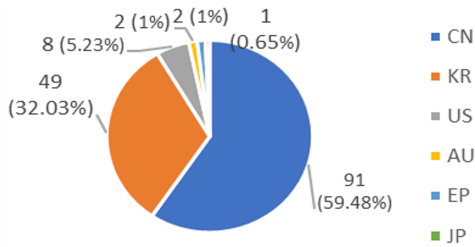


Fig. 2. The percentage of granted BIM patents.

4.1 Forecasting patent registration trends using time series methods

To forecast the trend of patent registrations, statistical methods were employed, including MA, SES and ARIMA. The Moving Average windows of 3, 6, and 9 periods were selected to represent short-term, medium-term, and long-term smoothing effects. The SES smoothing parameter ($\alpha = 0.1-0.9$) was systematically varied to evaluate different weighting schemes for recent observations. For ARIMA model selection, an automated grid search procedure was conducted by examining various combinations of autoregressive (p), differencing (d), and moving average (q) parameters. Candidate models were further screened based on preliminary stationarity assessment and low-order model specifications appropriate for the relatively small dataset. The results showed that ARIMA(0,2,1) yielded the lowest forecasting error ($MSE = 1.12$) and an AIC value of 64.46.

Table 2 presents a comparison of forecasting methods using MSE, RMSE,

Table 1. Comparison of forecasting methods.

Method	MSE	RMSE	MAPE
MA (3)	19.41	4.41	15.18
MA (6)	92.13	9.59	36.61
MA (9)	212.50	14.58	55.65
SES (0.1)	311.9	17.66	67.29
SES (0.2)	156.93	12.52	67.29
SES (0.3)	82.61	9.09	33.71
SES (0.4)	47.08	6.86	24.80
SES (0.5)	29.94	5.47	19.11
SES (0.6)	21.49	4.63	15.58
SES (0.7)	17.23	4.15	13.47
SES (0.8)	14.99	3.87	12.23
SES (0.9)	13.75	3.71	11.48
ARIMA (0,2,1)	1.12*	1.06*	3.13*

and MAPE. Annual patent registration counts from 2010 to 2025 were used as the forecasting dataset. The data were divided into a training set (2010–2022) and a testing set (2023–2025), with a forecasting horizon of three years. The comparison results indicate that ARIMA(0,2,1) achieved the lowest forecasting error ($MSE=1.12$). This finding suggests that the patent registration series exhibits a non-stationary trend structure that can be effectively addressed through second-order differencing ($d = 2$). In addition, the moving average component ($q = 1$) captures the influence of past forecast errors, enabling the model to better represent the underlying temporal dynamics of patent registration activity.

A total of 724 BIM-related patent documents were included in the topic modeling analysis. The parameter settings were selected based on commonly adopted BERTopic configurations reported in previous studies and the official BERTopic implementation guidelines. Specifically, UMAP parameters ($n_neighbors = 15$ and $n_components = 5$) were chosen to balance local and global document structures, while HDBSCAN parameters ($min_cluster_size = 5$ and $cluster_selection_method = "eom"$)

Table 2. Top Keywords of BIM.

Topic	Count	Top Keywords	Representative Patent
-1	17	Pile, carbon, method, emission	carbon method emission
0	633	BIM, building, model, construction, management	Building Information Modeling System for Construction Management
1	32	virtual, reality, visualization, augmented, immersive	Virtual Reality System for BIM Visualization
2	30	digital, twin, monitoring, platform, infrastructure	Digital Twin Platform for Smart Building Monitoring
3	12	interior, decoration, design, furniture, layout	Interior Decoration Design System
SES (0.3)	82.61	9.09	33.71
SES (0.4)	47.08	6.86	24.80
SES (0.5)	29.94	5.47	19.11
SES (0.6)	21.49	4.63	15.58
SES (0.7)	17.23	4.15	13.47
SES (0.8)	14.99	3.87	12.23
SES (0.9)	13.75	3.71	11.48
ARIMA (0,2,1)	1.12*	1.06*	3.13*

were used to ensure stable topic formation and avoid excessive fragmentation of topics. The analysis identified four major technology topics within the BIM patent landscape, as discussed in the following sections.

4.2 Weight assessment using the AHP method

4.2.1 Topic 0: BIM-based methods and systems

This topic constitutes the most substantial cluster in terms of patent volume and exhibits a pronounced upward trajectory. This trend unequivocally signals the accelerated development and widespread application of BIM technologies across construction systems, project management frameworks, and rigorous process control mechanisms.

4.2.2 Topic 1: Augmented Reality (AR) and Virtual Reality (VR)

The notable growth in patent activity within this technology area after 2020 suggests an increasing integration of Augmented Reality (AR) and Virtual Reality (VR) technologies with BIM systems.

4.2.3 Topic 2: Digital twin

This thematic area demonstrates a gradual, yet consistent expansion following 2021. The core focus lies on the progressive development of high-fidelity digital representations of physical assets. This effort is aimed at providing robust support for strategic planning, optimized maintenance protocols, and advanced project management functionalities throughout the asset lifecycle.

4.2.4 Topic 3: Interior and Decoration Design with BIM

The significant increase in patent activity after 2022 suggests that BIM applications are expanding beyond traditional structural engineering into specialized areas such as interior design and architectural detailing.

As shown in Fig. 3, the horizontal and vertical axes represent UMAP Dimension 1 and UMAP Dimension 2, respectively, which visualize the dimensionality reduction of high-dimensional textual data into a two-dimensional space. This transformation enables the observation of topic distribution and clustering patterns. Each dot represents a single document plotted in the UMAP space, providing an overview of how BIM-related texts are clustered by BERTopic based on semantic similarity. The green cluster, corresponding to Topic 0, is the largest group, indicating that a substantial number of documents are classified under the primary topic (e.g., general BIM systems or BIM-based management). The

remaining clusters represent more specialized documents, such as those related to AR/VR, Digital Twin, or interior design applications. The spatial distribution of the documents suggests moderate topic separation, although some overlap between clusters can be observed due to the interconnected nature of BIM-related technologies.

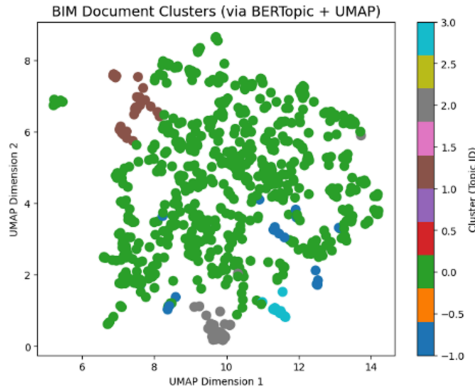


Fig. 3. BIM document cluster.

As shown in Fig. 4, results of topic analysis generated by BERTopic use dimensionality reduction to visualize the spatial relationships among the identified topics derived from patent data. Topic 0 appears as the largest cluster, indicating that it contains the highest number of documents or patents, and is located near the center of the map. This central position suggests that Topic 0 represents a general or overarching topic whose content is highly related to, or encompasses, other topics. In contrast, Topics 1, 2, and 3 are represented by smaller circles, reflecting fewer documents, and are positioned farther from Topic 0 in different directions. This spatial separation indicates that these topics are more specialized, focusing on specific technologies and exhibiting substantial differentiation from Topic 0. Topics positioned closer to each other imply greater content similarity, whereas topics

located farther apart indicate relatively distinct thematic content.

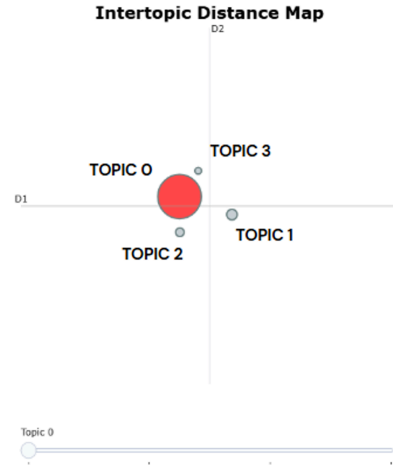


Fig. 4. BERTopic intertopic distance map.

Table 3 presents the topic quality evaluation results based on coherence score, silhouette score, and topic diversity. The BERTopic model achieved a coherence score of 0.364, indicating acceptable semantic consistency among keywords within each topic. The silhouette score of 0.121 suggests moderate separation between topic clusters, reflecting some overlap among BIM-related technology themes. In addition, the topic diversity score of 0.67 indicates a relatively high degree of keyword uniqueness across topics, supporting the distinctiveness of the identified thematic clusters.

Table 3. Topic modeling validation.

Top Keywords	Value
Coherence Score	0.364
Silhouette Score	0.121
Topic Diversit	0.667

As shown in Fig. 5, the temporal evolution of the topics provides critical insight into their development momentum. These trends explicitly indicate the rate of

increase or decrease for each thematic area throughout the study period. Topic 0 (BIM-based Methods and Systems), represented by the blue line, emerges as the dominant thematic area, sustaining the highest volume of documents across the entire observation period. It demonstrates a pronounced upward trajectory commencing after 2019 and ultimately peaking in 2024, recording more than 250 documents. This pattern unequivocally signifies the burgeoning attention and escalating research focus directed towards BIM-based methods and systems.

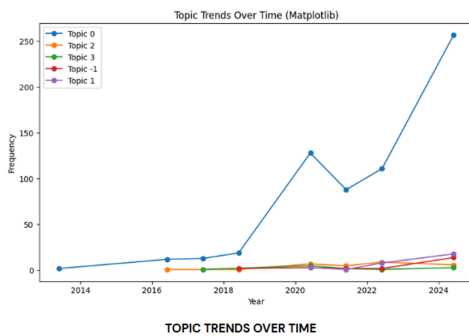


Fig. 5. Topic trends over time.

The remaining topics exhibit the following growth profiles:

- Topic 1 (Purple Line): Shows modest growth acceleration after 2020.
- Topic 2 (Green Line): Illustrates a gradual but consistent accrual of documents following 2021.
- Topic 3 (Orange Line): While possessing a comparatively smaller document count, this topic initiates a noticeable increase post-2022.
- Topic -1 (Red Line) denotes the outlier patents (noise), which are documents that do not possess sufficient

semantic coherence to be successfully assigned to any of the primary topic clusters.

The findings suggest that BIM technology has experienced sustained growth and continues to play an important role in construction-related innovation. Trend technologies such as AR/VR and Digital Twin also appear to be gaining increasing attention, particularly in applications related to model development and interactive simulation. The BERTopic analysis enabled the identification and classification of major technological themes, providing insights into recent patterns of BIM-related innovation. Among the forecasting models evaluated, ARIMA(0,2,1) achieved the lowest forecasting error, suggesting that it may be more suitable for modeling the patent registration series examined in this study. Overall, the results contribute to a better understanding of BIM technology development and emerging technological themes, while providing preliminary evidence that may support future research, technology planning, and policy discussions within the construction sector.

5. Conclusion

This study aims to analyze the development trends of Building Information Modeling (BIM) technology and exploring technologies in the construction industry, using patent data from Google Patents as the primary data source. Patent trend forecasting related to BIM technology was conducted using the Moving Average, Single Exponential Smoothing, and ARIMA methods. The results indicate that ARIMA(0,2,1) yields the lowest mean squared error (MSE) of 1.12, demonstrating the highest forecasting accuracy for patent trend prediction.

For technology topic classification (topic modeling), the analysis using the BERTopic model shows that BIM-related patents can be grouped into four main topics: BIM-based Methods and Systems, Augmented Reality (AR) and Virtual Reality (VR), Digital Twin and Interior/Decoration Design with BIM. Among these, BIM-based Methods and Systems have the largest number of patents, exhibiting a continuous upward trend after 2019 and peaking in 2024. The findings suggest that AR/VR and Digital Twin technologies may represent promising areas of future BIM-related innovation, although additional longitudinal data would be required to confirm long-term trends.

Future research should consider expanding the patent dataset by incorporating records from additional patent databases and validating the robustness of the findings across multiple patent information sources. Such efforts would improve the comprehensiveness and global representativeness of the analysis. Furthermore, the application of Dynamic BERTopic could provide finer temporal insights into topic evolution and emerging technological trajectories. Integrating citation network analysis with topic modeling may also offer a more comprehensive understanding of knowledge flows, technological convergence, and innovation diffusion within the Building Information Modeling (BIM) ecosystem.

References

- [1] Babič NČ, Podbreznik P, Rebolj D. Integrating resource production and construction using BIM. *Autom Constr.* 2010;19:539–543.
- [2] Chen W, Chen K, Cheng JCP, Wang Q, Gan VJL. BIM-based framework for automatic scheduling of facility maintenance work orders. *Autom Constr.* 2018;91:15–30.
- [3] Tang W. Application of BIM technology in the reinforcement and renovation of existing building inspection projects. *Alex Eng J.* 2023;82:240–247.
- [4] Weng KC, Chen JH, Cho RLT, Wei HH, Yu T. Exploring BIM technology development using patent citation analysis: USA case study. *J Civ Eng Manag.* 2025;31(2):129–143.
- [5] Pan X, Zhong B, Wang X, Xiang R. Text mining-based patent analysis of BIM application in construction. *J Civ Eng Manag.* 2021;27(5):303–315.
- [6] Park YN, Lee YS, Kim JJ, Lee TS. The structure and knowledge flow of building information modeling based on patent citation network analysis. *Autom Constr.* 2018;87:215–224.
- [7] Zhang D, Wu X, Lui P, Qin H, Zhou W. Identification of product innovation path incorporating the FOS and BERTopic model from the perspective of invalid patents. *Appl Sci.* 2023;13(13):7987.
- [8] Bamakan SMH, Bondarti AB, Bondarti PB, Qu Q. Blockchain technology forecasting by patent analytics and text mining. *Blockchain Res Appl.* 2021;2(2):100019.
- [9] Mazlumi SHH, Kermani MAM. Investigating the structure of the Internet of Things patent network using social network analysis: application in construction. *IEEE Internet Things J.* 2022;9(15):13458–13469.
- [10] Charmanas K, Mittas N, Angelis L. Topic and influence analysis on technological patents related to security vulnerabilities. *Comput Secur.* 2023;128:103128.

- [11] Wang L, Zhao D. Cross-domain function analysis and trend study in Chinese construction industry based on patent semantic analysis. *Technol Forecast Soc Change*. 2021;162:120331.
- [12] Mun HJ, Lee S. Collective dynamics in the awakening of sleeping beauty patents: a BERTopic approach. *Appl Sci*. 2025;15(19):10395.
- [13] Kim C, Lee J. Discovering patterns and trends in customer service technologies patents using large language model. *Heliyon*. 2024;10(14):e34701.
- [14] Jiang J, Ying F, Dhuny R. Unveiling technological evolution with a patent-based dynamic topic modeling framework: a case study of advanced 6G technologies. *Appl Sci*. 2025;15(7):3783.
- [15] Lu C, Zhu L, Xie Y, Xu W, Zhang C, Zhao Y, Cao Y. Advancing genetics of agriculture and biomolecules: a BERTopic-based analysis of research evolution in leading agricultural universities. *Mol Cell Biomech*. 2025;22(3):901.
- [16] Xian J, Wu JG, Lee SM. Modeling the evolution of virtual reality in nursing education: a BERTopic-based analysis of research trends and future directions. *BMC Nurs*. 2025;24:1332.
- [17] Google. Google Patents [Internet]. Mountain View (CA): Google; 2024 [cited 2025 Nov 10]. Available from: <https://patents.google.com/>