AI-Driven Insights into Construction Progress Monitoring

Ashkan Ebadi *, Yuhao Chen, Farzad Jalaei, Daniel Mao, Alexander Wong

Posted Date: 3 March 2026

doi: 10.20944/preprints202603.0245.v1

Keywords: construction progress monitoring; scientometrics; natural language processing; topic modelling; machine learning

 doi:10.20944/preprints202603.0245.v1

Article

AI-Driven Insights into Construction Progress Monitoring

Ashkan Ebadi [1,2,*], Yuhao Chen [2], Farzad Jalaei [3], Daniel Mao [1] and Alexander Wong [2]

[1] Digital Technologies Research Centre, National Research Council Canada, Toronto, Ontario, Canada
[2] Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada
[3] Construction Research Centre, National Research Council Canada, Ottawa, Ontario, Canada
* Correspondence: Ashkan.Ebadi@nrc-cnrc.gc.ca

Abstract

Construction progress monitoring is vital for effective project management, as it provides essential information that empowers managers to make timely and informed decisions, thereby ensuring successful project completion while preventing delays and cost overruns. The integration of advanced technologies, such as drones, the Internet of Things, and artificial intelligence (AI), offers promising techniques for automated monitoring, transforming traditional manual processes into efficient, data-driven systems that enhance accuracy and reliability in tracking project progress. This paper comprehensively analyzes the key research topics within the field of construction progress monitoring by integrating machine learning with large language models. We utilize AI techniques to extract, interpret, and synthesize key thematic clusters from a corpus of scientific publications spanning the years 2000 to 2024. Our findings reveal three major thematic areas, underscoring the interdisciplinary nature of construction research. The study demonstrates the dynamic evolution of topics, reflecting shifts in research focus and the growing influence of technological innovations. This research not only advances understanding in construction progress monitoring but also showcases the transformative potential of AI-driven methods in uncovering insights from large-scale data.

Keywords: construction progress monitoring; scientometrics; natural language processing; topic modelling; machine learning

1. Introduction

In construction project management, the effective monitoring and control of project progress are crucial tasks [1] as they provide critical information that empowers managers to make timely and informed decisions [2]. Successful project completion relies on the coordinated efforts of the project team, guided by the project manager who must oversee the project network and monitor it for cost, time, and quality deviations. The manager depends heavily on a reliable monitoring system to promptly identify real or potential issues. Therefore, inadequate progress monitoring can lead to a loss of project control, causing delays and escalating costs [1].

Traditional construction progress monitoring necessitates intensive manual data extraction and entry, which is time-consuming and susceptible to human error [3]. In addition, collecting and analyzing such information requires the expertise of specialized personnel [4]. On the other hand, in large-scale construction projects, managers often face significant pressure, as they are tasked with managing an overwhelming workload that demands timely completion and submission [5].

However, recent progress in digital technologies and computer science has facilitated automated progress monitoring in various domains [6]. The concept of automated monitoring in construction projects dates back to the late 20th century, with the introduction of Computer-Aided Design (CAD) and Geographic Information Systems (GIS) into the industry [6]. Advancements in Information and communication technology (ICT), Wireless Sensor Networks (WSN), Radio-Frequency Identification

 doi:10.20944/preprints202603.0245.v1

(RFID), and Global Positioning Systems (GPS) in the early 21st century further opened up new opportunities for the automated monitoring of construction projects [6]. As one of the most influential processes in automated construction monitoring, Building Information Models (BIM) has recently attracted significant interest from researchers [7]. This comprehensive ecosystem can compare the 3D geometry of components [8], creating a digital environment that represents the physical and functional properties of construction projects through advanced technologies, which is instrumental in design, construction, and operational processes [9]. It should be noted that BIM still relies on manually supplied data and updates, and by itself is not sufficient for automated progress monitoring and needs to be coupled with other advanced technologies, e.g., artificial intelligence (AI) [10].

The exponential growth in visual data captured through smartphones, drones, and cameras provides unique opportunities for digitally recording and analyzing the entire construction lifecycle [11]. AI and computer vision (CV) have recently emerged as promising advanced digital technologies for enhancing automated progress monitoring and quality inspection [12]. Innovations in digital technologies are revolutionizing monitoring and oversight capabilities. Computer vision, integrated with BIM, is pivotal in transforming physical-to-digital and digital-to-physical processes within the Construction 4.0 framework [13]. This integration enables the development of 3D reality models that enhance project control, safety inspection, quality assessment, and productivity analysis [7]. In addition, advanced technologies such as Unmanned Aerial Vehicles (UAVs), sensors, and machine/deep learning techniques offer alternative solutions to traditional human-centred monitoring processes [14]. These automated approaches, as evidenced by their success in manufacturing (e.g., [15]) and other industries, have the potential to greatly enhance construction site efficiency, boost safety, increase productivity, and improve quality control [14].

However, implementing and integrating these advanced digital technologies in the construction sector faces several challenges. The substantial expense associated with acquiring necessary equipment and initiation, along with considerable training costs, poses a genuine hurdle [16]. Furthermore, the rapid evolution of digital technologies may discourage construction companies from adopting them, as it necessitates dedicating time to learn new techniques, even though they offer significant potential for long-term returns on investment [14]. Additional challenges specific to construction projects, compared to manufacturing, for instance, include managing indoor and outdoor environments, dealing with weather-related issues, and operating in a dynamic and unstructured setting, which contrasts with the more controlled environment of manufacturing. Despite the challenges, companies are being motivated to embrace digital technologies to sustain or enhance their performance [17].

In this work, we leverage natural language processing (NLP), large language models (LLMs), and machine learning (ML) to conduct an extensive scientometric review of automated progress monitoring (APM) within the construction sector, spanning the period from 2000 to 2024. By mining thousands of scientific publications, we present a robust analytical framework designed to elucidate the research landscape and trace the temporal evolution of its principal research topics. This scientometric study distinguishes itself from existing literature reviews by employing advanced text analytics and AI methodologies to systematically mine and analyze a vast corpus of scientific publications, allowing for a data-driven exploration of thematic trends and evolutions over an extended period. Unlike traditional reviews, this approach provides a comprehensive, quantitative framework that uncovers underlying patterns and insights, offering a more objective and dynamic perspective on the research landscape. Our findings reveal patterns, providing decision-makers with fresh insights into the APM field, detailing its characteristics, focal areas, and developmental trajectory. Among the important findings, the study highlights the emergence of AI-driven methodologies as a transformative force in construction progress monitoring and the increasing importance of integrating real-time data analytics for enhanced decision-making processes. We aim for our results to enhance practitioners' and researchers' understanding of the current state of APM research, facilitating the identification of existing research domains and shedding light on potential future directions.

 doi:10.20944/preprints202603.0245.v1

This study is structured as follows: Section "Data and Methodology" outlines the data collection and methodology. Section "Results" showcases the findings. In the Section "Conclusion", the paper concludes. Finally, imitations and some future research directions are presented in the Section "Limitations and Future Work".

2. Data and Methodology

2.1. Data

2.1.1. Data Collection

This study encompasses all publications related to automated progress monitoring in construction, available via OpenAlex [18], a fully open scientific knowledge graph, and Elsevier's Scopus. We considered multiple data sources of different natures to thoroughly characterize the landscape of the APM research within the construction sector. Data were collected on May 12, 2025, by searching titles and abstracts of publications for construction progress monitoring.

2.1.2. Data Filtering

We excluded the data of 2025 as it did not cover the whole year, filtering in data for the period of [2000,2024]. The collected data included comprehensive metadata for each extracted publication available on the data sources' websites. The data encompassed various elements about papers, including titles, abstracts, publication dates, authors, and their affiliations. Only English-language papers were included in the analysis, and records with no publication date available were removed. Publications lacking a title or abstract were also excluded. The dataset contained 6,418 papers, with 4,019 sourced from OpenAlex and 2,399 obtained from Scopus. Figure 1 shows the data distribution.

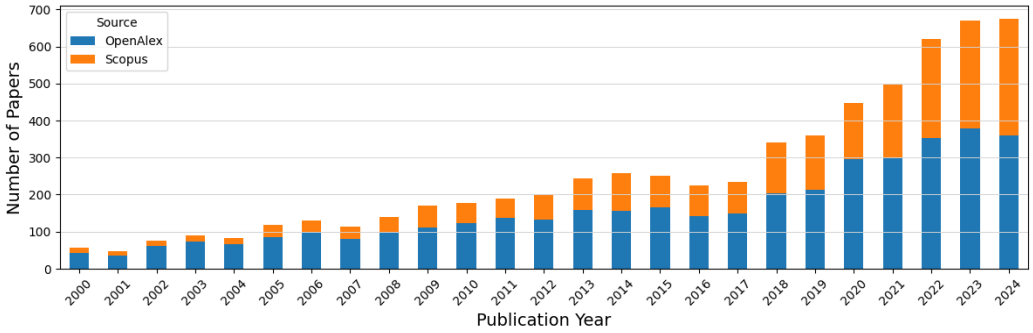


Figure 1. Distribution of publications over time.

2.1.3. Data Preprocessing

We first combined the title and abstract of each publication, generating a new feature, named "text". While the title offers a glimpse into the publication's content, the abstract delivers a more comprehensive and detailed summary. Therefore, we combined them to obtain a more representative summary of the publication's content. Several preprocessing steps were applied to "text", such as converting it to lowercase, removing stop words, and punctuations. We removed entries that had duplicate titles.

2.2. Methodology

Figure 2 shows the high-level conceptual flow of the analyses. The methodology and its components are explained in detail in this section. The analytical pipeline was developed using Python and executed on a cluster powered by NVIDIA GeForce RTX 3090 GPUs.

 doi:10.20944/preprints202603.0245.v1

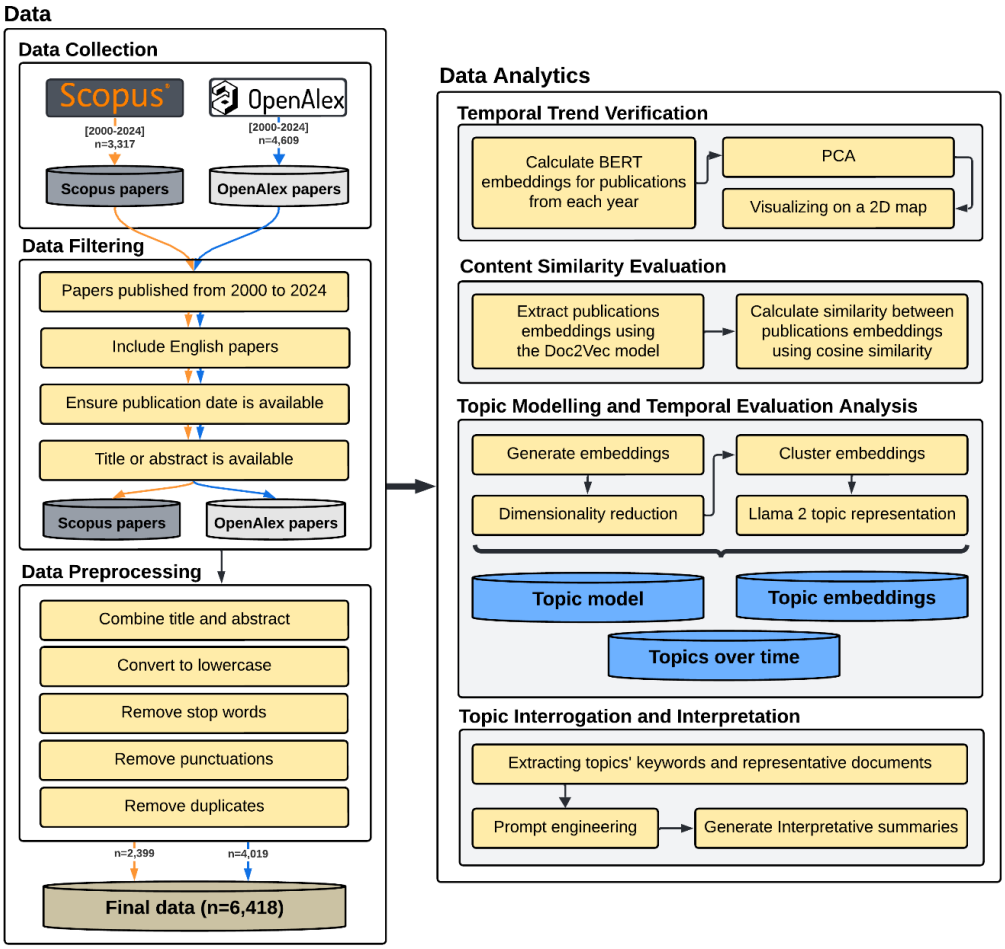


Figure 2. Overview of the high-level conceptual framework for the analyses, illustrating the systematic approach to exploring the APM landscape. After collecting, filtering, and preprocessing the data, patterns are extracted using NLP, LLM, and ML techniques, leading to the identification of thematic trends and the temporal evolution of research topics. Finally, findings are interpreted to provide stakeholders with a comprehensive understanding of the research landscape.

2.2.1. Temporal Trend Verification

To verify the presence of a temporal trend within the collected publications, we implemented a three-step approach. Initially, we used the Bidirectional Encoder Representations from Transformers (BERT) model [19] to calculate embeddings for publications from each year within the study period, denoted as $\{E_{2020}, E_{2021}, \ldots, E_{2024}\}$, where E_i represents the BERT-generated embedding vector corresponding to year i. This step involved leveraging the BERT model to transform textual data into high-dimensional vectors that encapsulate semantic information, providing a robust representation of the publication content for each year. Next, we employed principal component analysis (PCA) [20] to reduce the dimensionality of these embeddings. PCA serves to distill the embeddings into principal components that capture the most significant variance within the data, thereby simplifying the complex, high-dimensional space into more manageable forms without losing essential information. Finally, we visualized these reduced-dimensional embeddings on a 2D map, enabling us to analyze patterns and trends over time. This visualization assists in revealing shifts and clusters that might indicate changes in research focus or thematic evolution across the examined years.

2.2.2. Content Similarity Evaluation

To explore the progression of research terminology over time, we conducted an analysis of textual similarity among publications across various years. For this purpose, we initially computed

 doi:10.20944/preprints202603.0245.v1

embeddings for publications across various years. Given the varying lengths of text in publications, traditional fixed-length feature vector techniques, such as bag-of-words, may not be ideal for extracting embeddings. These methods have two significant limitations: first, they disregard the order in which words appear, and second, they overlook the semantic meaning of the words [21]. To overcome these limitations, we used the Doc2Vec model, also called the Paragraph Vector [21], to extract embeddings. The Doc2Vec model is an unsupervised algorithm designed to learn fixed-length feature representations from texts of varying lengths, such as the publications in our dataset. By employing this algorithm, we represented publications from each year as dense vectors, capturing the nuanced features of the text. Next, we employed cosine similarity to assess the similarity between the embedding vectors corresponding to different years.

2.2.3. Topic Modelling and Temporal Evaluation

We utilized BERTopic [22], a robust topic modelling technique that leverages advanced natural language processing capabilities, to extract coherent topics from the publications and examine their temporal evolution. Topic modelling is an unsupervised machine learning approach that summarizes extensive textual data by uncovering latent semantic themes. This technique has found widespread application across numerous fields, including technology foresight [23] and the study of new diseases and pandemics [24], among others.

BERTopic [22] combines BERT [19] embeddings with clustering algorithms to accurately capture the semantic nuances in the text data. By using BERT, BERTopic benefits from a deep understanding of language context and meaning, enabling it to identify topics that are not only statistically relevant but also contextually rich. Once topics are extracted, BERTopic facilitates the evaluation of their temporal evolution, allowing for tracking of how these topics develop and shift over time. This is particularly useful in analyzing trends within a corpus, offering insights into how the focus of research changes, helping to uncover emerging themes and the dynamics of interest in specific areas. BERTopic offers distinct advantages that are particularly beneficial for our study, setting it apart from other conventional topic modelling approaches. One of the primary benefits of BERTopic is its capability to learn coherent language patterns [22], making it adept at handling a variety of tasks beyond the limitations of models that typically specialize in a single domain [25]. This versatility ensures that BERTopic can effectively capture the diverse themes present within our corpus of publications. Moreover, BERTopic's architecture decouples the document embedding process from topic representation, granting us the ability to employ different techniques. This flexibility enhances our ability to adapt the model to the specific nuances of our data, ensuring that the extracted topics are both accurate and contextually relevant.

BERTopic [22] operates through a straightforward yet effective five-step process. The first step involves embedding documents using BERT, which captures the semantic richness of the text by generating high-dimensional vector representations. These embeddings encapsulate the context and nuances of the language within the documents. Following this, the dimensionality of the embeddings is reduced to simplify the data while preserving its essential structure and relationships. This step is crucial for managing computational complexity and enhancing the efficiency of subsequent analyses. In the third step, the reduced embeddings are clustered, which groups documents based on their similarity, identifying natural clusters within the data. Each cluster represents a potential topic, characterized by the shared themes of its constituent documents. Next, documents within each cluster are tokenized, breaking them down into individual words or tokens to facilitate detailed analysis. Finally, BERTopic extracts the best-representing words for each cluster, identifying the keywords that most accurately reflect the underlying theme of the documents grouped within that cluster. To enhance topic representations, we utilized the clusters and topics generated by BERTopic and integrated them with the Llama 2 large language model [26]. The Llama 2 model was fine-tuned to generate meaningful representations and labels for the extracted topics. BERTopic's topic creation capabilities, integrated with Llama 2's advanced topic representation features, enabled the improvement of the quality and clarity of the representations.

 doi:10.20944/preprints202603.0245.v1

2.2.4. Topic Interrogation and Interpretation

Utilizing large language models (LLMs), we developed an interactive approach in Python to enhance the interpretation of the extracted topics, enabling end-users to communicate effectively with these topics to gain deeper insights. In the first step, the pipeline extracts key components associated with each topic, including the main keywords and representative document samples. These components are aggregated to form a comprehensive semantic representation of each topic. Next, the extracted information acts as contextual input to guide the generation of interpretive summaries via prompt engineering. The following structured prompt template is then instantiated for each topic:

prompt = """You have a research topic that contains the following main documents: [DOCUMENTS]
The top-10 keywords of this topic are: [KEYWORDS]
Based on the information about the topic above, in two paragraphs, describe the importance and trend of this topic within the period from 2000 to 2024."""

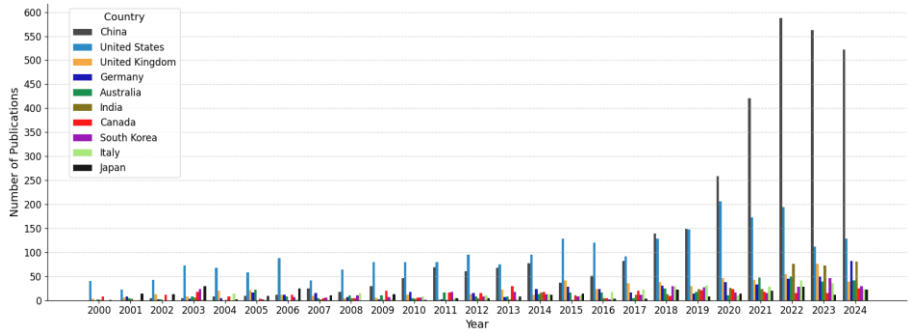
In the above template, "[DOCUMENTS]" and "[KEYWORDS]" refer to the representative documents and the associated set of keywords for a given topic, respectively, which are automatically retrieved from the BERTopic model. These prompts are submitted to a Llama 2 model [26]. The model processes the input and returns a coherent, human-readable summary that describes the characteristics and the latent semantic theme of the topic within the examined period. Users can further communicate with the LLM if needed. This workflow enhances the explainability of topic modelling results by integrating transformer-based language understanding, providing a scalable approach for interpreting high-dimensional topic spaces while retaining alignment with the original document corpus.

3. Results

3.1. Leading Nations

Figure 3a illustrates the top ten countries leading in research publications on construction progress monitoring from 2000 to 2024. These countries include China, India, South Korea, and Japan from Asia (n=4); the United Kingdom, Germany, and Italy from Europe (n=3); Australia; and the United States and Canada from North America (n=2). As shown in the figure, China and the United States have published significantly more than the other leading countries, especially in recent years. The United States held the top position until 2017, after which China surpassed it. As seen in Figure 3b, since 2021, the gap between China and the United States has widened notably, with China further solidifying its leading position. In addition, the figure indicates that China's interest in this field of research increased after 2006.

a) Top-10 countries.



b) Top-3 countries.

 doi:10.20944/preprints202603.0245.v1

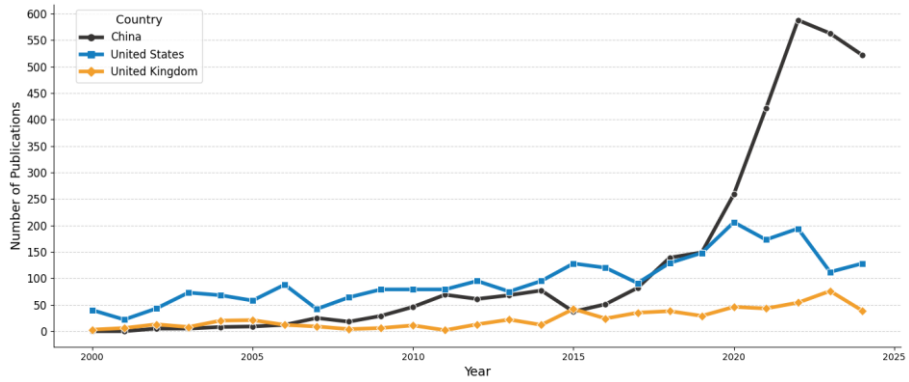


Figure 3. Historical trends of leading countries in publishing research on construction progress monitoring.

3.2. Temporal Trend Verification

As detailed in Section 2.2.1., we initiated the verification of temporal trends by performing PCA on the BERT embeddings of publications from each year. The outcome of this analysis is depicted in Figure 4, where the point colours gradually shift from light blue, representing the initial years of the study period, to dark blue, indicating the later years. This colour gradient visually conveys the progression and changes over time within the dataset. As seen, papers from the early, middle, and later years are clustered closely together, reflecting the evolving nature and progression of publication content over time, hence indicating the existence of a temporal trend. Moreover, a comparison between the initial and final years suggests a noticeable shift in research terminology and/or focus, indicating a significant evolution in the thematic aspects of the publications over time.

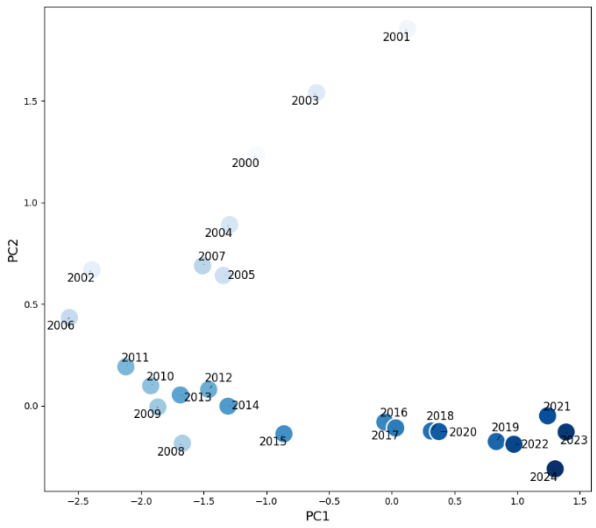


Figure 4. Verification of temporal trends through PCA on BERT embeddings of publications from each year. Point colours transition from light blue for the early years in the study period to dark blue for the later years, illustrating the progression over time.

3.3. Content Similarity Evaluation

As outlined in Section 2.2.2., we employed the Doc2Vec model to extract embeddings from the publications, followed by the use of cosine similarity to measure the similarity between these embedding vectors. This methodology allowed us to thoroughly examine the content similarity of publications across various time periods. The results are illustrated in Figure 5. Observations indicate a general increase in similarity among publications over time, with those from more recent years exhibiting the highest similarity to each other. Our analysis suggests that this upward trend in

 doi:10.20944/preprints202603.0245.v1

similarity began to surface around 2013. More specifically, since 2017, the landscape of research publications in this field has seen a notable increase in similarity. This may indicate that, as the field has matured, researchers have increasingly converged on a set of focused topics and methodologies that have proven to be fruitful, leading to a more unified research approach. Or, influential papers published around this time may have set new standards, prompting subsequent works to align closely with these groundbreaking findings. We will further investigate the research topics in the following sections. In addition, publications from the early years demonstrate the lowest levels of similarity, with research from 2001 being particularly underrepresented in subsequent years. This finding aligns with Figure 4, where publications from 2001 are notably more distant from those of other years, highlighting a potential divergence in research focus during that period.

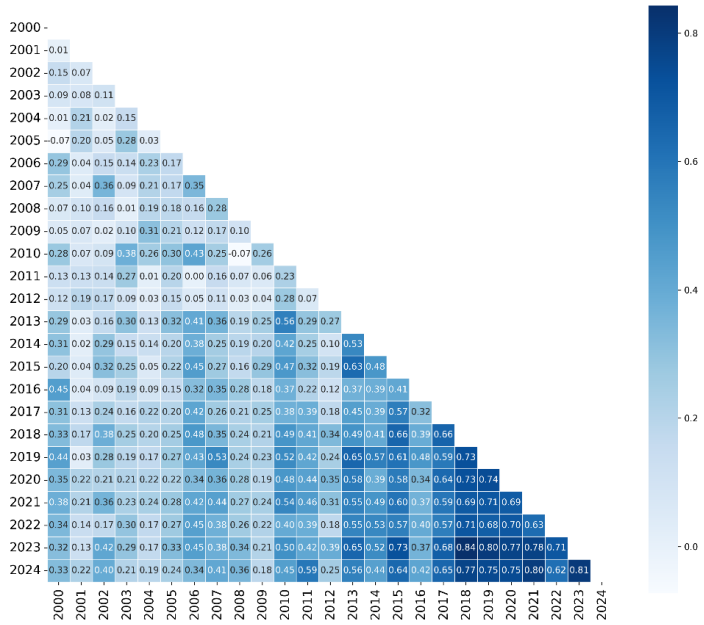


Figure 5. Content similarity heatmap of publications. Darker shades indicate greater similarity.

3.4. Topic Modelling and Temporal Evaluation

3.4.1. Extracted Topics

Table 1 presents the extracted topics by the BERTopic model, accompanied by their top three keywords determined using Maximal Marginal Relevance (MMR), along with the Llama 2-generated label for each topic. MMR is a technique that selects keywords by balancing relevance and diversity to ensure comprehensive topic representation.

Table 1. Extracted topics.

No	MMR (top-3)	Llama 2 Label
1	construction, monitoring, progress	Construction Progress Monitoring
2	tunnel, excavation, monitoring	Tunnel Deformation Monitoring
3	concrete, monitoring, structural	Bridge Monitoring and Damage Assessment
4	construction, project, monitoring	Construction Project Monitoring and Control
5	monitoring, soil, sensing	Satellite Sensing and Erosion Mitigation
6	water, china, ecological	Environmental Monitoring, Conservation and Restoration
7	energy, performance, housing	Sustainable Practices in the Industrial Sector
8	gender, housing, economic	Sustainability in Social and Economic Contexts
9	project, management, cost	Project Management and Cost Control
10	dam, construction, grouting	Construction Simulation and Monitoring of Rock-Fill Dams
11	monitoring, automated, cloud	Construction Progress Monitoring with Computer Vision

 doi:10.20944/preprints202603.0245.v1

12	technology, monitoring, agricultural	Intelligent Agricultural Monitoring System
13	beam, nuclear, security	Nuclear Safety and Security
14	river, conservation, wildlife	Wildlife Research and Species Protection
15	concrete, bridge, sensors	Advanced Concrete Optical Remote Sensing
16	fluorescent, nanoparticles, carbon	Nanotechnology Applications in Energy and Environmental Sectors
17	co2, project, township	CO2 and CO Emissions Management
18	electricity, network, monitoring	Smart Grid Monitoring
19	uav, construction, drone	Automation of Infrastructure Construction
20	global, crisis, citizenship	Global Economic Crisis and Its Impacts
21	building, civil, engineering	Building Services and Energy Efficiency
22	monitoring, transmission, electric	IoT Applications in Power Management
23	construction, bim, robotics	Robotics
24	reaction, synthesis, catalyzed	Organic Reactions and Catalysis
25	internet, law, media	Politics and Media Influence
26	telescope, galileo, Huygens	Mission Planning and Execution for Space Exploration
27	spacecraft, messenger, solar	NASA

3.4.2. Topics Hierarchy

Figure 6 presents a hierarchical clustering of the 27 topics extracted from the corpus. Close linkage distances within each cluster in the figure suggest strong thematic cohesion, while the longer linkage distances between clusters indicate broader topical divergence. Three major clusters are observed based on thematic similarities. The first cluster (green) includes, but is not limited to, topics related to environmental science, sustainability, agriculture, and bridge and tunnel monitoring. The first cluster (green) encompasses topics such as environmental science, sustainability, agriculture, and bridge and tunnel monitoring, highlighting a focus on ecological and structural themes. The second cluster (red) centres on progress monitoring and project management in construction, emphasizing the use of various technologies in these areas. The third cluster (cyan) covers a range of societal and high-tech themes, including the global economic crisis, space exploration, and nanotechnology applications. The blue line illustrates how these three major clusters are grouped. The diversity of topics may highlight the interdisciplinary nature of the field.

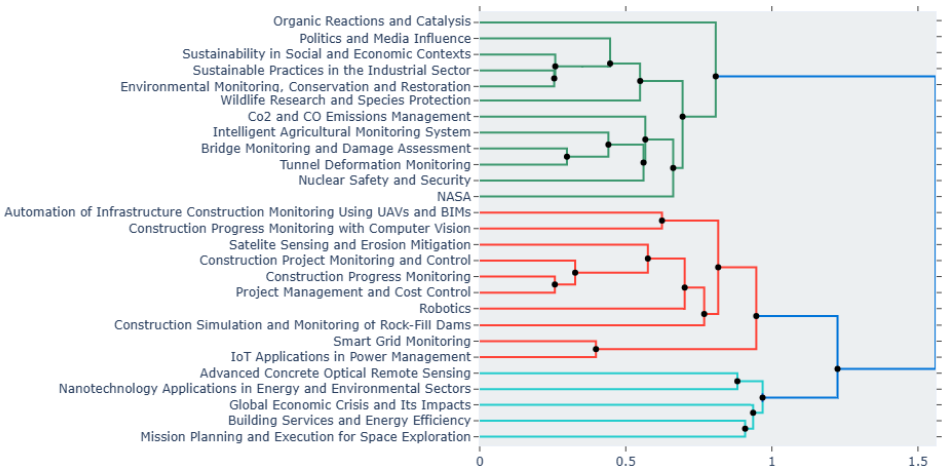


Figure 6. Hierarchical clustering of the extracted topics. Three major clusters are observed. The green cluster focuses on ecological and structural themes, the red cluster emphasizes construction progress monitoring and project management technologies, and the cyan cluster spans societal and high-tech topics.

 doi:10.20944/preprints202603.0245.v1

3.4.3. Topics Similarity

Figure 7 visualizes the pairwise semantic similarity scores between the 27 extracted topics. Stronger similarities are indicated by darker blue shades, where closely related topics such as "Construction Progress Monitoring", "Tunnel Deformation Monitoring", and "Bridge Monitoring and Damage Assessment" exhibit high mutual similarity. These topics form a dense block of high similarity in the top-left region, suggesting a coherent thematic cluster around infrastructure monitoring. In contrast, topics such as "NASA", "Politics and Media Influence", and "Mission Planning and Execution for Space Exploration" show lighter shades when compared with most others, indicating lower similarity and greater thematic distinctiveness. This aligns with the dendrogram's clustering (see Figure 6), highlighting a division between construction/environment-focused topics and more isolated or multidisciplinary themes. Overall, the topic similarity heatmap confirms the existence of well-defined clusters and supports the segmentation observed in the hierarchical clustering.

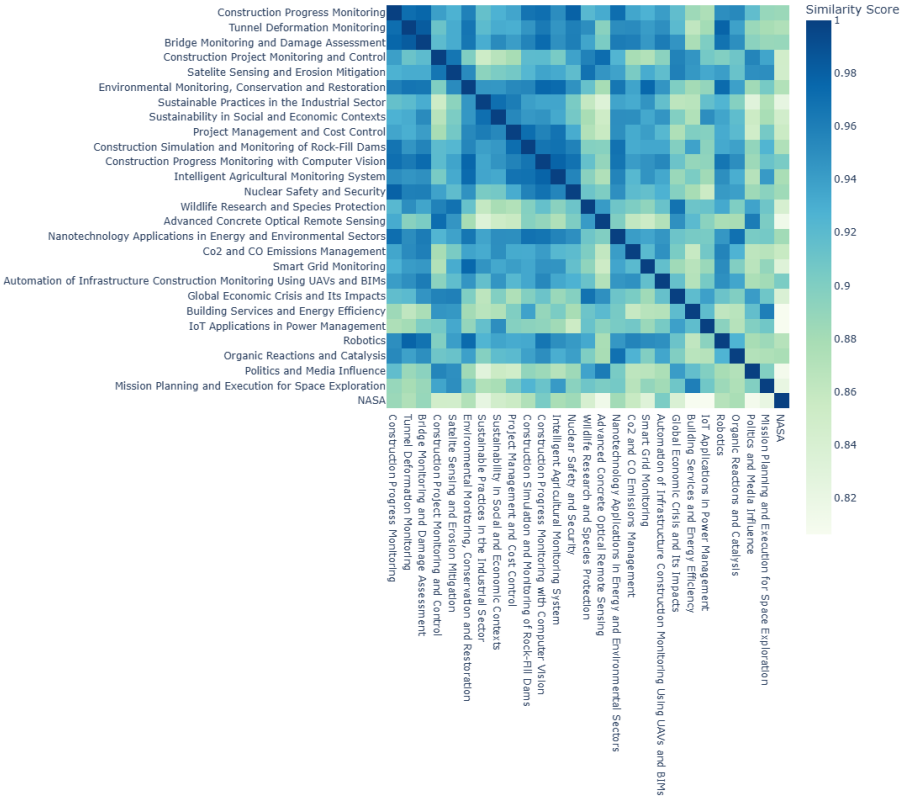


Figure 7. Pairwise semantic similarity between the extracted topics, with darker blue indicating higher similarity.

3.4.4. Temporal Evolution of Topics

Figure 8 illustrates the temporal evolution of the extracted topics, with line colours corresponding to their respective clusters in the hierarchical dendrogram (see Figure 6). The temporal evolution plots show notable differences in how the 27 topics have developed over time. In the green cluster, topics such as "Tunnel Deformation Monitoring", "Environmental Monitoring, Conservation and Restoration", "Sustainability in Social and Economic Contexts", "Sustainable Practices in the Industrial Sector", "Intelligent Agricultural Monitoring System", and "Bridge Monitoring and Damage Assessment" display a clear upward trajectory, suggesting increasing attention in the literature and growing relevance in research agendas. Topics such as "CO2 and CO Emissions Management" also show recent gains, aligning with global climate concerns. Meanwhile, topics such

 doi:10.20944/preprints202603.0245.v1

as "NASA", show minimal frequency and lack a sustained trend, implying they are either interdisciplinary or peripheral in this corpus.

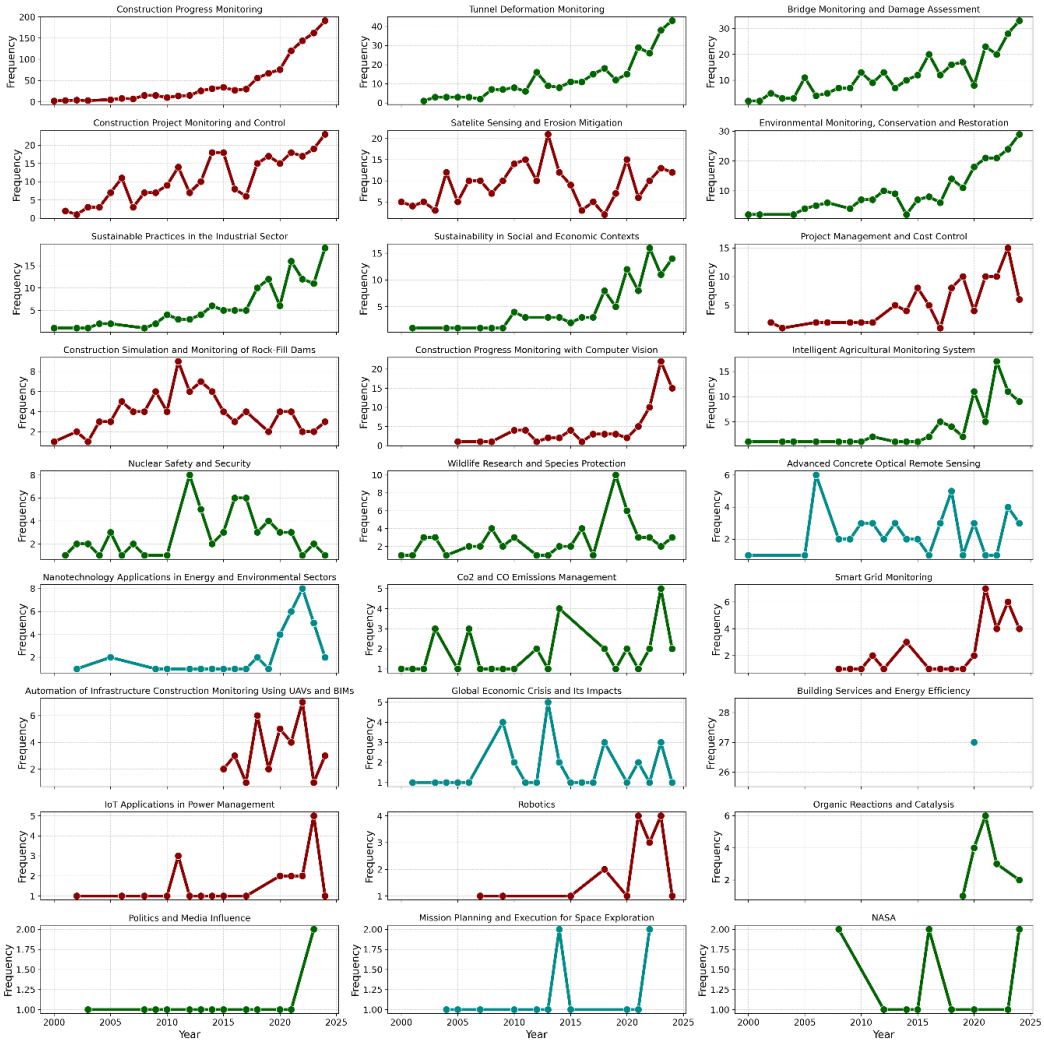


Figure 8. Temporal evolution of 27 extracted topics from the corpus [2000–2024]. Each subplot shows the annual frequency of a topic over time, with line colours representing their respective clusters identified via hierarchical clustering.

The red cluster, associated with construction monitoring and automation, demonstrates a different pattern. Many of these topics—such as "Construction Progress Monitoring", "Project Management and Cost Control", and "Construction Project Monitoring with Computer Vision"—exhibit either early sustained interest or a marked increase around post-2015. This likely reflects technological advances in AI, computer vision, and digital technologies and their impact on the construction sector. The cyan cluster, encompassing more niche or less-related topics such as "Nanotechnology Applications in Energy and Environmental Sectors", and "Global Economic Crisis and Impacts", shows sporadic or peaking interest. In addition, the "Building Services and Energy Efficiency" topic only appeared in 2020, lacking a trend. As seen, in general, the figure highlights how some research areas are experiencing consistent growth, while others remain episodic or specialized in scope.

3.4.5. Topic Interrogation and Interpretation

As detailed in Section 2.2.4., we implemented a programmatic integration of LLMs with BERTopic to facilitate interaction with the outputs of the topic model. Below, for demonstration

 doi:10.20944/preprints202603.0245.v1

purposes, we showcase the output of the topic interpretation component for the first topic (see Table 1), i.e., "Construction Progress Monitoring":

The importance of monitoring construction progress has grown significantly since 2000, driven by advances in technology and the increasing complexity of construction projects. The use of Building Information Models (BIM) has become widespread, providing a digital representation of the building process that can be easily monitored and analyzed. As a result, there has been a growing interest in developing methods for automated monitoring of construction progress, which can provide real-time data on project status and help identify potential issues before they become major problems. In addition, the increasing use of drones and other remote sensing technologies has enabled the creation of detailed 3D models of construction sites, allowing for more accurate monitoring and analysis of progress. Overall, the trend towards greater automation and digitization in construction monitoring has the potential to revolutionize the way construction projects are managed and delivered.

Looking ahead, the importance of monitoring construction progress is likely to continue to grow, driven by the increasing demand for efficient and sustainable construction practices. As construction projects become more complex and globalized, the need for accurate and reliable monitoring and analysis tools will only increase. Furthermore, the integration of artificial intelligence (AI) and machine learning (ML) technologies into construction monitoring is likely to become more prevalent, enabling even more sophisticated analysis and prediction of construction progress. For example, AI-powered predictive analytics can help forecast potential delays or issues before they occur, allowing project managers to take proactive steps to mitigate them. Overall, the trend towards greater digitization and automation in construction monitoring is likely to continue, with significant implications for the way construction projects are managed and delivered in the coming years.

This LLM-generated narrative also supports our findings depicted in Figure 8, monitoring construction progress has gained considerable momentum over the past two decades. Topics such as "Construction Progress Monitoring", "Construction Project Monitoring and Control", and "Construction Progress Monitoring with Computer Vision" all show a significant upward trend, particularly in recent years. This growth corresponds with the increasing adoption of digital tools such as Building Information Models (BIM), which facilitate real-time tracking and integration of construction data.

4. Conclusion

Using AI and large language models, we analyzed the domain of construction progress monitoring within the period of 2000 to 2024. Our methodology allowed for a robust extraction and representation of topics, revealing key thematic research clusters and insights into the evolution of the field. The integration of LLMs provided an additional layer of depth to the analysis, enabling the synthesis of complex data into coherent narratives. Our approach not only enhances the comprehension of the research landscape but also highlights the potential of AI-driven decision support tools in revealing patterns and insights within large-scale data.

Our main findings highlight three major thematic clusters (see Figure 6). These clusters align with existing literature, which emphasizes the growing importance of sustainable practices in construction [27] and the integration of advanced technologies [28]. In addition, the identification of these clusters underscores the interdisciplinary nature of construction research. By drawing connections between these topics, our study contributes to a more nuanced understanding of the interrelations within construction research.

Moreover, the study's findings reveal a dynamic evolution of topics (Figure 8), indicating shifts in research focus and emerging areas of interest, e.g., the growing influence of high-tech and digital solutions in addressing contemporary challenges within the construction industry. The surge in frequency for these topics may suggest that research and practice have aligned with industry needs for more efficient and scalable solutions. Moreover, as another example, the rise of "Automation of Infrastructure Monitoring Using UAVs and BIMs" reinforces the idea that the field is embracing novel sensing technologies, including drones and remote imaging. Looking forward, the observed

 doi:10.20944/preprints202603.0245.v1

trends point to a sustained and possibly accelerated interest in digitized and AI-enhanced construction monitoring. The recent spike in publications around computer vision applications indicates a shift toward automation, where machine learning algorithms are not only analyzing visual data but also predicting potential project delays and optimizing workflows. As construction projects scale in complexity and global reach, such tools will become indispensable. The temporal dynamics thus reflect not just past and present research priorities but also a trajectory that is likely to continue upward, especially as AI, robotics, and IoT technologies mature and integrate further into construction ecosystems. This has important implications for the future of construction management, pointing toward a more proactive, intelligent, and resilient approach to infrastructure delivery. Additionally, the importance of developing adaptable user interfaces and utilizing realistic training data is emphasized as crucial for effectively bridging the gap between research and industrial applications.

To summarize, the interdisciplinary nature of AI and computer vision highlights its versatile applications across various domains, particularly in construction, where it automates labour-intensive tasks such as progress monitoring, safety inspections, and quality control. Case studies demonstrating the practical benefits of these technologies have shown substantial returns on investment through time and cost savings, while also identifying ongoing research opportunities and challenges. Notably, there is a need for large-scale databases to train machine learning models and the integration of synthetic data to improve recognition tasks, emphasizing the continued evolution and potential of computer vision in industrial applications.

5. Limitations and Future Work

This study leveraged AI and LLMs to analyze key research topics within scientific publications, specifically focusing on the area of construction progress monitoring. Several limitations exist that warrant consideration. Although we considered multiple data sources, one limitation is the reliance on the quality and scope of the corpus, which may not fully represent the breadth of the field. Another limitation of the study is the lack of access to the complete text of the scientific publications, which restricts the depth of analysis and may result in incomplete topic extraction and interpretation. Additionally, the integration of LLMs, while enhancing the depth of analysis, may introduce biases inherent in pretrained models. For future research, expanding the corpus to include a more diverse array of documents could provide a more comprehensive view of the field. And, developing methods to mitigate biases in LLM outputs would enhance the objectivity and reliability of topic interpretations.

Author Contributions: Conceptualization, A.E.; methodology, A.E.; validation, A.E., Y.C., and F.J.; formal analysis, A.E.; investigation, A.E.; resources, A.E.; data curation, A.E.; writing—original draft preparation, A.E.; writing—review and editing, A.E., Y.C., F.J., D.M., and A.W.; visualization, A.E.; supervision, A.E.; funding acquisition, A.E., and A.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Construction Sector Digitalization and Productivity (CSDP) program of the National Research Council of Canada.

Data Availability Statement: Data can be provided upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Z. A. Memon, M. Z. A. Majid, and M. Mustaffar, "A SYSTEMATIC APPROACH FOR MONITORING AND EVALUATING THE CONSTRUCTION PROJECT PROGRESS," vol. 67, no. 3, 2006.
2. V. K. Reja, K. Varghese, and Q. P. Ha, "Computer vision-based construction progress monitoring," *Automation in Construction*, vol. 138, p. 104245, Jun. 2022, doi: 10.1016/j.autcon.2022.104245.

 doi:10.20944/preprints202603.0245.v1

3. J. Teizer, "Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 225–238, Apr. 2015, doi: 10.1016/j.aei.2015.03.006.
4. M. Pazhoohesh and C. Zhang, "Building on Our Growth Opportunities May 27 – 30, 2015 Miser sur nos opportunités de croissance REGINA, SK," 2015.
5. R. B. Z. Jawa Anak Gara, "The Development of Real-Time Integrated Dashboard: An Overview for Road Construction Work Progress Monitoring," *Journal of Hunan University Natural Sciences*, vol. 48, no. 5, Art. no. 5, Jun. 2021, Accessed: May 12, 2025. [Online]. Available: https://jonuns.com/index.php/journal/article/view/590
6. A. S. Rao *et al.*, "Real-time monitoring of construction sites: Sensors, methods, and applications," *Automation in Construction*, vol. 136, p. 104099, Apr. 2022, doi: 10.1016/j.autcon.2021.104099.
7. C. Zhang and M. Pazhoohesh, "Construction progress monitoring based on thermal-image analysis," presented at the The International Conference on Construction Management., Apr. 2017.
8. R. Sacks, C. Eastman, G. Lee, and P. Teicholz, *BIM Handbook: A Guide to Building Information Modeling for Owners, Designers, Engineers, Contractors, and Facility Managers*. Hoboken, New Jersey: Wiley, 2018.
9. D. Sarkar, D. Dhaneshwar, and P. Raval, "Automation in Monitoring of Construction Projects Through BIM-IoT-Blockchain Model," *J. Inst. Eng. India Ser. A*, vol. 104, no. 2, pp. 317–333, Jun. 2023, doi: 10.1007/s40030-023-00727-8.
10. A. H. Dalir, Z. Pezeshki, M. Ravanshadnia, E. Krinitsky, and I. A. Sultanguzin, "Automatic Monitoring in Construction Projects: Scientometric Analysis and Visualization of Research Activities," *Hum-Cent Intell Syst*, vol. 5, no. 1, pp. 21–43, Mar. 2025, doi: 10.1007/s44230-025-00089-3.
11. J. J. Lin and M. Golparvar-Fard, "Visual and virtual progress monitoring in Construction 4.0," in *Construction 4.0*, Routledge, 2020.
12. A. Ettalibi, A. Elouadi, and A. Mansour, "AI and Computer Vision-based Real-time Quality Control: A Review of Industrial Applications," *Procedia Computer Science*, vol. 231, pp. 212–220, Jan. 2024, doi: 10.1016/j.procs.2023.12.195.
13. J. J. Lin and M. Golparvar-Fard, "Construction Progress Monitoring Using Cyber-Physical Systems," in *Cyber-Physical Systems in the Built Environment*, C. J. Anumba and N. Roofigari-Esfahan, Eds., Cham: Springer International Publishing, 2020, pp. 63–87. doi: 10.1007/978-3-030-41560-0_5.
14. M. A. Musarat, A. M. Khan, W. S. Alaloul, N. Blas, and S. Ayub, "Automated monitoring innovations for efficient and safe construction practices," *Results in Engineering*, vol. 22, p. 102057, Jun. 2024, doi: 10.1016/j.rineng.2024.102057.
15. C.-Y. Cheng, P. Pourhejazy, C.-Y. Hung, and C. Yuangyai, "Smart Monitoring of Manufacturing Systems for Automated Decision-Making: A Multi-Method Framework," *Sensors*, vol. 21, no. 20, Art. no. 20, Jan. 2021, doi: 10.3390/s21206860.
16. A. Waqar, M. Houda, A. M. Khan, A. H. Qureshi, and G. Elmazi, "Sustainable leadership practices in construction: Building a resilient society," *Environmental Challenges*, vol. 14, p. 100841, Jan. 2024, doi: 10.1016/j.envc.2024.100841.
17. M. Kor, I. Yitmen, and S. Alizadehsalehi, "An investigation for integration of deep learning and digital twins towards Construction 4.0," *Smart and Sustainable Built Environment*, vol. 12, no. 3, pp. 461–487, Mar. 2022, doi: 10.1108/SASBE-08-2021-0148.
18. J. Priem, H. Piwowar, and R. Orr, "OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," arXiv.org. Accessed: May 16, 2025. [Online]. Available: https://arxiv.org/abs/2205.01833v2
19. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv.org. Accessed: Jun. 03, 2025. [Online]. Available: https://arxiv.org/abs/1810.04805v2
20. M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," *Nat Rev Methods Primers*, vol. 2, no. 1, pp. 1–21, Dec. 2022, doi: 10.1038/s43586-022-00184-w.
21. Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *In International conference on machine learning. PMLR*, 2014, pp. 1188–1196.

 doi:10.20944/preprints202603.0245.v1

22. M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv, Mar. 2022. doi: 10.48550/arXiv.2203.05794.
23. A. Ebadi, A. Auger, and Y. Gauthier, "WISDOM: An AI-powered framework for emerging research detection using weak signal analysis and advanced topic modeling," arXiv.org. Accessed: Jun. 19, 2025. [Online]. Available: https://arxiv.org/abs/2409.15340v1
24. A. Ebadi, P. Xi, S. Tremblay, B. Spencer, R. Pall, and A. Wong, "Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing," *Scientometrics*, vol. 126, no. 1, pp. 725–739, Jan. 2021, doi: 10.1007/s11192-020-03744-7.
25. R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Frontiers in Sociology*, vol. 7, 2022, Accessed: Sep. 26, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fsoc.2022.886498
26. H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv.org. Accessed: Jun. 19, 2025. [Online]. Available: https://arxiv.org/abs/2307.09288v2
27. T. D. Moshood, J. O. Rotimi, and W. Shahzad, "Enhancing sustainability considerations in construction industry projects," *Environ Dev Sustain*, Apr. 2024, doi: 10.1007/s10668-024-04946-2.
28. L. Zhang, Y. Li, Y. Pan, and L. Ding, "Advanced informatic technologies for intelligent construction: A review," *Engineering Applications of Artificial Intelligence*, vol. 137, p. 109104, Nov. 2024, doi: 10.1016/j.engappai.2024.109104.