Enhancing data accessibility in built asset digital twins with neural language models and immersive technologies

by

Mehrzad SHAHINMOGHADAM

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, DECEMBER 24, 2024

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Ali Motamedi, Thesis supervisor
Department of Construction Engineering, École de Technologie Supérieure

Mr. Rafael Menelau Cruz, Chair, Board of Examiners
Department of Software Engineering and IT, École de Technologie Supérieure

Mr. Erik Poirier, Member of the Jury
Department of Construction Engineering, École de Technologie Supérieure

Mrs. Rachel Pottinger, External Independent Examiner
Department of Computer Science, University of British Columbia

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON "DECEMBER 10, 2024"

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

As Pascal once said, "The last thing you figure out in writing [a book] is what to put first." So I begin with what matters most to me: expressing my gratitude.

I would like to express my gratitude to my supervisor, Dr. Ali Motamedi, whose approach cultivated both independence and confidence throughout this journey. I am sincerely appreciative of your support throughout my PhD journey. I would like to also extend my gratitude to my thesis jury committee: Professor Rachel Pottinger, Dr. Erik Poirier, and Dr. Rafael Menelau Cruz, whose insightful comments added clarity and depth to this work.

I would like to thank Dr. Worawan Natephra, Dr. Erik Poirier, Dr. Mohammad Mostafa Soltani, Dr. Mohamed Cheriet, Dr. Samira Ebrahimi Kahou, and Mr. Saman Davari, co-authors in my PhD research publications. I am particularly grateful to Dr. Cheriet and Dr. Ebrahimi Kahou for offering their expertise and mentorship in the field of Artificial Intelligence. Your guidance gave me the confidence and clarity to make meaningful contributions in my interdisciplinary research.

I would like to thank Mr. Phil Simpson, Mr. Jon Proctor, Mr. Richard Kelly, and Dr. Erik Poirier, whose support was instrumental in securing permission from National Building Specification (NBS, UK) and buildingSMART International to publicly redistribute the datasets generated throughout this research.

I would like to acknowledge Beslogic Inc. and MITACS for providing an internship opportunity. My sincere thanks go to Yannick Bessette and Cristian Liciu for their trust and openness in sharing the outcomes of our research. Special thanks to my friends, Hamed and Atena, for introducing me to this internship opportunity.

To my friends and my two brothers: despite the geographical distance, your presence in my thoughts pushed me forward. Asad, my brother, I was fortunate to have you nearby during these years, especially during the COVID-19 pandemic. Sarah, I do not know if I would be writing these final lines without your unwavering support and patience over the last three years. To my

mother and father—my greatest source of strength, inspiration, and resilience—, you have taught me the most valuable lessons in the school of life. I am grateful for your love and guidance.

Améliorer l'accessibilité des données dans les jumeaux numériques de biens construits grâce à des modèles de langage neuronaux et à des technologies immersives

Mehrzad SHAHINMOGHADAM

RÉSUMÉ

La transformation numérique rapide dans les secteurs de l'architecture, de l'ingénierie, de la construction et de la gestion des installations a suscité un intérêt croissant pour le jumelage numérique des actifs bâtis - créant des représentations virtuelles basées sur les données qui reflètent les états et les comportements des composants et systèmes d'actifs construits du monde réel. Pour actualiser le plein potentiel des jumeaux numériques - en tant que représentations dynamiques et axées sur les données des actifs physiques - il est essentiel d'intégrer des flux de données en temps réel à des informations statiques. Dans cette optique, les systèmes de modélisation des informations du bâtiment (BIM) et de l'internet des objets (IoT) peuvent servir de composants fondamentaux des jumeaux numériques des actifs bâtis, en combinant les données géométriques statiques et les données sémantiques riches de la BIM avec les données des capteurs IoT en temps réel. Cependant, malgré des recherches approfondies et des avancées dans les cadres d'interopérabilité et d'échange de données, en particulier ceux centrés sur le paradigme BIM ouvert, la complexité de l'accès et de l'interprétation des données sur les actifs bâtis reste un obstacle important aux interactions intuitives des utilisateurs avec les interfaces de jumeaux numériques.

Cette thèse vise à aborder la question critique de l'accessibilité des données dans les interfaces de jumeaux numériques en étudiant les potentiels pratiques de deux approches synergiques : la réalité virtuelle (RV) pour améliorer la représentation des données et les modèles de langage neuronaux pour améliorer la recherche d'informations. En particulier, la thèse examine le potentiel des environnements immersifs de RV pour représenter intuitivement les données complexes des actifs construits, offrant aux utilisateurs un moyen plus interactif et spatialement cohérent de naviguer et d'interpréter les informations multimodales provenant de diverses sources. Grâce à une étude de cas axée sur la surveillance du confort thermique en temps réel, la recherche démontre comment les interfaces basées sur la RV peuvent fournir un support efficace pour une représentation transparente des données BIM et IoT, ce qui peut améliorer de manière significative le processus de prise de décision basé sur les données pour les praticiens dans des scénarios complexes du monde réel.

En parallèle, la thèse étudie l'application de modèles de langage neuronaux, y compris les réseaux neuronaux profonds formés sur mesure et les grands modèles de langage (LLM) préformés, pour améliorer la recherche d'informations dans les systèmes de jumeaux numériques centrés sur le BIM. Motivé par le potentiel prometteur des techniques avancées de traitement du langage naturel pour améliorer les systèmes de recherche en incorporant la sémantique, ce travail évalue l'efficacité des techniques de modélisation du langage neuronal de pointe pour la tâche spécifique d'extraction d'entités à partir de requêtes d'utilisateurs. Les résultats expérimentaux, qui comparent les architectures d'apprentissage profond traditionnelles et émergentes, offrent de

nouvelles perspectives pour la recherche et la pratique. Ils soulignent l'importance critique des stratégies d'adaptation au domaine pour une performance efficace du modèle dans ce contexte spécialisé.

En outre, pour répondre aux lacunes des repères généraux existants dans l'examen de la transférabilité des capacités des LLMs pré-entraînés pour les tâches en aval, telles que celles liées à la recherche d'informations, la thèse présente un repère complet des modèles de pointe. La thèse est accompagnée de la publication des ressources de référence proposées, y compris des ensembles de données à grande échelle et de haute qualité provenant de sources reconnues par l'industrie. Compte tenu de l'intérêt croissant pour l'application des LLM à diverses applications dans la recherche sur l'environnement bâti, il s'agit d'une contribution particulièrement opportune et essentielle dans ce domaine qui évolue rapidement. La thèse conclut en soulignant les principales limites et les défis rencontrés, accompagnés de recommandations pour la recherche future avec le thème central de favoriser des interactions plus intuitives entre les utilisateurs et les jumeaux numériques du patrimoine bâti.

Mots-clés: Jumeau numérique, modélisation des informations de construction, Internet des objets, réalité virtuelle, traitement du langage naturel, recherche d'informations, reconnaissance d'entités nommées, apprentissage profond, intégration de texte, modélisation neuronale du langage, modèles de langage de grande taille, banc d'essai

Enhancing data accessibility in built asset digital twins with neural language models and immersive technologies

Mehrzad SHAHINMOGHADAM

ABSTRACT

The rapid digital transformation within the architecture, engineering, construction, and facilities management industries has led to a growing interest in digital twinning for built assets—creating virtual, data-driven representations that mirror the states and and behaviors of real-world built asset components and systems. To actualize the full potential of digital twins—as dynamic, data-driven representations of physical assets—it is essential to integrate real-time data streams with static information. In this light, Building Information Modeling (BIM) and Internet of Things (IoT) systems can serve as foundational components of built asset digital twins—combining BIM's static geometrical and rich semantic data with live IoT sensor data. However, despite extensive research and advancements in interoperability and data exchange frameworks, particularly those centered around the open BIM paradigm, the complexity of accessing and interpreting built asset data remains a significant barrier to intuitive user interactions with digital twin interfaces.

This thesis sets out to address the critical issue of data accessibility in digital twin interfaces by investigating the practical potentials of two synergistic approaches: virtual reality (VR) for enhancing data representation and neural language models for improving information retrieval. In particular, the thesis examines the potential of immersive VR environments to intuitively represent the complex built asset data, offering users a more interactive and spatially coherent means of navigating and interpreting multimodal information that comes from various sources. Through a case study focused on real-time thermal comfort monitoring, the research demonstrates how VR-based interfaces can provide an effective medium for seamless representation of BIM and IoT data, which can significantly improve the data-driven decision-making process for practitioners in complex real-world scenarios.

In parallel, the thesis investigates the application of neural language models, including custom-trained deep neural networks and pre-trained large language models (LLMs), for enhancing information retrieval in BIM-centered digital twin systems. Motivated by the promising potential of advanced natural language processing techniques for improving search systems by incorporating semantics, this work evaluates the effectiveness of state-of-the-art neural language modeling techniques for the specific task of entity extraction from user queries. The experimental results, comparing traditional and emerging deep learning architectures, provide novel insights for both research and practice. They underscore the critical importance of domain adaptation strategies for effective model performance in this specialized context.

Moreover, to address the shortcomings of existing general-purpose benchmarks in examining the transferability of the pre-trained LLMs' capabilities for downstream tasks, such as those related to information retrieval, the thesis presents a comprehensive benchmark of state-of-the-art models. The thesis is accompanied by the public release of the proposed benchmark resources,

including large-scale, high-quality datasets curated from industry-renowned sources. Given the surge of interest in applying LLMs to various applications within the built environment research, this is a particularly timely and essential contribution in this rapidly evolving field. The thesis concludes by highlighting key limitations and challenges encountered, accompanied by recommendations for future research with the central theme of fostering more intuitive user interactions with built asset digital twins.

Keywords: Digital twin, Building information modeling, Internet of things, Virtual reality, Natural language processing, Information retrieval, Named entity recognition, Deep learning, Text embedding, Neural language modeling, Large language models, Benchmark

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AEC-FM	Architecture, Engineering, Construction, and Facilities Management
BERT	Bidirectional Encoder Representations from Transformers
BIM	Building Information Modeling
bsDD	buildingSmart Data Dictionary
DSR	Design Science Research
GPT	Generative Pre-trained Transformer
HMD	Head Mounted Display
HTTP	Hypertext Transfer Protocol
IFC	Industry Foundation Classes
IoT	Internet of Things
LLMs	Large Language Models
LSTM	Long Short-Term Memory
MRT	Mean Radiant Temperature
NER	Named Entity Recognition
NLP	Natural Language Processing
PMV	Predicted Mean Vote
PPD	Predicted Percentage of Dissatisfied
RGB	Red, Green, Blue; Pixel intensity
RNNs	Recurrent Neural Networks
VR	Virtual Reality

INTRODUCTION

0.1 Context

The accelerating pace of digital transformation within the Architecture, Engineering, Construction, and Facilities Management (AEC-FM) industries has significantly driven the growing interest in the digital twinning of built assets, be it for buildings or infrastructure. Digital twins are envisioned as dynamic, data-driven representations that reflect the past, present, and projected future states and behaviors of physical assets—e.g., a digital twin of a building's HVAC system may integrate real-time sensor data, historical energy usage patterns, and a 3D digital model to predict energy consumption and optimize maintenance schedules (Motamedi & Shahinmoghadam, 2021). Achieving this vision necessitates the continuous collection, processing, and integration of a wide range of data, from structural and environmental conditions to operational performance. To provide such a comprehensive and holistic view, digital twins require seamless access to vast amounts of diverse and complex data generated across different stages and systems within the built asset lifecycle. In this light, Building Information Modeling (BIM) and similar technological paradigms, such as common data environments, are relevant to digital twinning as they share a common goal: to serve as a single source of truth, supporting various stakeholders throughout the entire lifecycle of built assets—from initial design to end-of-life (Davari, Shahinmoghadam, Motamedi & Poirier, 2022).

Given the need for providing a comprehensive view of built asset information, integrating BIM with the Internet of Things (IoT) systems offers a synergistic potential for digital twinning by incorporating rich semantic representations provided within BIM models with real-time data streams captured by IoT devices. This combination creates dynamic digital representations of built assets that reflect not only the static characteristics of assets (e.g., geometry, material type)

but also their live conditions and operational states and behaviors (Tang, Shelden, Eastman, Pishdad-Bozorgi & Gao, 2019; Shahinmoghadam & Motamedi, 2019).

The rising demand for data-rich built asset digital models has motivated extensive research and development efforts aimed at enhancing interoperability across numerous systems and data formats (Moretti, Xie, Merino Garcia, Chang & Kumar Parlikad, 2023). From leveraging the continuous development of open international data standards such as Industry Foundation Classes (IFC) (buildingSMART, 2020) to ontology-based mediation strategies for domain-specific information (Shahinmoghadam & Motamedi, 2021), the body of knowledge is rich with methods and techniques for improving data exchange at both the syntactic and semantic levels. While significant attention has been devoted to improving interoperability, the complexity of data accessibility from a user perspective remains largely underexplored. As interoperability advances and as more sources of information are programmatically accessible within the backend of digital twin interfaces, the question remains: How well are we addressing the challenge of making vast and complex integrated datasets accessible in a manner that is intuitive and consumable by end-users?

A central argument of this thesis is that regardless of how advanced digital twins become, the role of domain experts will remain indispensable. As we design and refine the digital twin technology stack for the built environment, it is essential that data access not only be programmatically available but also intuitive and tailored to the needs of human users. Without proper retrieval and representation, the inherent complexity and interconnectedness of built asset data—coupled with the diversity of data sources and formats—can hinder the seamless flow of information, preventing digital twins from fully realizing their potential in real-world scenarios.

Given that data accessibility is vital to enabling experts to make timely and informed decisions (Horvitz & Mitchell, 2020), this thesis aims at improving data accessibility in digital twin

interfaces from two critical perspectives: how users request the necessary data (retrieval), and how the integrated data—static and real-time data—is represented to them (representation).

The extensive volume and complexity of data in built asset digital twins—ranging from 3D geometric representations to real-time sensor data—urges for enhanced mechanisms for effectively navigating the data and extracting actionable insights. It is in this regard that immersive technologies, such as virtual reality, can play a transformative role in enhancing the intuitiveness of data representation in digital twin interfaces (Pirker, Loria, Safikhani, Künz & Rosmann, 2022; Stadtmann, Mahalingam & Rasheed, 2023). By immersing users in interactive, spatially accurate digital environments, a virtual reality interface enables them to visualize both static BIM data and dynamic IoT feeds, making the data more accessible and easier to interpret. Moreover, advancements in Natural Language Processing (NLP) present significant opportunities for improving the retrieval of information from complex, large-scale built asset datasets. The emergence and rapid evolution of neural language models, Large Language Models (LLMs) in particular, has revolutionized the field of information retrieval by extending the context-aware language understanding capability of these models. While advanced neural language modeling techniques have shown promising results in enhancing data retrieval in various fields—from medical to food (see Section 4.1)—further research is needed to adapt them for effective use in the context of complex built asset data.

In light of the preceding context, the central focus of this thesis is framed by two primary approaches: virtual reality and neural language modeling. Given their synergistic potential, these two approaches offer promising solutions to bridge the gap between complex data systems and user interfaces in the context of built asset digital twinning.

0.2 Problem statement

The primary aim of this thesis is to investigate how virtual reality and neural language models can be effectively applied to enhance user interaction with built asset digital twin systems by developing more intuitive mechanisms for data navigation and retrieval. To further delineate the research problem and the central theme of this thesis, the following observations are of particular relevance to the scope of this thesis.

First, virtual reality presents a promising avenue for creating more intuitive and immersive interfaces for digital twins. By integrating static and real-time data within a virtual environment, users can experience a more natural and engaging interaction with the digital twin, facilitating enhanced navigation and interpretation of data. Despite this potential, research on the application of virtual reality within the context of digital twins for built assets remains limited, particularly in relation to the implementation of case studies that effectively reflect the complexity of real-world scenarios (Pirker *et al.*, 2022; Shahzad, Shafiq, Douglas & Kassem, 2022). Second, the rapid advancement of deep learning-based NLP techniques has introduced a complex landscape of design choices for adapting the capabilities of language models to specific domains or applications (Muennighoff, Tazi, Magne & Reimers, 2022; Ling *et al.*, 2023). As a result of such rapid advancement and the novelty of these techniques, existing studies investigating the effectiveness of language models within the built environment research remain limited in scope, often focusing on ad hoc tasks using small datasets (Wang, Issa & Anumba, 2022d; Zhang & El-Gohary, 2023; Forth, Berggold & Borrmann, 2024), which can introduce biases in assessing the models' performance for domain-specific language understanding. Third, the scarcity of publicly available datasets for a robust evaluation of language models in the context of built environment research raises significant concerns regarding transparency and reproducibility, thereby impeding meaningful comparisons between different approaches (Shahinmoghadam, Kahou & Motamedi, 2024). These observations highlight a critical knowledge gap concerning the

effective utilization of neural language models to achieve a contextually accurate understanding of specialized terminology within the domain of built asset information management.

Based on the highlighted gaps (discussed further in Chapter 1) and the scope of the research, this thesis addresses the following specific research questions:

- **RQ1:** What is the potential of virtual reality in improving the navigation and integration of BIM and IoT data for real-time monitoring and assessment in complex scenarios?
- **RQ2:** How can neural language modeling techniques be effectively applied to enable intuitive information retrieval in virtual environments (digital twin interfaces)?
- **RQ3:** How effective are state-of-the-art pre-trained language models in capturing and representing semantics specific to built asset terminology?

By addressing these research questions, this thesis offers key insights toward narrowing the gap between theoretical potentials and practical implications of the investigated systems, supported by empirical evidence presented in Chapters 2-4.

0.3 Research methodology

To address the research questions outlined in the previous section, the Design Science Research (DSR) methodology, a widely used approach in information systems research (Offermann, Levina, Schönherr & Bub, 2009), was adopted as the overarching methodological framework. In particular, this research follows the six-step methodology proposed by Peffers, Tuunanen, Rothenberger & Chatterjee (2007), which provides a structured and iterative process to propose solutions for the identified problems. The six steps of the adopted methodology, as applied in this thesis, can be summarized as follows:

- **Problem identification and motivation:** The research begins by identifying the challenges associated with improving data accessibility in built asset digital twin interfaces. An overview

of the problem space was presented in the previous sections. The underlying motivations of the thesis are further supported in Chapter 1.

- **Objective definition:** Based on the identified research questions, the specific objectives of the research are formulated as follows: (1) examine the practical utility of integrating BIM and IoT data within VR environments to support complex, real-world applications where access to both static spatial and dynamic sensor data plays a critical role, (2) examine the effectiveness of state-of-the-art language modeling techniques in facilitating natural language-based information retrieval from built asset databases, and (3) benchmark the comparative performance of state-of-the-art pre-trained language models in supporting information retrieval from built asset databases.
- **Artifact design and development:** At the core of the DSR methodology lies the design and development of artifacts, as the most critical practice rule in this methodology (Peffers *et al.*, 2007). This thesis presents several artifacts developed to address the research questions, including a VR prototype integrating live sensor data and static spatial data for (near) real-time monitoring purposes, a semantic tagging framework for natural language-based search, and a benchmarking framework to evaluate language models' performance in understanding and retrieving information relevant to the descriptions of built assets.
- **Demonstration:** Following the development of the artifacts, their practical utility in addressing the corresponding research objectives is demonstrated through targeted activities. These include case studies (see Sections 2.5.1 and 2.5.2) and experimentations (see Sections 3.4 and 4.3) to highlight the practical relevance and contribution of the developed artifacts.
- **Evaluation:** This thesis employs quantitative methods to assess the effectiveness of the developed artifacts in achieving the research objectives. The evaluation methods, metrics, and relevant analyses are described in detail in Sections 2.5, 3.4, and 4.4.
- **Communication:** The knowledge generated from this research—including the proposed methods, developed artifacts, evaluation results, and discussions on the research's importance,

novelty, limitations, and future recommendations—is communicated through this thesis and a series of scholarly publications (detailed in the next section). Additionally, the majority of the datasets and codebase produced as part of this work have been made available in public repositories, ensuring transparency, reproducibility, and accessibility of the results.

0.4 Contributions and thesis structure

This thesis sets out to address the overarching goal of enhancing the accessibility of built asset information in digital twin interfaces so that domain experts can more intuitively navigate and retrieve both static and dynamic data. The approach presented in this thesis entails two interrelated primary endeavors: first, developing intuitive and immersive visualization environments that fuse multiple data sources, and second, investigating an intuitive approach to retrieving built asset data through natural language queries. This section highlights the contributions of the work to facilitate a clear understanding of how each chapter contributes to the main goal of the thesis.

A visual summary of the thesis contributions is presented in Figure 0.1. The key contributions of this research (denoted in the figure by solid boxes) are presented across three journal articles, which constitute the core chapters of this thesis—Chapters 2, 3, and 4. The secondary contributions (denoted in the figure by dotted boxes), although not included in the body of the thesis, are recommended as supplementary reading to provide additional context and insights into the research motivations for the corresponding chapters.

In Chapter 1, a comprehensive review of related works is conducted to highlight the knowledge gaps and provide the necessary background and theoretical framework for addressing the key research questions of the thesis. The chapter begins by reviewing state-of-the-art research at the intersection of digital twinning and built asset management, with a particular focus on BIM, IoT, and virtual reality. Subsequently, the chapter presents the rapid advancements in NLP and neural

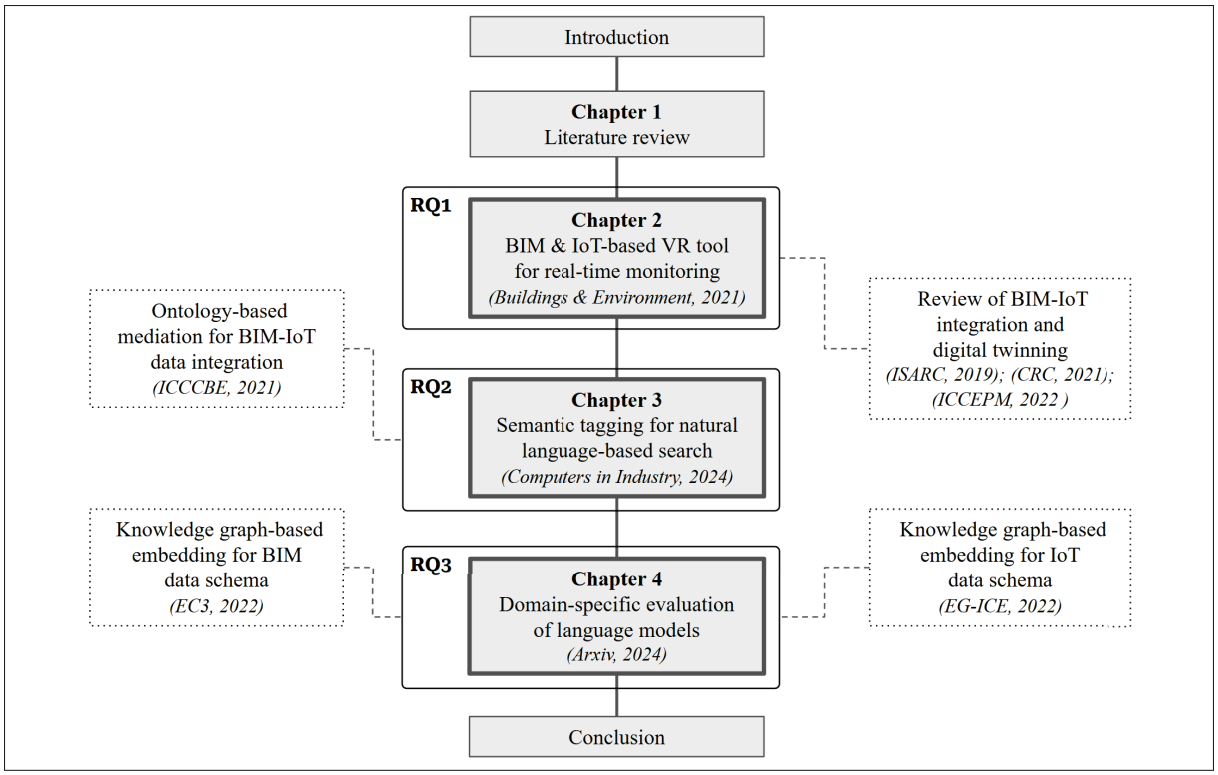


Figure 0.1 Visual summary of the thesis organization and contributions. Boxes with solid borders indicate the main contributions of the thesis (presented in Chapters 2-4 of the thesis). Secondary contributions made with respect to each research question are indicated by dashed boxes (not included in the thesis but suggested as supplementary readings)

language modeling techniques, particularly Large Language Models (LLMs), and their relation to the topic of information retrieval. The chapter identifies several key gaps that motivate the focus of the research presented in the following chapters. A detailed overview of the specific contributions is outlined below.

Main contributions

Chapter 2 tackles the representation component of data accessibility by presenting an immersive virtual reality system, integrated with BIM and live IoT data for real-time thermal comfort assessment. The rationale for selecting thermal comfort assessment as a case study lies in its

inherent complexity, which necessitates the integration of both spatial and live sensor data, thereby rendering it a representative scenario for a broad range of real-world applications. The developed system assists users in visualizing and interacting with dynamic environmental conditions (e.g., levels of indoor temperature and humidity), thereby enhancing data accessibility and supporting more informed decision-making processes. An additional contribution of the work is the development of a cost-effective IoT-enabled thermal imaging prototype using affordable sensors, making the proposed system practical for broader adoption in non-intrusive real-time building monitoring scenarios. Furthermore, the developed interface demonstrated the potential of performing real-time adjustments of monitoring variables, thereby facilitating the possibility of conducting remote what-if analyses for practitioners. These contributions collectively highlight how the integration of BIM and IoT within virtual reality interfaces can effectively address the challenge of making complex datasets accessible and actionable for the real-time monitoring of built assets.

Chapter 3 addresses the retrieval aspect of data accessibility by focusing on how advanced NLP techniques can be adapted to accurately parse and interpret users' natural language queries for built asset data. The work focuses on Named Entity Recognition (NER), which is a key task of semantic parsing in enabling natural language-based search. The study introduces a semantic annotation scheme rooted in IFC schema— an open, internationally-recognized built asset data classification standard—and compares traditional and more recent advanced deep learning architectures. The findings underscore the importance of domain-specific embedding learning and provide practical insights into the trade-offs between computational performance and prediction accuracy. The collective findings of this work contribute to the body of knowledge by advancing the understanding of how neural language models can be applied to facilitate intuitive information retrieval from complex, large-scale built asset datasets. Finally, by openly releasing data and code, the research promotes future advancements and collaborations, advocating for the development of larger, more diverse datasets.

The discussions presented in Chapter 3 highlight the significant influence of domain-specific text understanding by language models on the accuracy of user query interpretation. This is while the review of the relevant literature (see Section 1.3.2) reveals a critical research gap in the availability of robust datasets that can support a realistic evaluation of models' understanding of domain-specific terminology. These findings, discussed in detail in the concluding chapter of the thesis, served as a key motivation for the work presented in Chapter 4.

Chapter 4 extends the focus on leveraging language models to enable a more holistic evaluation of pre-trained language models—examining how well they capture the semantics of built asset–related textual input for tasks such as information retrieval. The work presents a comprehensive benchmark study of state-of-the-art text embedding models is conducted, with a particular focus on evaluating their effectiveness in capturing and representing the semantics of textual descriptions specific to built assets. The study examines 24 language models on proposed datasets comprising over ten thousand entries across three task categories: clustering, retrieval, and reranking. The results underscore the limitations of general-purpose benchmarks in adequately reflecting the unique semantic complexities and contextual dependencies inherent in texts related to built assets. Moreover, the analysis provides critical insights into the transferability of the evaluated models to this specialized domain, with particular attention to factors such as model size and pre-training strategies. In addition to contributing to the body of knowledge, another key contribution of this work lies in the development of public datasets, which were carefully curated from mature, reliable sources to ensure the quality and relevance of the data, facilitating broader research use and replication. By aligning the design of the proposed benchmark with widely recognized ones in the NLP research community and publicly releasing the developed datasets and software, this work facilitates robust evaluation practices for language models within built asset management, providing a foundation for future research and development in the field.

Secondary contributions

In addition to the main contributions outlined above, this research has produced several secondary scholarly contributions (see Figure 0.1). The results of an ongoing review of the state of research on digital twinning for built assets, with a focus on the role of enabling technologies such as BIM and IoT, have been presented at multiple academic conferences (Shahinmoghadam & Motamedi, 2019; Motamedi & Shahinmoghadam, 2021; Davari *et al.*, 2022).

Another notable contribution involved exploring the potential and challenges of using mediatory ontologies as global schemas to facilitate the integration of BIM and IoT databases (Shahinmoghadam & Motamedi, 2021). The findings from this study directly motivated the research presented in Chapters 3 and 4, by highlighting the inherent complexities of using expressive formal query languages such as SPARQL (Pérez, Arenas & Gutierrez, 2009), and the practical challenges of creating ontology-based global schemas (see Section 1.2.2).

Given the critical relevance of representation learning techniques—known as semantic embedding (see Section 1.3.2)—to this research, a parallel line of investigation explored knowledge graph embedding techniques for generating numeric, vectorized representations of built asset data. The corresponding results, presented in two conferences (Shahinmoghadam, Motamedi & Soltani, 2022a,b), relate directly to RQ3, and indirectly to RQ2, by demonstrating the potential of knowledge graph embeddings to facilitate various downstream tasks (including information retrieval) when using graph-based representations of built asset data. These two studies offer insights into the challenges of mapping the embeddings generated within disjoint embedding spaces, which influenced the refinement of the objectives of the research presented in Chapter 4.

CHAPTER 1

LITERATURE REVIEW

1.1 Harnessing the synergistic potentials of BIM, IoT, and VR

1.1.1 The need for complex case studies

Given their significant potential to improve various aspects of built asset design, construction, and operations management, digital twins have been receiving increasing popularity in both academia and industry (Rafsanjani & Nabizadeh, 2023; Poirier & Motamedi, 2024). Recent research suggests that the incorporation of both static and dynamic information plays an indispensable role in the effective digital twinning of built assets (Korotkova, Benders, Mikalef & Cameron, 2023). Accordingly, numerous studies have highlighted the growing interest in BIM-IoT data integration, suggesting that this combination can serve as a foundational element for built asset digital twins (Tang *et al.*, 2019; Shahinmoghadam & Motamedi, 2019; Eneyew, Capretz & Bitsuamlak, 2022). This combination leverages BIM's detailed spatial and semantic digital representations and IoT's real-time data collection capabilities (Shahinmoghadam & Motamedi, 2019).

In parallel, researchers have been investigating the potential of VR to leverage BIM's rich visualizations and real-time IoT data to create more immersive and interactive experiences for facility management (Natephra & Motamedi, 2019a,b; Baghalzadeh Shishehgarkhaneh, Keivani, Moehler, Jelodari & Roshdi Laleh, 2022). These integrations allow decision-makers to interact with digital twins of facility components and assets, offering opportunities for simulating alternative scenarios, real-time engagement, and the exchange of insights in an immersive, virtual environment, contributing to improved monitoring and planning in facility operations and maintenance (Casini, 2022).

While the synergistic potential of BIM, IoT, and VR in the digital twinning of built assets is evident, there is a notable knowledge gap in the existing body of research when it comes to practical, real-world applications of digital twins in the context of built assets. This gap can be attributed to

the relative novelty and technical complexities associated with the digital twin concept. Although the concept of BIM has benefited from extensive research and standardization efforts over the past three decades, digital twins remain a nascent area of study. Much of the progress in digital twins has been driven by industry practitioners and real-world implementations, with relatively fewer scholarly studies available to provide in-depth, empirical insights (Poirier & Motamedi, 2024).

A review of the existing literature reveals a large number of studies focusing on the definitions, theoretical frameworks, and potential enabling technologies of digital twins in built environment research (Davari *et al.*, 2022; Shahzad *et al.*, 2022; Pirker *et al.*, 2022; Baghalzadeh Shishehgarkhaneh *et al.*, 2022; Tuhaise, Tah & Abanda, 2023; Rafsanjani & Nabizadeh, 2023; Correia, Abel & Becker, 2023; Abdelrahman *et al.*, 2024). However, there remains the need for more studies that can adequately reflect the challenges and complexities of real-world scenarios, particularly where an orchestration of BIM, IoT, and VR is central to the implementation of digital twins. In this light, recent research has called for a more focused investigation into the applicability of digital twin systems that integrate BIM and real-time sensor data in extended reality environments for indoor monitoring of facilities (Wang, Gan, Hu & Liu, 2022e; Casini, 2022; Tuhaise *et al.*, 2023; Hu & Assaad, 2024).

The above-mentioned gap in the literature highlights the need for more case studies that capture the full complexity of integrating BIM and IoT in built asset management and representing the static and dynamic data within immersive and interactive virtual environments. It is within this context that the present thesis selects thermal comfort as the main theme for the case study presented in Chapter 2. Indoor thermal comfort assessment can present an effectively challenging use case because it requires detailed spatial and dynamic data, offering a scenario where the representation of BIM and IoT data plays a critical role in providing actionable insights to monitor and optimize indoor environmental conditions.

1.1.2 The need for human-centric interfaces

Developing user-friendly interfaces that support efficient data retrieval is crucial for bridging the gap between the technical capabilities of digital twins and the practical needs of end users. This includes the design of systems that can filter, prioritize, and present data in ways that are meaningful and actionable for users (Dingli & Haddod, 2019; Lee, Lee, Aucremanne, Shah & Ghahramani, 2023). However, a review of the existing literature on built asset digital twins reveals that the crucial role of human expertise in interpreting and utilizing the vast volume of built asset data, particularly in the integration of BIM data with real-time sensor observations, is often overlooked (Rafsanjani & Nabizadeh, 2023). A recent systematic review on urban digital twins aligns with the latter observation, revealing that while interoperability has been extensively discussed in the literature, the complexity of data and its impact on user interaction has received comparatively less attention (Lei, Janssen, Stoter & Biljecki, 2023).

This oversight is significant, as the inherent complexity and heterogeneity of data describing various aspects of built environment entities—including but not limited to structural details, environmental conditions, and operational performance—can pose considerable challenges for end users trying to access the data for various purposes. The extensive volume of data, combined with its diverse formats and origins, can create barriers to efficient data retrieval and usability in digital twin applications (Lei *et al.*, 2023). A recent review focused particularly on the topic of data management in digital twin research reveals that data search and discovery remains one of the least explored topics in this field (Correia *et al.*, 2023).

These studies underscore a critical gap: while a growing substantial volume of efforts has been devoted to investigating and developing the technical frameworks for data integration scenarios that support digital twin systems, far less attention has been paid to how end users—such as facility managers, engineers, occupants, and other decision-makers—can effectively locate and retrieve the relevant data within these systems. The subsequent section presents a detailed analysis aimed at narrowing down this gap in relation to the specific research questions posed in this thesis.

1.2 Interoperability and complexity of data in built asset digital twins

1.2.1 On the proliferation of local schemas

Data interoperability among information systems within the context of AEC-FM industries has been a widely investigated topic, with numerous studies addressing this issue at both syntactic and semantic levels. On the syntactic level, where the primary focus is on the technical aspects of linking systems and services (Panetto, 2007), the body of knowledge is rich with the examination of various data formats and exchange protocols. Application of open data representation and exchange standards falls in this category. In particular, open BIM tools and standards can play a crucial role in the ecosystem of digital twins as they can significantly facilitate structured interoperable data representations (buildingSmart International, 2024b). In this light, the IFC schema, with its recent formalization as an ISO standard (ISO 16739-1, 2024), plays a central role throughout this thesis. Detailed discussions on the roles of open BIM tools and standards in this thesis are provided in Section 3.2 and Section 4.1.

At the semantic level, where the meaning of the data being exchanged is of particular interest (Panetto, 2007), researchers have been exploring the potential of formal ontologies and semantic web technologies to tackle interoperability challenges across various AEC-FM disciplines (Farghaly, Soman & Zhou, 2023). Due to their promising capabilities, ontologies have been a focal point of research for over two decades (Farghaly *et al.*, 2023), maintaining their relevance and evolving under the broader framework of knowledge graphs, which have emerged as a recent trend in this area (Deng, Xu, Deng & Lin, 2022). Recent research suggests that the promising potential of ontology-based semantic modeling, or knowledge graphs in a generic sense, for integrating heterogeneous data models across various domains and lifecycle phases makes them highly relevant to digital twin research (Li, Rui, Zhu, Lu & Li, 2024).

As the research community continues to advocate for further investigation and development of ontologies or other forms of linked data representations (Tuhaise *et al.*, 2023), the proliferation of domain-specific data models raises concerns about the practical usefulness of these approaches.

The latter issue has been highlighted in various engineering applications (Hagedorn, Smith, Krishnamurty & Grosse, 2019), including IoT-enabled building monitoring (Esnaola-Gonzalez, Bermúdez, Fernandez & Arnaiz, 2020), where the application of ontologies remains limited to single-use, dataset-specific purposes and functions more as local schemas without broader interoperable semantics (Haller & Polleres, 2020).

1.2.2 On the complexity of leveraging global schemas

In response to the proliferation of ontologies as local schemas, researchers propose the use of global (top-level) ontologies, known as ontology-based mediation, to broaden the applicability of local schemas (Hagedorn *et al.*, 2019). Relevant to digital twins, it has been previously shown that a top-level ontology that provides a global view of local BIM and IoT data sources can facilitate programmatic access to data produced and stored in independent systems (Shahinmoghadam & Motamedi, 2021). Such approaches are relevant to the field of ontology alignment, which deals with identifying correspondences between different ontologies to enable interoperability and data sharing across various information systems (Trojahn, Vieira, Schmidt, Pease & Guizzardi, 2022).

However, there remains a lack of attention to the alignment of data models in built environment research (Farghaly *et al.*, 2023). From a technical perspective, the need for establishing entity correspondences between AEC-FM data models goes beyond domain-specific ontologies and reaches a wide spectrum, from generic data classification taxonomies such as IFC, to ad hoc project/firm-specific database schemas. The importance of such alignments can be highlighted from at least two key aspects.

First, it enables programmatic access to various database systems containing both static and dynamic data related to built assets (a key aspect of built asset digital twin functionality (Korotkova *et al.*, 2023)). However, in practice, such alignments are often established through manual interventions. This approach is seemingly neither scalable nor efficient in the context of built asset digital twinning, as these systems involve accessing a diverse range of heterogeneous

data sources, from IoT devices to legacy databases (Shahinmoghadam & Motamedi, 2021). A detailed discussion of the importance and current state of built asset information alignment along with the related knowledge gaps is presented in Section 4.1.

Second, it allows for more efficient and meaningful interaction with the vast array of data produced across the lifecycle of an asset. The importance of the latter aspect becomes more evident when considering the role of end users. End users are not just passive consumers of data; they are active participants in the ecosystem of information systems within a digital twin. As digital twins evolve into more sophisticated and complex information systems, providing intuitive access to data for end users—be they facility managers, engineers, decision-makers, or occupants—becomes increasingly challenging (Lei *et al.*, 2023). It is worth noting that the complexity of data retrieval can persist, even with a high degree of data interoperability. For example, training end users to use expressive query languages, such as SPARQL (Pérez *et al.*, 2009), introduces a steep learning curve. Not only must users become proficient in a sophisticated technical query language (Pérez *et al.*, 2009), but they must also be familiarized with the underlying data structures to be able to formulate accurate queries. A detailed discussion of the trade-offs between query expressivity and complexity in the context of BIM-IoT data integration can be found in an earlier work conducted within the framework of the present thesis (Shahinmoghadam & Motamedi, 2021).

Recent advancements in NLP research offer promising solutions for both the above-mentioned aspects. By providing a context-aware understanding of textual input, which can be schema elements and/or user queries, advanced NLP techniques have the potential to both facilitate semantic alignment between data models for programmatic access and simplify data retrieval through natural language interfaces. A detailed discussion of recent research trends and existing knowledge gaps regarding how these advancements can facilitate semantic alignment and data retrieval in the context of this research is provided in the next section.

1.3 Improving data retrieval and alignment with neural language models

1.3.1 Primer on neural language models and text representation

The field of NLP is a branch of artificial intelligence and linguistics that is focused on enabling computers to comprehend and interpret human language. NLP tasks span from basic applications such as text classification (Wang, Issa & Anumba, 2022c), machine translation (Devlin, Chang, Lee & Toutanova, 2018), and sentiment analysis (Naseem, Razzak, Musial & Imran, 2020) to more complex areas such as semantic search and question-answering systems (Onal *et al.*, 2018). Over the past decade, neural language modeling—estimating the probability of the next word's occurring given previous words with the help of artificial neural networks—driven by advanced deep learning techniques has revolutionized these tasks by allowing machines to understand more intricate patterns in language (Jurafsky & Martin, 2024; Bommasani *et al.*, 2021; Khurana, Koli, Khatter & Singh, 2023).

Recurrent Neural Networks (RNNs), which were popularized in the early 1990s (Salehinejad, Sankar, Barfett, Colak & Valaee, 2017) played a key role in the evolution of neural language models by introducing loops to handle sequential nature of the written text data. However, RNNs faced challenges such as the vanishing gradient problem, limiting their ability to capture long-term dependencies (Jurafsky & Martin, 2024). In response, architectures based on Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) have been studied to improve memory retention in sequence learning, though they remained computationally inefficient for long-range dependencies (Jurafsky & Martin, 2024). The introduction of the attention mechanism (Vaswani *et al.*, 2017) marked a significant advance by enabling models to focus on different parts of the input simultaneously rather than sequentially. This breakthrough laid the foundation for the transformer architecture, the backbone of early Large Language Models (LLMs) such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al.*, 2018) and GPT (Generative Pre-trained Transformer) (Radford, 2018) , significantly enhancing language models' ability to capture complex relationships across the text (Bommasani *et al.*, 2021; Jurafsky & Martin, 2024).

The significant recent breakthroughs in various NLP tasks have been driven by the powerful idea of transfer learning, which involves two key stages: pre-training—the process where large amounts of unlabelled text is used to train a deep neural network for a particular language modeling objective such as next token (word) prediction, thereby learning general language representations (Radford, 2018; Jurafsky & Martin, 2024),—and then fine-tuning of the pre-trained model—the process of adapting the learned representations to the specific downstream task of interest using a much smaller dataset (Bommasani *et al.*, 2021; Jurafsky & Martin, 2024). This adaptation process highlights the critical importance of text representation learning.

State-of-the-art text representation learning, often referred to as *"text embedding"*, involves training a deep neural network to convert natural language into numerical representations, typically in the form of dense vectors (Jurafsky & Martin, 2024). This process can occur as part of a standalone model or as a built-in component (neural network layer) within a larger language model (Pennington, Socher & Manning, 2014a; Lee *et al.*, 2024b). Early embedding models, such as word2vec (Mikolov, Sutskever, Chen, Corrado & Dean, 2013) and GloVe (Pennington *et al.*, 2014a) introduced static word embeddings, where each word had a single, fixed representation regardless of its context. While these models captured semantic relationships between words effectively, they struggled to address the dynamic semantics of words that depend on the context. Advancements in neural language modeling led to the advent of contextual embeddings, such as those produced by BERT (Devlin *et al.*, 2018) and GPT (Radford, 2018; Brown *et al.*, 2020) models, which was a transformative shift in language representation by dynamically adjusting word representations based on their surrounding words, enabling the model to capture deeper nuances in meaning (Liu, Kusner & Blunsom, 2020a; Bommasani *et al.*, 2021; Jurafsky & Martin, 2024).

As text embeddings represent the underlying meaning of words, phrases, or entire documents, researchers have been extensively investigating high-quality text representations, particularly in the fields of semantic search and information retrieval (Neelakantan *et al.*, 2022). Such research efforts are motivated by the fact that high-quality embeddings can improve search results by improving the accuracy of the results in various sub-tasks, such as understanding the intent

behind user queries in the intent classification step and connecting related concepts to the right entities in the database schema in Named Entity Recognition (NER) and entity linking steps (Jurafsky & Martin, 2024).

Building on the preceding introduction to neural language models, the following section presents an overview of the related research directions in built asset information management, while also identifying the knowledge gaps that this thesis aims to address.

1.3.2 Neural language models and built asset terminology

With the proliferation of pre-trained LLMs and advancements in transfer learning techniques, researchers from various domains have increasingly explored the development of domain-specific language models for a wide range of tasks, including text classification, clustering, and entity extraction. The built environment research has been no exception. Over the last few years, an extensive number of studies have investigated fine-tuning of general-purpose language models for various applications including but not limited to entity extraction from construction defect reports (Jeon, Lee, Yang & Jeong, 2022b), semantic parsing of regulatory documents (Fuchs, Dimyadi, Witbrock & Amor, 2024), classification of user queries over BIM models (Wang *et al.*, 2022d), cross-domain transformations of BIM models (Wang, Bergés & Akinci, 2024b), and material matching across digital material passports (Forth *et al.*, 2024). The review of the mentioned works and many other related studies shows that while the use of advanced language models in built environment research is gaining significant momentum, most of the works focus predominantly on reporting accuracy metrics from relatively small datasets that are often not publicly available. This is while several important underlying considerations remain largely overlooked within the literature.

A commonly overlooked issue is the imbalanced emphasis on model prediction accuracy while paying insufficient attention to computational efficiency. This oversight becomes particularly significant when considering prior findings that advanced neural language models, such as those built on transformer architectures, do not consistently surpass traditional models such as

LSTMs when applied to smaller, domain-specific datasets (Garcia-Silva, Berrio & Gómez-Pérez, 2019; Ezen-Can, 2020). The generally high accuracy scores reported for traditional deep learning architectures applied to BIM-related tasks align with the latter observation (Zheng, Lu, Chen, Zhou & Lin, 2022a; Shamshiri, Ryu & Park, 2024). While more recent LLMs such as Llama 3 family of models (Dubey *et al.*, 2024) have demonstrated superior performance on various benchmark tasks, there remains a lack of evidence in the built asset information management literature regarding the effectiveness of leveraging such emergent models in niche scenarios. In fact, this knowledge gap motivated the present study to compare LSTM and BERT architectures—two popular choices in the built asset literature—in Chapter 3. A detailed discussion of this issue along with a review of the related works are presented in Sections 3.2.4 and 3.4.4.

Given the predominant role of text as a central data modality within the built asset industry—from design specifications and construction records to maintenance logs and regulatory compliance reports—text embedding models have been increasingly applied in recent research (Zheng *et al.*, 2022a; Forth *et al.*, 2024; Chung, Kim, Baik, Chi *et al.*, 2024; Wang, Xiao, Bouferguene & Al-Hussein, 2024a; Jeon, Lee, Yang, Kim & Suh, 2024). These models are used not only for conventional NLP tasks (mentioned above) but also for emerging research areas such as knowledge graph embeddings for enhanced built asset information management (Shahinmoghadam *et al.*, 2022a; Yin *et al.*, 2023a).

However, the extensive and increasing availability of pre-trained embedding models has led to a proliferation of potential text embedding solutions. While this abundance offers extensive opportunities, it can also create confusion regarding model selection for different use cases, particularly in specialized domains (Enevoldsen, Kardos, Muennighoff & Nielbo, 2024; Muennighoff *et al.*, 2022; Lee *et al.*, 2024b). In this light, a careful evaluation of text embeddings becomes essentially critical.

Research has shown that evaluating text embeddings using a limited set of datasets from single tasks may not reflect their performance across a broader context (Muennighoff *et al.*,

2022). However, this issue is recurrent in the existing body of knowledge in built environment research, where the performance of language models is typically reported for single tasks using unpublished datasets. Such evaluations are inadequate for generalizing the strengths and limitations of state-of-the-art text embedding techniques within this highly specialized domain. Our comprehensive review of related works highlights the absence of robust benchmarks and a holistic approach to evaluating the capabilities of language models in effectively representing terminology specific to the built asset industry.

1.4 Summary

Given the growing demand for implementing and integrating BIM, IoT, and VR technologies, current literature calls for further case studies that adequately reflect the real-world complexities of these systems. This shortcoming underscores the need for further research to bridge the gap between existing academic knowledge and industry-driven advancements. In response, the present thesis focuses on live thermal comfort assessment by leveraging BIM and IoT data within a VR environment.

Moreover, the end-user experience—particularly regarding data navigation and retrieval—is frequently underemphasized in digital twin research, despite its critical role in interpreting highly complex, heterogeneous, and disparate built asset data. The review indicates that the growing body of work on localized, specialized data representation models, especially formal ontologies, offers promising solutions to address the complexities arising from the proliferation of such schemas. These observations highlight an important knowledge gap concerning data accessibility in digital twin systems, both in programmatic and intuitive retrieval contexts.

Finally, it was shown that recent advancements in neural language modeling offer promising solutions to address the mentioned data accessibility challenges. However, despite the increasing interest in utilizing advanced language models in built environment research, comparative evaluations of traditional and emergent models are missing from the literature. The provided discussions highlight the importance of text representation learning, known as text embedding.

Lastly, the chapter critically evaluates existing attempts to apply these models in built asset management research, highlighting the limitations in availability of robust evaluation frameworks that are applicable across a range of downstream applications in this field. In this light, there exists an essential need for a comprehensive benchmarking framework that can effectively evaluate the performance of text embeddings in tasks related to the specific domain of built asset information management, in particular, information alignment and retrieval. Given the scarcity of high-quality, publicly accessible datasets, establishing comprehensive public benchmarks is a crucial step toward enhancing transparency and fostering community engagement in this nascent yet fast-growing field of research.

CHAPTER 2

BIM- AND IOT-BASED VIRTUAL REALITY TOOL FOR REAL-TIME THERMAL COMFORT ASSESSMENT IN BUILDING ENCLOSURES

Mehrzad Shahinmoghadam[1] , Worawan Natephra[2] , Ali Motamedi[1]

[1] Department of Construction Engineering, École de Technologie Supérieure, 1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Digital Building Information and Virtual Environment Laboratory, Department of Architecture, Urban Design and Creative Arts, Mahasarakham University, Thailand

Article published in *Elsevier's Building and Environment*, in May 2021.

Abstract

Despite their vast potential for delivering rich and intuitive visualizations of live building monitoring data, digital twins have been rarely studied in the context of thermal comfort. To narrow this gap, this study investigates the synergistic benefits of Building Information Modeling (BIM), the Internet of Things (IoT) and Virtual Reality (VR) for developing an immersive VR application for real-time monitoring of thermal comfort conditions. A system architecture was proposed for live calculation of the PMV/PPD indices based on ASHRAE standard 55 and enrichment of BIM-based representations of building spaces in VR environments with live IoT-enabled monitoring data. Openly available software tools were used to make the geometric and sensory data accessible within a VR application and calculate the PMV/PPD indices. Using a semi-automated method, raw thermal images streaming from a cost-effective non-intrusive sensor were processed on an affordable edge computing device to enable near real-time calculation of Mean Radiant Temperature (MRT). A prototype of the system was implemented and used in a series of experiments where a dynamic thermal environment was created in a mechanically conditioned space. The results support the consistency between the system's output and the actual thermal sensations observed under various conditions.

2.1 Introduction

Over the last few decades, a diverse range of factors has contributed to an exponential increase in energy consumption globally. Meanwhile, the construction and operation of buildings account for nearly 40% of total global energy use (IEA, 2019). A major proportion of the energy consumption in buildings relates to space cooling/heating to achieve the occupants' thermal satisfaction (IEA, 2020).

Recently, the issue of monitoring indoor thermal comfort conditions has become a topic of interest to assess a building's performance (Rupp, Vásquez & Lamberts, 2015; Salamone *et al.*, 2018). The reason behind this is that the effective monitoring of thermal comfort indicators, which are accurate, real-time, visually rich, intuitive, and interactive, could lead to the optimization of building energy usage while achieving the desired indoor environmental conditions.

According to ASHRAE standard 55 (ASHRAE, 2017), one way to reflect the thermal sensation experiences of occupants is to represent them by two indices, namely Predicted Mean Vote (PMV) and Predicted Percentage of Dissatisfied (PPD). Compared to other models (e.g., adaptive models), the PMV-PPD model provides the possibility of predicting thermal comfort in a more numerical and rigorous way (ANSI, 2017). To measure PMV/PPD indices, certain environmental parameters need to be known, such as air temperature, relative air velocity, and Mean Radiant Temperature (MRT). With the rapid developments in sensing technologies, high-performance, cost-effective sensors are now readily available and able to monitor such thermal comfort parameters in real-time (Wang, Gluhak, Meissner & Tafazolli, 2013b). However, the real-time and accurate measurement of PMV/PPD values may prove difficult as it involves complex mathematical computations, particularly for the case of MRT, thereby making the live assessment of thermal comfort conditions considerably inefficient in practice (Sahari, Jalal, Homod & Eng, 2013).

Currently, Building Information Modeling (BIM) tools and processes offer vast opportunities to enhance building performance through visually rich and intuitive analyses and simulations. With the rapid developments in sensing technologies and cloud computing, there has been an

increasing interest in Internet of Things (IoT) solutions that take advantage of BIM platforms to provide a unified view of rich contextual building information and real-time sensory data (Shahinmoghadam & Motamedi, 2019). However, enriching BIM models with live streams of IoT data will be a challenging task due to the lack of sufficient interoperability between the two ecosystems (Shahinmoghadam & Motamedi, 2021; Tang *et al.*, 2019). Hence, the difficulties of integrating sensory data with building spatial data add to the challenge of live thermal comfort assessment in building enclosures.

The ability of Virtual Reality (VR) to mimic real-world experiences within digital platforms offers vast opportunities in terms of remote communications/collaborations and visual analytics (Moran, Gadepally, Hubbell & Kepner, 2015). Evidently, an interactive visualization of IoT data in a BIM-enabled immersive VR environment can significantly reduce the inherent complexities that impede the efficient monitoring of the thermal comfort indices in a real-time manner. However, to the best of the authors' knowledge, the application of BIM-IoT integrated systems for the live calculation of MRT and thermal comfort indices based on the ASHRAE 55 standard is missing from the current literature. On the other hand, most of the previous studies mainly focused on visualizing thermal comfort monitoring data using a color-spatial approach and 2D charts in BIM environments. However, by linking BIM and IoT data to immersive VR environments, more intuitive visualizations and higher levels of interactivity can be achieved for comfort assessment.

In light of all the aforementioned gaps, this paper proposes an integrated method by which a combination of BIM-based products and the live data streaming from environmental sensors can be represented in a VR environment to facilitate indoor thermal comfort assessment in mechanically conditioned spaces, using the PMV-PPD model. The specific objectives of this research are to: (1) investigate the integration of BIM geometric information and IoT-collected multi-environmental and thermography data within game engine environments; (2) develop cost-effective IoT prototypes for the live monitoring of thermal comfort parameters; (3) develop an immersive VR system with a user-friendly interface to visualize live IoT data and thermal

comfort indices; (4) conduct a real-world case study to test and validate the applicability of the proposed system.

This body of work builds on the authors' previously published studies (Natephra, Motamedi, Yabuki & Fukuda, 2017; Natephra & Motamedi, 2019a,b). However, the present work makes original contributions that are explained in detail in the subsequent sections. The paper shall also review related works to highlight the existing gaps. Finally, proposed systems, evaluation results, and conclusions are given.

2.2 Literature review and related works

2.2.1 Indoor thermal comfort assessment

Thermal comfort can be defined as "a condition of mind, which expresses satisfaction with the thermal environment and is assessed by subjective evaluation" (ASHRAE, 2017). The factors that influence thermal comfort inside buildings are of three types: measurable environmental, personal, and psychological (Grondzik & Kwok, 2019). Measurable environmental factors include air temperature (°C), air velocity (m/s), mean radiant temperature (°C) and relative humidity (%). Personal factors consist of metabolic rate (met) and level of clothing insulation (clo). Psychological factors include color, texture, sound, light, and aroma. The most commonly used factors to describe thermal comfort conditions in the literature are measurable environmental factors and personal factors.

Among the several existing models for predicting thermal comfort and thermal sensation, the PMV-PPD model is widely used and accepted not only for design but also in field assessments of comfort conditions (ANSI, 2017). Fanger et al. (Fanger, 1970) defined PMV as an index that predicts the mean votes of a large group of people on a thermal sensation scale. ASHRAE standard 55 recommends the following seven-point scale: +3: hot, +2: warm, +1: slightly warm, 0: neutral, −1: slightly cool, −2: cool, −3: cold (ASHRAE, 2017). To predict the building occupants' perception of a certain thermal condition using the PMV index, four environmental

factors, i.e., air temperature, MRT, air velocity, relative humidity, and two personal factors, i.e., clothing insulation level and metabolic rate, must first be determined. PMV can be calculated using the following equation (ISO *et al.*, 2004):

$$pmv = \left(0.303 \cdot e^{-0.036M} + 0.028\right) \cdot \left\{ \tag{2.1}$$

$$(M - W) - 3.05 \cdot 10^{-3} \cdot [5733 - 6.99 \cdot (M - W) - \rho_a] - 0.42 \tag{2.2}$$

$$- \left[(M - W) - 58.15 - 1.7 \cdot 10^{-5} \cdot M \cdot (5867 - \rho_a)\right] - 0.0014 \cdot M \tag{2.3}$$

$$\cdot (34 - t_a) - 3.96 \cdot 10^{-8} \cdot f_{cl} \cdot \left[(t_{cl} + 273)^4 - (t_r + 273)^4\right] \tag{2.4}$$

$$- f_{cl} \cdot h_c \cdot (t_{cl} - t_a) \Big\} \tag{2.5}$$

In the above equation:

- M is the metabolic rate (W/m^2)
- W is the effective mechanical power (W/m^2)
- I_{cl} is the clothing insulation (m^{2}K/W)
- f_{cl} is the clothing surface area factor
- t_{cl} is the clothing surface temperature (°C)
- t_a is the air temperature (°C)
- ρ_a is the water vapor partial pressure (Pa)
- V_{ar} is the relative air velocity (m/s)
- t_r is the mean radiant temperature (°C)
- h_c is the convection heat transfer coefficient (W/m^{2}K)

The other index, PPD, is a quantitative prediction of the percentage of occupants that will experience discomfort in a given thermal setting. The PPD method defines that according to a particular thermal sensation scale (i.e., +3: hot, +2: warm, −2: cool, −3: cold), how many

people will be dissatisfied with the thermal conditions (ASHRAE, 2017). PPD is expressed using the following equation:

$$PPD = 100 - 95 \cdot e^{(-0.03353 \cdot PMV^4 - 0.2179 \cdot PMV^2)} \tag{2.6}$$

Regarding personal factors, metabolic rate refers to the level of heat and mechanical work produced from the transformation of chemical energy through human metabolic activities (ASHRAE, 2017). The metabolic rate (measured in met units) varies depending on a person's physical activities (Grondzik & Kwok, 2019). The other personal factor, clothing insulation level, refers to the insulation layer of clothing that acts against human body heat loss (ASHRAE, 2017).

MRT is defined as the uniform surface temperature of the surrounding envelopes and surfaces of an enclosure that influences the human body's heat loss. MRT plays a major role in thermal comfort assessment, and previous studies have shown its significant effect on thermal comfort (Anderson, 2014). With reference to (Standard & Iso, 1998; Alfano, Dell'Isola, Palella, Riccio & Russi, 2013), MRT can be measured using a black globe thermometer, net radiometers, or radiation thermometers, or it can be calculated using following Equations 2.7-2.11. By considering the average temperature of surrounding surfaces (Ti) and the angle factors between the occupant's body and surrounding surfaces (F_{p-N}), the MRT can be evaluated using the following equation (ISO *et al.*, 2004):

$$T_{mrt}^4 = T_1^4 F_{p-1} + T_2^4 F_{p-2} + \cdots + T_i^4 F_{p-i} + \cdots + T_N^4 F_{p-N} \tag{2.7}$$

The angle factor between the person and the surface (F_{p-N}) can be calculated using the following equation:

$$T_{mrt}^4 = T_1^4 F_{p-1} + T_2^4 F_{p-2} + \cdots + T_i^4 F_{p-i} + \cdots + T_N^4 F_{p-N} \tag{2.8}$$

$$F_{p-i} = F_{max} \left[1 - e^{-(a/c)/\tau}\right] \left[1 - e^{-(b/c)/\gamma}\right] \tag{2.9}$$

$$\tau = A + B(a + b) \tag{2.10}$$

$$\gamma = C + D(b/c) + E(a/c) \tag{2.11}$$

in which "a" is the width of the surface above or below the center of the occupant's body; "b" is the height of the surface above or below the center of the occupant's body; "c" is the distance between the center of the occupant's body and the target surface; and coefficients, F_{max}, A, B, C, D and E describe angle factors for standing and sitting postures (Parsons, 2007).

2.2.2 Potential and challenges of live thermal comfort assessment

According to the above, predicting thermal comfort using the PMV-PPD model involves complex mathematical expressions. Evidently, approaching such equations manually will be considerably time-consuming and prone to human errors. Such drawbacks can be addressed by capturing various comfort-related parameters through a network of sensors and linking the monitoring data with computerized tools that have been developed based on existing standards, e.g., Berkeley's CBE Thermal Comfort Tool developed based on ASHRAE 55 (Tartarini, Schiavon, Cheung & Hoyt, 2020).

Among the mentioned thermal comfort parameters, MRT is one of the most complex ones to be analyzed (La Gennusa, Nucara, Rizzo & Scaccianoce, 2005; Walikewitz, Jänicke, Langner, Meier & Endlicher, 2015). Hence, a cost-effective method to capture MRT values in a real-time manner remains a challenge. This complexity can be explained by the fact that in addition to thermal radiation from the building's surfaces, the space geometry and the spatial relations between the occupant's body and the surrounding surfaces play a role in the measurement of MRT values.

Relevant to MRT measurement is infrared thermography which has been previously investigated in the context of thermal comfort to measure the temperature of indoor surfaces (Natephra *et al.*, 2017) or human skin (Aryal & Becerik-Gerber, 2019; Li, Menassa & Kamat, 2018). Moreover, researchers have been interested in integrating thermography data with 3D digital models (Lagüela, Díaz-Vilariño, Martínez & Armesto, 2013; Wang, Cho & Gai, 2013a; Ham & Golparvar-Fard, 2015; Sels *et al.*, 2019). In this regard, the limitations of two of the previous works are relevant to the present study. First, some of the previous methods involved considerable manual processes, such as manual stitching and interpretation of thermography data in (Natephra *et al.*, 2017). Second, the high cost of high-performance thermal cameras significantly limits their applicability in the context of IoT. Even though inexpensive thermal imaging modules (e.g., AMG8833 sensors) currently exist, they are less likely to meet the requirements for MRT calculations due to their low thermal sensitivity, limited pixel size, and field of view.

To address the aforementioned gaps, the following considerations were taken into account in this study. Regarding the first issue, a semi-automated method for real-time processing of thermal imaging data is proposed. Regarding the second issue, based on the cost-performance comparisons between the currently available thermal imaging sensors reported in (Li *et al.*, 2018; Mikkilineni, Dong, Kuruganti & Fugate, 2019), the FLIR Lepton module is selected in this study to meet the IoT's cost-effectiveness requirement. Although Lepton's effectiveness has been studied in the context of thermal comfort (Aryal & Becerik-Gerber, 2019; Cosma & Simha, 2019; Li, Menassa & Kamat, 2019), to the best of the authors' knowledge, no other study has previously investigated its potential for MRT measurement in building enclosures.

2.2.3 BIM, IoT, and VR for thermal comfort assessment

Recently, a growing amount of effort has been made to integrate BIM and sensor technologies for various purposes (Liu, Deng & Demian, 2018). Different from Wireless Sensor Networks (WSNs), an IoT system consists of a network of real-world objects equipped with embedded intelligence and sensing and actuation capabilities that are connected over the Internet (Cirani,

Ferrari, Picone & Veltri, 2018). Using IoT systems, live collection of data about the physical world objects and conditions will be enabled with the aid of wireless sensing technologies and different communication protocols. Consequently, a particular interest has emerged to integrate BIM databases with IoT environmental sensory data to aid in the prediction of indoor thermal comfort (Lu, Wang, Wang & Cochran Hameen, 2019; Underwood & Isikdag, 2011; Corry, Pauwels, Hu, Keane & O'Donnell, 2015; Ma, Liu & Shang, 2019).

According to the general architecture of IoT systems, a BIM-IoT deployment should consist of a cloud-enabled middleware within its architecture (Shahinmoghadam & Motamedi, 2019). While WSNs are most often developed to serve a single application, the concept of IoT is concerned with the common usage of data from sensor nodes by multiple applications (Minerva, Biru & Rotondi, 2015). Taking this requirement into account, cloud-based accessibility to IoT monitoring data through a widely used IoT platform was considered in the present study. Moreover, an efficient IoT-driven monitoring urges for using cost-effective sensors and devices with low rates of power consumption, which increases the risk of receiving inaccurate readings in return. Hence, the accuracy of the IoT sensing nodes proposed in this study is discussed against the measurement ranges and accuracy levels recommended in AHRAE standard 55.

In addition to BIM-IoT integration, the potential of using BIM data within game engine environments has been under investigation for a wide range of applications (Johansson, Roupé & Viklund Tallgren, 2014; Fukuda, Yokoi, Yabuki & Motamedi, 2019; Motamedi, Wang, Yabuki, Fukuda & Michikawa, 2017; Natephra *et al.*, 2017; Du, Zou, Shi & Zhao, 2018; Hosokawa, Fukuda, Yabuki, Michikawa & Motamedi, 2016). However, in the context of monitoring thermal comfort conditions, only a few studies have previously investigated BIM-IoT integration within VR applications. Examples of the most relevant existing studies that have previously investigated BIM integration with IoT and VR systems, along with their contributions and shortcomings are included in Table 2.1.

Table 2.1 Summary of the contributions and limitations of the previous relevant studies

Study	**BIM**	**IoT/ WSN**	**VR**	**System capabilities & measured parameters**	**Limitations**
(Chang, Dzeng & Wu, 2018)	✓	✓	×	• Measuring air temperature and relative humidity. • 3D Visualization of comfort parameters and PMV values in building models using a visual programming interface BIM tool.	• Not monitoring MRT and air velocity. • Local data storage. • Limited adaptability for performing in-depth analyses.
(Liu, Kuo, Shiu & Wu, 2014)	✓	✓	×	• Measuring air temperature and relative humidity. • PMV evaluation using sensor data. • Cloud-enabled storage of monitoring data. • Sensor data visualization enabled in 3D BIM models (using Revit API).	• Unclear how radiant temperature and air velocity data were monitored. • Local storage of monitoring data. • Limited adaptability for performing in-depth analyses.
(Marzouk & Abdelaty, 2014)	✓	✓	×	• Measuring air temperature and relative humidity. • Linking sensor readings to BIM-based models. • Environmental sensors were connected to Arduino microcontrollers.	• Not monitoring MRT and air velocity. • Thermal sensations not monitored. • Local storage of monitoring data.

Table 2.1 Summary of the contributions and limitations of the previous relevant studies (continued)

Study	**BIM**	**IoT/ WSN**	**VR**	**System capabilities & measured parameters**	**Limitations**
(Costa *et al.*, 2015)	✓	✓	✓	• Measuring air temperature, CO2, and humidity. • 3D immersive visualization of BIM and sensor data in a game engine.	• Very limited information given for system implementation. • Unclear how the system validation was performed.
(Natephra & Motamedi, 2019a)	✓	✓	✓	• Measuring air temperature, relative humidity, and light level. • Environmental sensors were connected to Arduino microcontrollers. • 3D immersive visualization of BIM and live sensor data in a game engine.	• Local storage of monitoring data.

This study addresses some of the shortcomings listed in Table 2.1, by proposing a system as explained in the next section.

2.3 Proposed method and system

Using BIM, IoT and VR, a novel system is proposed in this paper to monitor thermal comfort conditions based on the PMV-PPD model, as recommended in ASHRAE 55. The proposed system provides an intuitive and immersive way of visualizing the complex and dynamic data

about thermal comfort conditions in a real-time manner. Using the system, users can freely navigate through the virtual environment (as a realistic replication of building interiors) and monitor various information, including live environmental sensor data, MRT and PMV/PPD values, as well as thermal comfort charts. Moreover, the system maintains interactivity between the users and the virtual thermal comfort assessment tool by allowing them to modify the parameters of interest (e.g., level of clothing insulation, metabolic rate) from the VR user interface. This way, users can perform various what-if analyses to observe the prospective changes in thermal sensations based on a combination of the user-defined and the live actual data continuously streaming from the IoT sensors.

Although this body of work builds on the authors' previously published studies (Natephra *et al.*, 2017; Natephra & Motamedi, 2019a,b), the original contributions of the present work can be mentioned as follows. With respect to MRT calculations, this paper avoids the disadvantages of using high-performance expensive thermal cameras, as well as manual and offline processing of thermal images (e.g., extraction of RGB values and image stitching) as proposed in (Natephra *et al.*, 2017). This was realized by designing a cost-effective IoT-enabled thermal imaging prototype, and a semi-automated method for spatio-thermal analysis of thermography data, which enables near real-time monitoring of MRT values on the edge of the IoT network. Moreover, the present work makes new contributions by taking advantage of immersive and interactive experiences that can be delivered by enabling live thermal comfort monitoring within game engine environments. In this regard, visualization of live environmental sensory data and adaptive thermal comfort models in interactive 3D environments has been previously investigated by the authors (Natephra & Motamedi, 2019a,b). Yet, the current paper proposes a novel system architecture for developing VR applications that take all essential thermal comfort variables, from IoT nodes and user interactions, to monitor comfort conditions based on the PMV-PPD model.

The general architecture of the proposed system is shown in Figure 2.1. A detailed explanation of the delivery of the two main functionalities of the system, i.e., IoT-based live data collection and virtual thermal comfort assessment, are as follows.

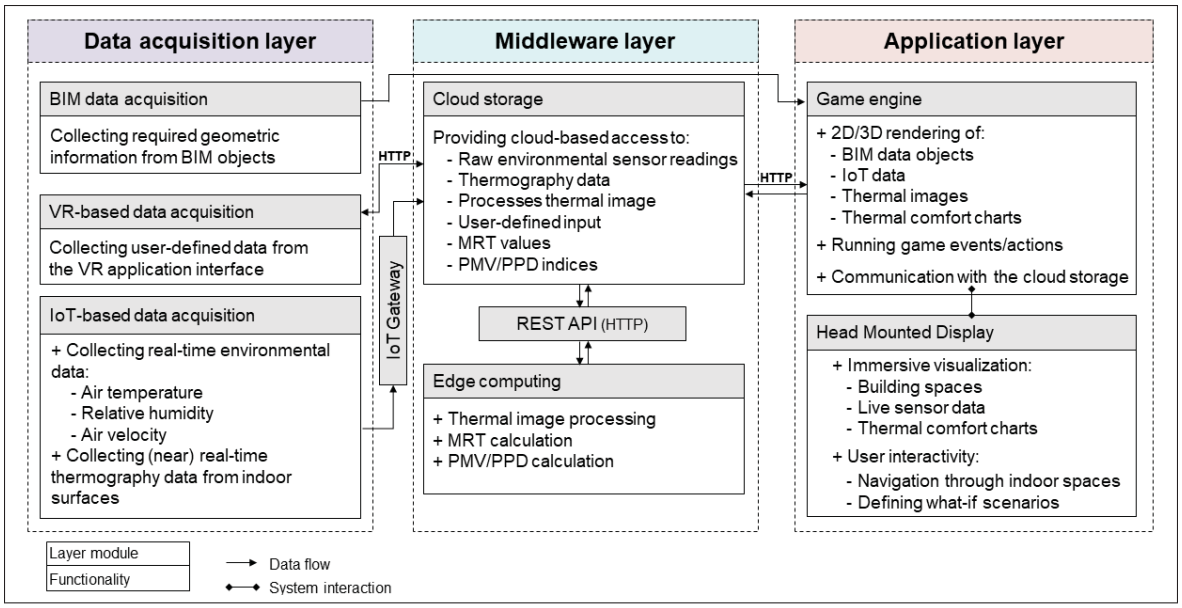


Figure 2.1 Proposed general system architecture

2.3.1 System architecture and functionalities

As can be seen from Figure 2.1, the proposed system architecture consists of three main layers, i.e., data acquisition, middleware, and application layers, delivering certain functionalities through the modules designated for each of them.

The data acquisition layer accumulates the essential sources of data to be transmitted to the other layers through three distinct modules: (i) BIM-based module to acquire 3D spatial objects from BIM models and convey them to the game engine environment within the application layer; (ii) VR-based module to obtain user-defined input and convey it to the middleware layer; (iii) IoT-based module to collect live environmental and thermography data, through a wireless network of IoT sensors, to be conveyed to the middleware layer.

The middleware layer delivers computation/analysis and data storage functionalities through the edge computing and cloud storage modules. The edge computing module performs the processing of raw thermal images (see Section 2.3.2), as well as the calculation of MRT and PMV/PPD indices (based on raw sensory data collected through the IoT-based module in the data acquisition

layer). The main advantage of delivering the aforementioned computations on the edge of the proposed IoT network is to avoid latency and bandwidth issues by processing the monitoring data as close as possible to their source of generation. The cloud-based storage module provides the application layer with access to IoT data (i.e., raw sensor data and processed thermal images) and thermal comfort monitoring data (computed on the edge of the network). Such accessibility was made possible through HTTP communications, which compared to the previously proposed methods (e.g., using serial communication interfaces in (Natephra & Motamedi, 2019a,b)) improves real-time connectivity between the physical building space and its virtual twin rendered in the game engine environment.

Finally, the application layer delivers visualization, immersion, and interaction functionalities. The game engine module performs 2D/3D rendering to create immersive visualizations of BIM-based spatial elements and thermal comfort monitoring data (stored in the cloud database) to be displayed in a Head Mounted Display (HMD) device. The HMD module delivers a sense of immersion for users and allows them to navigate through the virtual indoor spaces and interact with the digital twin of the building in two ways: First, with the help of rich immersive visualizations, users can easily monitor the average levels of environmental sensor outputs, thermal radiations from indoor surfaces, and PMV/PPD and bioclimatic charts, in a (near) real-time manner. Second, the current design of the system allows users to conduct different what-if scenarios to evaluate the occupants' thermal sensations, by providing the possibility of taking different input values for thermal comfort parameters (e.g., metabolic rate or clothing insulation) and observing the predicted outcomes on the thermal comfort charts. Such interactions can be actualized through the main user interface. For the present study, since one of the main objectives has been to provide end-users with a vivid sense of immersion when they are wearing the VR headset and navigating through the virtual building spaces, a first-person perspective was considered for the VR application development.

2.3.2 Thermography data processing

As mentioned previously, one of the objectives of this study is to collect, store, and retrieve thermography data, i.e., raw thermal images and pixel-wise temperature values, within the proposed system. However, due to the low spatial resolution of thermal images, especially in the case of low-resolution imaging modules such as the FLIR Lepton, the detection and extraction of visual features from thermal images will be quite impractical. An alternative solution could be to overlay thermal images with the corresponding visual images, as previously done for the Lepton module in references (Aryal & Becerik-Gerber, 2019; Li *et al.*, 2019). However, the use of visual cameras was deterred in the present proposal as it undermines the privacy-preserving aspect of the proposed IoT-enabled monitoring system. Thus, in this paper, a marker-based semi-automated registration method has been proposed, as depicted in Figure 2.2.

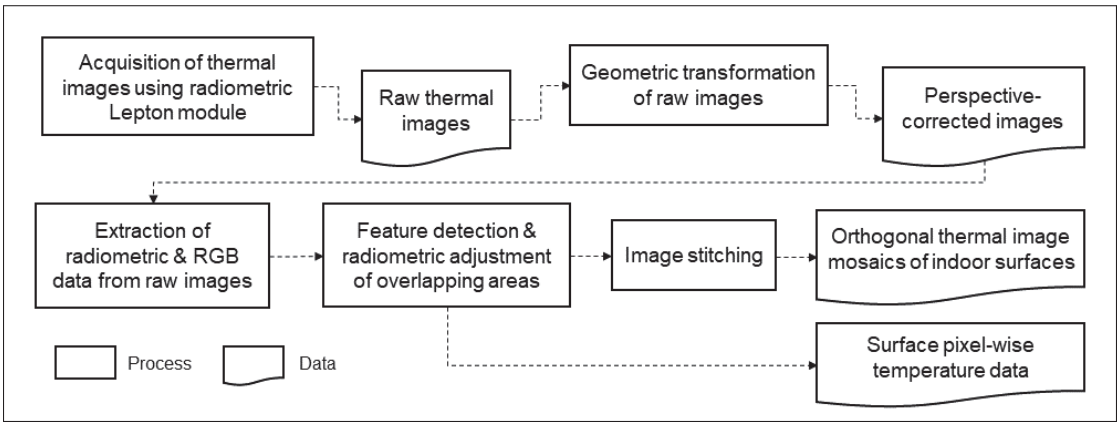


Figure 2.2 Proposed semi-automated method for thermography data processing

Using the proposed marker-based registration method, orthogonal thermal image mosaics can be created from multiple overlapping thermal images taken from indoor surfaces. The markers should be chosen in a way to create distinct visual features within the raw thermal images. This way, by knowing the exact location of each marker on the 3D digital model, the thermal images can be registered with reference to the location of the corresponding markers. However, to acquire orthogonal thermal images, whenever a new image is taken from a given orientation, the perspective distortions need to be corrected through a geometric transformation of the images.

This is an essential step in the process of capturing accurate temperature values in a pixel-wise manner. Once the perspective transformation step is completed, orthogonal thermal image mosaics along with their associated thermography data will be available to be transferred to the cloud server.

The next section gives a detailed account of the main steps and the technical considerations that were taken into account in this study to implement the general system architecture described above.

2.4 System implementation

In this study, implementation of the proposed system was done to evaluate its applicability. Figure 2.3 shows the development pipeline used to implement the proposed BIM and IoT-enabled VR system. The figure details the main process of transmitting the federated sources of building geometry and IoT data into the game engine environment.

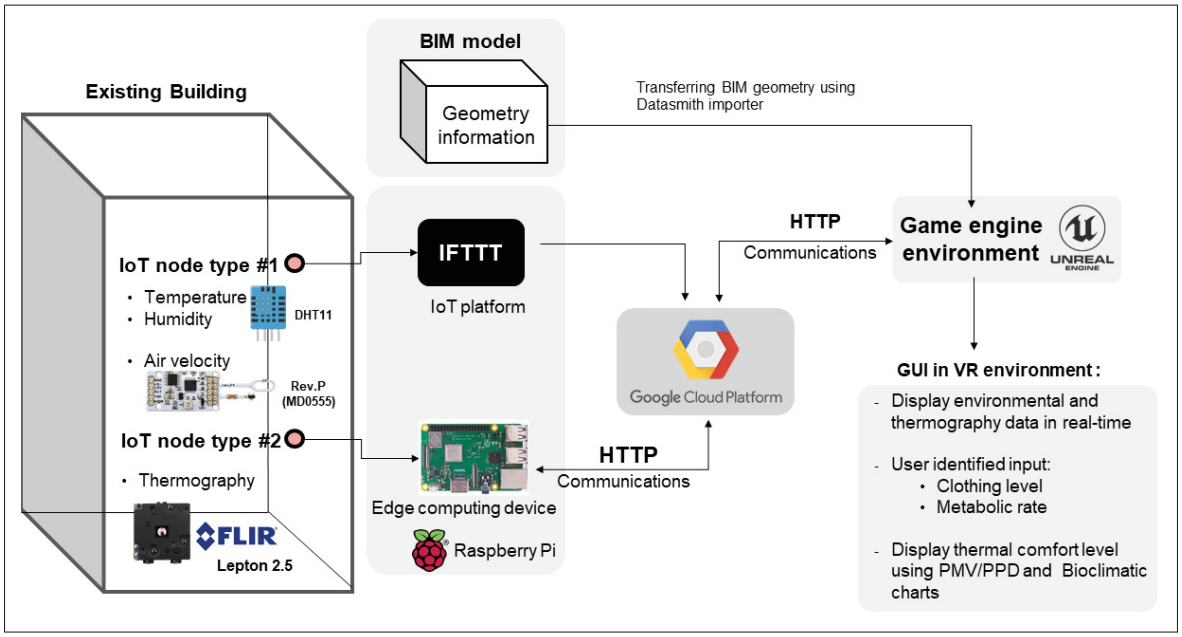


Figure 2.3 Proposed development pipeline

Regarding the IoT-based data collection module, two groups of wireless sensing nodes were designed and developed in this study. For the first group (IoT node type #1), DHT11 and Modern Device Rev.P wind sensors (MD0555 category) were considered to monitor air temperature and relative humidity, and air velocity, respectively. DHT 11 is a low-cost digital sensor ideal for conditions where the changes in air temperature and relative humidity remain in the ranges of 0–50 °C and 20–80%, respectively. The sensor's accuracy is ±2 °C for temperature readings and 5% of humidity readings. For air flows with less than 65 m/s speed that is not excessively variable in direction, Rev.P wind sensor measures average velocities with ±0.5 m/s accuracy without providing data about the direction of the airflow (Prohasky & Watkins, 2014). Although the measurement accuracy of the Rev.P wind sensor is less than the instrumentation accuracy recommended in the ASHRAE 55 standard, the purchase cost of it is extremely low (approximately US$24) (Carre & Williamson, 2018), thereby making it a potential candidate for IoT applications. Hence, the inclusion of this low-cost wind sensor within the proposed architecture is suggested in the present work, and discussions regarding the effect of its accuracy on the overall performance of the system are given later in Section 2.5.3.

The IoT sensors of this group were connected to ESP series microcontroller boards (ESP32 and ESP8266), which have gained marked popularity as efficient choices for IoT-based projects due to their relatively low cost and rate of power consumption.

For the second group (IoT node type #2), FLIR Lepton 2.5 thermal imaging module connected to Raspberry Pi microprocessor (Pi3 Model B+) was considered to measure surface temperatures to monitor MRT values at different spots within the building enclosure. A winning aspect of the Lepton module is that its 2.5 and 3.5 versions are radiometric-capable, meaning that they are able to capture accurate and calibrated temperature data in a pixel-by-pixel manner. Hence, the application of the Lepton module, version 2.5 (80×60-pixel size) has been considered in this paper.

The ESP and Raspberry Pi boards which were used in IoT node type #1 and Type #2, respectively, were powered using lithium polymer batteries and portable USB chargers, respectively. Images of the developed prototypes and the layout of the sensor network setup are shown in Figure 2.4.

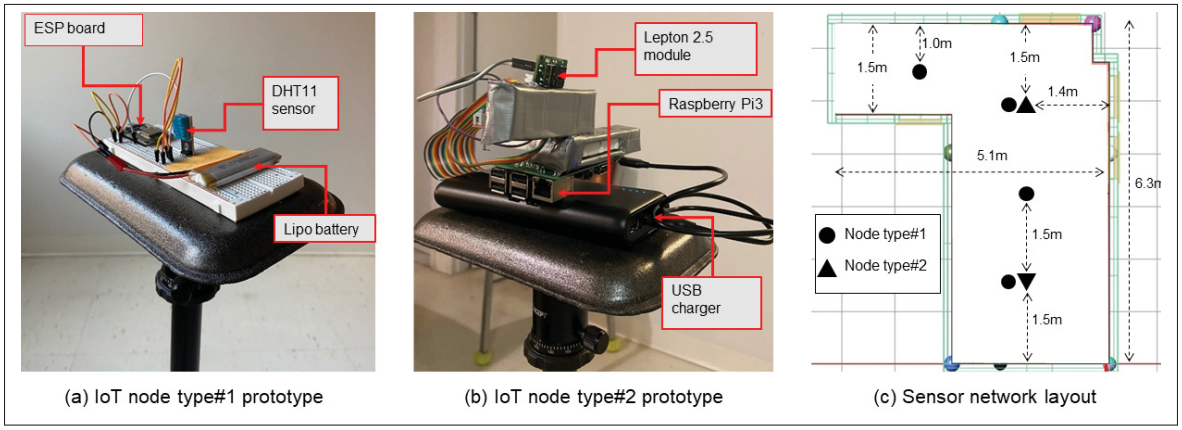


Figure 2.4 Implemented IoT-based data collection prototype

In order to ease the process of pushing sensor data to the cloud-based storage module, application of a widely used IoT platform, namely IFTTT (IFTTT, 2021), was considered for the proposed system. The ESP boards were programmed using C++ to establish HTTP communications with the IFTTT platform where the web requests to access the cloud services were processed.

As mentioned previously, in the proposed architecture, main computations occur on the edge of the network. In the current implementation, a set of Python scripts were deployed on a Raspberry Pi board used as the edge device. The same Python library (Tartarini & Schiavon, 2020) deployed on the backend of the CBE web-based thermal comfort tool was used to calculate MRT and PMV/PPD values. Another set of Python scripts were deployed on the edge device to automatically perform the processing of the thermal images (e.g., perspective transformation and detection of markers). The scripts were written based on the computer vision modules provided in the OpenCV library (OpenCV, 2021) and an openly available library designated to process the FLIR Lepton's output over SPI communications (PyLepton, 2021).

The thermal sensitivity of the FLIR Lepton 2.5 module is 0.05°C and its optimum operational temperature range varies from −10°C to +65°C. A critical requirement to consider when using

this module is its radiometric accuracy, since factors such as the size and emissivity of the target object may affect radiometric accuracy (FLIR, 2021). Among them, the effect of distance is of particular importance in the context of the present study. Hence, a pilot experiment was conducted to evaluate the accuracy of surface temperature measurements from distances that are highly probable within the indoor spaces of regular residential buildings. In this experiment, a body of hot water (approximate dimensions: $30cm \times 30cm \times 20cm$) with a known temperature ($\simeq$35°C, measured using a digital stick thermometer with 0.1°C accuracy) was used as the ground truth. Starting from a 2m distance, the average surface temperature was calculated for the region of interest (i.e., the hot body) for which radiometric data was extracted using the threshold function in OpenCV (thresholding range was manually acquired based on pixel intensity). The results showed an error close to 5% of temperature readings at a distance of $\simeq 5m$. The obtained rate of error complies with the information provided in (Li *et al.*, 2018) regarding the accuracy of FLIR Lepton 2.5 sensor. Hence, the accuracy test of Lepton's radiometric readings shows it should be used in spaces in which the distance of the thermal imaging apparatus from the room surfaces is less than $5m$.

In order to acquire the mean temperature values from the room surfaces, the proposed thermal image processing method (Section 2.3.3) was followed. In this study, LED lights, which can create evident visual features in thermal images, were selected as the image registration markers. The LED markers were installed in such a way as to form overlapping rectangles. Similar to the method proposed in (Natephra *et al.*, 2017), 30% overlap was considered for the thermal images taken of all the surfaces.

Example images of the LED markers and the corresponding thermal images used to create the thermal image mosaic for an example wall are shown in Figure 2.5a. For this example, the iteration steps to detect the markers using the binary threshold function in OpenCV are shown in Figure 2.5b. Based on the proposed method, once the perspective correction was completed, the average values of RGB and radiometric temperature data were calculated for the overlapping areas. As the thermography data corresponding to the overlapping areas were updated, all the single thermal images taken from each of the surfaces were automatically stitched together using

OpenCV tools, thereby creating orthogonal thermal image mosaics. To upload the most recent orthogonal thermal image mosaics along with their timestamp and corresponding radiometric data to the cloud storage module a Python script was deployed on the edge device. Subsequently, the thermography data stored in the cloud server could be made accessible to the edge device to perform the calculation of mean temperature values for each of the surfaces.

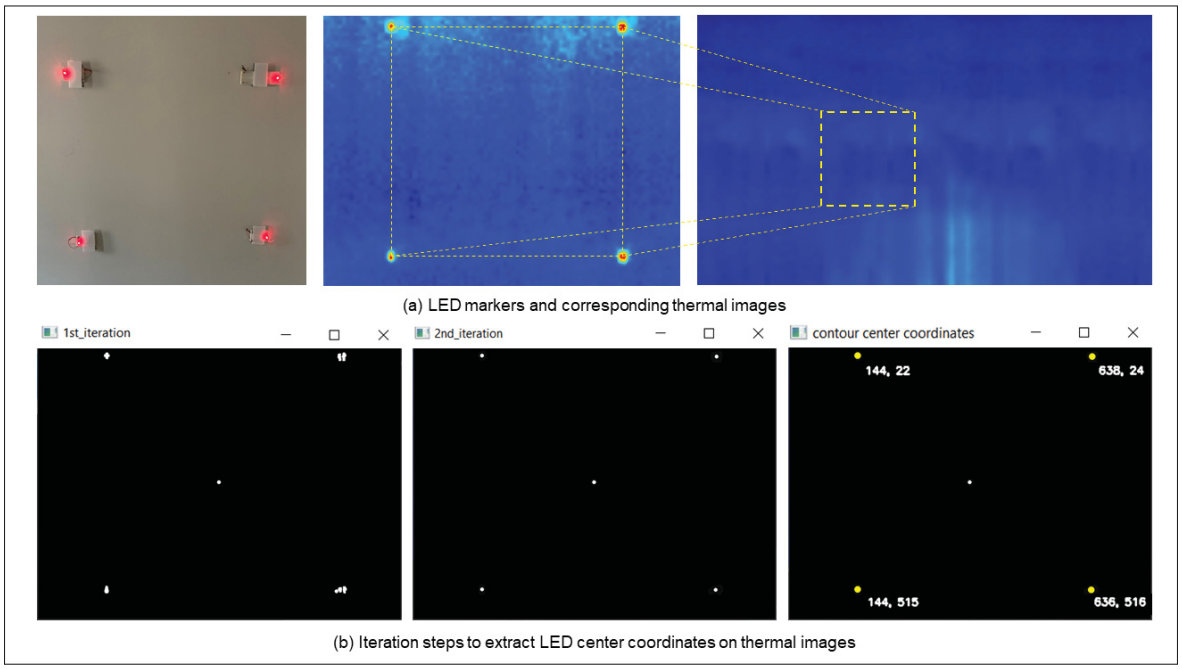
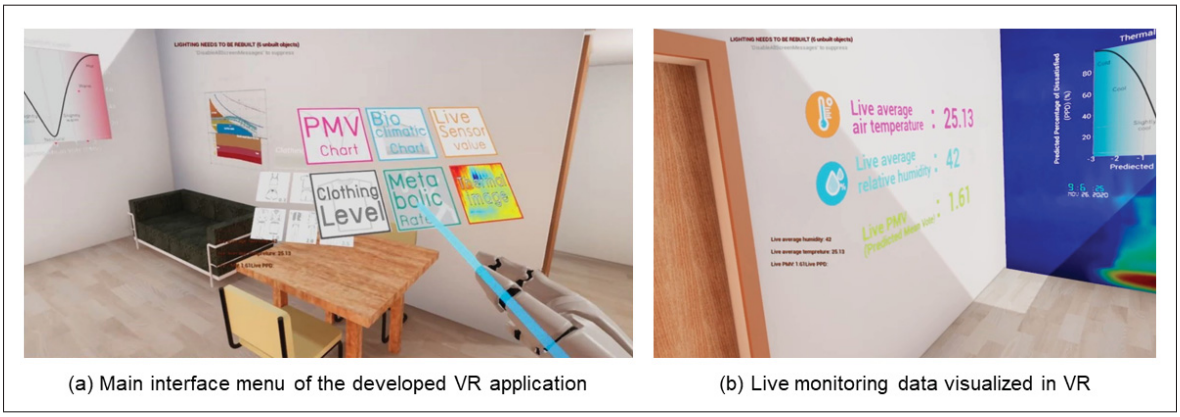


Figure 2.5 Marker-based detection of thermal image coordinates

To develop the VR thermal comfort assessment application, the following design choices were made. Regarding the hardware and software, the proposed system was developed for Oculus Rift S headsets (consisting of a head-mounted display and two controllers), using Unreal Engine 4 game engine (version 4.24.3) as the main development environment. To import the 3D geometric data representing building spaces into the selected game engine environment, Datasmith tool (Unrealengine, 2021a) was used, which in addition to direct data transfers from various BIM authoring tools (e.g., Autodesk Revit, SketchUp, Rhino 3D), enables the import of IFC (Industry Foundation Classes) files into the Unreal Engine environment.

Figure 2.6 shows runtime screenshots of the developed VR application. As it can be seen from Figure 2.6a, users can benefit from various features of the developed system through the main user interface menu. With the data collection apparatus set up in the experiment room, live IoT sensor data can be visualized in the virtual replication of the building spaces (Figure 2.6b).



(a) Main interface menu of the developed VR application

(b) Live monitoring data visualized in VR

Figure 2.6 VR thermal comfort application developed using Unreal Engine 4

Visualization and interactivity features provided within the developed VR application are depicted in Figure 2.7. Using the system, updated values of MRT, PMV and PPD indices that were automatically and continuously calculated on the edge of the network (using Python scripts written based on Equations 2.1-2.7) can be transmitted to the game engine environment (through HTTP requests) and the thermal comfort charts will be updated accordingly (see Figure 2.7). User-defined input, metabolism rate and level of clothing insulation in particular, could be selected from the system's interface menu (see Figure 2.7) and similarly communicated to the cloud server, and from there to the edge device through HTTP requests. To establish REST communications between the game engine and the cloud storage module, the open-source plug-in VaRest (Unrealengine, 2021b) was used in this study. Finally, using the built-in features of Unreal Engine, most recently uploaded thermal images could be downloaded to the workspace environment of the engine, thereby delivering updated visualization of surface thermography data throughout the application's runtime.

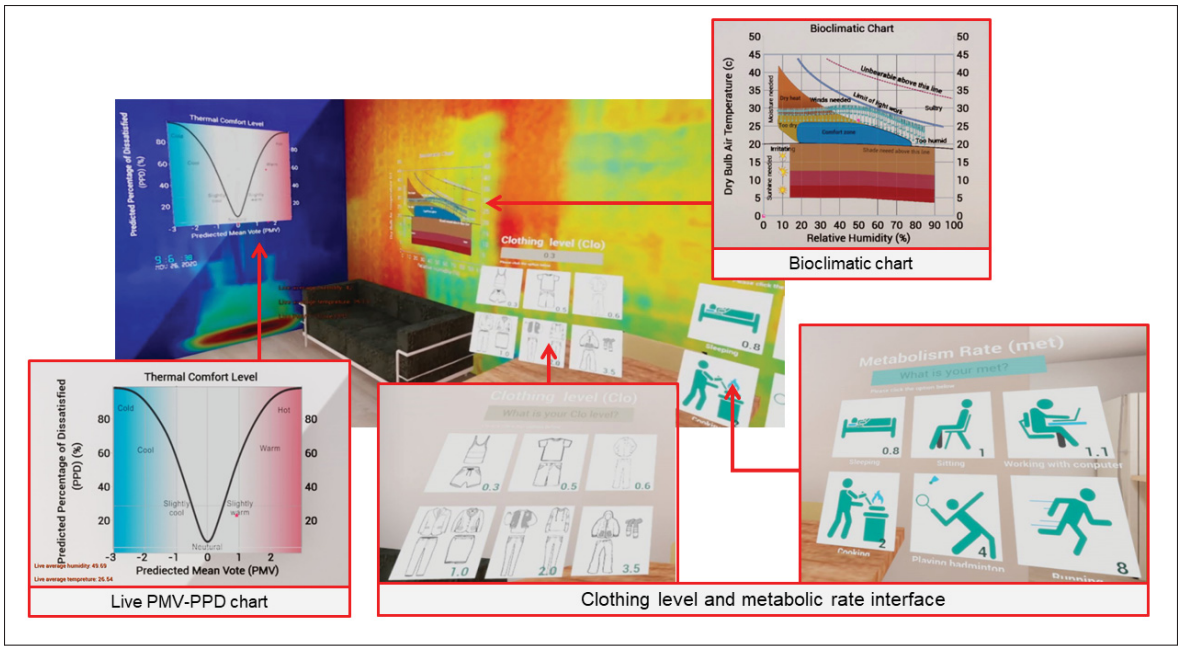


Figure 2.7 Visualization and interactivity features provided for the VR thermal comfort tool

2.5 Evaluation

2.5.1 Experimental setup

In this paper, a case study was designed to evaluate the overall performance of the proposed system for live and remote monitoring of thermal comfort conditions in regular residential building enclosures. The experiments were conducted in the living room of a two-bedroom apartment located in the student residences of École de Technologie Supérieure University, Montreal, Canada. The experiments took place in mid-November 2020, when the averages of the highest and lowest outdoor temperatures were approximately 6°C and −1°C, respectively.

Since the 3D BIM model of the residences building was not available, the actual dimensions of the experiment room were measured and the accurate 3D digital model of the experiment room was created using the Autodesk Revit software afterward. The visual images of the experiment room and its 3D BIM model are shown in 2.8. The 3D representations of the

spatial characteristics of the experiment room were imported to Unreal Engine to be used for scene creation in the VR application. For the experiments, an electrical baseboard heater ($1.25kw$ power consumption) and the kitchen's oven stove ($2.6kw$ power consumption for baking) were the main sources of heat generation, and a portable AC unit ($13{,}000BTU$; $1.3kw$ power consumption) was responsible for cooling and ventilation within the living room space (Figure 2.8c and d).

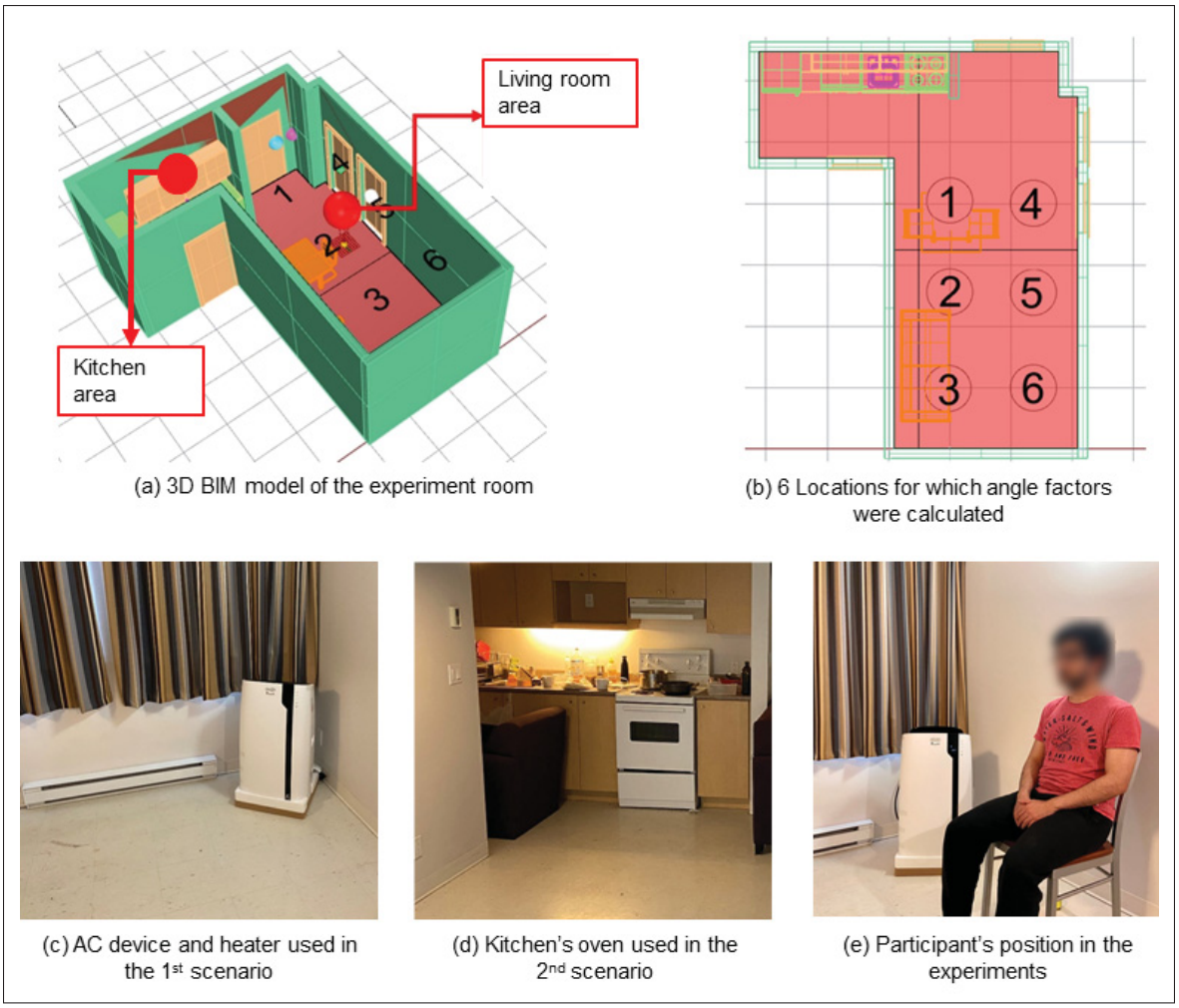


Figure 2.8 3D BIM model and visual images of the experiment room

For the case study, six different locations in the living room were considered in advance and angle factors were calculated manually for those locations to be compared with those calculated within the system. (Figure 2.8b). These locations were chosen so as to be situated at different

distances from the main sources of heat generation and cooling/ventilation. In the current implementation, the automated calculation of the angle factor with reference to the dynamic location of the virtual agent in the VR environment was only done for changing locations within the area covering the predetermined six locations. However, using the built-in feature of the game engine to track the location of the virtual agent according to the engine's local coordinate system, real-time calculation of the angle factor can be performed for all possible locations. To observe thermal comfort conditions within the experiment room, an actual occupant (male and 29 years old) was present in the apartment in which the tests were conducted. Although the participant is a graduate student at one of the departments in which the study was conducted, he had no background knowledge in the subject of thermal comfort assessment and related tools and standards.

2.5.2 Case description

To examine the performance of the proposed system in terms of reflecting the actual thermal sensations of the participant within the VR environment, in a real-time manner, two main scenarios were considered to imitate different thermal comfort conditions. In the first scenario, the initial room temperature was set at 23.5°C. Next, the baseboard heater located in the experiment room was set to 26°C to start generating heat within the experiment room. For this scenario, the participant was seated, relaxed (metabolic rate $\simeq$1), in a spot close to the heater (location "3" as depicted in Figure 2.8b), wearing an indoor clothing outfit typical for November in Montreal, Canada (i.e., sweatpants, long-sleeve sweatshirt, ankle socks, and underwear; clothing insulation $\simeq$0.74 clo). In the second scenario, to imitate relatively warmer conditions, the kitchen's oven temperature was set to 200°C while the initial room temperature was set to 25°C. With the same metabolic rate and clothing insulation level, this time thermal comfort conditions were observed at a spot close to the kitchen area (location "1" as depicted in Figure 2.8b).

For each scenario, the process of IoT-based data collection and monitoring of PMV and PPD indices took place within approximately 75 minutes. Data collection started 10min prior to

changing the thermal comfort conditions. Afterwards, a 25min period was taken to reach a steady state of discomfort conditions at the locations of interest. After this 25min period, the AC device was turned on and the sources of heat generation were set back to the initial conditions. Meanwhile, a 5min break was given to the participant followed by a 10min period of sitting relaxed to reach a steady metabolic rate. This total of 15min gave more time to the AC device to change the thermal comfort conditions for the next round of data collection. Finally, another 25min period was taken to monitor the changes in thermal comfort variables and the participant's sensations after turning the AC on and removing the sources of extra heat generation.

The essential details on how the experiments were conducted and the results were validated are given as follows. First, to evaluate the effectiveness of the proposed method for MRT calculation, it was necessary to remove the furniture from the experiment room. To test the effectiveness of the proposed method for thermography data processing, this assumption was made to eliminate the effect of large furniture that can change the pattern of heat radiation. Hence, during the tests, the only furniture remaining in the experiment room was the armless dining chair on which the participant was seated (Figure 2.8e). Second, since this study did not consider the effect of direct sunlight, the experiments were carried out between 7 and 10 p.m. This assumption was made due to the complexities involved in the real-time identification of the intensity of the solar radiations as it might go through significant changes depending on the dynamic position and orientation of the objects getting in the way of the sunlight (e.g., window blinds or external objects). The assumption only excludes the effect of asymmetric radiations for the cases in which the occupant is exposed to direct sunlight. Hence, the indirect effect of sunlight causing non-uniform radiations in the room can be captured using the proposed thermal imaging apparatus. Third, FLIR-ONE thermal cameras were used to capture surface temperature to cover the regions that fall beyond the field of view of the proposed thermal imaging prototype. The assumption here is that there would be no significant temperature changes for the surfaces located in a relatively far distance from the heating/cooling sources. Moreover, it should be mentioned that the sensor used for the FLIR-ONE device is the same type as the one used in this study. Hence, no difference in terms of thermography accuracy will be imposed by

using FLIR-ONE along with the proposed IoT prototype. Based on the latter assumption, the thermography data about the surfaces for which the MRT values remained relatively constant were collected and sent to the cloud server only once for the entire period of each round of the experiment. For surfaces that were undergoing noticeable changes in terms of MRT values, the required thermography data were collected and updated at 5min intervals.

2.5.3 Results

To verify the functional requirements of the implemented system, the following tests were carried out. The validity of HTTP responses, i.e., queries made within the game engine to retrieve live sensor data and calculated thermal comfort variables from the cloud database, was confirmed through numerous tests using the Postman (Postman, 2021) platform and concurrent observations over the cloud database interface and the corresponding visualizations within the virtual environment. Additionally, to inspect the quality of virtual representations, extensive walkthroughs were carried out within the VR environment to find possible flaws.

To evaluate the system's accuracy in terms of calculating PMV and PPD indices, as its core and most important functionality, numerous tests were conducted by comparing the records of PMV and PPD values computed on the edge device, with the values obtained from the CBE online thermal comfort tool. For those tests, the monitoring data (i.e., average values of room temperature, relative humidity, air velocity, and MRT) were manually extracted from the cloud storage and used in the CBE online tool. The results showed the PMV/PPD indices calculated within the runtime of the developed VR application are the same as the values calculated using the CBE web tool. An example of such comparisons is depicted in Figure 2.9.

To validate the consistency of the system output with actual thermal sensations, the participant was asked to describe his current thermal sensations, at 5min intervals (after each round of updating MRT values). The participant's ratings were obtained with reference to the seven-point ASHRAE thermal sensation scale described in Section 2.2.1. The participant's responses were used as reference points (ground truth) to validate the overall effectiveness of the system.

Figure 2.10, Table 2.2, and Table 2.3 represent the thermal comfort monitoring data for the first and second scenarios, which were observed during the experiments using the implemented system. The numbers in the figure and tables indicate the average values of each variable at each 5*min* interval. In both scenarios, the initial state begins with a neutral thermal sensation for the participant. The results in the tables show that by initiating the sources of heat generation, room temperature, and MRT values start to change more evidently compared to other parameters.

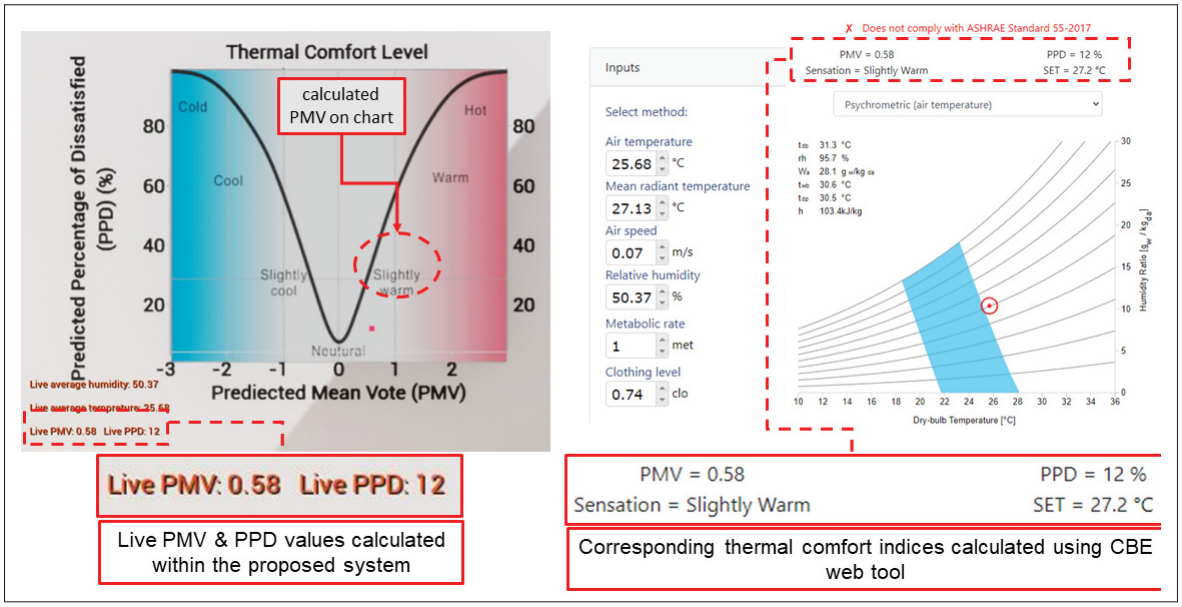


Figure 2.9 Comparison of system output with CBE tool calculated values

The changing values of air temperature and MRT parameters, and the corresponding calculated PMV values, as depicted in Figure 2.10, reveal a clear trend of continuous increase in PMV before the AC was turned on and the heater's thermostat was set back to its initial conditions. This is in complete accordance with what had been expected for each scenario. Hence, considering the apartment's small size and its design layout (the kitchen is open to the living room), both scenarios proved compelling in terms of producing slightly warm sensations before and restoring neutral sensations after turning on the AC.

In addition to the system output, Table 2.2, Table 2.3 include the data on the participant's ratings. As can be seen, the participant's responses showed convincing accordance between

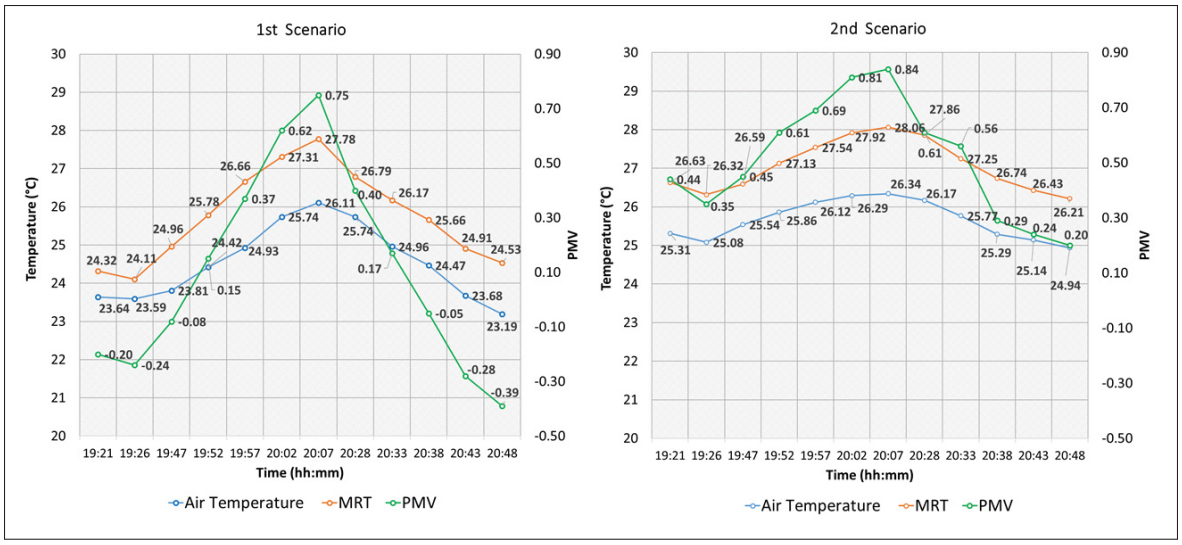


Figure 2.10 Monitored trends of changes in thermal comfort parameters and calculated values of PMV in both scenarios

the thermal sensations that he was actually experiencing during the experiment and the average PMV values computed at each $5min$ interval. In both scenarios, however, a case of incompliance was observed when the computed PMV values were close to the boundary value between the "neutral" and "slightly warm" sensation regions (i.e., PMV = 0.5).

2.5.4 Discussion and limitations

The observed inconsistencies between the participant's responses and the system output observed during the experiments can be explained by the aforementioned assumptions that were made for the sake of simplifying MRT calculations in this case study. Those incompliant cases can be also related to the accuracy of the sensors used in the proposed IoT architecture. For example, considering the $\pm 0.5 m/s$ accuracy of the wind sensor used in this case study (ASHRAE 55 recommendation: $0.05 m/s$), inaccurate readings of the air speed can significantly affect the reliability of the system's output for other scenarios where conditioning of the building space results in air flows with excessive variability in direction. Another important issue can be attributed to the radiometric accuracy of the low-cost thermal imaging module used for MRT calculations, which could be significantly problematic when the distance between the room

Table 2.2 Monitoring data and participant ratings for the first scenario

Experiment phase	**Time**	**Air Temp. ($^{\circ}C$)**	**MRT ($^{\circ}C$)**	**Air Velocity (m/s)**	**Relative Humidity (%)**	**PMV**	**Ratings**	**Compliancy**
Initial steady state (heater set at 23.5$^{\circ}C$	19:21	23.64	24.32	0.07	50.49	−0.20	neutral	✓
-	19:26	23.59	24.11	0.06	50.07	−0.24	neutral	✓
Heater thermostat set at 26$^{\circ}C$	19:47	23.81	24.96	0.06	49.93	−0.08	neutral	✓
-	19:52	24.42	25.78	0.06	50.11	0.15	neutral	✓
-	19:57	24.93	26.66	0.08	50.04	0.37	neutral	✓
-	20:02	25.74	27.31	0.07	49.97	0.62	slightly warm	✓
-	20:07	26.11	27.78	0.05	50.09	0.75	slightly warm	✓
Turning AC on at 22$^{\circ}C$	20:28	25.74	26.79	0.13	50.01	0.40	slightly warm	×
-	20:33	24.96	26.17	0.13	49.88	0.17	neutral	✓
-	20:38	24.47	25.66	0.15	49.58	−0.05	neutral	✓
-	20:43	23.68	24.91	0.14	49.63	−0.28	neutral	✓
-	20:48	23.19	24.53	0.13	49.52	−0.39	neutral	✓

surfaces and the module is close to or more than $5m$ (see Section 2.2.4). In this light, although the implemented system yielded relatively consistent results for the reasonably dynamic, yet moderate thermal conditions described in the case study, quantification of the system's output sensitivity to the bias of the readings obtained from the low-cost sensors suggested in this study (particularly for air speed measurement) is an important issue for future research. Accordingly, further evaluations under various experimental settings should be conducted to justify the system's accuracy in various prospective conditions.

To clarify the rationale behind the design of the experimental evaluations conducted with the participant, it should be remembered that the principal objective of this study (i.e., developing an immersive VR application which uses BIM and IoT data for live monitoring of thermal comfort conditions based on the PMV-PPD model) is based on the key assumption that the accuracy of the PMV/PPD indices are acceptable in terms of reflecting the thermal comfort experience of

Table 2.3 Monitoring data and participant ratings for the second scenario

Experiment phase	**Time**	**Air Temp. ($^{\circ}C$)**	**MRT ($^{\circ}C$)**	**Air Velocity (m/s)**	**Relative Humidity (%)**	**PMV**	**Ratings**	**Compliancy**
Initial steady state (heater set at 25$^{\circ}C$)	19:54	25.31	26.63	0.08	50.14	0.44	neutral	✓
-	19:59	25.08	26.32	0.08	49.88	0.35	neutral	✓
Turning kitchen oven on to 200$^{\circ}C$	20:21	25.54	26.59	0.09	50.11	0.45	neutral	✓
-	20:26	25.86	27.13	0.07	49.78	0.61	neutral	✓
-	20:31	26.12	27.54	0.09	49.92	0.69	slightly warm	✓
-	20:36	26.29	27.92	0.08	50.27	0.81	slightly warm	✓
-	20:41	26.34	28.06	0.07	50.02	0.84	slightly warm	✓
Turning AC on, set to 22$^{\circ}C$	21:02	26.17	27.86	0.14	49.86	0.61	slightly warm	×
-	21:07	25.77	27.25	0.12	49.99	0.56	neutral	✓
-	21:12	25.29	26.74	0.14	49.93	0.29	neutral	✓
-	21:17	25.14	26.43	0.13	49.87	0.24	neutral	✓
-	21:22	24.94	26.21	0.12	50.08	0.20	neutral	✓

occupants in regular building enclosures, i.e., indoor spaces with common geometric attributes with a moderately dynamic thermal environment. Hence, it is important to stress that the in-situ observations reported in Table 2.2, Table 2.3 were not meant to validate the accuracy of the PMV-PPD model. In fact, although PMV and PPD are two widely used thermal comfort indices recommended by the ASHRAE 55 standard, their performance remains open to dispute (the interested reader is referred to examples of the recent studies (Cheung, Schiavon, Parkinson, Li & Brager, 2019; Wang *et al.*, 2019; Ngarambe, Yun & Santamouris, 2020) in which the prediction accuracy of the PMV and PPD indices were discussed). In this light, it should be noted that for the present study the participant's responses were used as an indicator to check the consistency between the reflections of the thermal comfort conditions experienced by the participant within the physical environment and those calculated within the proposed system for

the virtual twin of the same physical space. In other words, the main objective of asking the participant about in-situ thermal experiences was to evaluate the consistency of the proposed tool by showing the degree to which the PMV-PPD values virtually represented within the immersive VR environment are consistent with personal reflections of the end-users who do not possess a profound background knowledge in thermal comfort assessment models/techniques (e.g., a building interior layout design specialist with an artistic background).

Finally, the authors acknowledge the fact that the overall validation of the system's consistency, i.e., including the validation of the used metrics, should be performed by including more participants with different backgrounds and individual characteristics. This is an important direction for future research because previous studies have shown the experienced levels of thermal comfort can be influenced based on individuals' characteristics such as age and sex (Wang *et al.*, 2019; Karjalainen, 2007; Cao, Lian, Du, Miyazaki & Bao, 2021). Hence, it is essentially important to further evaluate the validity of the system's consistency through future research efforts by including a diverse group of participants in the expert

2.5.5 Practical implications

Using the proposed system, facility operators and experts can intuitively monitor the complex and dynamic data that is associated with actual thermal comfort conditions. Without the need to be present in the physical space, the system allows users to navigate through the virtual spaces and observe actual thermal sensations at each spot in a real-time manner. By modifying the parameters from the user interface, they can perform different what-if scenarios and observe the outcomes based on the live monitoring data streaming from the IoT nodes installed within the building. A prospective scenario in which the system can be used in practice is for interior layout planning, where designers can be alerted to the spots that are flagged with discomfort signs.

One of the key advantages of using the proposed system lies in its capability to provide and integrated view of various building information available in BIM models (e.g., type of object material) and live IoT data. An interesting example that shows the value of the synergistic

benefits of combining BIM and IoT data in an immersive VR environment is when subsequent to identifying the spots with thermal discomfort, a building inspector wants to check for the possible heat sources located on the room surfaces causing the discomfort. Although thermal images can be used to immediately locate the spots with higher contrasts, an in-depth analysis of thermography data often requires contextual information about the scene. Without such additional information, in many circumstances, the visual contrasts in thermal images might lead to improper interpretations, as for example when the target object is made from a poor emitter material (e.g., steel, aluminum) and the visible contrasts corresponding to that object are caused by the reflections from surroundings. In such cases, using the proposed system the user can easily look for radiation abnormalities by walking through the surfaces overlaid with thermal images. Whenever needed, the user can remove the thermography layers to view the actual building objects and perform closer inspections by querying the contextual information available in the BIM data (e.g., type of material).

Thanks to the dropped cost of high-performance graphics processing units and availability of reliable development tools to import building geometry, as well as other information, such as materials, cameras, lights, etc., from various BIM authoring tools into VR development environments, will not be a serious barrier to the effective utilization of the proposed system.

However, the authors acknowledge the difficulty of setting up the monitoring environment, as a potential barrier to the efficient implementation of the proposed system in complex real-world settings. One issue in this regard refers to the placement of the sensors in the building spaces. While the system's applicability is less influenced in small spaces with simple geometric attributes (e.g., the experiment room in this study), the layout of the sensors' placement can leave a significant impact for larger monitoring environments containing more complex geometric attributes. Hence, for spaces that are more likely to yield variant distributions of thermal conditions, further research is needed with regard to the optimization of the sensor placement. BIM-based simulations and their combination with VR are promising in this regard as it has proved effective in previous studies (Motamedi *et al.*, 2017; Albahri & Hammad, 2017).

Another practical issue refers to the proposed method for the semi-automated registration of the thermal images. The difficulties of installing LED markers and extracting the markers' center coordinates through iteration steps, for the spaces with complex geometric attributes (e.g., rooms containing curved walls or ceilings with numerous indentations) would decrease the willingness of industry practitioners to use the system. To address this issue future studies will be conducted to investigate a fully-automated and non-intrusive registration of thermal images to the BIM's 3D objects that are represented in the game engine environment. In this respect, the study reported in (Sels *et al.*, 2019) provides promising insights as it suggested an automated method for the projection of 2D infrared images on 3D models, without the need to use additional hardware such as depth and RGB sensors.

2.6 Conclusion and future directions

By integrating BIM and live IoT data within game engine environments, this paper presented a novel method for real-time monitoring of thermal comfort conditions within the virtual twins of the regular building enclosures. To evaluate the overall applicability of the proposed system, a VR-enabled interface was developed for a widely available VR headset.

The main contributions of this research are as follows: (i) Investigating the synergistic benefits of BIM, IoT, and VR to effectively monitor different variables that influence thermal comfort for occupants based on PMV-PPD model; (ii) Proposing an IoT prototype based on a cost-effective thermal imaging sensor, as well as designing a semi-automated computer vision method for processing raw thermal images to calculate MRT values on affordable edge computing devices, in a (near) real-time manner; (iii) Evaluating the reliability of the existing open-source software tools and libraries for developing functional prototypes of the proposed system architecture, and providing a detailed description of the system implementation process.

The present study will serve as a base for further developments and improvements considered by the authors. In particular, based on the system architecture and development pipeline proposed in this work, the authors will investigate an augmented reality application by which thermal

comfort-related information can be overlaid on the real-world scenes of indoor building spaces. Moreover, to address some of the key challenges of integrating BIM and IoT data in an efficient and real-time manner, the authors will investigate a data integration framework that will be based on the principles of linked data and the application of formal ontologies and graph database systems.

CHAPTER 3

NEURAL SEMANTIC TAGGING FOR NATURAL LANGUAGE-BASED SEARCH IN BUILDING INFORMATION MODELS: IMPLICATIONS FOR PRACTICE

Mehrzad Shahinmoghadam[1] , Samira Ebrahimi Kahou[2] , Ali Motamedi[1]

[1] Department of Construction Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Department of Software Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article published in *Elsevier's Computers in Industry*, in December 2023.

Abstract

While the adoption of open Building Information Modeling (open BIM) standards continues to grow, the inherent complexity and multifaceted nature of the built asset lifecycle data present a critical bottleneck for effective information retrieval. To address this challenge, the research community has started to investigate advanced natural language-based search for building information models. However, the accelerated pace of advancements in deep learning-based natural language processing research has introduced a complex landscape for domain-specific applications, making it challenging to navigate through various design choices that accommodate an effective balance between prediction accuracy and the accompanying computational costs. This study focuses on the semantic tagging of user queries, which is a cardinal task for the identification and classification of references related to building entities and their specific descriptors. To foster adaptability across various applications and disciplines, a semantic annotation scheme is introduced that is firmly rooted in the Industry Foundation Classes (IFC) schema. By taking a comparative approach, we conducted a series of experiments to identify the strengths and weaknesses of traditional and emergent deep learning architectures for the task at hand. Our findings underscore the critical importance of domain-specific and context-dependent embedding learning for the effective extraction of building entities and their respective descriptions.

3.1 Introduction

The Architecture, Engineering, Construction, and Facility Management (AEC-FM) industry is currently undergoing a profound technological transformation, fueled by rapid advancements in Building Information Modeling (BIM) as a revolutionary paradigm for managing information regarding construction projects. Through BIM, an object-oriented digital representation of the physical and functional characteristics of a facility is created, that facilitates the exchange and interoperability of information across various digital platforms(Sacks, Eastman, Lee & Teicholz, 2018; Shahinmoghadam & Motamedi, 2019).

Thanks to the fast-growing transition to open BIM standards, notably the IFC schema, the industry has been witnessing an unprecedented exchange and interoperability of building information in digital formats (Jiang, Jiang, Han, Wu & Wang, 2019). However, with the abundance of data comes the challenge of effectively retrieving and utilizing the required information at different stages of the built asset lifecycle. Given the intricate and multifaceted nature of such data, there is a pressing need for intuitive and expressive mechanisms to facilitate information retrieval from building information models. An illustrative example in this context is the hypothetical scenario, where the energy efficiency ratings of the windows of particular types and dimensions that are located in specific spaces or levels within a high-rise residential complex are required to be assessed. Performing such queries within conventional BIM tools requires the user to be familiar with the underlying structure of object descriptions, as well as accurate naming conventions for the object types and property sets that are of interest to the user. As a result, the user, most often, needs to first manually navigate within the model to gain such information to be able to perform the desired queries. Even when the user is provided with such information ahead of time, the assignment of search filters often involves multiple manual steps, which can become notably time-consuming and prone to errors.

Natural Language Processing (NLP) offers a promising avenue to address this need by developing systems that can process user queries and extract pertinent information from vast building models. However, the rapid pace of advancements in the fields of NLP and deep learning has

resulted in a vast array of methodologies and architectures. This profusion of system design choices presents a critical challenge for AEC-FM researchers and practitioners looking to build NLP systems that are coupled with BIM environments. Moreover, given the relatively slower rate of technology adoption in the AEC-FM industry, it is crucial to select design choices that deliver an optimal balance between prediction accuracy and the computational costs associated with the chosen methods.

Recent reviews of the existing literature (e.g., (Zabin, González, Zou & Amor, 2022; Saka *et al.*, 2023)), point to the fact that despite the growing interests within the research community, the application of deep learning-based NLP in the realm of building information modeling is still in its early stages. This is while the unique structure and domain-specific terminologies that characterize building information models necessitate a carefully thought-out and adapted approach when implementing NLP techniques.

However, upon reviewing recent works that focused on the development of NLP-based information retrieval systems for BIM, it becomes apparent that little to no emphasis has been placed on addressing computational performance considerations alongside the prediction capabilities of the chosen architectures. In particular, researchers in the AEC-FM domain have increasingly gravitated toward the adoption of transformers (Vaswani *et al.*, 2017) and other emergent deep learning architectures for a variety of related tasks such as query classification (Wang *et al.*, 2022d; Zheng & Fischer, 2023) or entity extraction (Zhou, Zheng, Lin & Lu, 2022; Li *et al.*, 2021; Jeon, Lee, Yang & Jeong, 2022a). Nevertheless, the results reported in the AEC-FM body of research (e.g., (Wang *et al.*, 2022d)) confirm that in the context of information retrieval from building information models, traditional architectures such as Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997), appear to exhibit prediction performances that are in close proximity to their more recent, yet, significantly resource-intensive counterparts. Hence, a critical gap in the literature is an in-depth comparison of traditional and emergent NLP approaches that simultaneously weigh factors like model training challenges, inference latency, and model memory size. Consideration of these factors is of essential importance as they can significantly influence the applicability of the findings for real-world deployment

of these systems in the AEC-FM industry, which is known for its cautious pace in adopting innovative technologies.

Regardless of the techniques and methodologies adopted, a crucial step in developing an effective NLP-based system involves the task of Named Entity Recognition (NER) (Li, Sun, Han & Li, 2020), also known as semantic tagging or token classification. NER is concerned with identifying and classifying references to entities in text. These entities, in the context of building information, might be diverse building components/systems and their specific descriptors (e.g., object's location, quantity, or engineering specifications).

In response to the above-mentioned shortcomings, and with a particular focus on the NER task, the central research question that this study attempts to provide an answer to can be articulated as follows: Considering the typically concise nature of user queries, does the transformer architecture offer a solution that holds a discernible advantage over traditional architectures in tackling the problem at hand, i.e., extraction of semantic entities for building information retrieval? Further, if there is an advantage, is it substantial enough to justify the additional computational overhead that is inherently associated with more advanced and sophisticated deep learning architectures? Hence, the main objective of this study is to provide a comparative evaluation of LSTM and transformer-based deep learning architectures for a NER system that is tailored for natural language-based search in BIM. Moreover, the experiments conducted in this work demonstrate the impact of other related design decisions (e.g., embedding learning method) on the overall effectiveness of the intended NER system. Finally, to ensure broad adaptability across various disciplines and applications within the AEC-FM industry, this study investigates a semantic annotation scheme that is deeply rooted in the IFC schema, which is the most adopted open BIM standard.

3.2 Background

3.2.1 Open BIM, IFC schema, and NLP

Published and maintained by buildingSMART international organization, and registered as an ISO standard (ISO 16739-1, 2024), the IFC schema enjoys increasing adoption in industry practices. The object-oriented and non-proprietary nature of IFC makes it ideally suited to be utilized to serve a diverse array of applications across various domains (Motamedi, Soltani, Setayeshgar & Hammad, 2016).

The open nature of the IFC schema, coupled with its comprehensiveness, has stimulated the interest within the AEC-FM research community to investigate its potential in conjunction with data-driven artificial intelligence applications. To date, numerous studies have focused on leveraging the richness of built asset data represented in the IFC format for applications such as intelligent building design (Ghannad & Lee, 2021), defect detection (Sresakoolchai & Kaewunruen, 2021), point-cloud-based object detection (Koo, Jung, Yu & Kim, 2021; Seydgar, Motamedi & Poirier, 2022), and knowledge graph-based semantic representation (Shahinmoghadam *et al.*, 2022a).

Similarly, numerous studies have investigated IFC-based representations along with NLP techniques for various applications such as automated code checking (Zheng, Zhou, Lu & Lin, 2022b; Zhang & El-Gohary, 2023), facility information management (Xie *et al.*, 2019), processing of change requests (Dawood, Siddle & Dawood, 2019), and information retrieval or BIM model querying (Wang, Issa & Anumba, 2022b; Yin *et al.*, 2023b). Essentially, a common focus of these studies is to establish lexical and/or semantic correspondences between textual data (e.g., regulatory codes, engineering specifications) and the entities that are digitally represented within building information models. It is in this context that the IFC schema's well-structured classification taxonomies and rich semantic structure can be leveraged to draw meaningful connections between the elements of natural language text and the corresponding BIM entities.

3.2.2 The role of semantic entity extraction

To effectively establish semantic correspondences between the textual descriptions of the built asset elements and their properties on one hand, and the IFC classes and attributes on the other, a key preliminary step is to extract the pertinent entities, such as names of building components, materials, specifications, and locations, from the text input. This entity extraction process, referred to as the Named Entity Recognition (NER) task in the domain of NLP, is crucial for creating a semantically enriched interface between natural language representations and complex building information data structures. Through NER, unstructured text is analyzed, and specific segments are categorized into predefined classes, which can then be mapped to the corresponding elements within the IFC schema. Figure 3.1 highlights the role of the semantic entity extraction (tagging) step in a general NLP-based search pipeline for BIM databases. In the context of our study, the NLP-based pipeline is responsible for transforming raw natural language queries into formal statements, e.g., SQL queries, which can then interact with a conventional database system. As can be seen, the semantic tagging step serves as a foundational component of the search mechanism by identifying and classifying relevant entities from raw input queries. In this example, the comparison of the tokens highlighted within the user's raw input and the corresponding SQL query explains how the quality of extracted entities is pivotal in formulating valid formal query statements. Finally, it is worth noting that the theoretical pipeline denoted in Figure 3.1 is an example of the classical semantic parsing approach and emerging methods aim to directly translate natural language input into formal queries without explicit intermediate steps. However, the information provided by semantic tagging remains of great value even in these advanced pipelines as it contributes to the refinement of the generated query, e.g., when the generative model lacks familiarity with the explicit names of the entities within the IFC schema.

3.2.3 Deep learning for semantic sequence tagging

Deep learning has been the dominant approach in the landscape of NLP research, including NER. Historically, Recurrent Neural Networks (RNNs), in particular, LSTM networks and their bidirectional (Schuster & Paliwal, 1997) variants, i.e., BiLSTM, have been the architectures

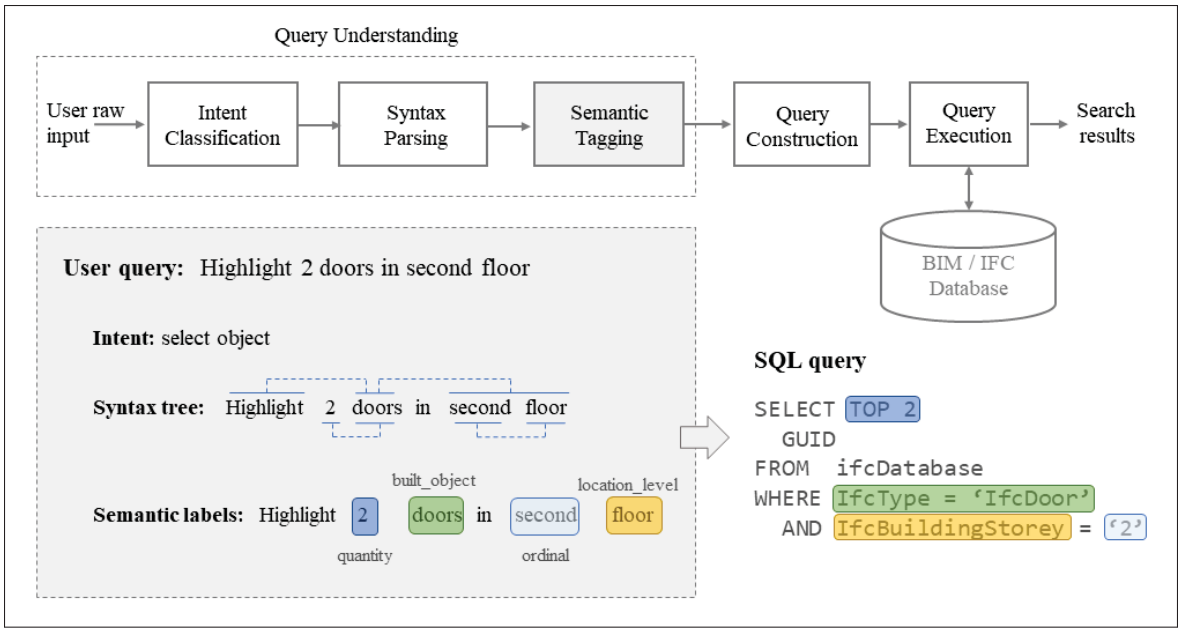


Figure 3.1 Example of a theoretical NLP-based search pipeline for BIM databases

of choice for tackling sequence processing tasks (Vaswani *et al.*, 2017). When employing RNN-based architectures for NLP tasks, it is a common practice to make use of pre-trained word embeddings. GloVe (Pennington *et al.*, 2014a) and word2vec (Mikolov *et al.*, 2013) are among the most widely used algorithms for unsupervised representation learning of words, i.e., encapsulating semantic and syntactic relationships between words in a continuous vector space.

To address the shortcomings of the RNN-based architectures, particularly their inefficiency in handling long-range dependencies, models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al.*, 2018) have been developed and become established as the state-of-the-art in NLP research. These emergent architectures are particularly proficient at capturing contextual information through attention mechanisms (Vaswani *et al.*, 2017). Transformer-based pre-trained language models such as BERT benefit from a built-in word embedding learning layer. The pre-trained embeddings, learned during the pre-training stage on a vast corpus of text, allow these models to start with a rich contextual understanding of the language.

3.2.4 Knowledge gap

Although the sophisticated architecture of transformer-based pre-trained models gives them an edge over the traditional RNN-based models, the higher demand for computational resources can limit their practical applications. Consequently, researchers have been applying techniques such as knowledge distillation (Hinton, Vinyals & Dean, 2015) to reduce computational requirements both for training (Sanh, Debut, Chaumond & Wolf, 2019) and inference (Liu *et al.*, 2020b). Within the AEC-FM domain, recent NLP research efforts have predominantly emphasized the predictive accuracy of transformer-based models, often overlooking the computational feasibility considerations (Wang *et al.*, 2022d; Jeon *et al.*, 2022a; Wu, Shen, Lin, Li & Li, 2021). Another key shortcoming of these studies is the lack of sufficient attention to the careful design of RNN-based competitor models. For instance, key considerations like the vocabulary size and vector dimensions of the embedding layer, or optimal sequence tokenization strategies, have not been given due importance. This oversight is particularly concerning given recent findings that suggest the superiority of transformer models over RNNs can be less definitive when dealing with smaller, domain-specific datasets (Ezen-Can, 2020). In this regard, a notable gap in the literature is an in-depth analysis of the trade-offs involved in adopting sophisticated deep learning architectures for natural language-based search and information retrieval systems that are tailored for BIM applications. Given this context, the central goal of this study is to undertake extensive experimentation and analysis to offer pragmatic insights on viable design choices that align with the practical needs and constraints of the AEC-FM industry. This involves a balanced consideration of both the predictive capacities and computational efficiencies of the competing neural network architectures.

3.3 Methods

The research framework and main steps taken to meet the objectives of this study are illustrated in Figure 3.2. To provide a comprehensive overview of the methods employed in this work, the remainder of this section elaborates on the description of the main components of the proposed framework.

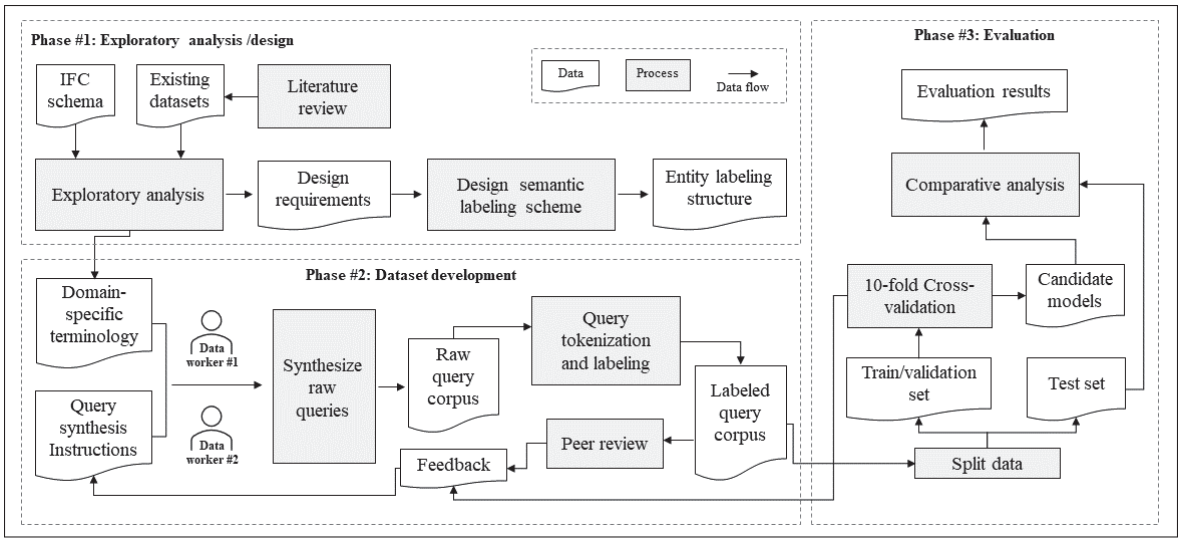


Figure 3.2 Research framework for semantic tagging of queries

3.3.1 Semantic labeling scheme

The term "semantic labeling scheme" refers to a structured method designed to categorize the elements (tokens) of a textual input (query statements) based on their semantics, i.e. meaning or context. This scheme serves as a robust foundation for systematic identification and annotation of those entities that hold semantic importance, thereby facilitating an accurate interpretation of the user queries. Given the critical role that this underlying structure plays in the successful implementation of a natural language-based search system, significant attention was devoted to the selection of reliable resources to guide the design process. In this light, our proposed labeling scheme is designed with reference to the following resources.

First, the IFC schema (version IFC-4.0.2.1 (buildingSMART, 2020)) is selected as the primary point of reference. The justification for this selection resides in the schema's comprehensiveness and maturity as a fully open-source software entity. Not only does the IFC schema provide extensive details on the formal representation of a vast array of information pertaining to building components, systems, and associated processes in a machine-readable format, but it also plays a pivotal role in facilitating interoperability, especially in the context of information retrieval from BIM models. Then, existing datasets are considered as the second category of guiding resources.

From the limited number of relevant datasets that are publicly available, the ones proposed in (Wang, Issa & Anumba, 2022a) and (Yin *et al.*, 2023a) serve as primary references for designing an effective semantic labeling scheme in the current study. Further elaboration regarding the necessary adjustments and improvements based on the existing data is provided in Section 3.4.

3.3.2 Dataset development

Following the formulation of the semantic annotation scheme, the subsequent phase entails the construction of a synthetic corpus encompassing user queries. The main deliverable of this step is a relatively large set of query statements that are represented in natural language (English). The query corpus is intended to encompass a wide array of common building objects, as well as descriptive mentions of those objects such as quantity and location information. More details regarding the scope of the developed dataset are given in Section 3.4.

Two data workers were involved in a four-month endeavor to build the base corpus of synthesized user queries. One of the data workers possessed a deep understanding of BIM concepts, with particular expertise in the IFC standard. This individual also had over five years of professional experience in the construction industry. Such expertise was a contributory element to ensure that the synthesized queries were relevant within the context of BIM. The other data worker was selected from a general engineering background, not specifically related to the AEC-FM domains. This decision was made intentionally to include input from an individual with less familiarity with BIM concepts and tools. This approach was aimed at resembling potential interactions that users with less technical expertise might have with a BIM model, thereby broadening the diversity and naturalness of the queries.

Each data worker was provided with general instructions, examples of queries, as well as a comprehensive list of building objects that were initially extracted from the IFC schema and then extended with synonyms and similar terms. To ensure the naturalness of the queries, it was encouraged to employ a conversational tone, incorporate a mix of formal and informal expressions, and avoid adhering to rigid patterns or templates. Moreover, it was communicated

that grammar rules do not need to be followed in a strict manner, as it is quite common in a real-world setting. In order to maintain a harmonious distribution of various semantic entities within the queries, periodic monitoring and analysis of data statistics are conducted throughout the corpus development phase. The final stage in the construction of the synthetic corpus involves adopting a systematic strategy to incrementally enhance the diversity and complexity of the dataset. This is vital in ensuring that the corpus is not only rich in content but also reflective of the multifaceted nature of real-world user queries. An elaborate explanation of the approach adopted to achieve this is provided in Section 3.4.

Upon finalizing the development of the query corpus, the next step involved assigning semantic labels to the entities of each sequence (query) present within the developed corpus, i.e., each entity in the sequence is tagged according to its role in the semantic structure that has been formulated based on the defined semantic labeling scheme. An essential preliminary step of this process is to transform each query into its constituent tokens. For this purpose, we utilize a pre-trained word tokenizer that is included within the NLTK suite (Bird, Klein & Loper, 2009). The choice of this specific tokenizer is based on its wide adoption and proven effectiveness in NLP tasks. To ensure the quality of the tokenization output, a manual check of all tokenized queries is conducted. This is an iterative process, aimed at aligning the tokenization output with the tokens embedded in the utilized embedding models, thereby minimizing the occurrence of out-of-vocabulary tokens in the subsequent stages of the study. Once the queries are tokenized, data workers proceed to the labeling step. In line with the majority of scholarly literature[1], our study employs the BIO notation used in CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003), which is a highly regarded benchmark dataset for the NER task. BIO notation discerns between the commencement, denoted by "B-" (short for beginning), and the continuation, denoted by "I-" (short for inside), of semantic entities, while the tokens with no semantic importance are represented by "O-" (short for outside). Finally, to ensure the robustness and reliability of the annotated dataset, each data worker was assigned the task of conducting peer reviews throughout the duration of the semantic labeling process.

[1]The body of literature related to NLP research.

3.3.3 Model development and evaluation

Drawing from the previous discussions concerning the selection of deep learning architectures in this study, the evaluation phase of the study focuses on the comparative effectiveness of RNN and transformer-based architectures. The main focus of the evaluation phase is to provide empirical evidence on the comparison between these architectures when deployed for the NER task within the context of natural language-based search in building information models. The BiLSTM architecture is selected as representative of RNN-based approaches, whereas the BERT architecture serves as the exemplar of transformer-based deep learning approaches. As opposed to standard LSTM architecture, BiLSTM has the advantage of understanding the context from both directions (left and right of a token) which makes it more similar to the bidirectional nature of the BERT-family models. While BERT-based models inherently benefit from built-in word embedding representations, implementation of an independent embedding learning layer is a prerequisite for BiLSTM models. To address this requirement, word2vec and GloVe algorithms are incorporated into our experimental framework. In addition to the embedding learning method, additional configuration parameters are taken into consideration during the design of the experiments. More details regarding these parameters will be provided in the following sections. The intention behind the inclusion of various parameters for embedding learning is to facilitate a thorough examination of how disparate design decisions for the embedding layer could impact the overall predictive performance, as well as the associated computational expense, for the subsequent NER task.

For evaluation, we use the exact-match evaluation approach (Li, Sun, Han & Li, 2022). The exact-match approach necessitates the precise determination of both the extent (boundaries) and the label (type) of an entity. With this approach, a named entity prediction is deemed correct if its outlined boundaries and its categorized type wholly correspond with the pre-established ground truth (Li *et al.*, 2022). The models' predictive performances are evaluated using three standard metrics for NER tasks, namely, precision, recall, and F1-score. Precision measures the percentage of correctly predicted entities out of all entities predicted by the model. Recall, on the other hand, measures the percentage of correctly predicted entities out of all actual entities

in the dataset. To account for label imbalance, we calculate the weighted average for each metric used (taking into consideration the number of instances for each true label in our evaluation dataset).

To present a thorough understanding of the advantages and disadvantages of RNN- and transformer-based architectures for the task at hand, the evaluations undertaken in this study entail two distinct forms of analysis. First, an assessment is carried out on the models' proficiency in accurately predicting labels across a variety of entity types, based on the semantic structure defined in this study. This evaluation facilitates the identification of areas where the models demonstrate strength, as well as potential areas for further refinement. Subsequently, a comparison of the computational performance of each architecture is undertaken by taking two important aspects: memory usage, and inference latency. This comparison enables an examination of the efficiency of each model, providing additional criteria for comparative evaluations within the context of this study.

3.4 Experiments and results

To support the reproducibility of the reported results and promote future research, all the data and code utilized in our experiments are openly available in a dedicated GitHub repository (Shahinmoghadam, 2023).

3.4.1 Data

3.4.1.1 Semantic labeling scheme

To develop the underlying semantic structure for our NER system, we first examine the IFC schema's high-level tree structure for the classification of building elements. From a quick glance at "ifcElement", which is one of the core classes within the schema, it can be observed that a wide range of common building entities are classified under two of its subclasses, namely, "ifcBuildingElement" and "ifcDistributionElement". A closer look at the element tree structure

reveals that a considerable number of subclasses are nested under a few subclasses. For example, the “IfcEnergyConversionDevice” class itself consists of twenty subclasses such as “IfcBoiler”, “IfcChiller”, and “IfcHumidifier”. Although such detailed taxonomy provides a clear and comprehensive classification for a vast array of building elements, it can increase the system design complexity, unnecessarily.

To meet an effective trade-off between the granularity level and simplicity of the system design, two main labels were considered for the underlying semantic structure that was developed in this study. The first label, “built_obj”, refers to all physical and tangible objects that are related to the structural and architectural aspects of a building, e.g. walls, doors, slabs, or stairs. The second label, “MEP_obj”, refers to a wide range of physical and tangible objects that are related to mechanical, electrical, and plumbing aspects of the building, e.g., heaters, pipes, sensors, or condensers. In addition to physical elements, we consider spatial elements such as location and containment information, which are frequently used when inquiring about building objects. Since the scope of our experiments is limited to queries within individual projects, we only consider floor (tagged with “loc_level”) and space (tagged with “loc_space”) types as named spatial elements in our tagging system. Next, a profound examination of two existing relevant datasets reveals a number of ambiguities within their semantic labeling scheme, as explained below.

First, in the labeling scheme proposed in (Wang *et al.*, 2022a), three of the used labels, namely, “ENTITY”, “TYPE”, and “OBJECT” bear significant semantic similarities which can cause confusions later throughout the system development pipeline. For example, in the case of posing a query such as: “How many casement windows exist on 2nd floor?”, based on the descriptions and examples provided in (Wang *et al.*, 2022a), it is not clear which of the three aforementioned labels should be assigned to “casement windows”. The reason for introducing “TYPE” as an independent tag, along with the “ENTITY” label, seems to be due to the fact that the researchers intended to use the information from “TYPE” entities to find corresponding element classes in the IFC schema through a synonym-based approach (Wang *et al.*, 2022b). Instead of introducing multiple labels for “TYPE” or “ENTITY” mentions, our proposition is to first identify all entity

mentions in the raw input (query) and then find correspondences with the entities as represented in the IFC schema based on a semantic similarity scoring mechanism. This approach assists in mitigating internal semantic ambiguities that could potentially emerge from an overlapping and imprecise definition of the semantic structure for the NER system. However, successful implementation of such semantic mappings requires profound investigations and falls outside the scope of the present work. The design and implementation of the mentioned semantic mapping system constitute a key aspect of our ongoing research project that is being conducted concurrently, and its output is set to be disseminated as an extension of the present work.

Second, it is observed that some general but essential semantic entities are missing from the datasets that are developed in (Wang *et al.*, 2022a) and (Yin *et al.*, 2023a). For example, the quantity of the objects and ordinal numbers are not considered to be labeled independently in the dataset. Although such descriptions can be considered as attributes of an object (i.e. providing object descriptions), it is important to extract them as independent semantic entities when identifying semantic roles of query tokens, since quantities or ordinal descriptions of an entity are not explicitly represented in the IFC schema (as pre-defined object attributes). To address such shortcomings, our developed dataset includes other semantic elements (labels) that play an essential role in the provision of an accurate understanding of natural language-based queries. These additional semantic elements are “quantity”, “ordinal number”, “number”, and “name”. An illustrative example highlighting the significance of an effective semantic labeling scheme is presented in the subsequent section.

3.4.1.2 Corpus development and annotation

Following the completion of the semantic labeling scheme, the query corpus development is undertaken. To delineate a tractable scope for query generation, first, the data workers were asked to add simple queries that intend to find, highlight, hide, or count objects of a single type. Subsequently, a systematic multi-stage increment in query complexity was initiated. In particular, at the first stage data workers were asked to incorporate new queries which include single descriptive phrases for the objects mentioned in the query. Such descriptions could refer

to the name, quantity, and location/containment information for a given object. In the next phase, new queries were added that encompassed references to multiple object types with the inclusion of simple descriptions within the query. In the final phase, more complex queries were added, featuring mentions of multiple object types, paired with diverse combinations of descriptions potentially referring to a singular or a subset of the object entities within a single query. Upon the conclusion of each phase, data workers are provided with the queries proposed by their counterparts. They are then prompted to integrate new queries that mirror the same intent but are articulated differently. The intention for this step is to collect as much variation in query phrasing as possible.

Upon the completion of the query corpus development, the semantic labeling process is undertaken over a span of one month. Using the pre-trained word tokenizer incorporated within the NLTK open-source package (Bird *et al.*, 2009), each query is broken down into individual tokens to which data workers manually assign semantic labels. Given that the creation of a custom tokenizer is not incorporated within the scope of this work, it is of essential importance to align the development of the dataset with the tokenizer that is intended to be utilized for the training of the NER model. By tailoring the dataset to fit the tokenizer, we ensure that the input delivered to the NER training algorithm corresponds with the input utilized during the embedding learning stage in the pre-training phase. Finally, as previously stated, the labeling process is subjected to a system of peer review. This cross-checking approach allows us to maintain high accuracy and consistency throughout the tokenization and dataset adjustment process.

The final version of the developed dataset contains 2380 query sentences, each of which labeled with the tags elaborated in Table 3.1. The dataset is developed to support the NER task within the context of information retrieval from BIM databases. The queries included in the dataset were designed to reflect realistic user interactions with the data related to “building” projects. The queries encompass three primary intents: selecting, deselecting, and counting building entities. Additionally, these queries involve descriptions of the attributes or spatial containment

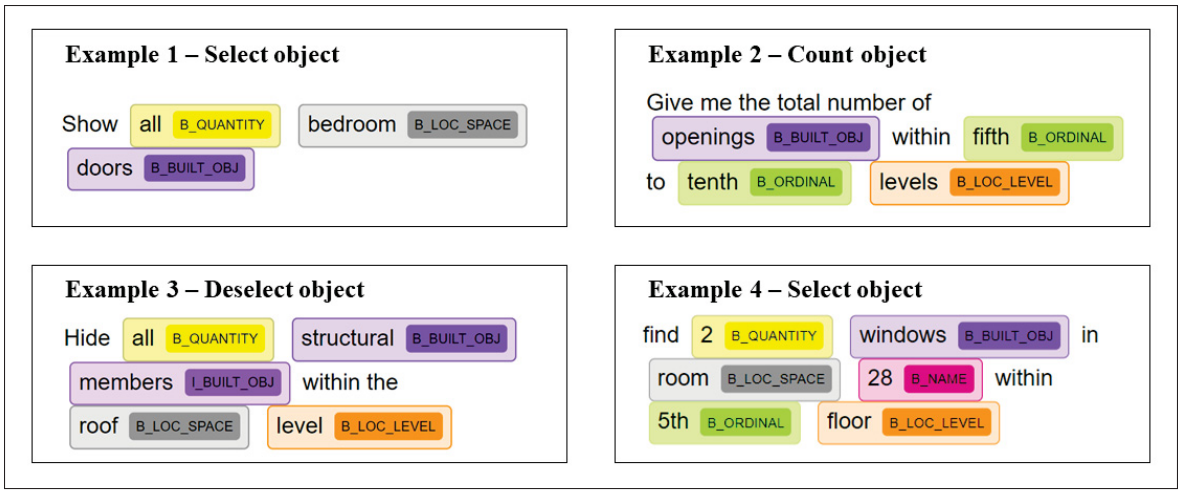


Figure 3.3 Examples of tagged queries

of the building entities. Descriptions of the semantic labels, as well as the frequency by which each label appeared in the dataset can be seen in Table 3.1.

To illustrate the application of each semantic element, examples of annotated queries are presented in Figure 3.3. "Example 1" represents queries with select intent where users seek to locate or filter specific building components or systems within a BIM model. This category of queries is foundational in user interactions with BIM databases, as it often forms the starting point for numerous other user requests. For example, inquiries about the properties of specific objects in a BIM database commence with retrieving the unique identifiers of the described objects and performing follow-up queries about the retrieved objects. "Example 2" represents queries with deselect intent that involve requests to temporarily remove certain elements from view in the BIM model. "Example 3" represents those queries that users might use to ask for the count of different building elements that meet the specific descriptions provided in the query. Finally, "Example 4" underscores the significance of devising an effective underlying semantic structure, by demonstrating a hypothetical case where numeric tokens, despite their pronounced lexical similarities, can carry diverse semantic roles in varying contexts. In this example, while the numeric token "2" refers to the quantity of an object, the token "28" describes the "name" field of the entity in the IFC schema structure. In particular, when translating this query into

Table 3.1 Tag description and frequency of appearances in the developed dataset

Semantic tag	Quantity	Tag description
B_mep_obj	1763	First token of a phrase specifying a mechanical, electrical, or plumping object.
B_quantity	1511	First token of a phrase specifying a quantity.
I_mep_obj	101	Subsequent tokens in a phrase specifying a mechanical, electrical, or plumping object.
B_loc_space	880	First token of a phrase specifying a spatial element (except levels/floors).
B_built_obj	810	First token of a phrase specifying an architectural or structural element.
B_name	700	First token of a phrase specifying the name attribute of an IFC entity, e.g., room name.
B_loc_level	625	First token of a phrase specifying an IFC level (IfcBuildingStorey).
B_ordinal	375	First token of a phrase specifying an ordinal number.
B_number	250	First token of a phrase specifying a number other than an object quantity or ordinal numbers.
I_built_obj	235	Subsequent tokens in a phrase specifying an architectural or structural element.
I_loc_space	101	Subsequent tokens in a phrase specifying a spatial element (except levels/floors).
I_ordinal	61	Subsequent tokens in a phrase specifying an ordinal number.

a formal language such as SQL, “28” should target the data field associated with the “name” attribute in the database. Meanwhile, “2” serves as a constraint in the SQL statement, specifying the number of objects to be retrieved.

3.4.2 Model development

Utilizing our developed dataset (comprising 2380 queries), seven groups of experiments are conducted as follows: In light of the dataset’s limited size, a 10-fold cross-validation strategy is adopted to reinforce the robustness of evaluation results. Moreover, to accommodate for potential random variations and facilitate the reproducibility of our results, the experiments are repeated using three distinct random seeds. In particular, for each random seed, the entire developed dataset is split into 10 distinct subsets. Subsequently, one of the folds is used as a test set, another one as a validation set, and the remaining folds serve as a training set, at each round of cross-validation. By repeating this process with a new test and validation fold, it is ensured that each data point would appear exactly once in the test set across the 10 folds. Thus, for each group of experiments, a total of 30 unique models are trained and evaluated using a distinct combination of initialization seed and data fold.

For all experiments, the model training phase is set for a potential 100 epochs. However, to reduce the over-fitting risks associated with smaller datasets, an early stopping mechanism anchored on the validation loss, is incorporated within the validation step. In the context of our experiments, validation loss refers to the disparity between predicted labels and true labels, for all the tokens within each validation fold. We utilize the cross-entropy loss function, as it is a widely used method for multi-class classification tasks such as token classification. In mathematical terms, the cross-entropy loss $\mathcal{L}$ for a single token at position t is computed as:

$$\mathcal{L}(y_t, \hat{y}_t) = -\log\left(\frac{\exp(\hat{y}_{t,y_t})}{\sum_j \exp(\hat{y}_{t,j})}\right) \tag{3.1}$$

Here, y_t and $\hat{y}_t$ denote the true label and the predicted label probabilities for the token at position t, respectively. $\hat{y}_{t,j}$ is the predicted probability for the label j at position t. The overall validation loss for each fold is then computed as the mean loss over all tokens within all query sequences in a given validation fold:

$$\mathcal{L}_{\text{val}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(y_{i,t}, \hat{y}_{i,t}) \tag{3.2}$$

In the above equation, N is the number of examples in the validation fold, and T is the number of tokens in each query sequence.

Our early stopping mechanism ensures that the training process is halted if no improvement in validation loss is observed across epochs. Given the small size of the dataset and to ensure model generalizability, we adopted a rigorous "zero patience" strategy: stopping at any epoch that does not contribute to the model's improvement. Taking one of the iterations in Experiment 7 as an example (see Figure 3.4), the training is halted before reaching 100 epochs, as validation loss stops decreasing by the end of the 26^{th} epoch. In fact, the evolution of training and validation losses for Experiment 7, as illustrated in Figure 3.5, shows that the average number of total training epochs across all 30 iterations is less than 11, with a total of 322 training epochs.

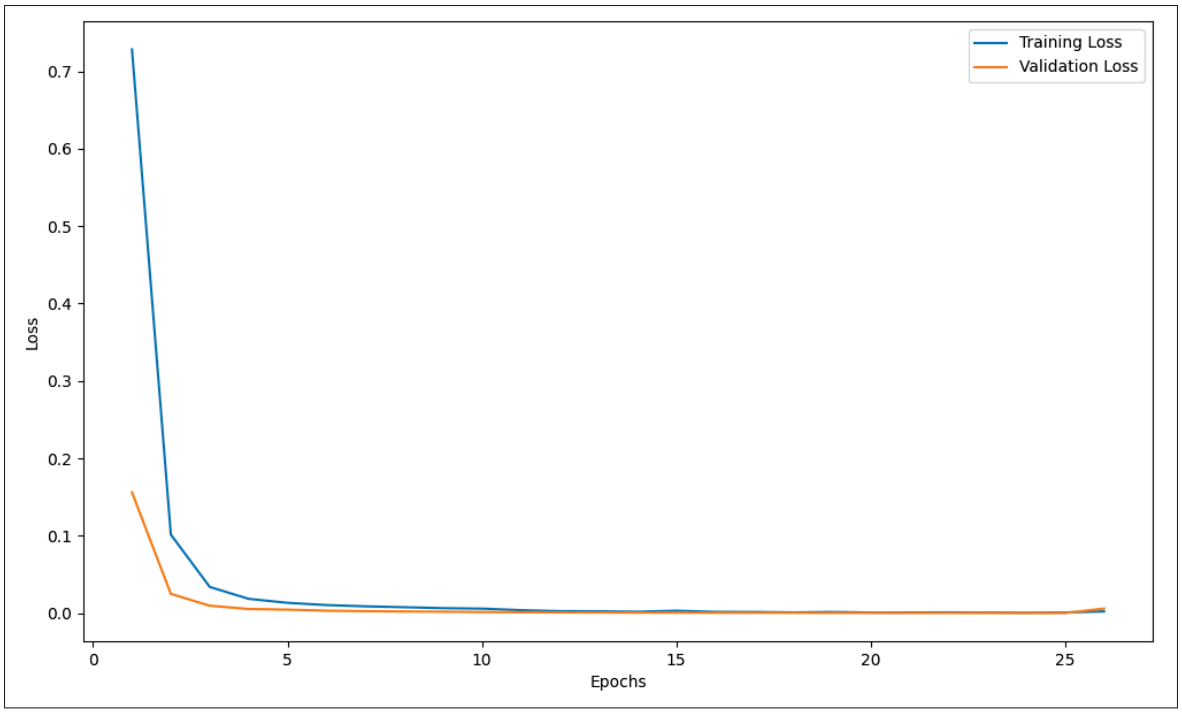


Figure 3.4 Example of early stopping (training stops after 26 epochs in Experiment 7, 7^{th} fold with 1^{st} random seed)

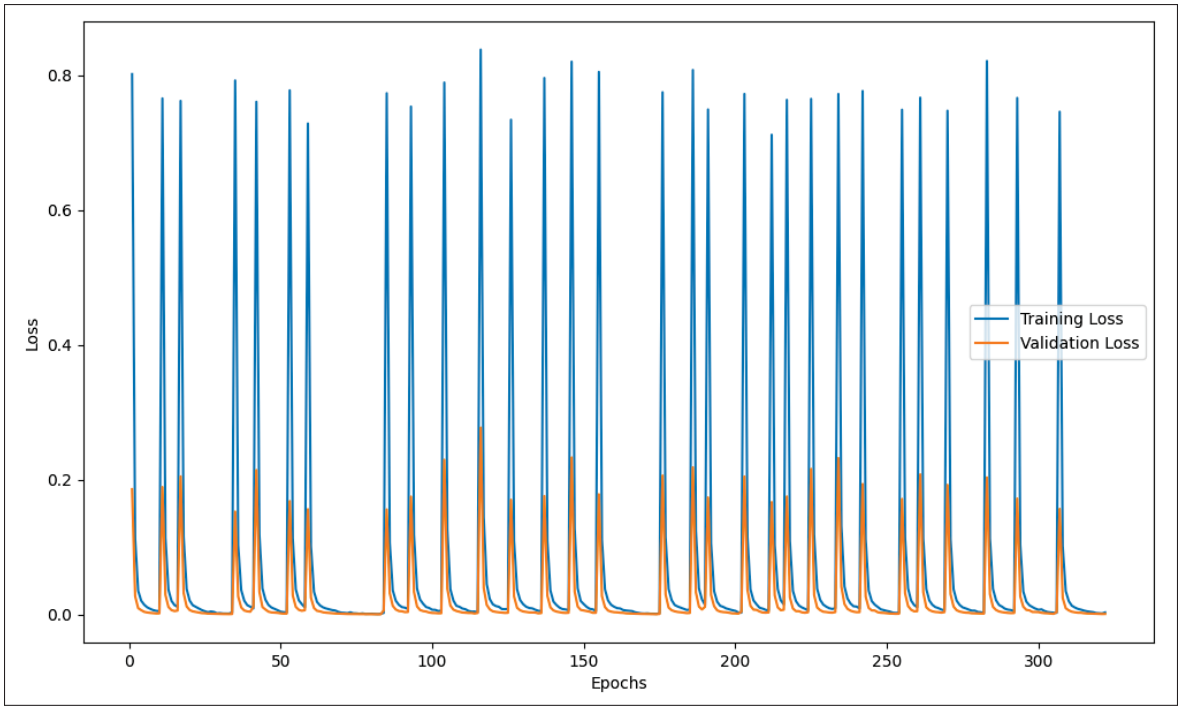


Figure 3.5 Example of the evolution of training and validation losses across all training epochs (Experiment 7)

Table 3.2 Summary of experimental results

Experiment	**Deep neural network architecture**	**Embedding learning model**	**Embedding vector dimension**	**Precision**	**Recall**	**F1-score**
Experiment 1	BiLSTM	word2vec	300	0.9915 (±0.0094)	0.9907 (±0.0113)	0.9910 (±0.0104)
Experiment 2	BiLSTM	Glove-01	100	0.9978 (±0.0073)	0.9974 (±0.0085)	0.9976 (±0.0079)
Experiment 3	BiLSTM	Glove-01	300	0.9991 (±0.0026)	0.9985 (±0.0050)	0.9988 (±0.0037)
Experiment 4	BiLSTM	Glove-02	300	**0.9997** (±0.0010)	**0.9991** (±0.0039)	**0.9994** (±0.0024)
Experiment 5	BiLSTM	Glove-03	300	0.9981 (±0.0036)	0.9978 (±0.0050)	0.9979 (±0.0042)
Experiment 6	BERT (Base)	built-in	768	0.9969 (±0.0020)	0.9947 (±0.0033)	0.9958 (±0.0026)
Experiment 7	BERT (DistilBERT)	built-in	768	0.9978 (±0.0019)	0.9960 (±0.0029)	0.9969 (±0.0023)

Based on the approach described above, seven sets of experiments are conducted, entailing the training of 30 distinct models for each architecture configuration (training a total of 210 models) to provide empirical evidence on the effect of various design decisions associated with the development of a NER system for natural language-based search in building information models. The hardware configuration comprises an NVIDIA RTX 3080 GPU, equipped with 16GB memory.

3.4.3 Results

Table 3.2 presents the summary of the results. Each value reported in the table represents the average and standard deviation of performance measures, computed based on the prediction results across all 30 evaluation iterations. Moreover, to provide a more nuanced understanding of the predictive performance of the trained models, evaluation scores corresponding to individual tags for the top three models (Experiment 4, 6, and 7) are presented in Table 3.3. More details on model configuration for each set of experiments are given as follows.

Table 3.3 Comparison of performance results for candidate models based on individual tags

		Built_obj	**loc_level**	**loc_space**	**MEP_obj**	**Name**	**Number**	**Ordinal**	**Quantity**
BiLSTM (Experiment 4)	Precision	**0.9996**	**1.0000**	**1.0000**	**0.9994**	0.9991	**1.0000**	**1.0000**	**1.0000**
	Recall	**0.9996**	**1.0000**	**1.0000**	**0.9972**	0.9990	**1.0000**	**1.0000**	**1.0000**
Model size: 63MB	F1	**0.9996**	**1.0000**	**1.0000**	**0.9983**	0.9990	**1.0000**	**1.0000**	**1.0000**
Inference time[a]: 647ms	Support	81	63	88	176	70	25	38	151
BERT (Experiment 6)	Precision	0.9988	**1.0000**	0.9996	0.9881	**1.0000**	0.9987	**1.0000**	**1.0000**
	Recall	0.9993	**1.0000**	**1.0000**	0.9794	**1.0000**	**1.0000**	**1.0000**	0.9998
Model size: 436MB	F1	0.9991	**1.0000**	0.9998	0.9837	**1.0000**	0.9993	**1.0000**	0.9999
Inference time[a]: 1410ms (millisecond)	Support	93	63	93	229	211	25	53	151
DistilBERT (Experiment 7)	Precision	0.9983	**1.0000**	**1.0000**	0.9920	**1.0000**	0.9976	**1.0000**	**1.0000**
	Recall	0.9990	**1.0000**	**1.0000**	0.9850	**1.0000**	**1.0000**	**1.0000**	0.9996
Model size: 265.5MB	F1	0.9986	**1.0000**	**1.0000**	0.9885	**1.0000**	0.9988	**1.0000**	0.9998
Inference time[a]: 761ms	Support	93	63	93	229	211	25	53	151

[a] The time taken for inference on the test set, which consists of 238 sequences of tokens (queries)

With respect to the implementation of the embedding layer for BiLSTM-based models, word2vec and GloVe algorithms are utilized in Experiment 1 and Experiment 2-5, respectively. As mentioned in our review of the existing literature, a significant number of similar previous studies did not pay sufficient attention to the configuration of the embedding layer for training a NER model when employing pre-trained word embeddings. To address this shortcoming, the following configuration parameters are taken into consideration for the design of our experiments.

First, the dimensionality of the embedding vectors is examined. Using GloVe-01 pre-trained embeddings, we experimented with 100 and 300 embedding vector dimensions for Experiment 2 and Experiment 3, respectively. Second, the size of the vocabulary included in the embedding matrix is examined, i.e., altering the number of tokens included within the embedding vocabulary list to examine the effect of missing tokens. Table 3.4 presents the summary of missing token statistics for each of the pre-trained embedding models used in this study. Our key motivation

for this step is to enable a more equitable comparison between the utilized word2vec and GloVe embedding models. To this end, we manually change the embedding vocabulary of GloVe-03 embeddings (used in Experiment 5) to have the same vocabulary as the word2vec embeddings used in Experiment 1. By employing identical vocabulary sets and embedding dimension sizes, we are able to make a direct comparison of the effectiveness between these two pre-trained embedding models. However, it is important to note that this comparison cannot be extended to the underlying embedding algorithms, as these models were trained on distinct datasets.

Table 3.4 Number of tokens missing from the embedding models

Semantic tag	**word2vec**	**GloVe-01**	**Glove-02**	**GloVe-03**
B_Name	700	514	432	700
B_Quantity	638	0	0	638
B_Ordinal	193	7	0	193
B_Number	131	7	0	131
O	0	76	0	0
B_MEP_obj	15	0	0	15
B_Loc_space	8	8	0	8
Total misses	1685	606	432	1685

Finally, with respect to the transformer-based architectures, we utilize the BERT base model and a distilled version of it, i.e., DistilBERT, that are available in the Transformers open-source library (Wolf *et al.*, 2020). When using pre-trained language models such as BERT family models, it is important to ensure that the tokens included within the training and evaluation data used for the new task at hand are in accord with the data upon which the language model has been originally trained for the tokenization task. In this light, for the experiments reported in this work, great care is taken to ensure that the developed dataset is adapted to the configuration of the used models. This is done by tailoring our developed dataset to the model's original tokenizer. That is, for each raw query, first, the model's tokenizer is used to tokenize the input, then labels are extended to the corresponding sub-word tokens. For example, in the case of using the BERT base model, the terms "heaters" and "exchanger" are tokenized as "heat ##ers" and "exchange ##r", respectively. By taking this critical step, we ensure that the input delivered

to the NER training algorithm corresponds with the input utilized during the pre-training stage of the utilized models.

3.4.4 Discussion

Given the shortcomings of similar past studies, the main objective of the conducted experiments is to provide a comparative examination that evaluates both predictive performance and computational overhead associated with various model design choices for the task at hand. The rest of this section discusses the contrasting attributes of BiLSTM and transformer models, emphasizing their respective advantages and disadvantages[1].

3.4.4.1 Predictive performance

From the standpoint of predictive performance, as can be observed from Table 3.2, the BiLSTM-based model trained in Experiment 4 demonstrated superior results compared to those using the transformer architecture. This finding appears to diverge from a prevailing trend within the AEC-FM research community, where transformer models have often been reported to demonstrate a *significant* superiority in terms of prediction capabilities (compared to RNN-based models). In this light, our primary argument is that the outcomes of this study, as well as those of related research, should be interpreted cautiously due to the following critical reasons.

First, this parity in prediction performance may not necessarily hold under the condition when the models are subjected to evaluation on a substantially larger and more diverse dataset. In fact, the limited size and scope of the developed dataset is the key limitation of the present and past similar studies, which might inadvertently be favoring one architecture over another, thereby providing a potentially skewed perspective on its comparative performance (see discussion of limitations in section 3.5). Second, the literature on this topic frequently lacks detailed information concerning the training and evaluation of RNN-based models for sequence or token classification, especially with regard to crucial aspects such as the configuration of the

[1] See the further discussion provided in the concluding chapter of the thesis.

embedding layer. In support of the latter argument, the comparison of the scores obtained from Experiments 1 through 5 reveals that an effective embedding layer can enhance the overall prediction performance of BiLSTM-based models.

3.4.4.2 Computational performance

Although the BiLSTM architecture allows understanding the context from both directions (left and right of a token), the BERT architecture benefits from an additional set of distinct advantages. Primarily, being a transformer-based model, BERT leverages attention mechanisms that enable it to deal more effectively with contextual dependencies in a textual input, a challenge often faced by RNN-based architectures (see section 3.4.4.3). However, transformer-based language models like BERT embed a significantly large number of parameters that are pre-trained on a large corpus of text, equipping the model with a rich understanding of diverse language semantics. While the base BERT model contains about 110 million parameters, our largest BiLSTM model contains less than a million parameters. However, this considerable parameter count in BERT and other transformer-based models correlates with increased computational demands.

While transfer learning approaches can reduce the training duration and associated costs through a fine-tuning process for domain-specific downstream tasks, the extensive number of parameters in pre-trained large models can contribute to a considerable computational burden during inference time, i.e., when the final model is utilized in a production setting. Such higher computational demands can directly affect the applicability of the model due to issues such as memory consumption and inference latency. For instance, as depicted in Table 3.3, our experiments revealed that the average time required for inference on the test set was more than twice as long for the fine-tuned BERT base model in comparison to the BiLSTM model. Furthermore, due to the more substantial model size, the memory utilization of the BERT model notably exceeded that of our candidate BiLSTM model (see Table 3.3). Nonetheless, as previously discussed, knowledge distillation techniques can be exploited to lessen the computational complexity while maintaining a high level of predictive performance. Our empirical analyses substantiate that

the distilled version of the BERT model manifests considerably lower computational overhead while preserving the predictive performance, as indicated in Table 3.3.

3.4.4.3 Contextual word representation

Another advantageous attribute of the transformer-based models in terms of contextual understanding of the textual sequences, lies in their inherent ability to generate context-dependent embeddings. In fact, traditional embedding models such as word2vec and GloVe are trained to generate static word embeddings, which essentially means that the same vector is used to represent a particular word, regardless of its contextual usage. Such static embeddings can lead to inherent ambiguities in the natural language utterances. For example, within the context of the IFC schema, the term "roof" can denote two distinct types— either "ifcRoof" or "ifcSpace". Depending on the particular context in which the user query is formulated, the former refers to an independent building element while the latter refers to a building space associated with the roof object. Although the distinction may seem subtle, discerning between these two types is critical, as they represent fundamentally different entities in a building information model. As another illustrative example, consider the following two queries:

- **Query 1:** "*On the first floor, select all windows.*"
- **Query 2:** "*First, select all windows then export them.*"

From a semantic lens, the term "first" in "Query 1" serves as a crucial component in describing the location of the objects of interest, i.e., the windows. In contrast, in "Query 2", the same term holds no substantial relevance to describing the window objects. However, it should be noted that the tone of the queries embodied in our developed dataset is similar to that of "Query 1". Hence, although the BiLSTM model's prediction scores for "Ordinal" entities are the same as those of BERT models (see Table 3.3), lower scores can be expected if the evaluation dataset were to include a broader range of conversational tones.

In real-world settings, there exist numerous scenarios in which the ability of the NER system to dynamically interpret the word semantics remains essential. Context-aware generation of

embeddings is a feature that is more readily available in transformer-based language models than in traditional embedding methods such as word2vec and GloVe. The context-aware embeddings generated by models such as BERT enable a more accurate and nuanced understanding of the semantic structures within the query, thereby facilitating the correct classification and representation of elements based on the rich context in which they are embedded. This aptitude for contextual discernment is particularly advantageous in specialized domains such as search in building information models where precision and clarity are vital.

Additionally, it is worth noting that in previous similar studies, e.g., (Wang *et al.*, 2022d; Zheng *et al.*, 2022b), researchers have predominantly employed pre-trained word embeddings, which had been derived from general text corpora such as Wikipedia dumps. This approach can be problematic due to the fact that domain-specific terminologies, which are crucial in specialized fields such as building information modeling, may be inadequately represented or altogether absent from the training datasets upon which these embeddings are generated. This is why it is important to examine the comprehensiveness of the tokens embodied within the embedding matrices that are used for the training of the NER classifier. Hence, as our experimental design encompasses the exploration of different sizes of vocabularies within the embedding matrix, we examine the coverage of pre-trained embeddings contained within each embedding matrix utilized in our experiments. The number of cataloged instances of tokens that were not present in each subset derived from the original embedding matrices is presented in Table 3.4 for reference.

3.4.4.4 Error analysis

To further investigate the results, an error analysis is undertaken by examining the evaluation metrics for individual token classes. The scores presented in Table 3.3 indicate that the BiLSTM model examined in Experiment 4, shows slightly lower performance in classifying “Name” entities when compared to its BERT-based counterparts. Moreover, the BiLSTM model under investigation demonstrates lower accuracy in predicting two other types of entities: “Built_obj” and “MEP_obj’, relative to the rest of the entity types predicted by the same model. Upon closer inspection of the scores in Table 3.3, an intriguing pattern is observed: For all three mentioned

types where the BiLSTM model demonstrates predictive shortcomings, there exists a gap in the number of support instances compared to those for the BERT and DistilBERT models. This gap, which suggests variations in tokenization output, is most notable in the number of supports for the "Name" category. This observed pattern leads us to hypothesize that the tokenization process might be at the root of the BiLSTM model's underperformance for these three classes. To further highlight the implications of this hypothesis, a detailed discussion of the tokenization procedures employed in this study follows below.

As outlined in the methods section, tokenization of raw queries for training BiLSTM models is conducted at the word level, utilizing a pre-existing tokenizer. However, this approach can lead to issues with segmenting input queries, particularly for domain-specific terminology or phrases that may not be adequately represented for the pre-training of the tokenizer. As illustrated in Table 3.3, the support instances for entities of "Name" type are markedly different between the models, with BERT and DistilBERT accounting for an average of 211 instances as opposed to BiLSTM's 70 instances. This substantial gap suggests that BERT's tokenization mechanism possesses a more nuanced capability to segment the input sequence's words and phrases into granular sub-elements (tokens). In other words, BERT's tokenizer appears to be more adept at segmenting textual input into constituent parts that can be semantically processed. The summary of missing tokens reported in Table 3.4 supports our argument, as the name phrases embedded in our dataset were adopted from real-world building information models. The absence of such tokens within the embedding layer ultimately leads to the model's incapability to discern the contextual relevance of such strings. A name such as "R-011", in the phrase "room R-011", encodes contextual information which, when analyzed at a word level, may cause the tokenizer to interpret the alphanumeric portion "R-011" as a mere number. Our analysis of erroneous predictions by the BiLSTM model in Experiment 4, confirms that tokens representing the "Name" class are misclassified as belonging to the "Number" and "Ordinal" categories. This is an illustrative example of how critical information can be lost or misunderstood due to the limitations of the tokenizer when it lacks adaptation to the domain-specific lexicon and contextual syntax.

3.4.4.5 Summary

These points highlight crucial issues within the realm of building information processing, as many phrases, such as names and engineering specifications, comprise acronyms and alphanumeric combinations, with dynamic contextual variability. In this light, it is important to take an adaptive input sequence processing strategy, one which can sufficiently accommodate the nuances inherent in building information terminology, and can effectively segment and encode user queries based on their both particular syntactic nature and contextual relevance. This is where the word-piece tokenization and context-aware embedding features of advanced transformer-based models such as BERT are advantageous. With their capacity to operate on subword units and adaptively construct context-aware embedding representations, even for unfamiliar terms, transformer-based pre-trained language models offer a compelling advantage, particularly when access to a large and high-quality corpus of text related to building information is limited.

3.5 Practical implications, limitations, and future directions

Table 3.2 and Table 3.3 show that the BiLSTM model exhibits prediction performance that is slightly superior to that of the transformer-based models. This marginal superiority in predictive performance can make RNN-based architectures particularly appealing for practical applications, given their lower resource consumption. However, in light of the discussion points outlined in the previous section, the findings of this study, as well as those of other works utilizing limited datasets, should be approached with a measured level of skepticism. Hence, when selecting architectures for practical purposes, we recommend considering the characteristics of the dataset at hand and the downstream task, in addition to predictive performance.

This cautious approach becomes more relevant when considering the additional challenges associated with implementing traditional architectures. In particular, in contrast to advanced architectures such as BERT, opting for traditional architectures such as BiLSTM necessitates the distinct design and implementation of a tokenization layer, followed by embedding learning for the new tokens that are underrepresented or missing within existing pre-trained embeddings.

These preliminary but essential steps can pose significant challenges, given the domain-specific nature of the data involved. When dealing with building information, textual data is not only highly specialized but also scarce. The process of curating large datasets for unsupervised embedding learning, as well as labeled datasets for training deep learning models for query tokenization, will be labor-intensive and costly and demands domain expertise to ensure the accuracy and relevance of the data. Under these constraints, developing an effective tokenizer tailored for domain-specific language can become a daunting task.

Finally, it is important to recognize that real-world applications often deal with extensive and heterogeneous data related to the design, engineering, construction, and maintenance of the built assets. Under such circumstances, the inherent advantages of transformer-based models like BERT in capturing long-range dependencies and context-rich representations might substantiate and justify the increased computational costs.

While the current tags included within our developed NER dataset cover a wide range of buildings and other facilities components and systems, future developments should include a broader range of entity descriptions. At present, the dataset covers three main categories of entity descriptions: quantity, name, and location, providing a foundational framework for identifying and classifying key elements within building information models. The planned developments aim to devise an effectively granular level of object property tags that are critical in the detailed analysis and management of built assets. These include material type, element profile, geometrical dimensions, status (e.g., installed, inspected), and physical properties related to design specification, safety, and compliance aspects of building components. Regarding the effective size of the data, the current statistics for labeled entities included in our dataset (see Table 1) can provide initial clues for future developments. However, the more expansive and diverse the dataset, the more likely it is to encompass the nuanced variations and edge cases that are likely in real-world situations. This recognition underscores the importance of larger-scale interdisciplinary collaborations, involving experts from various backgrounds. The hope is expressed that the open release of our data and code will foster these types of collaborations.

Given the limitations imposed by our dataset size, future assessment necessitates evaluation attempts based on larger, more diverse datasets that encompass a broader spectrum of terminologies and linguistic structures representative of the building information modeling domain. In light of the mentioned limitations, along with the discussions concerning the benefits of formulating a semantic labeling structure that is grounded in the IFC schema, it is advisable for future research to develop larger datasets that encompass a broad spectrum of entities and properties as represented within the IFC schema. Expanding the dataset not only reinforces the robustness of the NER system in the face of diverse and complex user queries, but also broadens its applicability across a spectrum of use cases that necessitate the identification of semantically significant entities within domain-specific textual content. For example, it can be used to extract building entities from design specifications, requests for information, or building codes, and then find the corresponding entities within the digital model of the building.

3.6 Conclusion

In the specific context of building information retrieval, where queries are usually brief but densely packed with domain-specific terms and expressions, the choice of deep neural network architecture remains a critical question. The effectiveness of the chosen architecture in accurately identifying and classifying entities within these queries can have a profound impact on the quality of information retrieval, as well as associated computational costs. In this study, we investigated various token classification systems in the context of identifying references related to building entities and their respective descriptors. By structuring our semantic labeling scheme on the basis of an open and comprehensive schema, i.e. IFC, our proposed system can process diverse user queries and identify and categorize a diverse set of essential entity types. This approach facilitates a dynamic application of our system, allowing it to be easily modified and extended. Moreover, the open and inclusive nature of the IFC schema facilitates the flexibility and adaptability of the system.

In this study, we conducted a series of experiments to set forth a comparative analysis on both predictive performance and computational demands of utilizing traditional versus transformer-

based architectures for the identification of mentioned building entities and their attributes which are described within naturally articulated user queries. In contrast to the dominant narrative in contemporary AEC-FM literature, our experimental findings suggest that NER models trained on traditional architectures can demonstrate competitive predictive capabilities when compared to transformer-based architectures. However, our study advocates for large-scale and diverse datasets as limitations inherent in small datasets can introduce elements of uncertainty. As a result, the authors of this study strongly urge the research community to prioritize the creation and utilization of expansive, high-quality benchmark datasets.

The discussion outlined in this work reveals that regardless of the overall lower predictive performance scores, transformer-based models can excel at capturing domain-specific nuances, alphanumeric combinations in particular, within the textual data specific to the building information domain. Hence, despite the higher computational overhead, the inherent context-sensitive embedding representation and tokenization capabilities enabled by transformer architecture can mitigate the resource expenditure associated with the development of tailor-made models for domain-specific downstream tasks, such as the one studied in this work. Finally, the inherent adaptability of our proposed system extends its applicability beyond the specific use case studied in the present work (i.e., natural language-based search in BIM), thereby enabling conformity to a variety of applications across disparate disciplines. This is an important advantage of our proposed semantic labeling structure as the future evolution of natural language-based applications for digital models of the built environment will necessitate a high degree of flexibility to accommodate varying domains and requirements.

CHAPTER 4

BENCHMARKING PRE-TRAINED TEXT EMBEDDING MODELS IN ALIGNING BUILT ASSET INFORMATION

Mehrzad Shahinmoghadam[1] , Ali Motamedi[1]

[1] Department of Construction Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3

Article submitted to *Nature's Scientific Reports*, in October 2024.

Abstract

Accurate mapping of the built asset information to established data classification systems and taxonomies is crucial for effective asset management, whether for compliance at project handover or ad-hoc data integration scenarios. Due to the complex nature of built asset data, which predominantly comprises technical text elements, this process remains largely manual and reliant on domain expert input. Recent breakthroughs in contextual text representation learning (text embedding), particularly through pre-trained large language models, offer promising approaches that can facilitate the automation of cross-mapping of the built asset data. However, no comprehensive evaluation has yet been conducted to assess these models' ability to effectively represent the complex semantics specific to built asset technical terminology. This study presents a comparative benchmark of state-of-the-art text embedding models to evaluate their effectiveness in aligning built asset information with domain-specific technical concepts. Our proposed datasets are derived from two renowned built asset data classification dictionaries. The results of our benchmarking across six proposed datasets, covering three tasks of clustering, retrieval, and reranking, highlight the need for future research on domain adaptation techniques. The benchmarking resources are published as an open-source library, which will be maintained and extended to support future evaluations in this field.

4.1 Introduction

Asset management plays a pivotal role in ensuring optimal performance and extended life span of the built environment through a systematic process of monitoring and maintaining various facilities and equipment. The rapid advancement of digital technologies has led asset owners to increasingly demand enriched digital twins at project handover to support real-time operations and maintenance of the built assets (Love & Matthews, 2019). Simultaneously, the growing awareness of the benefits of digitized asset management highlights the essential need for federated access to built asset data (Moretti *et al.*, 2023). This requires aligning extensive data sources and their underlying schema with established data models, classification systems, or taxonomies to facilitate data accessibility for diverse stakeholders and improve interoperability across various software environments. However, aligning built asset data with pre-defined classification systems poses significant challenges in practice. A key challenge stems from the multi-source and multi-disciplinary nature of built asset data, which leads to the use of diverse formats and terminologies across different projects and stakeholders. For example, the terminology that architects utilize to describe the specifications for a particular building component or system can vastly differ from those used by structural engineers or subcontractors. Moreover, the structures of domain-specific classifications used in different disciplines often vary in granularity. For instance, the detailed engineering descriptions of an HVAC system provided by mechanical engineers may be far more comprehensive than those required and used by operations and maintenance teams. Finally, variations in local regulations and standards can further complicate the alignment process, particularly for large-scale or international projects. These issues, combined with the dynamic and evolving nature of built asset data throughout an asset's lifecycle, lead to potential inconsistencies when integrating this data into a unified digital asset management environment.

In response, there have been several initiatives aimed at facilitating the digital delivery of built asset information while ensuring its conformity with predefined or standardized descriptions (data models, taxonomies, etc.). One major initiative is buildingSMART Data Dictionary (bSDD) (buildingSmart International, 2024a), an international and ongoing effort whose

main objective is to create shared definitions for describing the built environment. This is achieved through a collection of interconnected data dictionaries that are both human-readable and machine-readable (buildingSmart International, 2024a). Although making various data dictionaries programmatically accessible will facilitate access to agreed and consistent terms, the complexity and dynamic diversity of the built asset terminology necessitate robust data mapping strategies to accommodate various data descriptions and updates (Forth *et al.*, 2024). As a result, the asset information alignment process remains predominantly manual, relying heavily on the expertise of domain specialists to accurately map complex technical data (Roberts, Pärn, Edwards & Aigbavboa, 2018). The significant challenges associated with the manual alignment process, including high costs, time consumption, and potential for human error, highlight the need for more automated and reliable data mapping solutions.

The central thesis of our research builds upon the argument that recent advancements in natural language processing/understanding research can significantly enhance automated data mapping processes. In particular, the rich and contextualized representation of textual inputs as numeric vectors, commonly known as text embedding (Pennington, Socher & Manning, 2014b; Lee *et al.*, 2024b), provides advanced capabilities for machines to understand the semantics of the intricate terminologies. Earlier methods such as word2vec (Mikolov *et al.*, 2013) and GloVe (Pennington *et al.*, 2014b) relied on static embeddings, i.e., generating fixed representations of numerical vectors for each word based on their co-occurrence in large corpora. However, recent neural language models, dominantly built on top of the transformer architecture (Vaswani *et al.*, 2017), can generate dynamic, context-sensitive embeddings. The capability of recent embedding models in adapting the representation of words (or sub-word tokens) based on their surrounding context has motivated researchers and practitioners across diverse fields to leverage the power of contextual text embeddings to drive advancements in their respective domains. From traditional databases integration (Cappuzzo, Papotti & Thirumuruganathan, 2020) to public figure perceptions in social science studies (Cao & Kosinski, 2024), the increasing volume of encouraging reports on leveraging text embedding models to deliver a more nuanced text understanding in various specialized domains (Rasmy, Xiang, Xie, Tao & Zhi, 2021; Ostendorff

et al., 2021; Rouhizadeh *et al.*, 2024; Wilkho, Chang & Gharaibeh, 2024; Cao & Kosinski, 2024) reinforces the relevance of these models in automating data alignment in the domain of built asset information management.

Based on the observation that built asset data predominantly exists in textual form (Wu *et al.*, 2022), we argue that state-of-the-art text embedding models present promising opportunities to refine the automated alignment of built asset information. However, the extensive and increasing availability of pre-trained language models has led to the proliferation of potential text embedding models, creating confusion regarding model selection for different use cases (Muennighoff *et al.*, 2022). Moreover, recent research indicates that general-purpose text embedding models often struggle to maintain consistent performance across diverse tasks and domains (Lee *et al.*, 2024b). This is while most previous studies utilizing pre-trained or fine-tuned language models in built environment research have been significantly limited in scope, primarily focusing on ad-hoc downstream tasks with small evaluation datasets (Shahinmoghadam *et al.*, 2024; Jung, Hockenmaier & Golparvar-Fard, 2024; Wang *et al.*, 2024b; Forth *et al.*, 2024; Jeon *et al.*, 2024). Such limitations can result in a potentially skewed perspective on the overall domain-specific text understanding of these models (Shahinmoghadam *et al.*, 2024). Additionally, scarce public access to the datasets used in previous works poses another important challenge to the transparency and reproducibility of the reported results. This motivates us to examine the extent to which existing language models can be directly leveraged to deliver contextually accurate mappings of domain-specific terminology within the context of built asset information management. In this work, we present a comprehensive benchmark of state-of-the-art text embedding models to evaluate their effectiveness in capturing and representing the semantics of textual descriptions related to built assets. Through this evaluation, we aim to identify the strengths and limitations of existing language models in enhancing data alignment practices within the built asset domain. Our proposed benchmark is aligned with the Massive Text Embedding Benchmark (MTEB) (Muennighoff *et al.*, 2022), a benchmark recognized extensively in both academic and practical contexts for its robustness and utility. We benchmark 24 text embedding models on our developed datasets that amount to a total of more than ten thousand data entries across six tasks, making

our evaluations the most comprehensive ones in this specialized field to date. By making our datasets and benchmark software publicly available, we encourage future research to build upon our work, contributing to continuous improvements in this domain.

4.2 Methods

4.2.1 Data sources

Given the built environment's multidisciplinary nature, the datasets included in the benchmark must encompass an expansive spectrum of sub-domain subjects, including architectural, structural, mechanical, and electrical systems. To ensure a diverse coverage of built products in our benchmark, we carefully examined the selection of data sources used for creating task-specific datasets. A detailed description of the corpus development and data extraction processes is provided below.

The initial step in creating the benchmark's task-specific datasets is the development of a consistent corpus of built products. Based on the requirements of the tasks within our benchmark, the core corpus needed to include the following key information for each product: name or title, description, and corresponding labels (group categories). The two primary sources used to develop the built product corpora are as follows:

- **Industry Foundation Classes (IFC):** Published and maintained by buildingSMART International (buildingSmart, 2024), IFC is an open international data model offering comprehensive digital descriptions of various aspects of building and infrastructure projects. Originally designed to facilitate interoperability and information exchange among different software applications and stakeholders, IFC provides a comprehensive representation of various aspects of built asset entities. We utilize IFC version 4.3.2.0 (buildingSmart International, 2024c), recently approved as an ISO standard (ISO 16739-1:2024).
- **Uniclass:** Developed and maintained by the National Building Specification (NBS) (NBS, 2024a), Uniclass is a unified classification system for the built environment. We utilize version 1.33 of the Uniclass Pr Product Table (NBS, 2024b). Uniclass has extensive coverage,

encompassing over 8,000 product types, making it one of the most recognized and widely adopted classification systems in the built asset industry.

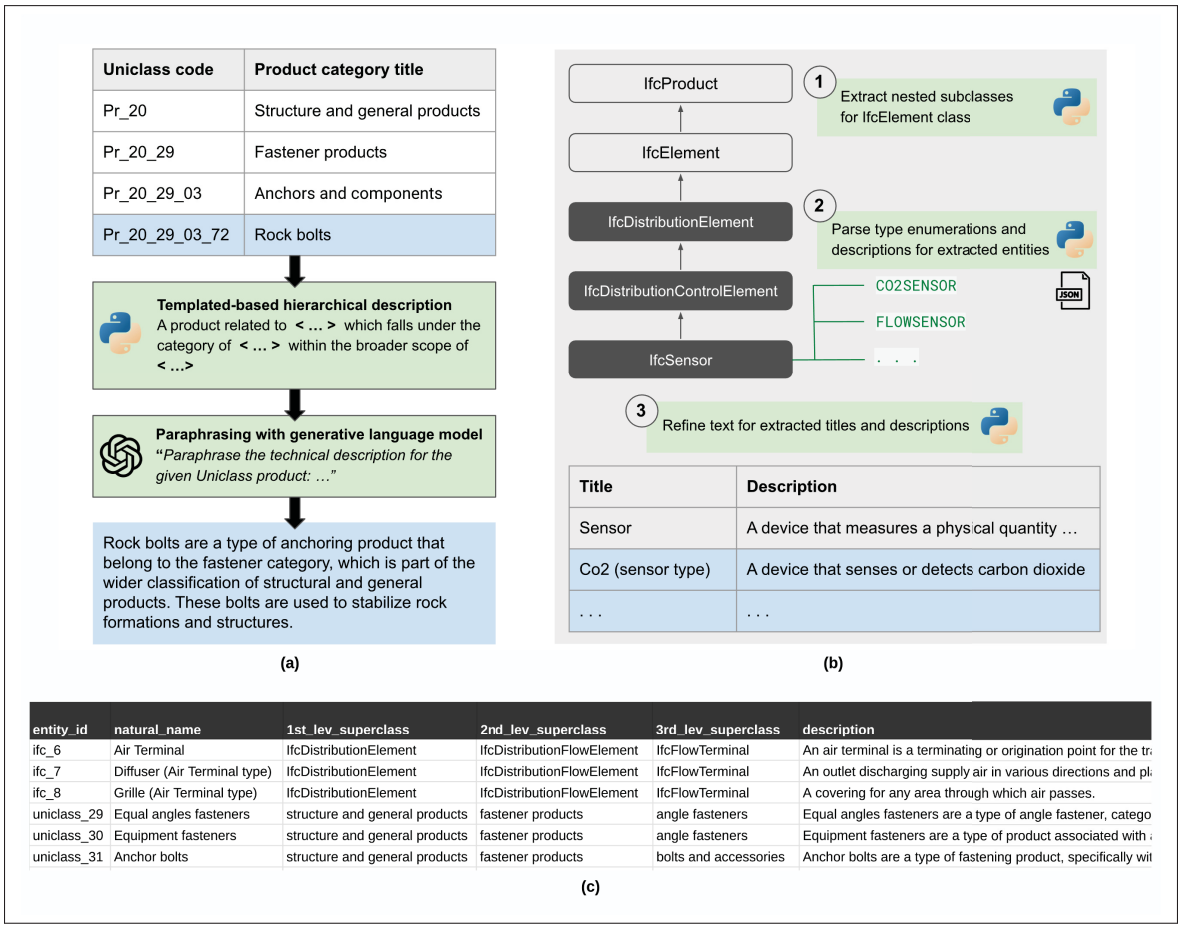


Figure 4.1 Overview of the main steps in developing the built product corpus: (a) Example of extracting categories and synthesizing entity descriptions from raw Uniclass entries; (b) Example of hierarchical relation extraction for main entities and their enumerated types from the IFC schema; (c) Sample records from the developed corpus, containing product titles, descriptions, and categories with three levels of granularity

4.2.2 Data extraction

To create a corpus of products with corresponding names, descriptions, and labels, we undertook the following steps: For Uniclass, we utilize the publicly-available CSV format of the products table, which comprises over 8,000 products categorized into three hierarchical levels. Product

names were directly extracted from the table, while product categories were inferred from the numeric codes associated with hierarchical categories (see Figure 4.1). To automatically extract the corresponding textual labels for each product, we developed a script to scrape the table programmatically. As the original table does not include product descriptions, we propose a method (detailed in the subsequent subsection) to synthesize a description for each product. We retained only those products that have labels for all three classification levels. After applying this filtering process, the Uniclass corpus comprises 4,234 instances, which remains sufficiently large for our benchmarking purposes.

Regarding the IFC schema, we parse the official schema content by utilizing resources from an open-source Python library (IfcOpenShell, 2024) that enables programmatic access to IFC entities. Initially, we extracted entities of interest from a JSON-formatted file containing the complete list of IFC entities, their type enumeration, and their definition (derived from IFC's official documentation). An analysis of the "IfcProduct" class within the IFC schema indicated that a significant majority of product entities are classified under the "IfcElement" class. Therefore, we focused exclusively on the "IfcElement" subclasses. After removing IFC entities with missing descriptions (less than 1% of total "IfcElement" entities), we developed a script to extract each entity's top three parent classes to serve as the product category labels. In addition to entity superclass groups, we use the domain-specific schemas (e.g., structural, HVAC, building control) from IFC's official documentations (buildingSmart International, 2024c) as an additional source for entity label assignment. The resulting IFC corpus comprises 977 entities (total of parent entities and type enumerations).

4.2.3 Data augmentation and curation

The process of generating textual descriptions for Uniclass entities is depicted in Figure 4.1(a). Initial entity descriptions are synthesized by sequentially concatenating the entity's category titles, progressing from the most specific to the most general. An example of the synthesized descriptions is provided in Figure 4.1(a). These concatenated descriptions are then paraphrased using a generative language model to create more nuanced and natural descriptions, relaxing the

text from the rigid template initially employed. We generated paraphrased descriptions using the most advanced version of the GPT-4 model available at the time of conducting the experiments (gpt-4-turbo-2024-04-09). Although the prompts used for generating paraphrased descriptions (Shahinmoghadam, 2024) were designed to prevent the alteration or addition of facts, it was essential to manually review all generated descriptions due to known potential inaccuracies of generative language models. The review is carried out by two domain experts, each with over ten years of experience in the field. Each expert cross-checked the issues identified by the other, and the final decisions were made based on mutual agreement. The following curation steps are undertaken to ensure the accuracy and consistency of the extracted product names and descriptions. We preprocess native IFC entity names and convert them into a readable form (e.g., "IfcHeatExchanger" to Heat Exchanger; see examples in Figure 4.1(b) and (c)). For IFC class enumeration types, where the enumeration name alone might be ambiguous, we append the parent class type in parentheses. For example, the enumeration WATER, a subclass of "IfcBoilerTypeEnum", is represented as "WATER (Boiler Type)" (see examples in Figure 4.1(c)). Following the same logic, we enrich the product descriptions by concatenating the product's name at the beginning of the description for both Unicalss and IFC entities. This step reinforces contextual clarity, as the natural entity names carry significant semantic information. Finally, we manually review and modify the entity descriptions that contain inconsistent information, such as notes related to the schema version history or future depreciation notes.

4.2.4 Sampling

To ensure a robust entity selection when creating task-specific datasets, we implemented the following sampling strategies: For positive sampling, we adopt a semantic diversity approach. Given a targetted subset of built products, we generate text embeddings for all corresponding text inputs, i.e., product names and descriptions. Embeddings are generated using a state-of-the-art text embedding model ("mxbai-embed-large-v1" (Li & Li, 2023)). From this set of embeddings, we randomly choose an initial sample as a starting point. Subsequently, we iteratively select additional samples by identifying those that exhibit the slightest similarity to the most recently

selected sample, as determined by cosine similarity scores, i.e., the cosine of the angle between two embedding vectors. This process repeats until the desired number of samples is achieved. This method ensures that the samples selected for a particular subset (e.g., products of the same category) yield diverse representations within the embedding space by selecting inputs that are semantically dissimilar to the ones already chosen. For negative sampling, we prioritize the selection of product samples that yield closer semantic similarity to a given query (a product name or description) but belong to a different class. We compute the cosine similarities between the query and negative samples using the same embedding model used in the semantic diversity sampling and select samples with higher similarities. By selecting more similar candidates as negative samples, the dataset can better benchmark the model's capability to capture the subtle differences between closely related classes. This method, commonly known as hard negative sampling, is particularly effective for evaluations involving fine-grained classifications, such as differentiating between closely related categories in IFC and Uniclass classification hierarchies. In all sampling methods, including plain random sampling, once a sample is selected, it is only reused in another subset once all samples included within the pool have been exhausted. This way, we maximize the utilization of available samples and maintain diversity within the datasets.

4.3 Benchmark

4.3.1 Tasks overview

Evaluating text embeddings across different tasks is crucial for assessing the transferability of their capabilities to various downstream applications. Hence, our proposed benchmark covers three main tasks: clustering, retrieval, and reranking. In addition to domain coverage and cross-task adaptability, evaluating text embedding models requires careful consideration of input text length. To ensure the coverage for varying input lengths, the text entities included in our datasets fall into two categories: (a) sentences, which are derived from product titles/names, and (b) paragraphs, which are derived from product descriptions/definitions. Accordingly, each task-specific dataset in our benchmark is grouped into one of the following categories:

- **Sentence to Sentence (S2S):** Utilizing product titles as input text.
- **Paragraph to Paragraph (P2P):** Utilizing product descriptions (which can be concatenated with the product name) as input text.
- **Sentence to Paragraph (S2P):** Comparing product titles against product descriptions.

Our proposed benchmark follows MTEB (Muennighoff *et al.*, 2022) for reporting text embedding performance scores. Hence, various metrics are implemented within our benchmark, which can be computed with flexible parameter configurations. The primary metrics, which serve as default scores for task-specific as well as overall comparisons reported in this study, are outlined in each task's description.

4.3.1.1 Clustering

Clustering tasks involve grouping similar built products into meaningful clusters based on their similarities in textual representation. Our proposed tasks include S2S and P2P categories, where product names and descriptions act as input text for each dataset type, respectively. Each clustering task dataset is comprised of various subsets, covering diverse subdomain subjects and different levels of granularity. To create the subsets within each clustering dataset, we first select a subset of product labels from one of the three levels of product hierarchy, either from one specific corpus or across both corpora. We then apply the previously described diversity-based sampling method to sample product names (S2S datasets) or descriptions (P2P datasets) for selected labels.

To ensure the quality of the subsets, we evaluate the baseline scores using two embedding models, one for the upper threshold ("mxbai-embed-large-v1" (Li & Li, 2023)) and one for the lower threshold ("paraphrase-multilingual-MiniLM-L12-v2" (Reimers & Gurevych, 2019)). A subset is included in the dataset only if its score with the upper threshold model is below 0.8 and greater than 1/N with the baseline model, where N is the number of unique labels. The upper and lower thresholds are set to maintain task difficulty and ensure the task performs better than

random guessing, respectively. Subsets meeting these criteria are shuffled to eliminate order bias before being added to the dataset.

We compute V-measure scores (Rosenberg & Hirschberg, 2007) by training a mini-batch k-means model using vector embeddings, with k set to the number of unique labels in each clustering subset. The V-measure, ranging from 0 to 1 (higher is better), represents the harmonic mean of two distinct metrics: homogeneity and completeness. Here, homogeneity measures the extent to which clusters contain only products from a single category, while completeness indicates how well all products from a given category are grouped into the same cluster. More details regarding the calculation of V-measure can be found in (Rosenberg & Hirschberg, 2007).

4.3.1.2 Retrieval

Retrieval tasks aim to identify relevant documents, i.e., product textual descriptions, in response to a given query. Our proposed retrieval datasets are framed as S2P and P2P tasks, where built asset descriptions serve as the corpus (the documents to be retrieved), and product titles and descriptions act as queries for the S2P and P2P tasks, respectively. The query-document relevancy ground truth is derived from existing mappings that identify the alignment between IFC and Unicalss product entities. These mappings, validated and published by NBS (NBS, 2024a), can be found in the official Unicalss table release (NBS, 2024b).

First, we encode all queries and product descriptions into corresponding embedding vectors. These embeddings are then used to calculate the pairwise similarity between a given query and all product descriptions using cosine similarity. Subsequently, product descriptions included in each retrieval dataset are ranked according to descending cosine similarity scores. Finally, we compute nDCG@10 (Normalized Discounted Cumulative Gain (Järvelin & Kekäläinen, 2002) at rank 10) as the primary metric. This score, which can range between 0 and 1 (higher is better), reflects the relevancy of the ranked products based on their positions within the top 10 ranks by applying a logarithmic discount factor to penalize results that appear lower.

4.3.1.3 Reranking

In our reranking tasks, the aim is to rank a set of product descriptions with reference to their relevance to a product query. Similar to retrieval tasks, reranking tasks are framed as S2P and P2P types, and pairwise similarity between query and product description embeddings is computed based on cosine similarity. The primary distinction between retrieval and reranking tasks lies in their scope and focus. While our retrieval tasks involve ranking the entire product corpus, reranking narrows the focus to a smaller set of positive and negative subsets, which are selected using the methods outlined in the previous section to ensure diversity and difficulty (avoiding very high scores from overfitting) within the dataset. Positive and negative samples are selected using the methods described in the previous section, thereby maintaining the diversity and difficulty of the dataset. By concentrating on a smaller and more challenging group of product descriptions, our reranking tasks aim to provide a more fine-grained evaluation of the model's ability to rank relevant items accurately.

Similar to retrieval tasks, we use cosine similarity to compute pairwise similarity between a given query and product descriptions included in corresponding positive and negative sets. Subsequent to ranking the descriptions based on the cosine similarity scores, we compute MAP (Mean Average Precision) as our primary metric. MAP provides an averaged measure of precision across all relevant products, ranging between 0 and 1, with higher values indicating better performance. It is worth noting that retrieval metrics reflect overall ranking quality while reranking metrics focus on how early relevant products appear in the list.

4.4 Results

Table 4.1 provides a comprehensive summary of the dataset statistics across the three main tasks in our benchmark. The unique number of sample entries in our clustering datasets shows that more than half of the samples available from the combined product corpora could pass the quality thresholds explained in the methods section. In the retrieval and reranking task, the same retrieval and reranking document corpus is shared between the subtasks of each task category.

Table 4.1 Summary of dataset statistics per benchmark task

Clustering tasks	No. of subsets	Unique/total samples	Avg. sample length	Total No. of unique labels	Avg. unique label per subset
Clustering-s2s	18	2545/3815	28.04	31	5
Clustering-p2p	20	3067/4577	207.91	35	5
Retrieval tasks	No. of queries	Avg. query length	No. of documents	Avg. document length	No. of document per query (Avg.)
Retrieval-s2p	977	30.35	2761	312.75	8
Retrieval-p2p	977	128.5	2761	312.75	8
Reranking tasks	No. of queries	Avg. query length	No. of positives (unique/total)	No. of negatives (unique/total)	Avg. samples length
Reranking-s2p	179	27.89	1253/1253	2281/3759	310.15
Reranking-p2p	179	140.44	1253/1253	2241/3759	309.66

This design enables a comparative analysis of model performance on different query types, with S2P focusing on shorter product names and P2P targeting longer product descriptions. We applied a 1:3 positive-to-negative sampling ratio to create a balanced yet challenging evaluation set, ensuring that models must distinguish effectively between relevant and irrelevant documents.

To outline the distinctions between our newly constructed datasets and existing ones, we conducted a thematic semantic similarity comparison between our clustering datasets and those from MTEB benchmark. Using the "stella-en-400M-v5" model, which is the most performant small-sized model in our evaluations (see Table 4.2), we generated embeddings for 200 randomly selected samples and averaged them within each dataset. Figure 4.2 depicts the cosine similarity matrix as a heatmap, where darker shades indicate higher content similarity. The high similarity scores between our proposed subtasks confirm strong internal consistency within our benchmark. Moreover, moderate to high similarities with StackExchange, Reddit, and Arxiv datasets reflect thematic overlaps with broader domain content. A discussion of the observed similarities is provided in the next section.

In our benchmarking experiments, we evaluated models across a broad range of sizes, from relatively small models with 33 million parameters to significantly larger models exceeding

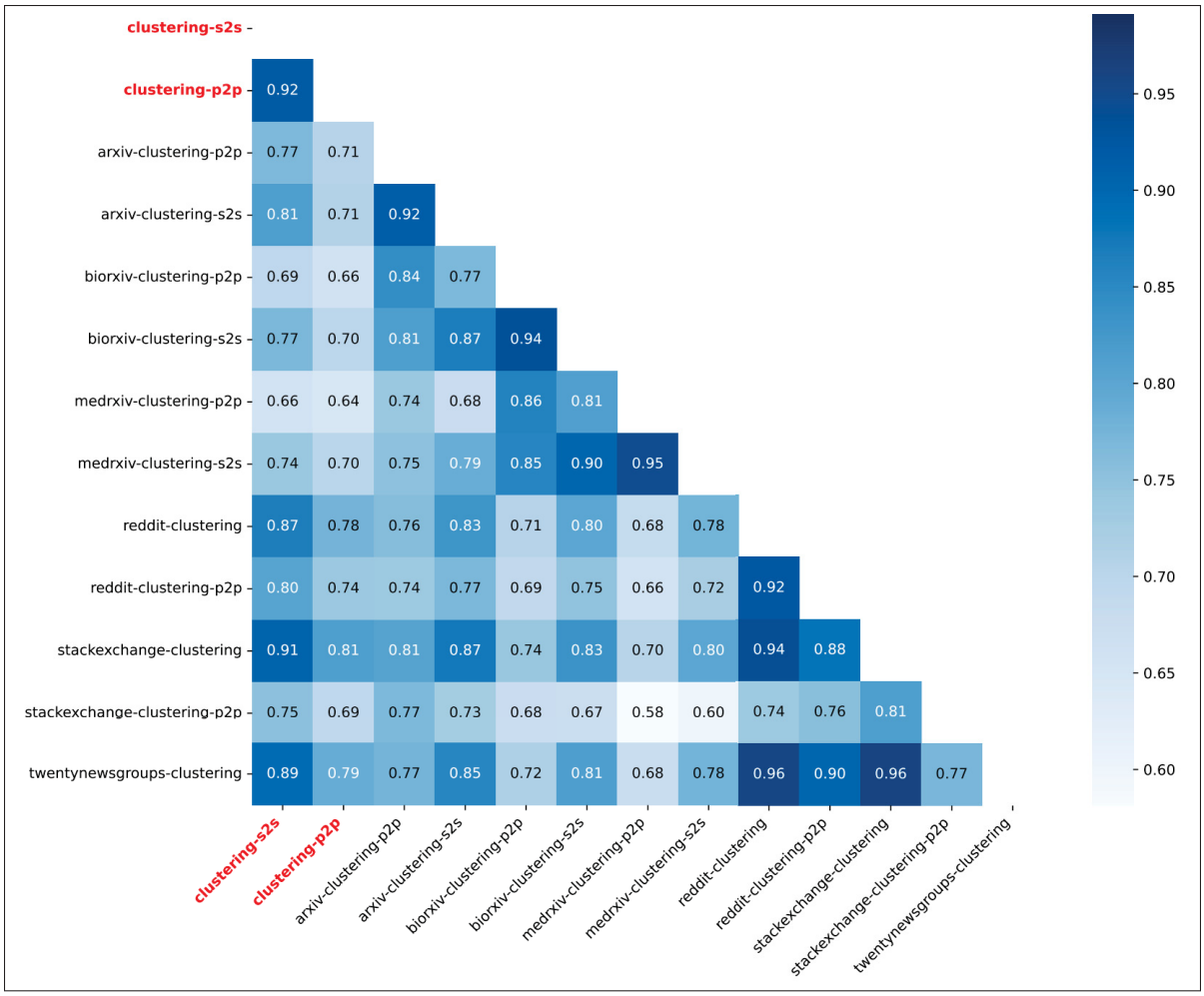


Figure 4.2 Thematic similarity heatmap between our proposed clustering tasks and those from MTEB. Average embeddings are derived from 200 random samples per dataset, encoded using the "mxbai-embed-large-v1" model (Li & Li, 2023). Datasets from our proposed benchmark are highlighted in red

seven billion parameters. However, due to computational constraints, the majority of models tested have less than one billion parameters. The selected models span various positions on the most recent record of MTEB leaderboard (as of September 21, 2024), ranging from first place (i.e., "NV-Embed-v2" (Lee *et al.*, 2024a)) to 136th place (i.e., "paraphrase-multilingual-MiniLM-L12-v2"). For models that are pre-trained with instruction-based data, we used built-in or recommended prompts as provided in the model card's official web page or associated research papers, when available. For example, "mxbai-embed-large-v1" requires custom prompts only

Table 4.2 Average scores of benchmarked models per task, based on the task-specific metrics mentioned in the task descriptions. The first and second highest scores for each task are highlighted in bold and underlined, respectively. MTEB ranks are sourced from records as of September 21, 2024

Tasks (→)	Clustering		Retrieval		Reranking		Avg.	Param.	MTEB
Models (↓)	s2s	p2p	s2p	p2p	s2p	p2p	-	(mil)	Rank
Pre-trained without task instructions									
gte-base-en-v1.5	48.38	51.83	79.98	59.42	66.54	66.73	62.15	137	39
gte-large-en-v1.5	43.42	51.05	83.32	63.27	**72.76**	70.15	64.00	434	24
bge-base-en-v1.5	43.00	51.78	82.56	61.65	67.01	63.38	61.56	109	43
bge-large-en-v1.5	46.69	52.41	82.60	64.86	68.44	65.47	63.41	335	35
UAE-Large-V1	45.45	49.53	83.32	66.42	70.04	68.53	63.88	335	29
GIST-Embedding-v0	46.43	49.96	82.82	62.78	68.81	65.75	62.76	109	41
GIST-large-Embedding-v0	47.97	47.91	84.01	67.06	69.53	68.03	64.08	335	34
e5-base-v2	42.59	50.24	80.83	61.46	69.11	62.91	61.19	109	64
e5-large-v2	42.11	49.45	81.95	64.63	68.61	64.58	61.89	335	55
multilingual-e5-large-instruct	48.01	52.82	80.35	64.55	67.85	65.90	63.25	560	42
multilingual-e5-small	42.98	48.16	76.38	55.03	64.78	62.34	58.28	118	112
all-MiniLM-L12-v2	42.00	46.52	79.97	58.81	66.20	63.97	59.58	33	117
paraphrase-multilingual-MiniLM-L12-v2	37.60	45.70	69.01	49.90	61.23	59.15	53.77	118	136
gte-base	45.96	51.55	82.91	62.95	68.97	66.26	63.10	109	51
gte-large	48.54	55.24	84.32	66.08	70.94	69.25	65.73	335	47
gte-small	44.31	55.55	82.37	60.55	68.82	65.23	62.80	33	70
Pre-trained with task instructions									
gte-Qwen2-7B-instruct	50.19	<u>62.39</u>	**86.28**	73.20	69.47	67.51	68.17	7069	6
mxbai-embed-large-v1	47.49	52.45	83.51	66.60	70.10	69.66	64.97	335	28
multilingual-e5-large-instruct	48.10	59.43	82.91	64.42	70.53	69.23	65.77	560	42
NV-Embed-v2	**58.61**	**67.34**	<u>85.23</u>	**77.02**	66.67	70.34	**70.87**	7851	1
stella-en-1.5B-v5	<u>53.60</u>	54.57	84.18	71.21	71.57	<u>71.77</u>	67.82	1545	3
stella-en-400M-v5	53.39	55.78	84.60	70.00	69.58	69.36	67.12	435	7
Proprietary embedding APIs									
text-embedding-3-small	49.72	49.72	79.97	65.68	65.33	66.99	62.90	-	58
text-embedding-3-large	49.75	55.48	84.99	<u>75.38</u>	<u>71.93</u>	**72.46**	<u>68.33</u>	-	30

for retrieval and reranking tasks, while "NV-Embed-v2" needs specific task-based prompts for clustering tasks as well. For models without built-in task instructions, we applied a general set of prompts to ensure consistency across tasks (prompts are available at the project's public GitHub repository (Shahinmoghadam, 2024)).

The top-ranked model in our benchmark, "NV-Embed-v2", also holds first place on the latest MTEB leaderboard. However, it does not consistently outperform all other models across all tasks. In fact, a closer examination reveals variability in model size and performance relationship. For example, "gte-small", the smallest model in our evaluation with 33 million parameters,

delivers competitive scores, nearly matching the average scores of models ten times its size and even outperforming larger models in specific tasks. Despite the previously reported strong correlation between model size and performance (Muennighoff *et al.*, 2022), our experiments show that superior performance associated with larger models is only evident at the extreme upper end of the parameter scale. This observation supports the growing emphasis on developing and deploying smaller, more efficient models for both research and real-world applications in this specialized field.

Motivated by the hypothesis that existing datasets with similar thematic content would yield comparable performance evaluations, we examined the consistency of relative model performances as follows: Given the observed thematic similarity between our clustering datasets and specific MTEB datasets, particularly "StackExchange" and "Reddit" (see Figure 4.2), we compared the rankings of model performance across both our datasets and the selected MTEB datasets. As it can be seen from Table 4.3, the comparative evaluation of the relative rankings indicates a notable variation in model performances, notably in the case of "multilingual-e5-large-instruct", "gte-small", "stella_en_1.5B_v5", and "text-embedding-3-small". These observed variabilities further highlight the limitations of relying on general-purpose benchmark datasets, even when relatively high thematic similarities are present, underscoring the importance of domain-specific evaluations.

While our benchmarking experiments primarily focused on open-source models, we also included the proprietary text embedding models from OpenAI, both the small and large versions. The inclusion of the proprietary models is motivated by a recent study where closed-source models tend to achieve relatively higher performance when embedding text in underrepresented languages (Enevoldsen *et al.*, 2024). We hypothesize that built asset text, as an underexplored domain, might be similarly better represented by proprietary models. Notably, text-embedding-3-large ranks second in our benchmark, performing nearly on par with the top-ranked model. In contrast, the smaller model performed more moderately, ranking in the middle of our benchmark. While the former observation aligns with the findings of Enevoldsen *et al.* (2024), the latter is in line with the latest MTEB leaderboard results where closed-source commercial embedding

APIs generally underperform compared to their open-source counterparts. These observations raise questions about the underlying factors. However, the lack of knowledge about the key characteristics of proprietary models, such as their size and diversity in training data, prevents us from offering a detailed, conclusive account of their relative performance.

Our benchmarking results reveal a notable difference in performance between shorter and longer text inputs in different tasks. In particular, across the board, models consistently show lower performance in the S2S clustering task compared to the P2P one. This observation can be attributed to the limited presence of contextual clues given the significantly short length of the input text in the S2S clustering task (see Table 4.1). On the other hand, in reranking and retrieval tasks, the majority of the models yield moderately higher scores in S2P tasks. The likely explanation for the latter observation is that the shorter length of the sentences (product names) in S2P tasks can lead to a lower amount of irrelevant information (noise) in the input query. Since product names tend only to encapsulate the critical information about the target product, they can yield more precise and discriminative text (query) representations for similarity matching.

Table 4.3 Comparison of model rankings across datasets with high thematic similarity (see Figure 4.2)

	NV-Embed-v2	gte-Qwen2-7B-instruct	multilingual-e5-large-instruct	stella_en_400M_v5	gte-small	text-embedding-3-large	gte-large	stella_en_1.5B_v5	mxbai-embed-large-v1	bge-large-en-v1.5	gte-base-en-v1.5	bge-base-en-v1.5	gte-base	gte-large-en-v1.5	e5-base-v2	GIST-Embedding-v0	text-embedding-3-small	UAE-Large-V1	e5-large-v2	multilingual-e5-small	GIST-large-Embedding-v0	all-MiniLM-L12-v2	paraphrase-multilingual-MiniLM-L12-v2
clustering-p2p (ours)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
stackexchange-clustering	1	3	10	4	19	7	9	2	16	14	8	20	11	6	18	15	5	12	17	21	13	22	23
reddit-clustering	4	1	16	3	18	10	6	2	9	15	14	21	12	11	17	13	5	8	19	23	7	20	22

4.5 Discussion

Our benchmarking results offer critical insights into the effectiveness of state-of-the-art pre-trained text embedding models in aligning built asset information. One of the key findings of our study is the variability in performance across tasks, even among top-performing models. Our results suggest that model effectiveness is not strongly correlated across model sizes, emphasizing that size alone is not a reliable predictor of model performance in the specialized domain of built asset information management. The interpretation of the relationship between model size and embedding effectiveness is further complicated by the performance gap observed when comparing models pre-trained with and without instruction tuning. Instruction-tuned models showed higher performance in the majority of our benchmark tasks. Considering the larger size of the instruction-tuned models included in our experiments, the latter observation raises an important question for future research: To what extent can instruction-tuning help smaller models adapt to the specialized domain of the built environment? This opens a promising line of investigation into how task-specific training with instruction-based data can better align a model's understanding with the intricate semantics of built asset data, particularly for models with smaller sizes. Finally, in addition to the variability in model performance across different tasks and text input lengths, the results of our comparative examinations highlight the limited transferability of evaluations based on general benchmarks. Our experiments indicate that, even with relatively high thematic similarity, general-purpose benchmarks remain inadequate in capturing the unique semantic complexity and contextual dependencies present in the textual descriptions of the built asset.

The above-mentioned points highlight the critical need for tailored benchmarking datasets to examine the effectiveness of various domain adaptation strategies in this field of research. Our work contributes to the body of research by laying a robust foundation for future evaluations and providing a benchmark that is carefully constructed to reflect the complexities of built asset data. Our proposed datasets cover diverse subdomains and exhibit varying levels of granularity, mirroring real-world scenarios where built products are required to be mapped across various data dictionaries. The datasets can be used not only for evaluating new or

fine-tuned text embedding models for cross-mapping built asset data but also as a contextually rich text corpus to support the training of task-specific language models for other downstream tasks, such as information extraction. Finally, this work contributes to the broader discourse on the transferability of the general-purpose language models' capabilities by focusing on built asset data as a representative example of niche and underexplored domains.

One key limitation of our study is that the text sources used in our work are exclusively in English, limiting the generalizability of our findings to other languages. Another significant challenge was identifying data sources that were both of high quality and could be redistributed as public datasets. In this light, although the developed datasets proved sufficient for our current analysis, future work could benefit from larger-scale datasets and introduce training and validation splits to support new tasks. It is recommended to prioritize exploring more extensive and diverse text sources to include multiple languages and new tasks where the availability of large training splits plays a crucial role, such as text classification or reranking based on cross-encoder architectures. Finally, through the public release of our benchmarking resources, which are openly accessible at a public GitHub repository (Shahinmoghadam, 2024), in alignment with the MTEB's open-source software, we aim to ensure the reproducibility and extendability of our work through community-driven enhancements.

CONCLUSION AND RECOMMENDATIONS

Discussion of findings

The primary aim of this thesis was to investigate the potential of virtual reality and neural language models to enhance user interaction with built asset digital twin systems, focusing on developing intuitive mechanisms for data navigation and retrieval to enhance data accessibility in built asset digital twins. Upon the review of the relevant previous works, elaborated in Chapter 1, this thesis was set out to address three objectives. First, to investigate the potential of virtual reality for enhancing navigation and integration of BIM and IoT data for real-time monitoring and assessment in complex indoor monitoring scenarios. Second, to investigate the application of neural language modeling techniques to enable extraction of entities of interest from user queries, thereby facilitating intuitive information retrieval within virtual digital twin interfaces. Third, to evaluate the effectiveness of state-of-the-art pre-trained language models in capturing and representing semantics specific to built asset terminology.

In Chapter 2, a VR system was developed that integrates BIM spatial data with live IoT data to enable real-time thermal comfort assessment. Thermal comfort was chosen as the case study due to its complexity, requiring both spatial and dynamic sensor data, making it representative of broader real-world applications. The system allows users to visualize and interact with dynamic environmental conditions, improving data accessibility and supporting real-time decision-making. A notable contribution of the work is the development of a cost-effective IoT-enabled thermal imaging prototype, using affordable sensors to provide a practical solution for non-intrusive real-time building monitoring. It was shown how the proposed system can enable what-if analyses, facilitating real-time adjustments to monitoring variables.

The choice of thermal comfort monitoring as a case study was significant due to its complexity, requiring spatial and dynamic sensor data. This choice made the research findings relevant to a broader range of real-world applications that rely on the integration of BIM and IoT data. The

demonstrations presented in Section 2.5 highlight how VR interfaces, when combined with BIM and IoT data, can make complex datasets more accessible and actionable for real-time monitoring, allowing end users to monitor dynamic indoor conditions without being physically present, navigate through virtual spaces, and set up different scenarios to assess outcomes. However, the study was limited in its scope of evaluation. The validation relied on a small number of participants and did not take into account individual factors such as age, sex, and personal preferences that can influence thermal comfort perception. Further research is needed to validate the system's consistency with a larger and more diverse group of participants.

The research efforts presented in Chapter 2 collectively contribute to the challenge of making large and multi-faceted data accessible in a visually coherent and immersive environment. By demonstrating how BIM-IoT integrated data is visualized and interacted with, the chapter contributes an essential component of the overall thesis goal of improving data accessibility. However, from the perspective of *"intuitive"* data accessibility—a key theme of this thesis—, the VR prototype developed in response to RQ1 showed that despite the intuitive representations of integrated data, the scope of accessibility remained bound to predefined user interface components (e.g., nested static menus and drop-down lists), which may not scale well to increasingly large and complex sources of digital twin systems. Such dependence on preconfigured access mechanisms motivated the research in Chapter 3. To address the shortcomings of traditional search methods in immersive interface settings where physical controls and attention spans differ from conventional interfaces, the next phase of the research focused on harnessing language models to enable more intuitive and flexible information retrieval through user queries.

Addressing how neural language modeling techniques can be effectively applied to enable intuitive information retrieval in virtual environments, Chapter 3 focused on the task of semantic tagging of user queries—also known as token classification or named entity recognition task. One of the contributions was structuring a reusable and extensible semantic labeling scheme on

the widely recognized IFC schema, demonstrating the ability to handle a variety of building entities within diverse user queries. This flexibility, combined with the open nature of the IFC schema, ensures that the system can be reused, extended, or modified for broader applications such as semantic parsing of design specifications or regulatory documents.

A key focus of the study presented in Chapter 3 was on the comparison of traditional and more recent deep learning architectures for the token classification of building-related queries, based on the hypothesis that given the specific nature of these queries—often short and packed with domain-specific terms—, the choice of architecture can significantly affect both the accuracy of information retrieval and the computational demands. The findings suggest that traditional deep learning architectures, such as BiLSTM, *"might"* achieve competitive task-specific performance in predictive accuracy while maintaining lower computational costs. Yet, a closer look at the scores presented in Table 3.2 and Table 3.3 reveals a critical observation that seems to be *"recurrent"* in the literature (see Section 3.2.4). A key argument of this thesis is that the close proximities of prediction performance scores presented in Chapter 3 and observed in previous similar works, suggest that the evaluation datasets are limited in quantity or not sufficiently representative. This critical observation leads to the primary limitation of the work presented in Chapter 3: reliance on a relatively small dataset, as the most probable cause for the observed performance parity between BiLSTM and transformer models. A noteworthy remark in this regard is that at the early stage of designing the experiments, it was initially hypothesized that the fine-tuned transformer-based models, with their sophisticated subword tokenization strategies (see Section 3.4.3), should demonstrate a notable superiority at capturing more nuanced domain-specific language (e.g., alphanumeric combinations). This expectation was driven by transformer-based models' strength in contextual embedding, and subword tokenization which leads to fewer out-of-vocabulary tokens, compared to using traditional word embedding models such as word2vec and GloVe (see Table 3.4). Yet, the conducted experiments fall short in providing strong evidence to support the latter hypothesis.

The above discussion point highlights a critical gap in this rapidly expanding body of the literature: the need for more comprehensive and in-depth evaluations of the language models' domain-specific text understanding, particularly in the context of built asset terminology. It was in light of this gap that the next phase of this thesis prioritized on proposing a robust evaluation framework.

Chapter 4 presented the benchmarking study of state-of-the-art pre-trained language models in capturing and representing semantics specific to built asset terminology. The central focus of this research effort was on evaluating the quality of the *"text embeddings"* generated from *"pre-trained"* language models. The rationale behind choosing this particular focus is twofold. First, as discussed in Section 4.1, the quality of the embeddings generated from text input significantly influences the quality of the language model's output in downstream tasks such as information retrieval. Second, despite the significant research opportunities made available by the accelerating proliferation of open-weight pre-trained language models (Sections 1.3.2 and 4.1), this abundance can create a *"paradox of choice"*: making it challenging to identify which models are most suitable for a given domain-specific task. Hence, this study aimed to provide the essential artifacts within a reusable and extensible framework to facilitate the systematic benchmarking of text embedding models, assisting both researchers and practitioners in navigating this increasingly crowded decision space.

The benchmarked models were assessed on three types of tasks: clustering, retrieval, and reranking, which support the main steps of semantic parsing studied in Chapter 3 (see Figure 3.1). In particular, six datasets were developed, comprising over ten thousand domain-specific entries, covering a wide range of topics represented in both short and long textual descriptions. The findings highlight the limitations of general-purpose benchmarking resources in reflecting the model's capability to capture the intricate semantic nuances inherent to built asset descriptions, further underscoring the need for domain-specific evaluations. The empirical results suggest that

models performing well on generic benchmarks may not necessarily perform well on capturing the subtle, domain-specific tasks. The study also reveals variability in model performance, influenced by factors such as model size and pre-training strategies, with larger models generally performing better yet at higher computational costs.

Although not included within the scope of this thesis, an immediate future work is to leverage the findings and resources presented in Chapter 4 to propose enhancements in relation to RQ2, which addresses information retrieval in built asset digital twins systems. An example scenario is as follows: using the benchmarking results to shortlist candidate models based on model size and/or pre-training strategy, and then leveraging the proposed text corpus development and sampling methods to fine-tune candidate models using supervised contrastive learning techniques (Khosla *et al.*, 2020), where even small number of high-quality samples can result in notable predictive performance improvement.

The author of this thesis believes the study reported in Chapter 4 makes a timely and essential contribution to the body of the relevant literature by the development and public release of high-quality and large datasets, curated from reliable sources. These datasets not only facilitate further research in the field but also promote transparency and reproducibility, two essential elements for advancing the evaluation of language models in this domain. Moreover, the presented benchmarking experiments not only enhance our understanding of the capabilities and limitations of current models but also contribute to the broader ongoing efforts focused on leveraging language models to address challenges in built asset information management. Furthermore, it offers encouraging evidence for leveraging pre-trained language models to address challenges related to information alignment in the built asset industry—a research prospect with significant practical value. Given the distributed nature of built asset information, which often resides in multiple sources that are organized according to local structures, explicit integration of all data in a single system may not be feasible. However, the connectivity of

dispersed data can be enhanced through the semantic understanding capability of language models, paving the way for efficient and integrated programmatic access to built asset information. Building on these insights, a potential avenue for further research is highlighted in the next section.

Taken together, the work presented in this thesis underscores the potential of combining immersive interfaces with advanced neural language models in the context of built asset information management. It was demonstrated that VR-enabled BIM-IoT integration provides a powerful means for stakeholders to navigate complex data intuitively, while neural language models can enable more flexible and intuitive query-based retrieval of information. The evaluation of language models on domain-specific benchmarks highlights the need for continued refinement and domain adaptation strategies that can capture the intricate semantics of built environment terminology.

The limitations acknowledged in the previous chapters of the thesis highlight crucial areas for future research. From the broader perspective of this thesis, several recommendations for future research directions are provided below.

Future works

This thesis has opened several promising avenues for future research, summarized as follows:

- **Semantic parsing and end-to-end translation for search:** Regarding the work related to RQ2, this research primarily focused on the entity tagging task (see Chapter 3). Future studies should investigate other critical components of the conventional semantic parsing pipeline, such as intent classification and formal query construction. Additionally, given the promising results of transformer-based models in end-to-end translation of natural language queries directly into formal query representations such as SQL or SPARQL, a comparative evaluation between traditional semantic parsing and end-to-end techniques is recommended.

- **Multi-source search:** While the integration of BIM and IoT data was studied as part of addressing RQ1, the experiments related to RQ2 and RQ3 primarily focused on BIM databases. Future work should investigate effective natural language-based search mechanisms for linked and heterogeneous data sources. Extending the research presented in Chapter 3 beyond structured data in BIM models to include unstructured data sources (e.g., textual documents), will also be essential for creating more comprehensive information retrieval systems. Given the variability of database systems, future research could explore the challenges of building adaptable, flexible systems that can translate queries into various formal query languages such as SQL for relational databases and SPARQL or Cypher for knowledge graph stores.
- **Retrieval from knowledge graphs:** Related to the subject of multi-source search, a particularly promising direction is to leverage the proposed ontology-based mediation framework (Shahinmoghadam & Motamedi, 2021), which aims at unifying programmatic access through a global ontology. In this light, the research on knowledge graph embeddings for question answering (Huang, Zhang, Li & Li, 2019) and its synergy with research on LLMs (Pan *et al.*, 2024) is of particular interest. The text embedding benchmarking artifacts presented in Chapter 4 and earlier experiments on BIM-based knowledge graph embeddings (Shahinmoghadam *et al.*, 2022a,b) can be leveraged to facilitate future works in this direction.
- **Data model/ontology alignment:** Related to multi-source search and knowledge graph-based approaches mentioned above, an essential direction for future research is to focus on the language model-supported "entity linking" task (Sevgili, Shelmanov, Arkhipov, Panchenko & Biemann, 2022). The importance of this research direction in the context of this thesis is twofold. First, semantic entity mapping is a critical task in conventional semantic parsing pipelines—matching the extracted entities with the exact terms used in the database schema/ontology (see the discussion of semantic mapping provided in Section 3.3.1). Second, from the broader perspective of data accessibility, it can facilitate the automated or semi-automated mapping of local and global schemas, thereby improving programmatic

access to diverse data sources in digital twin systems (see the discussion of global schemas provided in Section 1.2.1).

- **Multi-modal embedding:** An interesting future direction is to explore joint representations of different data modalities, including text, images, 3D geometry, and audio. Recent advancements in utilizing pre-trained single-modality models as backbones for multimodal systems, rather than training such models from scratch (Bordes *et al.*, 2024), highlight a promising approach. The benchmarking framework introduced in Chapter 4 provides a solid foundation for selecting pre-trained language models as the language backbone of a multimodal system (e.g., a VR interface with the capability of performing information retrieval based on the user's audio input, in an end-to-end manner).
- **Comprehensive system evaluation:** Related to RQ1, future research should prioritize conducting more extensive system evaluations, particularly for the case of thermal comfort monitoring, while also expanding to include a broader range of diverse, real-world applications. These evaluations should investigate data accessibility challenges in the context of multi-source search systems, as discussed above. Another promising avenue lies in leveraging multimodal capabilities not only for information retrieval (e.g., end-to-end speech-based retrieval) but also for enabling more intuitive interactivity within immersive environments. For example, context-aware assistance could be achieved by integrating multimodal models that provide real-time insights through the combination of live IoT data streams, spatial context, and the user's specific intents.
- **Extension of the benchmarking framework:** A key limitation of the proposed benchmarking datasets (see Chapter 4) is the exclusive reliance on English-language text describing built asset products. This limitation is rooted in the scarcity of high-quality and publicly redistributable sources. Future research should focus on expanding the evaluation datasets to include more diverse, extensive, and multilingual text sources. Moreover, building on the potential for multimodal understanding of user input discussed above, future benchmark datasets should

also include tasks that assess the integration of multiple modalities such as text, audio, image, and spatial data.

BIBLIOGRAPHY

Abdelrahman, M., Macatulad, E., Lei, B., Quintana, M., Miller, C. & Biljecki, F. (2024). What is a Digital Twin Anyway? Deriving the Definition for the Built Environment from over 15,000 Scientific Publications. *arXiv preprint arXiv:2409.19005*.

Albahri, A. H. & Hammad, A. (2017). Simulation-based optimization of surveillance camera types, number, and placement in buildings using BIM. *Journal of Computing in Civil Engineering*, 31(6), 04017055.

Alfano, F. R. d., Dell'Isola, M., Palella, B. I., Riccio, G. & Russi, A. (2013). On the measurement of the mean radiant temperature and its influence on the indoor thermal environment assessment. *Building and Environment*, 63, 79–88.

Anderson, K. (2014). *Design energy simulation for architects: Guide to 3D graphics*. Routledge.

ANSI, A. (2017). ASHRAE Handbook—Fundamentals. ASHRAE.

Aryal, A. & Becerik-Gerber, B. (2019). A comparative study of predicting individual thermal sensation and satisfaction using wrist-worn temperature sensor, thermal camera and ambient temperature sensor. *Building and Environment*, 160, 106223.

ASHRAE. (2017). ASHRAE standard 55 Thermal environmental conditions for human occupancy.

Baghalzadeh Shishehgarkhaneh, M., Keivani, A., Moehler, R. C., Jelodari, N. & Roshdi Laleh, S. (2022). Internet of Things (IoT), Building Information Modeling (BIM), and Digital Twin (DT) in construction industry: A review, bibliometric, and network analysis. *Buildings*, 12(10), 1503.

Bird, S., Klein, E. & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E. et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B. et al. (2024). An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.

buildingSMART. (2020). Industry Foundation Classes, IFC standard 4.0.2.1. buildingSMART. Retrieved from: https://standards.buildingsmart.org/IFC/RELEASE/IFC4/ADD2_TC1/HTML/.

buildingSmart. [Accessed: 2024-06-24]. (2024). buildingSmart International. Retrieved from: https://www.buildingsmart.org/.

buildingSmart International. [Accessed: 2024-06-24]. (2024a). buildingSMART Data Dictionary (bSDD). Retrieved from: https://www.buildingsmart.org/users/services/buildingsmart-data-dictionary/.

buildingSmart International. (2024b). Digital Twins and the Systems Perspective. buildingSmart International.

buildingSmart International. [Accessed: 2024-06-24]. (2024c). IFC 4.3 Documentation. Retrieved from: https://standards.buildingsmart.org/IFC/RELEASE/IFC4_3/.

Cao, T., Lian, Z., Du, H., Miyazaki, R. & Bao, J. (2021). Differences in environmental perception of gender and sleep quality in self-regulating sleep thermal environment. *Indoor and Built Environment*, 30(9), 1568–1579.

Cao, X. & Kosinski, M. (2024). Large language models know how the personality of public figures is perceived by the general public. *Scientific Reports*, 14(1), 6735.

Cappuzzo, R., Papotti, P. & Thirumuruganathan, S. (2020). Creating embeddings of heterogeneous relational datasets for data integration tasks. *Proceedings of the 2020 ACM SIGMOD international conference on management of data*, pp. 1335–1349.

Carre, A. & Williamson, T. (2018). Design and validation of a low cost indoor environment quality data logger. *Energy and Buildings*, 158, 1751–1761.

Casini, M. (2022). Extended reality for smart building operation and maintenance: A review. *Energies*, 15(10), 3785.

Chang, K.-M., Dzeng, R.-J. & Wu, Y.-J. (2018). An automated IoT visualization BIM platform for decision support in facilities management. *Applied sciences*, 8(7), 1086.

Cheung, T., Schiavon, S., Parkinson, T., Li, P. & Brager, G. (2019). Analysis of the accuracy on PMV–PPD model using the ASHRAE Global Thermal Comfort Database II. *Building and Environment*, 153, 205–217.

Chung, S., Kim, J., Baik, J., Chi, S. et al. (2024). Identifying issues in international construction projects from news text using pre-trained models and clustering. *Automation in Construction*, 168, 105875.

Cirani, S., Ferrari, G., Picone, M. & Veltri, L. (2018). *Internet of things: architectures, protocols and standards*. John Wiley & Sons.

Correia, J. B., Abel, M. & Becker, K. (2023). Data management in digital twins: a systematic literature review. *Knowledge and Information Systems*, 65(8), 3165–3196.

Corry, E., Pauwels, P., Hu, S., Keane, M. & O'Donnell, J. (2015). A performance assessment ontology for the environmental and energy management of buildings. *Automation in Construction*, 57, 249–259.

Cosma, A. C. & Simha, R. (2019). Using the contrast within a single face heat map to assess personal thermal comfort. *Building and Environment*, 160, 106163.

Costa, A. A., Lopes, P. M., Antunes, A., Cabral, I., Grilo, A. & Rodrigues, F. M. (2015). 3I buildings: Intelligent, interactive and immersive buildings. *Procedia Engineering*, 123, 7–14.

Davari, S., Shahinmoghadam, M., Motamedi, A. & Poirier, E. (2022). Demystifying the definition of digital twin for built environment. *International conference on construction engineering and project management*, pp. 1122–1129.

Dawood, H., Siddle, J. & Dawood, N. (2019). Integrating IFC and NLP for automating change request validations. *Journal of Information Technology in Construction*, 24, 540–552.

Deng, H., Xu, Y., Deng, Y. & Lin, J. (2022). Transforming knowledge management in the construction industry through information and communications technology: A 15-year review. *Automation in Construction*, 142, 104530.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dingli, A. & Haddod, F. (2019). Interacting with intelligent digital twins. *Design, User Experience, and Usability. User Experience in Advanced Technological Environments: 8th International Conference, DUXU 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21*, pp. 3–15.

Du, J., Zou, Z., Shi, Y. & Zhao, D. (2018). Zero latency: Real-time synchronization of BIM data in virtual reality for collaborative decision-making. *Automation in construction*, 85, 51–64.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A. et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Enevoldsen, K., Kardos, M., Muennighoff, N. & Nielbo, K. L. (2024). The Scandinavian Embedding Benchmarks: Comprehensive Assessment of Multilingual and Monolingual Text Embedding. *arXiv preprint arXiv:2406.02396*.

Eneyew, D. D., Capretz, M. A. & Bitsuamlak, G. T. (2022). Toward smart-building digital twins: BIM and IoT data integration. *IEEE access*, 10, 130487–130506.

Esnaola-Gonzalez, I., Bermúdez, J., Fernandez, I. & Arnaiz, A. (2020). Ontologies for observations and actuations in buildings: A survey. *Semantic Web*, 11(4), 593–621.

Ezen-Can, A. (2020). A Comparison of LSTM and BERT for Small Corpus. *arXiv preprint arXiv:2009.05451*.

Fanger, P. O. (1970). Thermal comfort. Analysis and applications in environmental engineering.

Farghaly, K., Soman, R. K. & Zhou, S. A. (2023). The evolution of ontology in AEC: A two-decade synthesis, application domains, and future directions. *Journal of Industrial Information Integration*, 36, 100519.

FLIR. [Accessed: 12th Jan 2021]. (2021). Retrieved from: https://www.flir.com/globalassets/imported-assets/document/flir-lepton-engineering-datasheet.pdf.

Forth, K., Berggold, P. & Borrmann, A. (2024). Domain-specific fine-tuning of LLM for material matching of BIM elements and Material Passports. *Proc. of 2024 ASCE International Conference on Computing in Civil Engineering*.

Fuchs, S., Dimyadi, J., Witbrock, M. & Amor, R. (2024). Intermediate representations to improve the semantic parsing of building regulations. *Advanced Engineering Informatics*, 62, 102735.

Fukuda, T., Yokoi, K., Yabuki, N. & Motamedi, A. (2019). An indoor thermal environment design system for renovation using augmented reality. *Journal of Computational Design and Engineering*, 6(2), 179–188.

Garcia-Silva, A., Berrio, C. & Gómez-Pérez, J. M. (2019). An empirical study on pre-trained embeddings and language models for bot detection. *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pp. 148–155.

Ghannad, P. & Lee, Y.-C. (2021). Developing an advanced automated modular housing design system using deep learning and Building Information Modeling (BIM). In *Computing in Civil Engineering 2021* (pp. 587–595).

Grondzik, W. T. & Kwok, A. G. (2019). *Mechanical and electrical equipment for buildings*. John wiley & sons.

Hagedorn, T. J., Smith, B., Krishnamurty, S. & Grosse, I. (2019). Interoperability of disparate engineering domain ontologies using basic formal ontology. *Journal of Engineering Design*.

Haller, A. & Polleres, A. (2020). Are we better off with just one ontology on the Web? *Semantic Web*, 11(1), 87–99.

Ham, Y. & Golparvar-Fard, M. (2015). Mapping actual thermal properties to building elements in gbXML-based BIM for reliable building energy performance modeling. *Automation in Construction*, 49, 214–224.

Hinton, G., Vinyals, O. & Dean, J. (2015). Distilling the Knowledge in a Neural Network. Retrieved on 2023-06-22 from: https://arxiv.org/abs/1503.02531v1.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.

Horvitz, E. & Mitchell, T. (2020). From data to knowledge to action: A global enabler for the 21st century. *arXiv preprint arXiv:2008.00045*.

Hosokawa, M., Fukuda, T., Yabuki, N., Michikawa, T. & Motamedi, A. (2016). Integrating CFD and VR for indoor thermal environment design feedback. *CAADRIA proceedings*.

Hu, X. & Assaad, R. H. (2024). A BIM-enabled digital twin framework for real-time indoor environment monitoring and visualization by integrating autonomous robotics, LiDAR-based 3D mobile mapping, IoT sensing, and indoor positioning technologies. *Journal of Building Engineering*, 86, 108901.

Huang, X., Zhang, J., Li, D. & Li, P. (2019). Knowledge graph embedding based question answering. *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 105–113.

IEA. (2019). Global status report for buildings and construction 2019. International Energy Agency, Paris.

IEA. (2020). Tracking Buildings 2020. International Energy Agency, Paris.

IfcOpenShell. [Accessed: 2024-06-24]. (2024). IfcOpenShell. Retrieved from: https://github.com/IfcOpenShell/IfcOpenShell/.

IFTTT. [Accessed: 12th Jan 2021]. (2021). Retrieved from: https://ifttt.com.

ISO, E. et al. (2004). Ergonomics of the Thermal Environment–Analytical Determination and Interpretation of Heat Stress Using Calculation of the Predicted Heat Strain. International Organization for Standardization Geneva.

ISO 16739-1. (2024). Industry Foundation Classes (IFC) for data sharing in the construction and facility management industries. Retrieved from: https://www.iso.org/standard/84123.html.

Järvelin, K. & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4), 422–446.

Jeon, K., Lee, G., Yang, S. & Jeong, H. D. (2022a). Named entity recognition of building construction defect information from text with linguistic noise. *Automation in Construction*, 143, 104543.

Jeon, K., Lee, G., Yang, S. & Jeong, H. D. (2022b). Named entity recognition of building construction defect information from text with linguistic noise. *Automation in Construction*, 143, 104543. doi: 10.1016/j.autcon.2022.104543.

Jeon, K., Lee, G., Yang, S., Kim, Y. & Suh, S. (2024). Dynamic building defect categorization through enhanced unsupervised text classification with domain-specific corpus embedding methods. *Automation in Construction*, 157, 105182.

Jiang, S., Jiang, L., Han, Y., Wu, Z. & Wang, N. (2019). OpenBIM: An enabling solution for information interoperability. *Applied Sciences*, 9(24), 5358.

Johansson, M., Roupé, M. & Viklund Tallgren, M. (2014). From BIM to VR-Integrating immersive visualizations in the current design process. *Fusion-Proceedings of the 32nd eCAADe Conference-Volume 2 (eCAADe 2014)*, pp. 261–269.

Jung, Y., Hockenmaier, J. & Golparvar-Fard, M. (2024). Transformer language model for mapping construction schedule activities to uniformat categories. *Automation in Construction*, 157, 105183.

Jurafsky, D. & Martin, J. H. [Online manuscript released August 20, 2024]. (2024). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. Retrieved from: https://web.stanford.edu/~jurafsky/slp3/.

Karjalainen, S. (2007). Gender differences in thermal comfort and use of thermostats in everyday thermal environments. *Building and environment*, 42(4), 1594–1603.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C. & Krishnan, D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, 33, 18661–18673.

Khurana, D., Koli, A., Khatter, K. & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, 82(3), 3713–3744.

Koo, B., Jung, R., Yu, Y. & Kim, I. (2021). A geometric deep learning approach for checking element-to-entity mappings in infrastructure building information models. *Journal of Computational Design and Engineering*, 8(1), 239–250.

Korotkova, N., Benders, J., Mikalef, P. & Cameron, D. (2023). Maneuvering between skepticism and optimism about hyped technologies: Building trust in digital twins. *Information & Management*, 60(4), 103787.

La Gennusa, M., Nucara, A., Rizzo, G. & Scaccianoce, G. (2005). The calculation of the mean radiant temperature of a subject exposed to the solar radiation—a generalised algorithm. *Building and Environment*, 40(3), 367–375.

Lagüela, S., Díaz-Vilariño, L., Martínez, J. & Armesto, J. (2013). Automatic thermographic and RGB texture of as-built BIM for energy rehabilitation purposes. *Automation in Construction*, 31, 230–240.

Lee, C., Roy, R., Xu, M., Raiman, J., Shoeybi, M., Catanzaro, B. & Ping, W. (2024a). NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. *arXiv preprint arXiv:2405.17428*.

Lee, J., Dai, Z., Ren, X., Chen, B., Cer, D., Cole, J. R., Hui, K., Boratko, M., Kapadia, R., Ding, W. et al. (2024b). Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*.

Lee, K. S., Lee, J.-J., Aucremanne, C., Shah, I. & Ghahramani, A. (2023). Towards democratization of digital twins: Design principles for transformation into a human-building interface. *Building and Environment*, 244, 110771.

Lei, B., Janssen, P., Stoter, J. & Biljecki, F. (2023). Challenges of urban digital twins: A systematic review and a Delphi expert survey. *Automation in Construction*, 147, 104716.

Li, D., Menassa, C. C. & Kamat, V. R. (2018). Non-intrusive interpretation of human thermal comfort through analysis of facial infrared thermography. *Energy and Buildings*, 176, 246–261.

Li, D., Menassa, C. C. & Kamat, V. R. (2019). Robust non-intrusive interpretation of occupant thermal comfort in built environments with low-cost networked thermal cameras. *Applied energy*, 251, 113336.

Li, J., Sun, A., Han, J. & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1), 50–70.

Li, J., Sun, A., Han, J. & Li, C. (2022). A Survey on Deep Learning for Named Entity Recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1), 50–70. doi: 10.1109/TKDE.2020.2981314. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Li, R., Mo, T., Yang, J., Li, D., Jiang, S. & Wang, D. (2021). Bridge inspection named entity recognition via BERT and lexicon augmented machine reading comprehension neural model. *Advanced Engineering Informatics*, 50, 101416.

Li, T., Rui, Y., Zhu, H., Lu, L. & Li, X. (2024). Comprehensive digital twin for infrastructure: A novel ontology and graph-based modelling paradigm. *Advanced Engineering Informatics*, 62, 102747.

Li, X. & Li, J. (2023). AnglE-optimized Text Embeddings. *arXiv preprint arXiv:2309.12871*.

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X. et al. (2023). Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*.

Liu, C.-c., Kuo, W.-l., Shiu, R.-s. & Wu, I.-c. (2014). Estimating and visualizing thermal comfort level via a predicted mean vote in a BIM system. *Osaka University: Osaka, Japan*.

Liu, Q., Kusner, M. J. & Blunsom, P. (2020a). A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Deng, H. & Ju, Q. (2020b). Fastbert: a self-distilling bert with adaptive inference time. *arXiv preprint arXiv:2004.02178*.

Liu, Z., Deng, Z. & Demian, P. (2018). Integration of Building Information Modelling (BIM) and sensor technology: a review of current developments and future outlooks. *Proceedings of the 2nd International Conference on Computer Science and Application Engineering*, pp. 1–5.

Love, P. E. & Matthews, J. (2019). The 'how'of benefits management for digital technology: From engineering to asset management. *Automation in Construction*, 107, 102930.

Lu, S., Wang, W., Wang, S. & Cochran Hameen, E. (2019). Thermal comfort-based personalized models with non-intrusive sensing technique in office buildings. *Applied Sciences*, 9(9), 1768.

Ma, G., Liu, Y. & Shang, S. (2019). A building information model (BIM) and artificial neural network (ANN) based system for personal thermal comfort evaluation and energy efficient design of interior space. *Sustainability*, 11(18), 4972.

Mikkilineni, A. K., Dong, J., Kuruganti, T. & Fugate, D. (2019). A novel occupancy detection solution using low-power IR-FPA based wireless occupancy sensor. *Energy and Buildings*, 192, 63–74.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Minerva, R., Biru, A. & Rotondi, D. (2015). Towards a definition of the Internet of Things (IoT). *IEEE Internet Initiative*, 1(1), 1–86.

Moran, A., Gadepally, V., Hubbell, M. & Kepner, J. (2015). Improving big data visual analytics with interactive virtual reality. *2015 IEEE high performance extreme computing conference (HPEC)*, pp. 1–6.

Moretti, N., Xie, X., Merino Garcia, J., Chang, J. & Kumar Parlikad, A. (2023). Federated data modeling for built environment digital twins. *Journal of Computing in Civil Engineering*, 37(4), 04023013.

Motamedi, A. & Shahinmoghadam, M. (2021). BIM-IoT-integrated architectures as the backbone of cognitive buildings: Current state and future directions. *BIM-enabled Cognitive Computing for Smart Built Environment*, 45–68.

Motamedi, A., Soltani, M. M., Setayeshgar, S. & Hammad, A. (2016). Extending IFC to incorporate information of RFID tags attached to building elements. *Advanced Engineering Informatics*, 30(1), 39–53.

Motamedi, A., Wang, Z., Yabuki, N., Fukuda, T. & Michikawa, T. (2017). Signage visibility analysis and optimization system using BIM-enabled virtual reality (VR) environments. *Advanced Engineering Informatics*, 32, 248–262.

Muennighoff, N., Tazi, N., Magne, L. & Reimers, N. (2022). MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Naseem, U., Razzak, I., Musial, K. & Imran, M. (2020). Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, 113, 58–69.

Natephra, W. & Motamedi, A. (2019a). BIM-based live sensor data visualization using virtual reality for monitoring indoor conditions. *Intelligent & Informed—Proceedings of the 24th CAADRIA Conference; Haeusler, M., Schnabel, MA, Fukuda, T., Eds*, pp. 191–200.

Natephra, W. & Motamedi, A. (2019b). Live data visualization of IoT sensors using augmented reality (AR) and BIM. *36th International symposium on automation and robotics in construction (ISARC 2019)*.

Natephra, W., Motamedi, A., Yabuki, N. & Fukuda, T. (2017). Integrating 4D thermal information with BIM for building envelope thermal performance analysis and thermal comfort evaluation in naturally ventilated environments. *Building and Environment*, 124, 194–208.

NBS. [Accessed: 2024-06-24]. (2024a). National Building Specification. Retrieved from: https://www.thenbs.com/.

NBS. [Accessed: 2024-06-24]. (2024b). Uniclass. Retrieved from: https://uniclass.thenbs.com/.

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C. et al. (2022). Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

Ngarambe, J., Yun, G. Y. & Santamouris, M. (2020). The use of artificial intelligence (AI) methods in the prediction of thermal comfort in buildings: Energy implications of AI-based thermal comfort controls. *Energy and Buildings*, 211, 109807.

Offermann, P., Levina, O., Schönherr, M. & Bub, U. (2009). Outline of a design science research process. *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*, pp. 1–11.

Onal, K. D., Zhang, Y., Altingovde, I. S., Rahman, M. M., Karagoz, P., Braylan, A., Dang, B., Chang, H.-L., Kim, H., McNamara, Q. et al. (2018). Neural information retrieval: At the end of the early years. *Information Retrieval Journal*, 21, 111–182.

OpenCV. [Accessed: 12th Jan 2021]. (2021). Retrieved from: https://github.com/opencv/opencv.

Ostendorff, M., Ash, E., Ruas, T., Gipp, B., Moreno-Schneider, J. & Rehm, G. (2021). Evaluating document representations for content-based legal literature recommendations. *Proceedings of the eighteenth international conference on artificial intelligence and law*, pp. 109–118.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J. & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Panetto, H. (2007). Towards a classification framework for interoperability of enterprise applications. *International Journal of Computer Integrated Manufacturing*, 20(8), 727–740.

Parsons, K. (2007). *Human thermal environments: the effects of hot, moderate, and cold environments on human health, comfort and performance*. CRC press.

Peffers, K., Tuunanen, T., Rothenberger, M. A. & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3), 45–77.

Pennington, J., Socher, R. & Manning, C. (2014a). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

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

Pérez, J., Arenas, M. & Gutierrez, C. (2009). Semantics and complexity of SPARQL. *ACM Transactions on Database Systems (TODS)*, 34(3), 1–45.

Pirker, J., Loria, E., Safikhani, S., Künz, A. & Rosmann, S. (2022). Immersive virtual reality for virtual and digital twins: A literature review to identify state of the art and perspectives. *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 114–115.

Poirier, E. A. & Motamedi, A. (2024). An Extensible, Service Oriented Digital Twinning Framework to Enable Implementation Planning and Capability Development in the Built Asset Industry. *Proc. of the CIB W78 Conference 2024*.

Postman. [Accessed: 12th Jan 2021]. (2021). Retrieved from: https://www.postman.com/.

Prohasky, D. & Watkins, S. (2014). Low cost hot-element anemometry verses the TFI Cobra. *19th Australasian Fluid Mechanics Conference*.

PyLepton. [Accessed: 12th Jan 2021]. (2021). Retrieved from: https://github.com/groupgets/pylepton.

Radford, A. (2018). Improving language understanding by generative pre-training.

Rafsanjani, H. N. & Nabizadeh, A. H. (2023). Towards digital architecture, engineering, and construction (AEC) industry through virtual design and construction (VDC) and digital twin. *Energy and Built Environment*, 4(2), 169–178.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1), 86.

Reimers, N. & Gurevych, I. (2019, 11). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Retrieved from: http://arxiv.org/abs/1908.10084.

Roberts, C. J., Pärn, E. A., Edwards, D. J. & Aigbavboa, C. (2018). Digitalising asset management: concomitant benefits and persistent challenges. *International Journal of Building Pathology and Adaptation*, 36(2), 152–173.

Rosenberg, A. & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 410–420.

Rouhizadeh, H., Nikishina, I., Yazdani, A., Bornet, A., Zhang, B., Ehrsam, J., Gaudet-Blavignac, C., Naderi, N. & Teodoro, D. (2024). A Dataset for Evaluating Contextualized Representation of Biomedical Concepts in Language Models. *Scientific Data*, 11(1), 455.

Rupp, R. F., Vásquez, N. G. & Lamberts, R. (2015). A review of human thermal comfort in the built environment. *Energy and buildings*, 105, 178–205.

Sacks, R., Eastman, C., Lee, G. & Teicholz, P. (2018). *BIM handbook: A guide to building information modeling for owners, designers, engineers, contractors, and facility managers*. John Wiley & Sons.

Sahari, K. M., Jalal, M. A., Homod, R. & Eng, Y. (2013). Dynamic indoor thermal comfort model identification based on neural computing PMV index. *IOP conference series: Earth and environmental science*, 16(1), 012113.

Saka, A. B., Oyedele, L. O., Akanbi, L. A., Ganiyu, S. A., Chan, D. W. & Bello, S. A. (2023). Conversational artificial intelligence in the AEC industry: A review of present status, challenges and opportunities. *Advanced Engineering Informatics*, 55, 101869.

Salamone, F., Belussi, L., Currò, C., Danza, L., Ghellere, M., Guazzi, G., Lenzi, B., Megale, V. & Meroni, I. (2018). Integrated method for personal thermal comfort assessment and optimization through users' feedback, IoT and machine learning: A case study. *Sensors*, 18(5), 1602.

Salehinejad, H., Sankar, S., Barfett, J., Colak, E. & Valaee, S. (2017). Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*.

Sanh, V., Debut, L., Chaumond, J. & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schuster, M. & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11), 2673–2681.

Sels, S., Verspeek, S., Ribbens, B., Bogaerts, B., Vanlanduit, S., Penne, R. & Steenackers, G. (2019). A CAD matching method for 3D thermography of complex objects. *Infrared Physics & Technology*, 99, 152–157.

Sevgili, Ö., Shelmanov, A., Arkhipov, M., Panchenko, A. & Biemann, C. (2022). Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3), 527–570.

Seydgar, M., Motamedi, A. & Poirier, É. (2022). Performance Assessment of Deep Neural Networks for Classification of IFC Objects From Point Cloud With Limited Labeled Data. *Transforming Construction with Reality Capture Technologies*.

Shahinmoghadam, M., Motamedi, A. & Soltani, M. (2022a). Leveraging Textual Information for Knowledge Graph-oriented Machine Learning: A Case Study in the Construction Industry. *Proceedings of the 29th International Workshop on Intelligent Computingin Engineering (EG-ICE)*.

Shahinmoghadam, M. [Accessed: November 22, 2023]. (2023). ifc-neural-parser. Retrieved from: https://github.com/mehrzadshm/ifc-neural-parser.

Shahinmoghadam, M. [Accessed: 2024-10-20]. (2024). built-bench-paper (GitHub repository). Retrieved from: https://github.com/mehrzadshm/built-bench-paper.

Shahinmoghadam, M. & Motamedi, A. (2019). Review of BIM-centred IoT Deployment–State of the Art, Opportunities, and Challenges. *Proceedings of the 36th International Symposium on Automation and Robotics in Construction (ISARC 2019)*, pp. 1268–1275.

Shahinmoghadam, M. & Motamedi, A. (2021). An ontology-based mediation framework for integrating federated sources of BIM and IoT data. *Proceedings of the 18th International Conference on Computing in Civil and Building Engineering: ICCCBE 2020*, pp. 907–923.

Shahinmoghadam, M., Motamedi, A. & Soltani, M. M. (2022b). Enabling downstream machine-learning over the textual information contained in building knowledge graphs. *EC3 Conference 2022*, 3, 0–0.

Shahinmoghadam, M., Kahou, S. E. & Motamedi, A. (2024). Neural semantic tagging for natural language-based search in building information models: Implications for practice. *Computers in Industry*, 155, 104063.

Shahzad, M., Shafiq, M. T., Douglas, D. & Kassem, M. (2022). Digital twins in built environments: an investigation of the characteristics, applications, and challenges. *Buildings*, 12(2), 120.

Shamshiri, A., Ryu, K. R. & Park, J. Y. (2024). Text mining and natural language processing in construction. *Automation in Construction*, 158, 105200.

Sresakoolchai, J. & Kaewunruen, S. (2021). Integration of building information modeling and machine learning for railway defect localization. *IEEE Access*, 9, 166039–166047.

Stadtmann, F., Mahalingam, H. P. & Rasheed, A. (2023). Data Integration Framework for Virtual Reality Enabled Digital Twins. *2023 IEEE 9th World Forum on Internet of Things (WF-IoT)*, pp. 1–6.

Standard, I. & Iso, B. (1998). Ergonomics of the thermal environment—instruments for measuring physical quantities. *International Organization for Standardization*.

Tang, S., Shelden, D. R., Eastman, C. M., Pishdad-Bozorgi, P. & Gao, X. (2019). A review of building information modeling (BIM) and the internet of things (IoT) devices integration: Present status and future trends. *Automation in construction*, 101, 127–139.

Tartarini, F. & Schiavon, S. (2020). pythermalcomfort: A Python package for thermal comfort research. *SoftwareX*, 12, 100578.

Tartarini, F., Schiavon, S., Cheung, T. & Hoyt, T. (2020). CBE Thermal Comfort Tool: Online tool for thermal comfort calculations and visualizations. *SoftwareX*, 12, 100563.

Tjong Kim Sang, E. F. & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147. Retrieved from: https://aclanthology.org/W03-0419.

Trojahn, C., Vieira, R., Schmidt, D., Pease, A. & Guizzardi, G. (2022). Foundational ontologies meet ontology matching: A survey. *Semantic Web*, 13(4), 685–704.

Tuhaise, V. V., Tah, J. H. M. & Abanda, F. H. (2023). Technologies for digital twin applications in construction. *Automation in Construction*, 152, 104931.

Underwood, J. & Isikdag, U. (2011). Emerging technologies for BIM 2.0. *Construction Innovation*, 11(3), 252–258.

Unrealengine. [Accessed: 12th Jan 2021]. (2021a). Retrieved from: https://www.unrealengine.com/en-US/datasmith.

Unrealengine. [Accessed: 12th Jan 2021]. (2021b). Retrieved from: https://www.unrealengine.com/marketplace/en-US/product/varest-plugin.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Walikewitz, N., Jänicke, B., Langner, M., Meier, F. & Endlicher, W. (2015). The difference between the mean radiant temperature and the air temperature within indoor environments: A case study during summer conditions. *Building and Environment*, 84, 151–161.

Wang, C., Cho, Y. K. & Gai, M. (2013a). As-is 3D thermal modeling for existing building envelopes using a hybrid LIDAR system. *Journal of Computing in Civil Engineering*, 27(6), 645–656.

Wang, H., Gluhak, A., Meissner, S. & Tafazolli, R. (2013b). Integration of BIM and live sensing information to monitor building energy performance. *Proceedings of the 30th CIB W78 International Conference*, 30, 344–352.

Wang, N., Issa, R. R. A. & Anumba, C. J. (2022a). Named Entity Recognition Algorithm for iBISDS Using Neural Network. *Proc. Construction Research Congress*, pp. 521–529. doi: 10.1061/9780784483961.055.

Wang, N., Issa, R. R. A. & Anumba, C. J. (2022b). NLP-Based Query-Answering System for Information Extraction from Building Information Models. *Journal of Computing in Civil Engineering*, 36(3), 04022004. doi: 10.1061/(ASCE)CP.1943-5487.0001019. Publisher: American Society of Civil Engineers.

Wang, N., Issa, R. R. A. & Anumba, C. J. (2022c). Transfer learning-based query classification for intelligent building information spoken dialogue. *Automation in Construction*, 141, 104403. doi: 10.1016/j.autcon.2022.104403.

Wang, N., Issa, R. R. & Anumba, C. J. (2022d). Transfer learning-based query classification for intelligent building information spoken dialogue. *Automation in Construction*, 141, 104403.

Wang, T., Gan, V. J., Hu, D. & Liu, H. (2022e). Digital twin-enabled built environment sensing and monitoring through semantic enrichment of BIM with SensorML. *Automation in Construction*, 144, 104625.

Wang, Y., Xiao, B., Bouferguene, A. & Al-Hussein, M. (2024a). Proactive safety hazard identification using visual–text semantic similarity for construction safety management. *Automation in Construction*, 166, 105602.

Wang, Z., Bergés, M. & Akinci, B. (2024b). Pre-trained language model based method for building information model to building energy model transformation at metamodel level. *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, 41, 17–25.

Wang, Z., Yu, H., Luo, M., Wang, Z., Zhang, H. & Jiao, Y. (2019). Predicting older people's thermal sensation in building environment through a machine learning approach: Modelling, interpretation, and application. *Building and Environment*, 161, 106231.

Wilkho, R. S., Chang, S. & Gharaibeh, N. G. (2024). FF-BERT: A BERT-based ensemble for automated classification of web-based text on flash flood events. *Advanced Engineering Informatics*, 59, 102293.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q. & Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Retrieved from: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M. & Yang, Z. (2022). Natural language processing for smart construction: Current status and future directions. *Automation in Construction*, 134, 104059.

Wu, H., Shen, G. Q., Lin, X., Li, M. & Li, C. Z. (2021). A transformer-based deep learning model for recognizing communication-oriented entities from patents of ICT in construction. *Automation in Construction*, 125, 103608.

Xie, Q., Zhou, X., Wang, J., Gao, X., Chen, X. & Liu, C. (2019). Matching real-world facilities to building information modeling data using natural language processing. *Ieee Access*, 7, 119465–119475.

Yin, M., Tang, L., Webster, C., Li, J., Li, H., Wu, Z. & Cheng, R. C. K. (2023a). Two-stage Text-to-BIMQL semantic parsing for building information model extraction using graph neural networks. *Automation in Construction*, 152, 104902. doi: 10.1016/j.autcon.2023.104902.

Yin, M., Tang, L., Webster, C., Xu, S., Li, X. & Ying, H. (2023b). An ontology-aided, natural language-based approach for multi-constraint BIM model querying. *Journal of Building Engineering*, 107066.

Zabin, A., González, V. A., Zou, Y. & Amor, R. (2022). Applications of machine learning to BIM: A systematic literature review. *Advanced Engineering Informatics*, 51, 101474.

Zhang, R. & El-Gohary, N. (2023). Transformer-based approach for automated context-aware IFC-regulation semantic information alignment. *Automation in Construction*, 145, 104540.

Zheng, J. & Fischer, M. (2023). BIM-GPT: a Prompt-Based Virtual Assistant Framework for BIM Information Retrieval. *arXiv preprint arXiv:2304.09333*.

Zheng, Z., Lu, X.-Z., Chen, K.-Y., Zhou, Y.-C. & Lin, J.-R. (2022a). Pretrained domain-specific language model for natural language processing tasks in the AEC domain. *Computers in Industry*, 142, 103733.

Zheng, Z., Zhou, Y.-C., Lu, X.-Z. & Lin, J.-R. (2022b). Knowledge-informed semantic alignment and rule interpretation for automated compliance checking. *Automation in Construction*, 142, 104524.

Zhou, Y.-C., Zheng, Z., Lin, J.-R. & Lu, X.-Z. (2022). Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking. *Computers in Industry*, 142, 103746.