UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Improving Request for Information (RFI) Processing in Construction Projects using Natural Language Processing (NLP) Techniques

Muneeb Afzal

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of
Associate Professor Johnny Kwok-Wai Wong
Associate Professor Alireza Ahmadian Fard Fini

University of Technology Sydney
Faculty of Design Architecture and Building

November 2024

Certificate of original authorship

I, Muneeb Afzal, declare that this thesis is submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy, in the School of Built Environment of the Faculty of Design, Architecture and Building at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Signature:

Production Note:
Signature removed prior to publication.

Date: 13 November 2024

Note on the thesis format

This thesis has been submitted in fulfillment of the requirements for the degree of Doctor of Philosophy and follows the thesis-by-compilation format. It adheres to the guidelines outlined in the "Graduate Research Candidature Management, Thesis Preparation and Submission Procedures 2023" of the University of Technology Sydney. The referencing style used throughout the thesis is based on the 7th edition of the American Psychological Association (APA).

Acknowledgement

I would like to extend my deepest gratitude to my exceptional supervisors, Associate Professor Johnny Wong and Associate Professor Alireza Fard Fini. Their research expertise and academic guidance have played an instrumental role in shaping my development as a doctoral student. Associate Professor Johnny Wong provided invaluable, constructive feedback throughout my journey. His unwavering support, kindness, and wealth of experience in research were key in helping me seamlessly achieve critical milestones without undue pressure. He was always available and deeply invested in my success, making the process enriching. Associate Professor Alireza Fard Fini, helped me refine the focus of my research, providing unique insights from various perspectives. His ability to pinpoint the core challenges I sought to address significantly influenced my approaches and ultimately shaped the direction of my research. The support, mentorship, and dedication of my supervisors have been crucial to my growth, and I am immensely grateful for their guidance.

I would like to extend my sincere gratitude to the School of Built Environment and the Faculty of Design, Architecture, and Building at UTS for the invaluable support. The essential research space and outstanding facilities provided by the School of Built Environment were crucial to the successful completion of my thesis. I am also deeply thankful to the DAB HDR Support team, for their consistent assistance. Additionally, I would like to express my appreciation to the Graduate Research School for their unwavering support throughout my candidature.

Furthermore, I am grateful for the wonderful friendships I formed at UTS over the past three years, particularly with my friends from the fishbowl tank—Ellie, Isabella, Flora, and Mehrafarin who made this journey more memorable. A special thanks to Mona, whose expertise in design management provided me with invaluable advice that significantly influenced my research.

My deepest gratitude goes to my better half, Rabiya, for her unwavering love and support. I would also like to thank my dear sons, Zaviyar and Salaar, who have been a constant source of joy during this period. Additionally, I am grateful to my parents for their kindness, support, and prayers, as well as for the upbringing that has shaped me into the person I am today.

Research Publications Acknowledgements

Publication Details	Status	Location in this thesis
Afzal, M., Wong, J. K. W., & Fini, A.A.F. (2024). Towards digital approach for managing request for information (RFI) in construction projects: a literature review. *Construction Innovation*.	Published and available online in Construction Innovation journal.	This review paper is incorporated in chapter 2 of thesis.
Afzal, M., Wong, J. K. W., & Fini, A. A. F. (2023, August). Unlocking Insights: Analysing Construction Issues in Request for Information (RFI) Documents with Text Mining and Visualisation. In *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)* (pp. 1-6). IEEE.	Published and available online in IEEE-CASE-2023 conference.	Some parts of this conference paper are incorporated in chapter 4 of thesis.
Afzal, M., Wong, J. K. W., Fini, A. A. F. & Sankaran, S (2025). From data to decisions: Automated RFI content analysis for efficient RFI management. *Project Management Journal*.	Article under review	This paper is incorporated in chapter 4 of thesis.
Afzal, M., Wong, J. K. W., & Fini, A. A. F. (202). A two-step deep learning-driven NLP pipeline for efficient information extraction from construction RFIs. *Smart and Sustainable Built Environment.*	Article under review	This paper is incorporated in chapter 5 of the thesis.

The extent of the contribution of the student and other authors:

The major part of the research, including the literature review, developed methodology, and verification of the results, has been done by the student in all papers. Other authors provided guidance and insights to improve the articles and, in some cases, revised the writing of the articles.

Signature of Author (Muneeb Afzal)	Production Note: Signature removed prior to publication.	**Date: 07/10/2024**
Signature of Co-Author (Johnny Kwok-Wai Wong)	Production Note: Signature removed prior to publication.	**Date: 30 Oct 2024**
Signature of Co-Author (Alireza Ahmadian Fard Fini)	Production Note: Signature removed prior to publication.	**Date: 29/10/2024**
Signature of Co-Author (Shankar Sankaran)	**Shankar Sankaran**	**Date:4-Nov-2024**

Table of Contents

List of tables

List of figures

Abbreviations

AEC	Architecture, Engineering, and Construction
AECO	Architecture, Engineering, Construction, and Operations
AI	Artificial Intelligence
ANN	Artificial Neural Networks
BEP	BIM Execution Plan
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
BIM	Building Information Modelling
BoW	Bag-of-Words
CBOW	Continuous Bag-of-Words
CDE	Common Data Environment
CLD	Causal Loop Diagram
CMR	Construction Management at-Risk
CNN	Convolutional Neural Network
CRF	Conditional Random Field
DB	Design-Build
DBB	Design-Bid-Build
DL	Deep Learning
DWG	Drawing Reference (Annotation tag)
EDMS	Electronic Document Management Systems
FN	False Negatives
FP	False Positives
GRU	Gated Recurrent Unit
IDF	Inverse Document Frequency
IPD	Integrated Project Delivery
KNN	k-Nearest Neighbours
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
LOC	Location (Annotation tag)
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
NB	Naïve Bayes
NER	Named Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OCR	Optical Character Recognition
PDS	Project Delivery System
POS	Parts of Speech
PRO	Problematic Component (Annotation tag)
RF	Random Forest
RFI	Request for Information
RNN	Recurrent Neural Networks
ROI	Return on Investment
SD	System Dynamics
ST	Systems Thinking
SVM	Support Vector Machine

TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
TN	True Negative
TP	True Positives
t-SNE	t-distributed Stochastic Neighbour Embedding
VDC	Virtual Design and Construction

Abstract

Request for information (RFI) document is an essential communication tool to seek clarifications across the project lifecycle. Preparing, evaluating, and responding to the RFIs consume resources such as time and cost. Moreover, a burst of RFIs, unresolved RFIs, or slow responses to RFIs may lead to project risks such as schedule delay and cost escalation. Finding effective ways of managing these risks associated with RFI documents can reduce their frequency and turnaround time, minimising their adverse impacts on projects. Previous research has tried to categorise the content of the RFIs to identify the root causes and develop best practices; however, their classification methods primarily depend on manual content analysis, which is inherently labour-intensive and time-consuming. Such processes lack effectiveness, and their outcomes tend to be error-prone and biased. Moreover, there is no existing automated framework in the body of knowledge to analyse the unstructured text data in the RFI document. Therefore, this research aims to develop a mechanism for automated, efficient, and unbiased text classification and entity extraction for efficient information extraction from the unstructured RFI statements.

Accordingly, this research leverages natural language processing (NLP) techniques to decode unstructured information within RFIs into easily understandable and actionable insights. The developed work consists of three primary models: (1) an NLP-based multiclass text classification model employing deep learning-based recurrent neural networks (RNNs) to categorise RFIs according to their project phase, supplemented with a topic modelling approach to visualise key topics and themes from the RFIs. (2) an NLP-based multiclass text classification model designed to identify the predominant issue within RFIs, and (3) an information extraction —named entity recognition (NER) model aimed at obtaining critical entities from RFIs. This comprehensive research enables the early detection of design, execution, procurement issues, and specific issues presented by an RFI, such as coordination, constructability, specification/scope, design/drawing discrepancies, and review/approval. The NER model facilitates the automated identification of problematic components, their locations, and relevant drawing references in an automated manner.

With this wealth of information readily available, project stakeholders can optimise the RFI process, potentially shortening review periods and streamlining the prioritisation of RFIs. The efficiency of the developed models was assessed through experimental investigation and

evaluation on datasets from actual construction projects. This evaluation focused on their information extraction and text classification abilities. The results demonstrated that the automated processing of RFIs through the NLP models not only outperformed human performance in terms of accuracy and efficiency but also showed strong performance on independent test sets. Specifically, the models achieved higher precision, recall and F-1 scores in identifying key entities and classifying issue types during RFI analysis. This strong performance refers to the models' ability to generalize well across unseen data, achieving consistent evaluation metrics further validating their effectiveness and reliability for real-world application in construction document analysis.

The significance of this research lies in the development of multiple text classification models powered by NLP and deep learning pipelines, enabling stakeholders to gain meaningful insights from RFIs and efficiently resolve issues. It enhances construction communication workflows and lays the foundation for intelligent, real-time RFI analysis. Furthermore, the research offers a detailed implementation roadmap, guiding researchers and practitioners on how to integrate these models into real-world construction settings. This roadmap serves as a practical tool to elevate current RFI management practices and improve the quality of future project documentation.

Chapter 1: Introduction

This chapter introduces the research and discusses related background, problem statement, and gap analysis. Additionally, it presents the research questions, aims and objectives, the study's significance, and the thesis's structure.

1.1 Introduction to the research

The construction industry is dynamic and complex, requiring efficient stakeholder communication from different disciplines. Accordingly, different formal channels are employed in traditional and Building Information Modelling (BIM)-enabled projects to support this communication. One such communication tool is the request for information (RFI) document. RFI is a document through which a contractor formally seeks clarification from designers regarding any issues related to contract documents (Abdel-Monem & Hegazy, 2013). With the emergence of RFI, a process is initiated that is overwhelming on a project and has repercussions for the project stakeholders. Hence, RFI is interchangeably an indicator of a project's productivity, success and interface health (Psomas & Alzraiee, 2020). RFIs are widely recognized as a prevalent challenge in the architecture, engineering, and construction (AEC) industry, and act as an indicator of poor contract documentation both in Australia (Gajendran and Brewer, 2012; Simpeh et al., 2011; AIB, 2005) and internationally (Tribelsky and Sacks, 2011; Andi and Minato, 2003). Even though RFIs are essential communication tools, they carry inherent risks (Hughes et al., 2013). Incomplete, lack of detail/clarity in specifications and lack of constructability in design packages may result in a burst of RFIs, impeding work progress (Hasan et al., 2018; Jarkas and Bitar, 2012). One of the significant issues with RFI is that they are unpredictable and cannot be included in a project's baseline schedule. Regardless, RFIs can become part of the critical path during construction, and untimely RFI response time can potentially lead to project delays (Kelly & Ilozor, 2020).

Reviewing RFI questions and proposing a suitable response to satiate the contractor's requirements is time-consuming. The person-hours required to resolve an RFI directly translate into a financial burden borne by the client or consultant. Different research studies have attempted to quantify the cost implications of reviewing and closing an RFI. Hughes et al. (2013) estimated the cost of reviewing and responding to RFIs for projects conducted between 2001 and 2012 at approximately US$859,680 per project, based on a review period of 6,368 hours. Aibinu et al. (2019) adopted the same methodology in their analysis. They estimated this cost in Australia to be US$682,500 per project (based on 650 RFIs per project, 4 hours of

administration at AU$100/hour and 4 hours of technical hours at AU$250/hour), using the 2019 AU$/US$ exchange rate. The average cost per RFI is an essential metric for comparing project performance. According to Hughes et al. (2013), there are approximately 796 RFIs per project, and it takes about nine days to respond to an RFI. This estimation can translate to a consumption of 13,535 person-hours per project. Sparksman (2015) estimates the cost to process a single RFI between $598 and $2,078 for a $1–10 million project. Additionally, the schedule delay can increase hidden costs for liquidated damages and general conditions (Papajohn & El Asmar, 2021). Ultimately, both the consultant and the contractor suffer in terms of cost performance, with contractors incurring expenses for drafting RFIs and gathering supporting documentation, and clients or consultants bearing the costs of preparing responses (Papajohn and El Asmar, 2021).

Beyond cost implications, the RFI process presents numerous additional challenges, particularly concerning delays in processing. According to Papajohn et al., (2018) delay in responding to RFI is a critical factor that negatively impacts project communication and labour productivity (Jarkas et al., 2014). Moreover, detecting informational errors across multiple drawings can be challenging, time-consuming, and error-prone for contractors or subcontractors (Filho et al., 2016a, 2016b). On consultants' end the delay in addressing RFIs promptly can be attributed to inadequate fees and a shortage of skilled workforce (Philips-Ryder et al., 2013). RFI process is also an indicator of project performance (Liao et al., 2020), which is negatively impacted by frequent inquiries and design clarifications from contractors. Another critical issue with the RFI process involves the potential misuse, where parties may abuse the system by submitting unjustified RFIs, later using them to justify delays or during litigation (Hanna et al., 2012). Ideally, RFIs should be need-based, considering the significant repercussions of initiating an RFI. When used appropriately, this process can enhance communication and strengthen trust between contracting parties (Zuppa et al., 2016; Philips-Ryder et al., 2013).

Considering the challenges associated with the RFI process and the risks posed by delays in processing RFIs, it is crucial to promptly draft responses and ensure timely closure (Afzal et al., 2024). Recognising these implications, researchers and industry stakeholders have proposed various solutions to streamline the RFI process. These solutions can be categorized into three types: a) industry-employed platforms, b) manual content analysis by researchers, and c) technological solutions developed by researchers. Different strategies have been

implemented in practice to address challenges in the RFI process and improve efficiency. Building Information Modelling (BIM) is a technology recognised for reducing errors in drawings and minimizing discrepancies in grid and column alignments (Sompolgrunk et al., 2021). However, BIM adoption introduces new challenges like modelling errors and unresolved clashes, leading to increased RFIs (Afzal et al., 2023). Common data environments (CDE) platforms such as Aconex, Autodesk Construction Cloud, and Procore have transformed RFI exchange from traditional methods (paper/letter/email-based RFI exchange) to streamlined, systematic and centralised RFI exchange, enhancing RFI management and tracking capabilities (Sandoval et al., 2023; Das et al., 2020). Yet, risks such as data loss and legal complications accompany their use (Afzal et al., 2023). Furthermore, although these platforms generate substantial data, their potential to leverage this data for improved decision-making and insightful analysis remains underutilized (Zawada et al., 2024).

Research aimed at improving the RFI process has developed two main categories of approaches: human-driven RFI codification and automated RFI assessments. Human-driven methods involve manual content analysis to categorize RFIs based on factors such as RFI type, property code, discipline code, work-element code, and the reasons behind each RFI, tailored to the expertise of researchers (Bhat et al., 2017). This approach helps address critical questions about why issues arise and how they can be resolved. However, it is time-consuming and resource intensive. In contrast, automated methods utilise advanced techniques such as machine learning and natural language processing (NLP). For example, Lee and Yi (2017) applied machine learning algorithms for pre-bid risk classification and topic modelling of RFIs. Shrestha et al. (2023) used artificial neural networks (ANN) to classify RFIs based on their impact levels, and Panahi et al. (2023) integrated computer vision and NLP to extract information from drawings and retrieve relevant RFIs. These automated approaches offer improved efficiency and scalability in managing RFIs compared to manual methods.

While these models represent a promising start, the overall research direction is still in its infancy, highlighting a pressing need for the development of NLP-driven pipelines to streamline the RFI process. Recently, the construction industry and its associated body of knowledge have begun embracing digital transformation through advanced data-driven methods (Musarat et al., 2021). This adoption is particularly crucial for the construction sector, which historically has been slow to adapt to emerging technologies (Bademosi et al., 2022), despite being a significant generator of textual data. These data-driven approaches are also

necessary for RFI process to generate actionable insights, which have the potential to streamline the process. Furthermore, developing these NLP models to support existing CDE platforms could greatly enhance their ability to utilise RFI datasets for efficient review and closure by stakeholders. The models will also provide researchers with a pathway that marks a significant shift from manual content analysis to efficient and automated information extraction from RFIs.

1.2 Gap analysis and problem statement

In identifying research gaps aimed at streamlining the RFI process, three key areas emerge. Firstly, there is a lack of comprehensive understanding regarding the RFI process itself, including its challenges, risks, and existing solutions for improvement. RFI is considered as necessary evil (Aibinu et al., 2019) of the construction sector. They are inherent to every project, and yet there exists a significant gap in fully deconstructing the RFI process across the prevalent themes within literature (Afzal et al., 2024). Addressing this gap is crucial for developing more effective, advanced and state-of-the-art solutions to mitigate RFI risks and enhance the overall process. Within the existing body of knowledge, there is a notable scarcity of NLP-driven studies focusing on construction documents (Wu et al., 2022), and even fewer that specifically apply text mining techniques to RFIs. Consequently, there is a lack of comprehensive evaluation utilising NLP techniques and supervised machine learning approaches to extract insights from the unstructured content of RFIs. Furthermore, the application of deep neural networks, including advanced algorithms like convolutional neural networks (CNN), recurrent neural networks (RNN), and transformers, remains untapped for RFI corpora. Given that RFIs present complex, unstructured queries, there is a necessity to integrate NLP techniques and deep learning approaches to extract insights from them. Currently, there is no single model available for issue classification from RFIs, and there is also a notable absence of named entity recognition (NER) applications tailored to RFIs. Therefore, there is a critical need to develop these NLP applications for RFIs, not only to establish baseline models for future research but also for practical integration into existing RFI tracking and management systems such as electronic document management systems (EDMS) and common data environments. The rationale for adopting NLP and deep learning approaches is further elaborated in Chapter 2 (Literature Review), where the limitations of traditional methods are critically evaluated, and the potential of AI-driven techniques in construction communication is highlighted. A comprehensive discussion on the selection of specific NLP techniques and deep learning algorithms for the RFI dataset is provided in Chapters 4 and 5.

While these deep learning and NLP-driven models will not replace existing EDMS or CDE platforms, they will serve as aids to help decision-makers derive informed decisions from actionable insights generated by them. Additionally, they will augment the functionalities and features provided by these platforms.

By evaluating gaps in existing studies, the primary research problem addressed in this thesis is **how NLP-driven pipelines can effectively extract insights from construction RFIs.**

1.3 Research questions

This research will seek answers to the following key questions

1. What are the current practices, challenges, and technological advancements in the RFI process, and how do they fit within the existing construction innovation landscape?
2. Which machine learning models, as well as NLP techniques, are suitable for automated phase-wise separation and topic modelling of RFIs?
3. How can deep learning and NLP methods be effectively applied to automatically classify issues and extract key entities within RFIs to enhance RFI management practices?

1.4 Research aims and objectives

The aim of this research is to develop domain-specific models that effectively analyse the unstructured content of RFIs through advanced NLP techniques for efficient information extraction. By leveraging these insights, decision-makers can effectively identify issues and uncertainties embedded in RFI content across the project lifecycle. This process aims to significantly reduce RFI processing times and facilitate the integration of learned information into future projects to enhance documentation quality. This is meant to be achieved through the following three objectives:

1. Synthesising and assimilating RFI literature to investigate shortcomings in the RFI process, the repercussions of delayed RFI resolution, and solutions proposed by both academic research and practical applications to streamline RFI process.
2. Developing and validating a domain-specific text classification model to categorise RFIs into construction phases for efficient routing and automated classification, supplemented by a topic model approach to cluster RFIs and uncover predominant themes.
3. Developing and validating a domain-specific issue classification model, followed by creating an NER model to efficiently extract key entities from RFIs for automated

information extraction.

Based on the above objectives, the research will conclude by providing a roadmap on how the above-mentioned text mining models can be effectively utilised within traditional construction and CDE-based RFI management offering guidance to industry stakeholders and setting a direction for future research and development.

1.5 Proposed approach and overview of research methodology

This thesis aims to leverage NLP techniques for efficient text classification and information extraction from unstructured RFI content, with the goal of accelerating the RFI review process and extracting valuable insights and lessons learned. NLP, a sub-field of artificial intelligence (AI), focuses on techniques that enable computers to analyse and process natural language, such as text or speech (Manning et al., 1999). NLP is widely used in applications such as language translation, speech recognition, information retrieval, and extraction. Previous literature on extracting information from construction documents (Afzal et al., 2024) has guided the use of machine learning-based NLP to also improve the RFI process. Accordingly, this research leverages NLP-based approaches for phase-wise classification of RFIs, issue identification, key entity extraction, and topic modelling.

Chapter 3 outlines the overall research methodology employed in this study, following the design science research framework. This methodological framework identifies a real-world problem and subsequently leads to the development of a tangible solution to address it. Accordingly, this research first carried a systematic literature review of the existing RFI process to understand the state-of-the-art, which helped identify research gaps, formulate the research problem, and develop research questions, aims, and objectives. This review examined the challenges and risks faced in the RFI process and evaluated available solutions for tracking and managing RFIs. A detailed feature review of existing CDEs was conducted, highlighting their limitations, particularly the lack of automatic insight generation. Accordingly, this research leveraged NLP-based approaches for phase-wise classification of RFIs, issue identification, key entity extraction, and topic modelling of RFIs. Figure 1.1 details the techniques employed on the RFI corpus and the extracted information and applications.

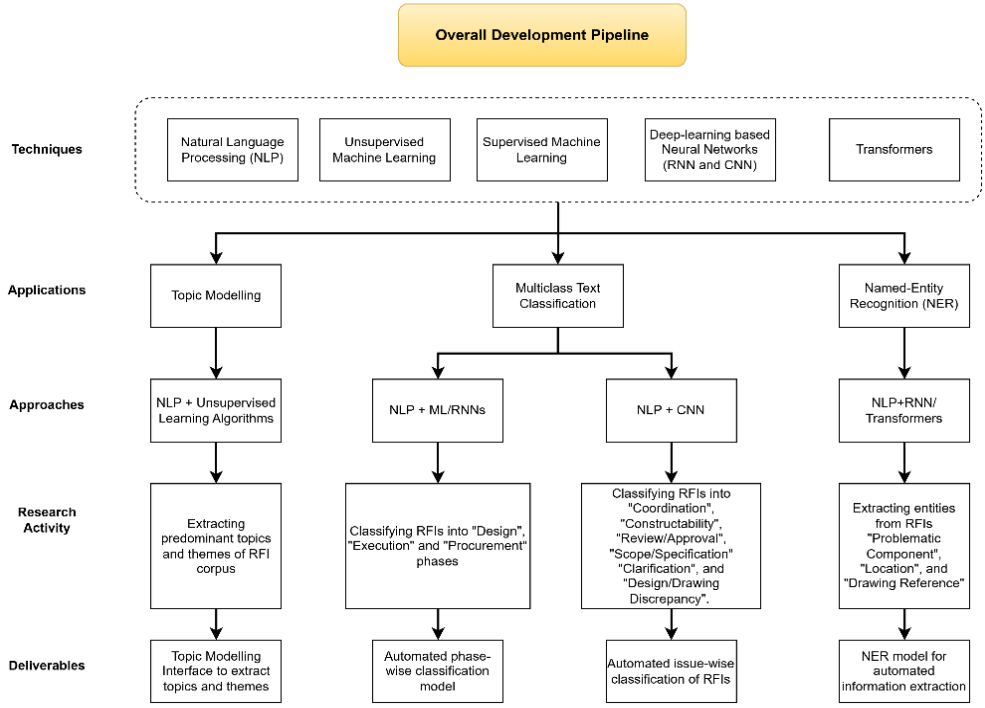


Figure 1-1. Overview of techniques, application and deliverables of each model.

Traditional machine learning algorithms and deep learning-based recurrent neural networks were utilised for development of multiclass text classification models. Additionally, a named entity recognition model was developed to extract key entities from RFIs. Topic modelling was achieved to obtain the visualisation of predominant topics/themes within the RFIs.

1.6 Significance of the research

This research makes substantial contributions to both academia and industry by addressing a critical gap in the management of RFI processes through advanced deep learning and natural language processing techniques. The study introduces three novel models—a phase-wise separation model, an issue classification model, and a named entity recognition model—built upon established architectures, but uniquely tailored for RFI analysis. These models, developed using a first-of-its-kind real-world RFI dataset, establish new benchmarks for performance in construction informatics, providing a foundation for future research. Theoretically, this work will lead to understanding of RFI patterns, issue severity, and mitigation strategies, while methodologically, it explores underutilized techniques such as feature representation and ensemble learning in construction NLP.

From a practical standpoint, the developed models offer transformative value to industry stakeholders by automating labor-intensive RFI classification and information extraction. The

phase-wise separation model enables targeted issue resolution within specific project phases, while the issue classification and NER models facilitate rapid identification of problem sources, components, and drawing references. These capabilities empower project teams to proactively address risks, reduce RFI backlogs, and improve decision-making. Furthermore, the proposed NLP pipeline demonstrates how existing common data environments and BIM platforms can leverage underutilized data for actionable insights, paving the way for integration via plugins. By providing a detailed roadmap for implementation, this research not only enhances current RFI management practices but also sets the stage for broader adoption of AI-driven automation in construction documentation, ultimately improving project efficiency and reducing delays.

1.7 Structure of the thesis:

The overall structure of the thesis is outlined to provide a clear roadmap for the reader. A brief overview of each chapter is presented in Figure 1.2, highlighting the progression and logical flow of the research:

Chapter 1: Introduction

The purpose of this chapter is to introduce the research to the readers and provide an overview of the chapters included in the thesis. This chapter presents the research background, summary of gaps, problem statement. Additionally, it outlines the research questions, objectives and aims, and an overview of the proposed approaches and research methodology employed. Finally, it presents an overview of the thesis structure, outlining each chapter in detail and providing a guide to help the reader navigate through the thesis effectively.

Chapter 2: Literature review

The purpose of this chapter is to review the state-of-the-art literature on RFI management. This review is conducted through scientometric analysis and a structured literature review. First, literature characterisation and bibliometric analysis are performed, followed by content analysis of the predominant themes related to the RFI process. From these themes, a critical analysis is presented, discussing the future trajectory of RFI management within the context of

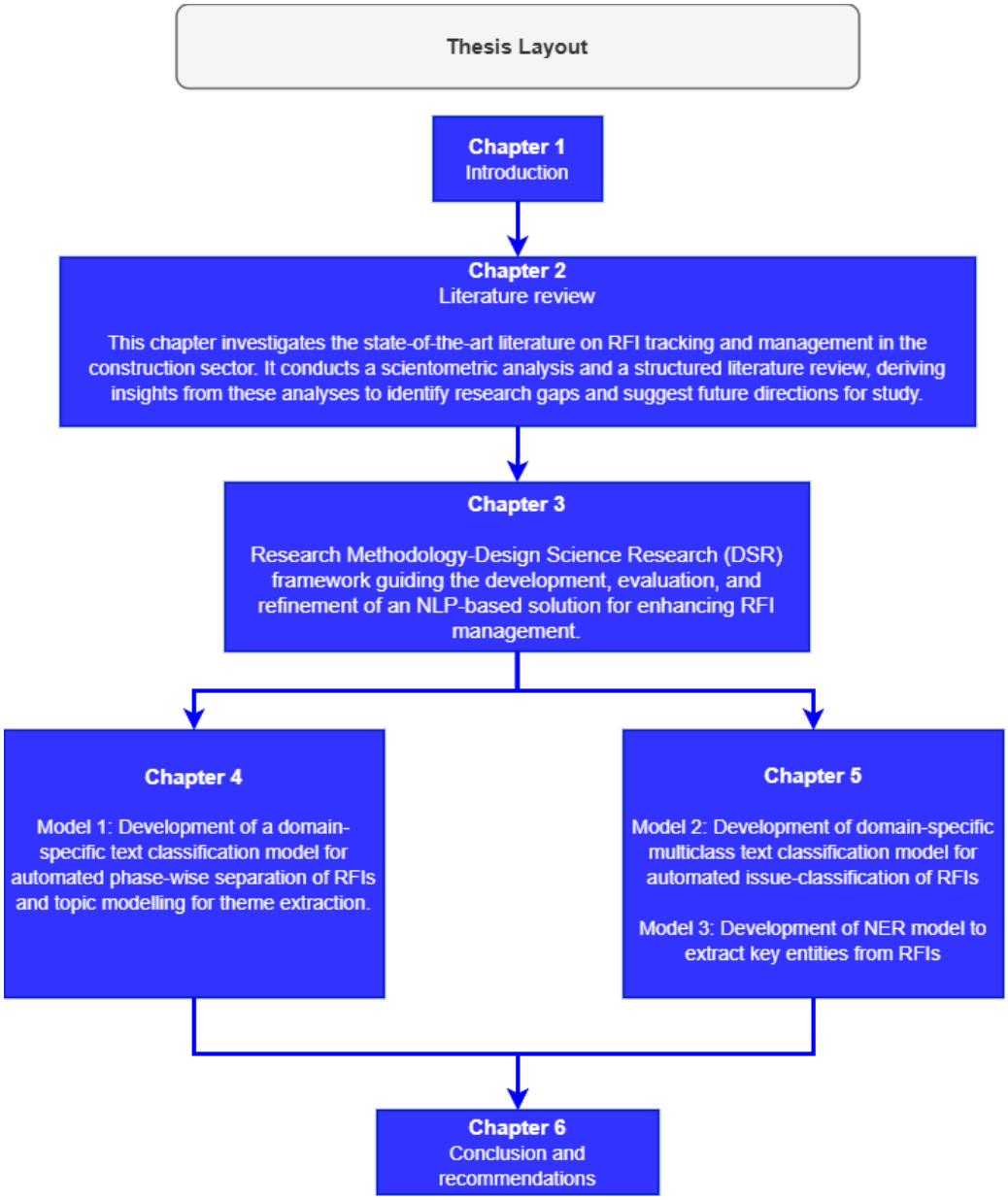


Figure 1-2. Proposed thesis structure.

construction innovation and NLP. Additionally, best practices for current RFI management are provided to assist industry stakeholders. This study extends the RFI literature by offering deeper insights into its impact and the issues arising from the process, enabling stakeholders to understand the RFI process holistically and incorporate best practices to minimise its negative impact. Lastly, the larger knowledge gap identified by the literature review, which necessitates the adoption of NLP for efficient RFI management, sets the agenda for the subsequent sections. These sections delve into the development and testing of our advanced NLP models for efficient RFI management.

Chapter 3: Research design and methodology

This chapter outlines the overall research methodology guiding the thesis and illustrates the connections between its chapters through a comprehensive diagram. It specifically emphasizes the use of the design science research framework, which forms the foundation for the development, evaluation, and refinement of the proposed NLP-based solution for RFI management in construction projects.

Chapter 4: Automated phase-wise separation and topic modelling of RFIs through natural language processing

The purpose of this chapter is twofold. First, it explores supervised learning methods to classify RFIs based on project phases. This involves investigating various NLP techniques and the comparison of and deep learning-based recurrent neural networks and traditional machine learning algorithms for the automated classification of RFIs into "design", "execution", and "procurement" phases. The research also incorporates ensemble learning strategies to enhance the classification results. The chapter includes an experiment demonstrating the efficacy of automated multiclass text classification in comparison with human performance. Second, the chapter introduces the application of topic modelling to the RFI dataset, using unsupervised clustering algorithms to extract key topics and themes, and providing advanced visualizations for stakeholders. Overall, the chapter focuses on comparing the effectiveness of deep learning versus machine learning in multiclass classification and applying unsupervised learning for topic extraction and visualization.

Chapter 5: Improving responsiveness in construction RFIs: leveraging natural language processing for efficient issue classification and key entity extraction

This chapter also serves a dual purpose, providing a detailed description of the development of two distinct NLP models for different applications. In the first phase, the chapter outlines the development of a multiclass text classification model using CNN algorithm for issue-wise classification of RFIs. The issues classified include "coordination," "constructability," "design/drawing discrepancy," "review/approval," and "scope/specification clarifications." In the subsequent phase, another model is developed for extracting key entities from RFIs such as "problematic component", "location", and "drawing reference". Here, bidirectional long short-term memory (BiLSTM) and bidirectional encoder representations from transformers (BERT) models are utilised to create a named entity recognition model. These models are then enhanced through an ensemble method by integrating a conditional random field (CRF) layer.

Chapter 6: Conclusion and recommendations

The purpose of this chapter is to conclude the research efficiently by presenting insights on how the developed models can be effectively integrated into the existing RFI workflow. Additionally, this chapter serves as a guide for future researchers on leveraging NLP to streamline RFI documentation. Furthermore, key findings from the research are presented, and the limitations of the study are also highlighted.

Chapter 2: Literature Review

This chapter reviews the literature related to the RFI process through scientometric analysis and systematic literature review. First, literature characterization and bibliometric analysis are performed, followed by content analysis of the predominant themes related to the RFI process. Accordingly, best practices for the RFI process are then established to improve RFI management. The study concludes by situating the current research within the realm of construction innovation, aiming to enhance the RFI process using state-of-the-art technologies. This study extends the RFI literature by providing deeper insights into its impact and the problems originating from the process. Finally, this literature review lays the foundation for the methodology adopted in this research endeavour, which utilises NLP to develop models for automated information extraction from RFIs.

2.1 Introduction

The request for information is an important document through which contracting parties seek clarifications in construction documents including but not limited to plans, drawings, specifications, and agreements (Abdul-Monem and Hegazy, 2013; McGraw-Hill, 2008). Between the period of design start-up and the call for tender, the design consultants may not have adequate time to develop full documents, and as a result, the tender documents are frequently unclear, incomplete, and may cause severe conflicts among various disciplines (Jarkas and Bitar, 2012). Incomplete or ambiguous technical specifications require clarifications, leading to consecutive disruptions to work progress (Jarkas and Bitar, 2012). Also, incomplete detail or lack of clarity in specifications, constructability issues in work packages, and differing technical interpretations may result in the burst of RFIs, impeding the smooth flow of work (Hasan et al, 2018; Jarkas and Bitar, 2012). These practices leave many issues undetected until construction begins and leads to downstream rework. Thus, the RFI process facilitates their timely resolution. Even though RFIs are essential communication tools, they may carry inherent risks (Hughes et al., 2013). RFI is an unavoidable administrative process and yet a ‘non-value adding’ activity for a construction project (Liao et al., 2020). The occurrence of RFI is unpredictable and cannot be included in a project's baseline schedule. Hence, RFIs have the potential to become part of the critical path, and late RFI response can lead to schedule delays (Kelly and Ilzor, 2020). The processing of reviewing RFI questions and producing a suitable response to satiate the contractor is time-consuming and costly (Love et al., 2014) and may trigger change orders which could become a source of disputes and claims

(Hughes et al., 2013). For these reasons, RFIs has been labelled as a 'necessary evil' (Aibinu et al., 2019), and therefore, management of RFI needs to be strategised throughout construction projects.

Reducing RFIs and minimising their response time has drawn increasing attention in both the academic research and industry over the last few decades. Despite the growing body of knowledge, RFI management process remains partially or fully reliant on human intervention, making them susceptible to inefficiencies and errors. RFIs produce useful, precise and integral project information. Analysing and discussing different themes related to these processes will help the researchers and industry practitioners understand the inherent risks associated with the RFI processes and devise best practices to mitigate those risks. Thus far, there is lack of a literature review in the body of knowledge that dissects the RFI processes to develop understanding. With the goal of fostering and directing further research on improving the management of RFI in construction projects, this chapter focuses on several streams. First, this review provides an in-depth insight into the RFI processes and summarises the current key research areas in leading peer-reviewed journal publications from different databases. Second, this review will provide directions for researchers in designing future studies and in introducing new research avenue to improve RFI management through best practices. The synthesis of the literature on RFI assists the industry in understanding and developing a more robust approach to simplify the document-intensive tasks by streamlining RFI processes and minimising human errors. In this chapter, the research and development of RFI process in the architecture, engineering, construction and operations (AECO) industry for the past two decades (i.e., 2000 to 2022). This time period ensures the coverage of advancement and progression of RFI process from traditional to more recent BIM or common data environment driven project lifecycles.

The remaining sections of this review are structured as follows. The next section provides an overview of RFI processes in construction projects, including the background of RFI and its significance in construction projects. Section 3 of this chapter reports the methodical approach adopted to identify and summarise relevant articles, and Section 4 characterises the literature and provides a bibliometric analysis of publications included in this review. Section 5 discusses key research themes, contributions of journal publications, and current trends in RFI studies. Section 6 identifies research gaps and outlines directions for future studies. The subsequent section discusses how this literature review fits within the current landscape of construction innovation. In the final section of this chapter, the conclusions and contributions are presented.

2.1.1 The RFI process

The exchange of RFI documents is a set of interrelated tasks (Aibinu et al., 2019) that facilitate communication among the project stakeholders, contributing to the resolution of issues. Typically, the RFI process (Fig. 2.1) starts when an RFI requestor, for instance, a contractor or sub-contractor, seeks clarifications to resolve a problem during the project's lifecycle. Upon receiving the RFI from the contractor, the client/design team registers it into the RFI log before furthering into the review process (Papajohn et al., 2018). The submitted RFIs can be queued and reviewed at once (Chin and Russel. 2008) or they may not queued and addressed independently (Papajohn et al., 2018; Nasrallah and Bou-Matar, 2008). Once RFI is received by the responder, further actions are taken depending on the clarification sought. Preparing a response to an RFI may trigger further communication sub-process as additional information from other team members and disciplines may be required to address the issue raised (Morales et al., 2022; Aibinu et al., 2019). Thus, responding to an RFI can be a hectic task and has cost and time implications. After analysing the RFI question, a response is formulated which may trigger further RFIs between the parties involved. Depending on the queries raised, RFIs can be classified into various categories. These RFI classifications may provide actionable insights to help stakeholders quickly review and close an RFI. These RFI classifications are discussed in section 5 of this chapter.

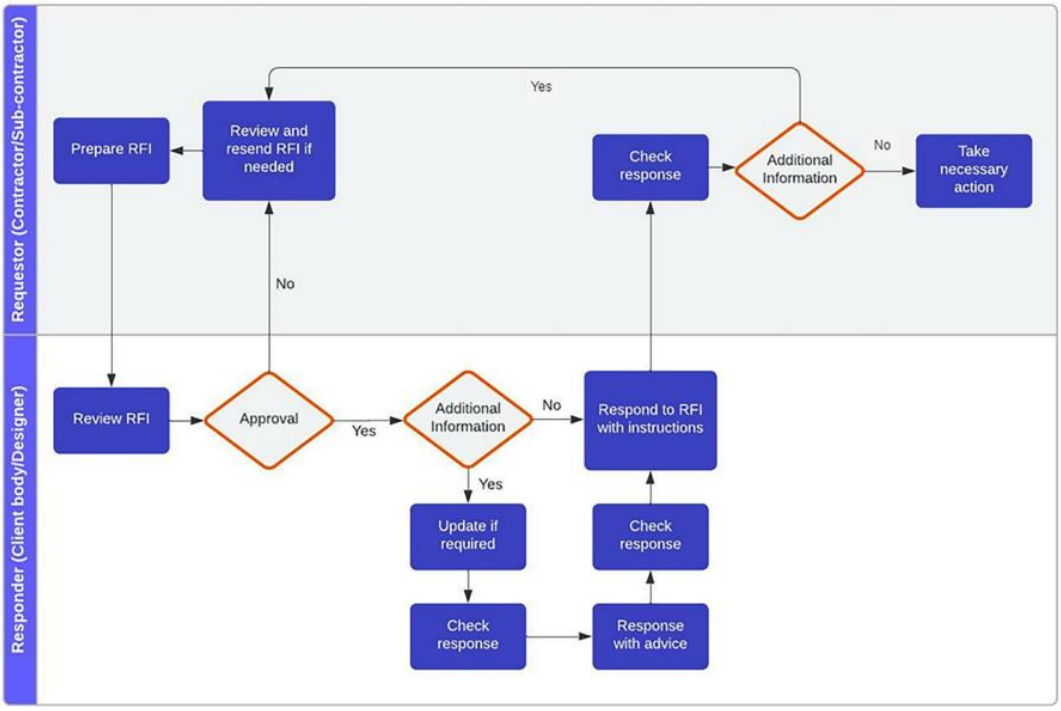


Figure 2-1. The RFI process in construction adapted from Morales et al., (2022).

2.1.2 Anatomy of a construction RFI

While different construction companies have developed different format of an RFI document most RFIs contain the same key elements. Fig. 2.2 provides a standardised diagrammatic representation of a typical RFI document utilised in a construction project. This representation has been adapted from the source (Mao et al., 2007) and further modified by examining different RFI samples acquired from the open-source RFI templates (Ellis 2022; Ramos 2022). The content of the RFIs is primarily divided into two parts: a) structured information and b) unstructured information body. The structured information includes details such as: the project name, RFI number, details of the stakeholders (both initiator and the responder), their company names, the necessary dates, and any supplementary document (CAD drawing or material-related information) in the following pages. Some organisations may add other essential details to the RFIs such as, impact on cost and schedule, priority levels, date till which response is required etc. The RFI document may also include any addendum documents/drawings/specifications associated with the sought query. The clarification sought by the contractor and the answer provided by the designer is considered unstructured information (Lee and Yi, 2017). Another document that aids in the RFI process is an RFI log. RFI logs serve as an administration tool (Papajohn and Asmar, 2021) that registers all the essential characteristics associated with an RFI in a spreadsheet. The RFI log is a crucial tool to track and monitor the status of an RFI to determine if the status of an RFI is still active or closed.

2.2 Literature review methodology

This study involves a systematic literature review using a methodology in line with the outline used in (Zhong et al., 2019) which consists of four main phases (Fig. 2.3); 1) data acquisition, 2) data processing, 3) data mapping and 4) themes and gaps analysis. Through this approach, the study identifies the existing themes in literature, proposes best practices, and pinpoints research gaps and future developments in RFI processes within the construction industry. This methodology section discusses ‘data acquisition’ (phase 1) and data processing (phase 2) steps in detail. The next section details bibliometric analysis and literature characterisation (phase 3) which involve five steps including time series analysis (i.e. annual publications since 2000), publications geographical distribution, distribution of articles across leading construction journals, co-occurrence analysis of RFI-related keywords, and co-authorship analysis. Such analyses are conducted to trace the development of the research (Merigó et al., 2015) expand the body of knowledge and provide researchers with a deeper understanding of the scope and

progress in the field (Liao et al., 2018). The bibliometric analysis informs the results and discussion section by highlighting the state-of-the-art RFI research themes, identifying existing gaps, and outlining directions for future research.

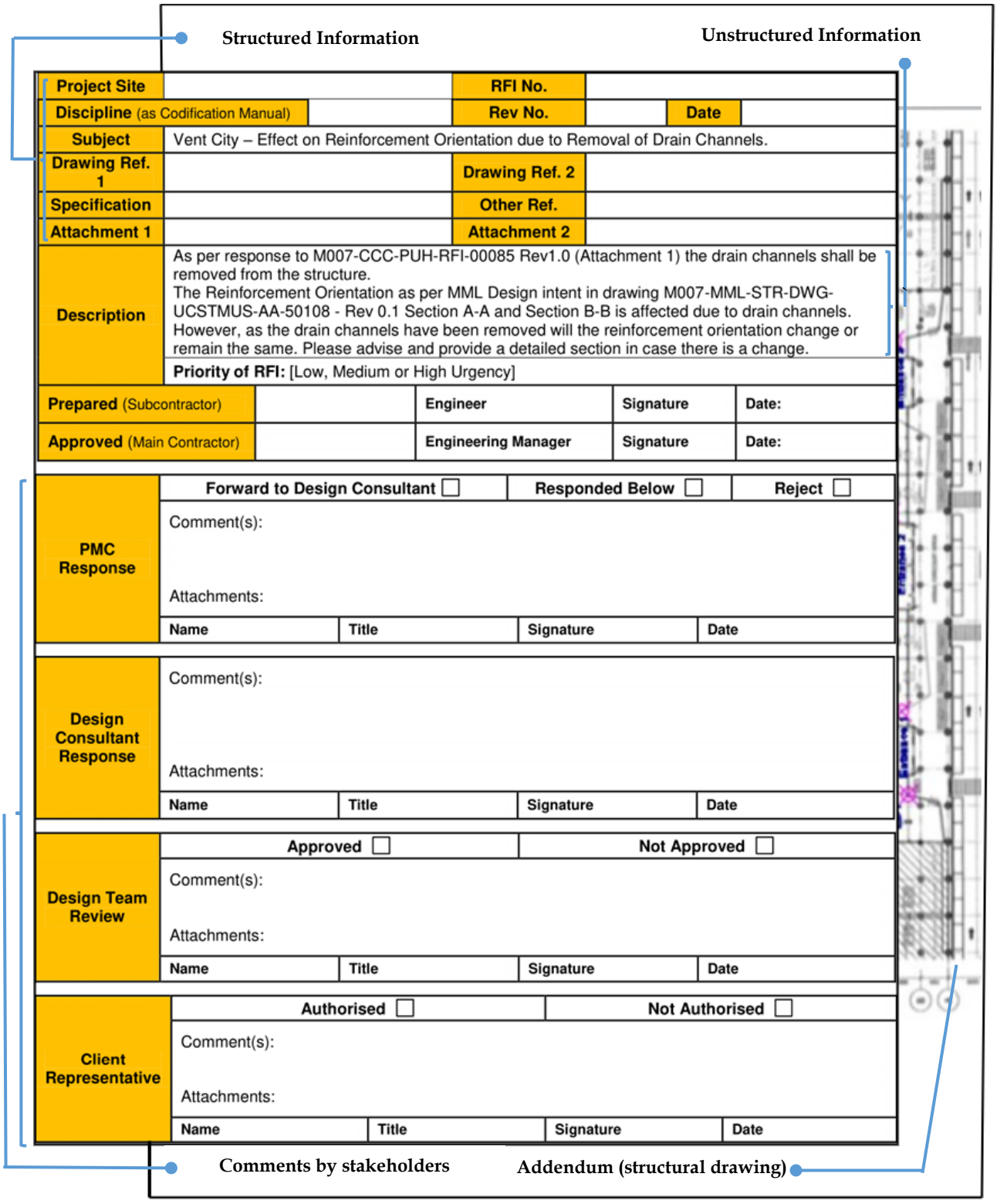


Figure 2-2. Anatomy of an RFI document in construction.

2.2.1 Literature search and screening

The literature search and screening framework followed in this study is in line with (Baek et al., 2021) and consists of three stages: 1) exhaustive literature search (Stage 1), 2) duplicate removal (Stage 2), and 3) filtering through content analysis (Stage 3). In Stage 1, Web of Science (WoS) and Scopus databases were searched for relevant literature. For this purpose, semantic search strings consisting of RFI-related keywords were developed to retrieve the relevant literature. Different combinations of the keywords joined by Boolean operators of "AND" and "OR" were incorporated. This resulted in 140 articles, with 99 from Scopus and 41 from WoS. From within Scopus results, the articles were restricted to engineering and computer science subject categories to maintain focus and retrieve relevant articles in line with the current research's theme. For WoS, the default categories were slightly different; engineering, construction building technology, computer science and business economics were the predominant categories from where the most relevant RFI literature was extracted. The period of these publications was restricted to post-2000, and the language of the articles was limited to English. This period is carefully elected as it encompasses the growing momentum in built environment research from traditional construction management systems towards more machine dominated or advanced technology-oriented solutions and smart platforms developed by academia and industry (Loosemore, 2014; Hooper and Haris, 2010; Brandon and Lu, 2008), ensuring maximum research coverage. The studies incorporated in this review included: original research articles, conference proceedings, book chapters, and review articles. With an initial investigation and comparison of the research titles in the next stage, 40 studies were removed due to duplication in both databases. In the final step, through the content analysis of the title, abstract, and manuscript, 11 articles were removed due to being outside our research scope, leaving behind a final count of 89 relevant articles for the present literature review. These stages of literature screening are illustrated in Fig. 2.4.

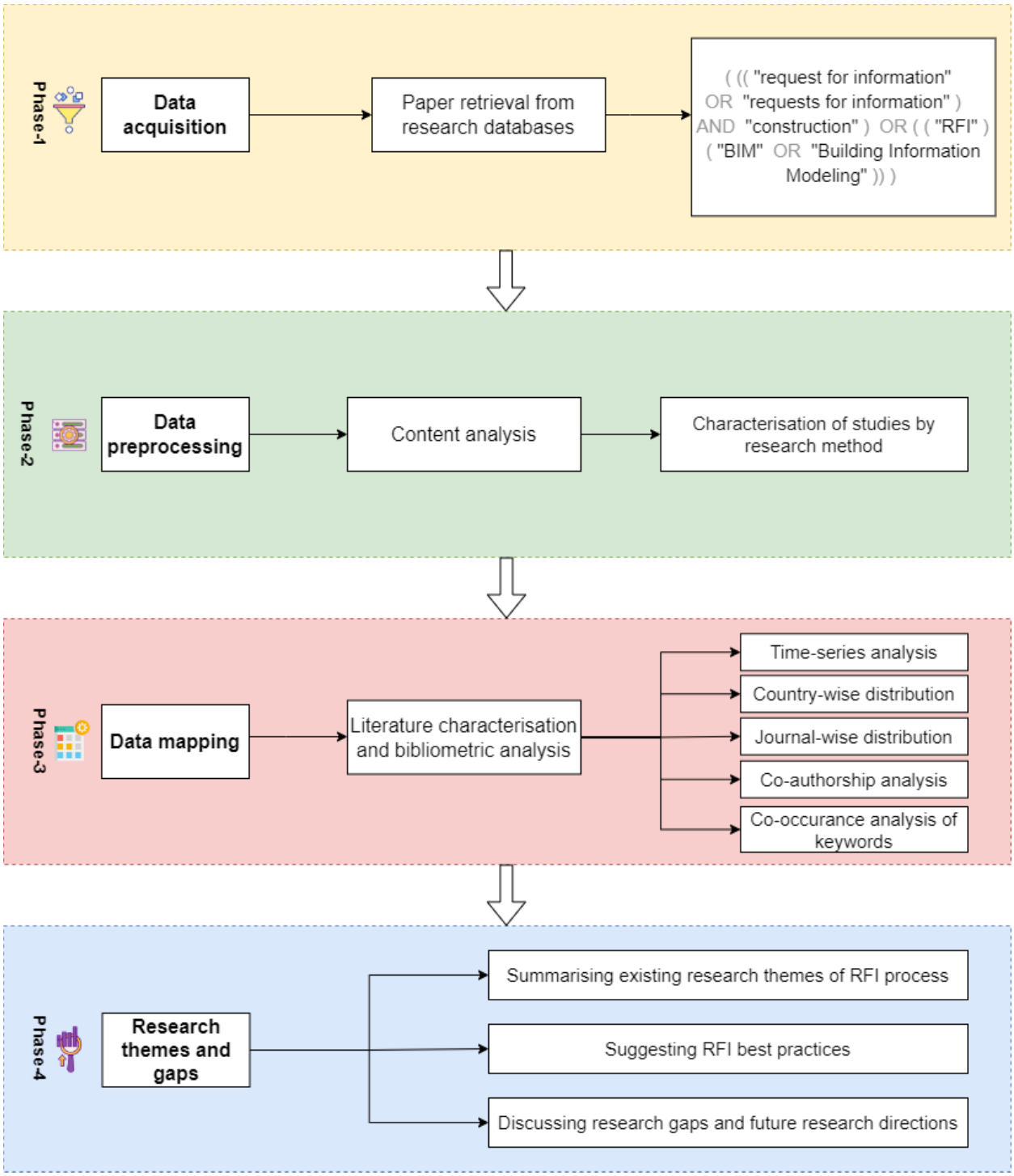


Figure 2-3. The outline of research design.

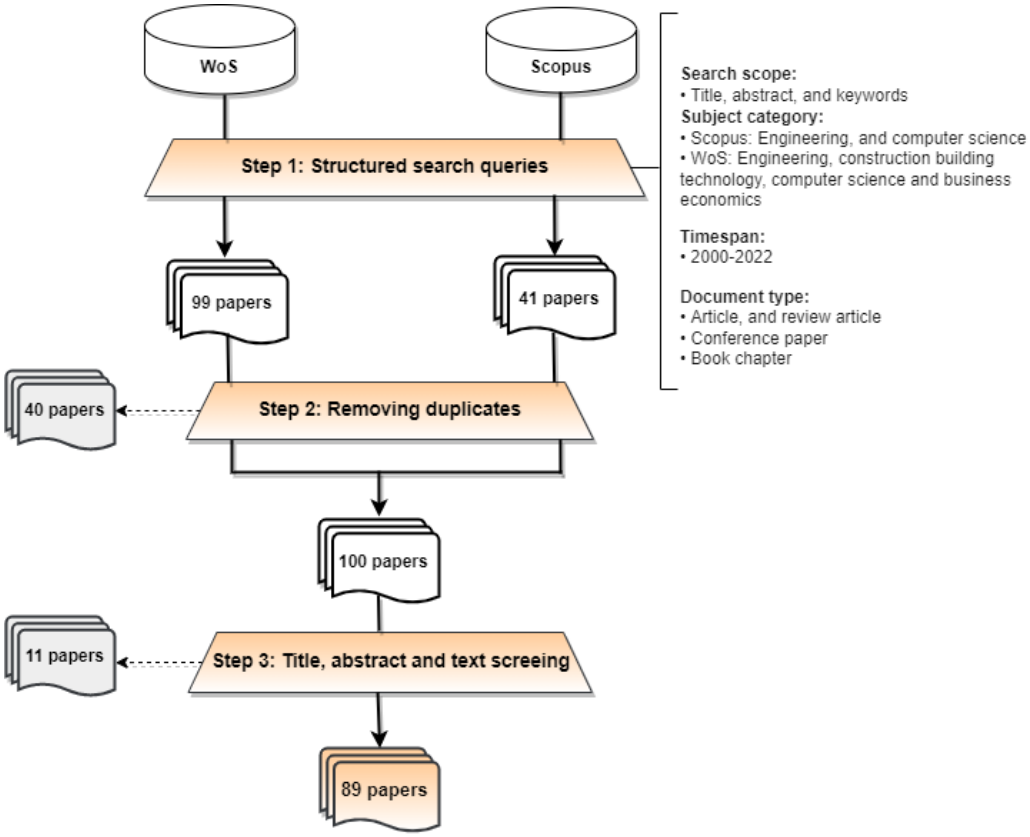


Figure 2-4. Literature screening procedure.

2.2.2 Content analysis

In this step, a total of 89 shortlisted articles regarding the RFI processes were rigorously reviewed. Each reviewed paper was examined and then characterised by its research method. These attributes are obtained from previous studies (Li and Kassem, 2021; Gao and Pishdad-Bozorgi, 2019). Within the 89 articles, 43% of the publications gave insight regarding different aspects related to the RFI processes and 35% undertook a case study approach and then validated them with some form of experimentation. These experiments can be hypothetical or mimic real-world scenarios, and can be demonstrated through existing technological provisions, novel system/prototype or a proposed computational algorithm. Within 31 case study-based publications, the validation of the research was achieved through three approaches, including: i) experiment with existing methods/tools (24 articles), ii) experiment with proposed software system/prototype (4 articles), and iii) experiment with proposed computational algorithm (3 articles). Also, there were 13% of the total articles aimed at proposing a framework and 7% of the publications incorporated interviews/focus groups/questionnaires/workshops as a research method. Among these articles, there are twelve research articles proposed a framework, and six publications validated the framework through

an experiment. There was only one proof-of-concept simulation article (Papajohn et al., 2018) and one literature review (Sompolgrunk et al., 2021). This review paper briefly discussed RFI as one of the metrics to measure the return on investment (ROI) for BIM implementation on construction projects. No comprehensive review has been done to focus on RFI-related research in the construction sector. It is, therefore, this study's aim to close this gap within the present literature review and investigate the RFI management processes and factors that influence it with a focused scope.

2.3 Literature characterisation and bibliometric analysis

This section characterises the literature and provides a bibliometric analysis of the 89 publications included in this review. Firstly, a time-series analysis was conducted to show the publication trends between 2000-2022. Next, an in-depth location analysis was used to highlight the country-wise research publications. Then, a detailed publication document analysis was conducted to identify the highest-cited articles and the number of journal articles published by each publisher. A detailed co-occurrence analysis was also performed to determine the research trends and inclination of research for RFI management. Finally, the co-authorship network depicting the collaboration among the authors was analysed to forward the research related to the RFI process in the construction sector.

2.3.1 Time series analysis of RFI-related studies

The growth of RFI related research between 2010 and 2022 is illustrated in Figure 2.5. Between 2000 to 2006, research in RFI grew slowly. The number of publications per year constituted a single digit, except in 2010 when 10 publications contributed to the body of knowledge pertaining to RFI process. Between 2014 and 2016, RFI-related articles increased 29%. This can be attributed to the emergence of BIM and its potential to better manage RFIs, triggering discussions (Bhat et al., 2017; Barlish and Sullivan, 2012) around RFI management through digital platforms. In 2017, a sharp decline was observed. However, the publication number started to climb until 2020, but it did not reach the previous record high in 2016. Nonetheless, the number of RFI-related publications have been fluctuating with another decline seen after 2020. Such trends may imply that RFI has not been in the focal of construction research while it has been always a major source of issues and disputes amongst project participants (Hughes et al., 2013).

It must be noted that RFI process still impacts construction projects in a plethora of ways, with both direct and indirect implications. There has not been considerable research on the influence

of emerging technologies to further streamline the RFI process and potentially avoid disputes down the line of a project. It still remains a challenge in BIM-driven RFI management, indicating the significance and relevancy of this research direction. Also, while other research areas in construction informatics are leading the trends, RFI-related research has observed moderate advancements due to several limitations in the light of the nature of data involved. Baek et. al (2021) blames the inability of researchers to acquire practical data such as RFIs and other unstructured documents because of lack of quantity and confidentiality leading to lack of research.

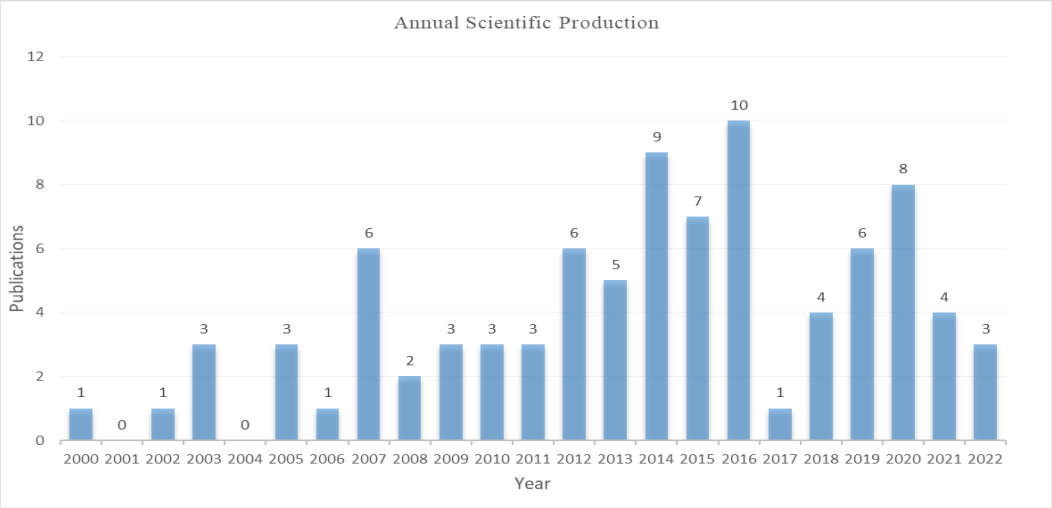


Figure 2-5. Yearly publications for RFI-related research.

2.3.2 Publications geographical distribution

The location analysis suggested that the United States has generated the most publications (45) and citations (1027). After the United States, Canada (15 publications and 83 citations) and Kuwait (6 publications and 402 citations) are the most prominent countries contributing to the research field (Figures 2.6 and 2.7). In addition to these countries, Australia, Chile, and China, have each contributed to three research papers each. The citations-wise breakdown of the countries indicates China, South Korea, and Chile have a higher number of citations (36, 34, and 18, respectively). In total, 18 countries are represented in the 89 publications included in this literature review. Based on the frequency of country-wise publications, it can be implied that the RFI-related research has been a subject of interest for many countries around the world, with many of them already embarking decent progress with regards to contribution to the research field.

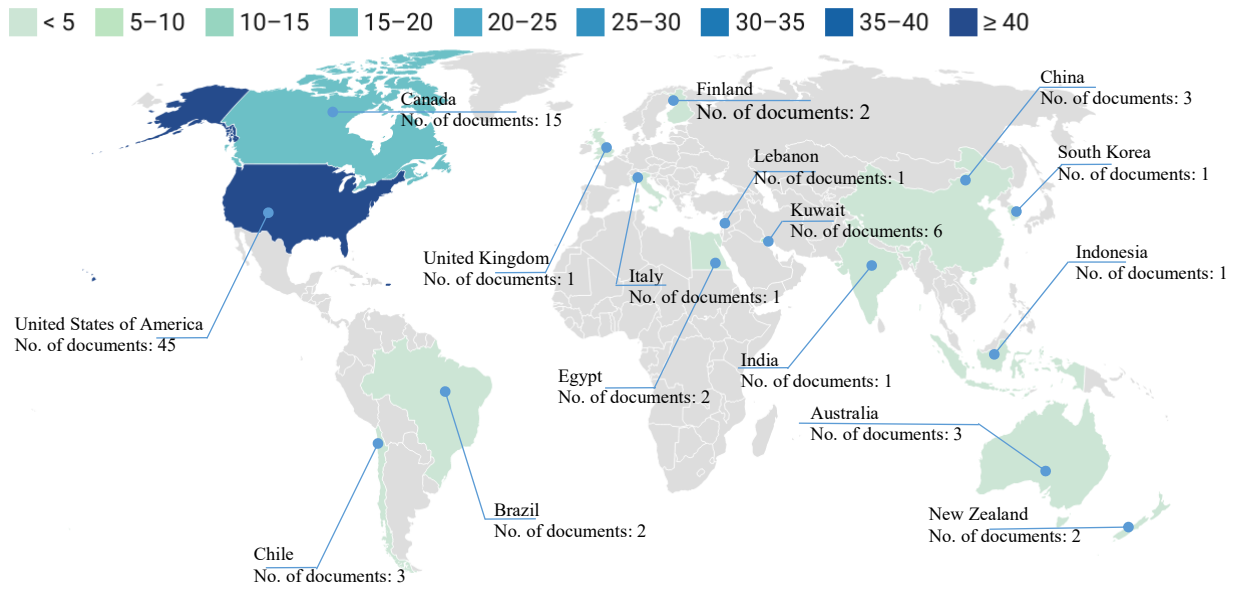


Figure 2-6. Publications distributed by country.

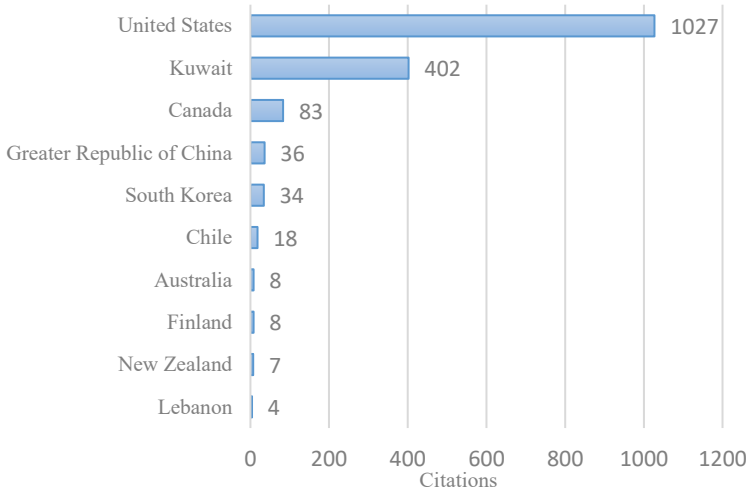


Figure 2-7. Top 10 most cited countries.

2.3.3 Distribution of articles across main construction journals

Among these 89 RFI-related publications, 53 articles were published in journal, 34 articles were published in conference proceedings, and 2 were published as chapter in books. Table 2.1 shows only journals having published at least two articles. Journal of Construction Engineering and Management of the American Society of Civil Engineers (ASCE) is the leading journal contributing to the eight publications in the research domain, followed by Automation in Construction and International Journal of Construction Management with 5 papers each respectively. With the RFI-related publications in the conference proceedings, ASCE is currently the prominent publisher with most conference contributions related to the topic.

Among all publications, the paper publised by Barlish & Sullivan (2012) received the highest citations. Overall, the RFI processes have been an integral research topic, and have been widely discussed both in management and technology oriented literature of architecture, engineering and construction studies.

Table 2-1. Distribution of articles by journal name.

Source Title	No. of Publications	Publisher	% Total Publications
Journal of Construction Engineering and Management	7	American Society of Civil Engineers (ASCE)	8
Automation in Construction	5	Elsevier	6
International Journal of Construction Management	5	Taylor and Francis	6
Journal of Information Technology in Construction	4	International Council for Research and Innovation in Building and Construction	4
Engineering, Construction and Architectural Management	3	Emerald	3
Journal of Management in Engineering	3	American Society of Civil Engineers (ASCE)	3
Construction Innovation	2	Emerald	2
Journal of Computing in Civil Engineering	2	American Society of Civil Engineers (ASCE)	2
Journal of Professional Issues in Engineering Education and Practice	2	American Society of Civil Engineers (ASCE)	2
Practice Periodical on Structural Design and Construction	2	American Society of Civil Engineers (ASCE)	2

2.3.4 Co-occurrence analysis of keywords

The keywords co-occurrence analysis provides informaion about the core intent of the publication. The keyword network in the analysis provides an insight into the knowledge domain, depicting prominent research topics and the intelligent arrangement between them (Lee and Su, 2010). In this study, the VOSviewer tool was adopted to conduct the keyword co-occurrence analysis. Before keyword mapping, all the keywords were made homogenous, for example, "BIM" and "Building Information Modelling" were given the same entry as "Building Information Modelling (BIM)". This ensured that VOSviewer categorises all the keywords with semantically similar meanings as one keyword. Following, irrelevant keywords, for example state of Qatar, students, Kingdom of Bahrain, Kuwait or BIM pilot project, with no semantic overlap with the research theme were manually removed. The criteria for the minimum keyword occurrence is set to two to ensure the holistic representation and optimal graphics of clustering results (Zhang et al., 2020). With this criteria, from within 658 keywords,

158 met the threshold and formed the research cluster map (Fig. 2.8), depicting the research on the RFI process in the construction sector.

The size of the nodes and the thickness of the lines in Fig. 2.8 represent the keyword co-occurrence and the relational sameness with other keywords, respectively. A thicker line between the two keywords presents a stronger relationship between their research areas (Zhang et al., 2020). As indicated in Fig. 2.8, 'request for information (RFI)' is the predominant keyword across the research studies. Other frequented keywords include 'Building Information Modelling (BIM)', 'project management, and 'architectural design'. The thicker lines from 'Request for Information (RFI)' to the abovementioned keywords represent their close relationships. Keywords such as ‘BIM’, ‘4-D scheduling’, ‘virtual design and construction’, ‘electronic data exchange’, ‘information technology’, indicate a prominent research theme in RFI literature that discusses the impact of digital tools in mitigation and overall better management of RFIs. Also, different construction processes such as ‘documentation’, ‘project control’, ‘communication’, ‘decision making’, ‘change management processes’, ‘project delivery’ and ‘lean construction’ are presented as prominent keywords interlinked with ‘request for information’, indicating the impact of RFIs on these processes. Overall, the keyword cluster and the connection between the keywords reinforce the notion that RFI is an integral component of the construction industry; however, the scientometric map also points to the absence of robust mechanisms for better management of RFIs.

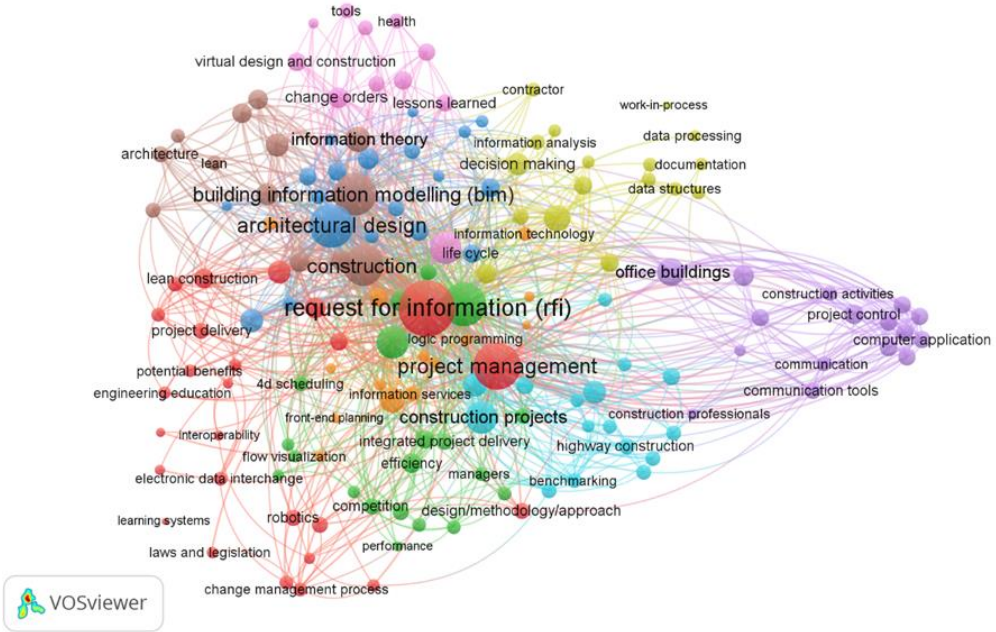


Figure 2-8. Network of co-occurring keywords.

2.3.5 Co-authorship analysis

From the 89 publications included in this review, 24 authors were identified as having published at least two papers each. Again, the VOSviewer tool was utilised to map and visualise the co-authorship analysis of these 24 authors to identify the frequency, strength, and relation between their co-occurrences (Fig. 2.9). The co-authorship diagram, as defined by Schuldt et al., (2021), consists of the three components: frame size (frequency), frame colour (relationship), and lines (co-occurrences) connecting them. Each frame represents an author, and the size of the frame represents the number of publications associated with the author's name. Similarly, the wider the lines and closer the distance between the frames, the more co-published research between the authors. The colours of the frames represent the clusters of associated authors. There are 11 clusters representing authors with at least 2 publications. This topic has garnered some international collaborations as well. For example, studies (Kim et al., 2021; Sompolgrunk et al., 2021; Mao et al., 2007) have been produced due to collaborative research between different international academic institutions.

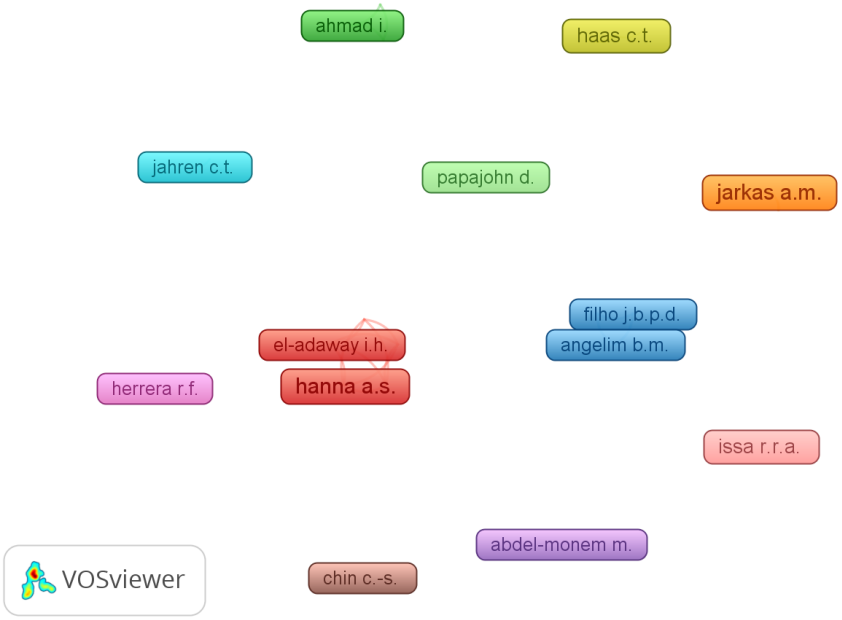


Figure 2-9. Co-authorship network analysis.

2.4 Results and Discussions

Content analysis is a commonly used research method to analyse and synthesise the research content. In the context of investigating the state of the art in RFI-related literature screening procedure within the construction sector, content analysis was applied to identify patterns, themes and relevant information. Content analysis allows for the collection and organisation

of information while revealing trends and patterns present in documents (Krippendorff, 2013). For this study, a combined approach of qualitative and quantitative content analysis was adopted to ensure comprehensive exploration. Qualitative content analysis involves categorising data to gain insights into contextual nuances related to the research. Conversely, quantitative content analysis focuses on deriving numerical values, like frequencies and rankings, by systematically counting occurrences of specific themes or topics (Chan et al., 2009). By using both qualitative and quantitative content analysis, this research seeks to gain a holistic understanding of RFI process in the construction industry. Through the content analysis of the RFI-related body of knowledge, five main research themes were identified, including:

- risk mapping;
- influence of project delivery methods;
- BIM and RFI management;
- other digital tools to aid RFI management; and
- classification frameworks within literature

2.4.1 Risk mapping

The RFI process is initiated with a contractor formally approaching the designer to seek clarifications. However, this process is characterised as problematic and may lead to project risks. To understand the problematisation associated with the RFI processes, it can be beneficial to apprehend its implications through the systems thinking (ST) approach. ST is a robust method for comprehensively evaluating complex processes (Xu and Coors, 2012). A causal loop diagram (CLD) (Figure 2.10) is developed to map the interdependencies among challenges mentioned in the literature arising from RFI communication. The development of the CLD involved a comprehensive identification of all causal factors or variables that contribute to the issuance of RFIs. Following the identification of causal factors from the literature, a thorough analysis was undertaken to determine their specific impacts or effects on the RFI process and, subsequently, their influence on the overall project. This step was crucial in understanding how each variable interacts within the system and how changes in one aspect can propagate throughout the project, affecting its outcomes and performance. By examining these relationships, a comprehensive understanding of the broader implications of RFI process on the project's success and efficiency was achieved. The arrows in the CLD indicate the

direction of causality, and the linking relationships, whether proportional or inverse, are presented through + or – signs at the tip of arrows, respectively.

Figure 2.10 shows most arrows concentrating on "RFIs generated" and most arrows emerging from "delay in RFI' response". This is due to researchers heavily relying on these indicators to compare functionality or performance of different project settings. The CLD illustrates that RFIs are generated due to "incomplete contract documents" and/or "mistakes in contract documents" (Gajendran and Brewer, 2012). Contractors and sub-contractors continually discover these discrepancies throughout the project lifecycle (Daoud and Allouche, 2003; Philips-Ryder et al., 2013). In particular, these issues include:

- imprecise construction documents;
- vague specifications;
- specification and plan contradictions; and
- unanticipated field conditions (Hanna et al., 2012).

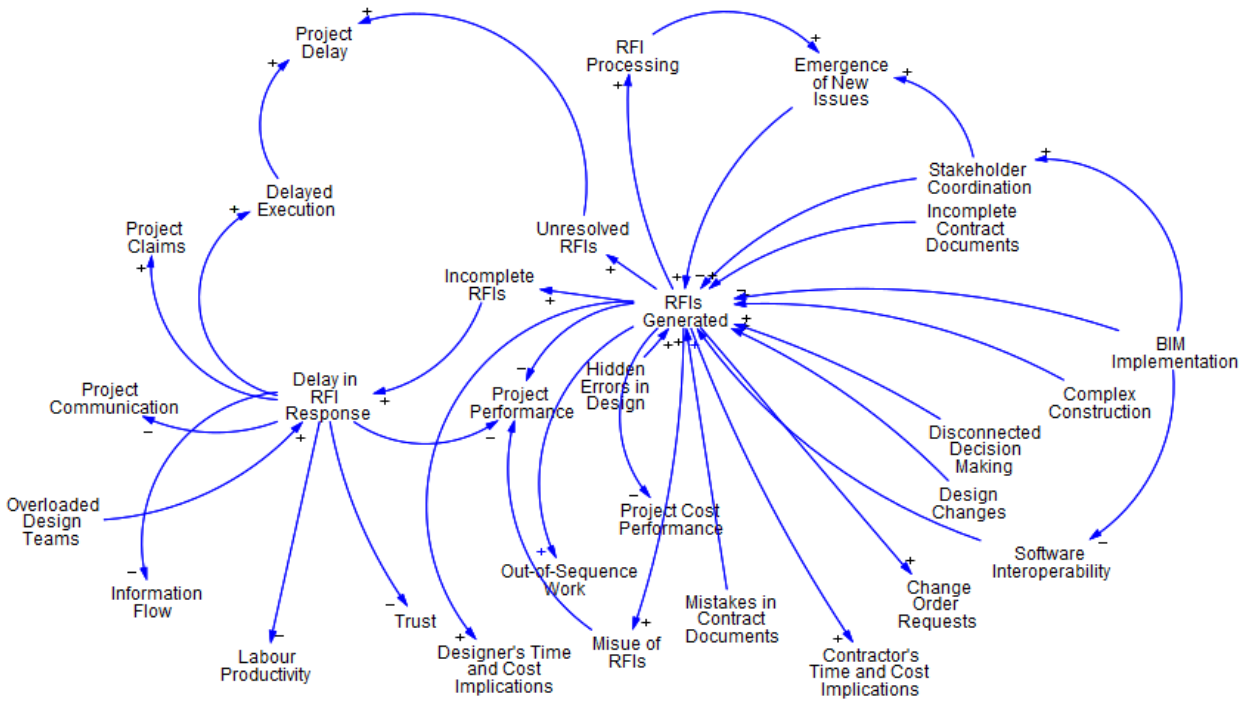


Figure 2-10. Causal loop diagram for RFI problematisation mapping.

If these issues are not dealt with timely, they appear as downstream risks in design omissions, complex designs, uncoordinated drawings, incompatible or inconsistent project specifications (Jarkas, 2015) and field execution problems (Issa et al., 2003). The "emergence of new issues" in CLD gives birth to another cycle of requests for information. The recurrent disruptions also

result in low productivity, extended time scales and escalated construction costs. The activities mentioned in the RFIs can lead to project delays and possible claims if they become part of a critical path (Papajohn and El Asmar, 2021). This usually happens due to untimely responses to the RFIs (Papajohn et al., 2018).

Moreover, Figure 2.10 reveals that "delay in RFI response" is a critical factor that has a detrimental effect on "project communication" (Papajohn et al., 2018) and "labour productivity" (Jarkas et al., 2014). The risks and uncertainties in RFIs, along with resulting schedule delays, can significantly impact the "project cost performance" for both the client and the contractor (Liao et al., 2020). The contractor bears expenses in drafting RFIs and gathering supporting documentation, whereas the client or consultant accrues costs in preparing the response to RFIs (Papajohn and El Asmar, 2021). Furthermore, it is difficult for contractors or sub-contractors to detect informational errors from multiple drawings as it is challenging, time-consuming and error-prone (Filho et al., 2016a, 2016b). Consultants blame inadequate fees and a shortage of skilled workforce to promptly address the RFIs (Philips-Ryder et al., 2013). The CLD also demonstrates that successful "BIM implementation" facilitates "stakeholder coordination" across disciplines during the design phase, resulting in a notable reduction of undetected errors (Filho et al., 2016a, 2016b). Conversely, subpar "BIM implementation" and insufficient "software interoperability" (Ali et al., 2014) contribute to a significant increase in RFIs and excessive paperwork burdens.

The developed CLD manifests RFI process as an indicator of "project performance" (Liao et al., 2020), which gets negatively affected by frequent enquiries and design clarifications raised by the contractors. The potential abuse and "misuse of RFIs" is important to note (Hanna et al., 2012). The response to the RFI may also lead to "out-of-sequence" activities (Ibrahim et al., 2020) during execution. Ideally, RFIs should be minimised; however, if used properly, the process may enhance communication and reinforce "trust" between the contracting parties (Zuppa et al., 2016; Philips-Ryder et al., 2013). Overall, this CLD exposes a lack of relationships defined in the literature to develop reinforcing and balancing loops, indicating limited research about RFI problematisation. More sophisticated modelling techniques can be incorporated to create better assessments.

2.4.2 Influence of project delivery methods

Project delivery systems (PDSs) are defined as the relation between different contracting agencies and the time involved in that relation (Hanna, 2016). The scope of this search includes

integrated project delivery (IPD) and traditional PDSs, including design-build (DB), design-bid-build (DBB) and construction management at-risk (CMR). Each delivery method has unique features and varying strengths and discrepancies (Papajohn and El Asmar, 2021). Various researchers have incorporated RFI as an essential metric while comparing multiple project delivery systems. Among other factors, the RFI-related metrics include RFIs issued per million, RFI processing time (Hanna, 2016) and an average cost of an RFI (Kim et al., 2021). Table 2.2 summarises all the studies in the literature that incorporate different RFI metrics to compare various project settings.

Research study by Bilbo et al. (2015) compared CMR and IPD. The CMR project issued 466 RFIs throughout the project, whereas 40 RFIs were issued for the IPD project. El-Asmar et al. (2013) report that traditional project delivery methods contain ten RFIs per $1m and a two-week processing time, compared to an IPD project, which has two RFIs per $1m and one-week processing time. Higgins et al. (2012) mention that DBB projects lead towards more RFI generation, having less incentive to resolve RFIs compared to methods where contractors are involved right from the design phase. DBB projects fail to achieve an integrated approach because of the incapability to get all the key stakeholders (client, consultants and contractors) to collaborate in the project's planning or design phases (American Institute of Architects, 2007). Contrary to this, (Kelly and Ilozor, 2020) reported no significant differences in RFI metrics for traditional and IPD-based project settings. In the case of PDS with a DB setting, fewer RFIs are expected as they are resolved readily by the DB contractor's engineer [Maryland DOT (MDOT), 2013]. Overall, the RFI metrics are believed to perform better in IPD settings than the traditional PDSs.

Another reason for fewer RFIs in the IPD setting is an agreement between stakeholders to resolve all the issues in a team meeting rather than raising an RFI (Bilbo et al., 2015). Hence, the challenges still exist but are tackled through a different communication channel. In the IPD setting, only issuing an RFI if necessary is encouraged. These measures keep the overall RFI metrics lower. However, there are some disadvantages to the project members in this approach. RFIs serve as an essential tool for record-keeping. In the case of limited RFI, bringing all the project details forward for crucial matters such as approvals and periodic payments becomes challenging. Bilbo et al. (2015) refer to a case study wherein IPD setting, more RFIs were produced at the end of the project to generate records and track changes. Furthermore, IPD only effectively reduced the number of formal RFIs; however, the project delivery method did not affect the quantum of issues/disruptions raised on the project. Hence, project clients should

also reassess their unrealistic expectations of IPD as a solution to construction problems, as the trade and qualitative literature suggested.

Irrespective of the project setting, RFI communication is an administrative process controlled and processed within an office setting. Hence office-based RFI communication is critical to on-site execution, and improving office-related processes can streamline RFI metrics (Alves et al., 2016) and enhance process management. Communication sequences for RFIs and submittals must be adequately managed and evaluated based on project characteristics, constraints and needs to avoid delays (Alves et al., 2016). To intelligently manage RFIs, it is necessary to understand actions that might impede communication sequences and avoid automating wasteful practices. IT provisions for RFI processes should be linked with construction documents and schedule activities to provide managers with a more comprehensive view of their RFIs and necessary information to close active RFIs. The system can highlight which RFIs need urgent action or remain overdue and send alerts to those accountable for closing the RFIs.

Finally, managing RFIs during construction projects is an inherent challenge irrespective of the project delivery method adopted and the literature reports using digital design technologies facilitating the RFI process. The following section discusses the merits and demerits of BIM for RFI management.

Table 2-2. RFI metrics in traditional PDSs and IPDs.

Delivery method (s)	Project Type	Research scope (RFI related)	Total no. of Projects	BIM-enabled	RFI-related metric	Unit	Findings	Citation
CM/GC	Highway	Determining impact of CM/GC project delivery on RFI response time	1	N/A	RFI response time	# of construction days delayed	During the project, 66 constructions days were taken to respond to RFIs The days delayed may or may not translate to project delay	(Papajohn et al., 2018)
DBB; CM/GC; DB	Highway	Comparing the impact of DBB, CM/GC, and DB on RFI response time	17	N/A	RFI response time	# of construction days delayed	No. of days delayed to produce response during the project are below: • DBB: 139 days • CM/GC: 125 days • DB: 380 days The days delayed may or may not translate to project delay.	(Papajohn and Asmar, 2021)
IPD	Energy	Developing IPD implementation framework for energy project	1	✓	RFI frequency	# of RFIs	No mention of no. of RFIs The framework reduced the project cost by 20% and cut project duration by 25%	(Psomas and Alzraiee, 2020)
IPD; CMR	Healthcare	Comparing the impact of IPD and CMR on RFI frequency	2	IPD (✓); CMR (N/A)	RFI frequency	# of RFIs	For both projects the number of RFIs were lower than expected average due to great working relations between the teams • CMR: 466 RFIs9 • IPD: 40 RFIs	(Bilbo et al., 2015)
IPD; traditional PDSs (DBB, CM GMP, CM cost + and DB)	Varying	Comparing impact of IPD and traditional PDSs on RFI frequency	93 (18 IPD)	N/A	RFI frequency	# of RFIs	Overall, at the 95% confidence level, no significant RFI differences between IPD and traditional PDSs IPD healthcare projects had significantly fewer RFIs than non-IPD healthcare projects.	(Kelly and Ilzor, 2020)
IPD; Non-IPD (CMR, DBB, and DB)	Infrastructure	Comparison of IPD and non IPD on RFI frequency and response time	N/A	N/A	RFI frequency; RFI response time	# of RFIs/million dollars; construction weeks	RFI response time • IPD: 1.4 weeks; Non-IPD: 1.5 weeks RFI frequency • IPD: 3.9 RFIs/million dollar; Non-IPD: 8 RFIs/million dollar	(Hanna, 2016)
DBB; DB	Infrastructure	Comparing cost performance of DBB and DB with RFIs	2	N/A	Avg. cost/RFI	$	Average cost/RFI: Design–build method • Owner: $1360; Builder: $4018 Design–bid–build method • Owner: $23,555.60; Builder: $783.02	(Kim et al., 2021)

2.4.3 Building information modelling and request for information management

2.4.3.1 Strengths of BIM to support request for information process.

Reduced number of RFIs in BIM-enabled projects is one of the top-mentioned benefits in the body of knowledge (Barlish and Sullivan, 2012). BIM implementation can improve the overall design by significantly improving interdisciplinary design coordination, which aids in reducing the number of requests for information (Filho et al., 2016a, 2016b). Contrary to a BIM-enabled lifecycle, in a non-BIM setup, construction professionals cannot readily visualise conflicts through 2D drawings, resulting in an administrative burden to resolve (Filho et al., 2016). Unmanaged conflicts at the design stage led to rework and a burst of RFIs, which inhibit project productivity. BIM pinpoints the likely problems before the beginning of construction work providing a robust mechanism to deal with RFIs. Filho et al. (2016a, 2016b) explain the BIM-based RFI management process. First, BIM-based models relevant to different disciplines are generated. Second, constructability analysis is performed through a 3D-coordinated/federated model containing all designs; a report with all RFIs is generated. The design teams perform the reviews and release the updated version of the models. The coordinated model and latest design are examined in the final step to ensure that pointed problems have been solved. In this stage, knowledge exchange between the design and virtual design and construction (VDC) team occurs to ensure that the design has been validated and most RFIs have been resolved. After modelling the design changes, a final report is released containing comments unresolved. Giel & Issa (2013) explain in detail how BIM-enabled design can reduce the number of RFIs:

- Reduced dimensional inconsistencies: BIM helps reduce discrepancies and rounding errors in 2D drawing sheets.
- Reduced document discrepancies between disciplines: The information segregation during 2D design drawing preparation in each discipline leads to errors that are more significant. BIM ensures smooth design coordination across different disciplines reducing the number of RFIs.
- Reduced 2D errors and omissions: BIM provides a virtual model at the pre-construction stage, which provides a chance to minimise the clarifications that otherwise become RFIs.
- Reduced grid or column alignment issues: During design, column and grid alignment contributes to major source of conflicts in a project; however, in BIM-enabled design, these issues are resolved during pre-construction, reducing the occurrence of RFIs.

- Reduced direct clashes: Advanced visualisation abilities in BIM resolve conflicts between systems and disciplines early. These conflicts may include clashes between architectural, structural, HVAC and other building systems.

2.4.3.2 Challenges BIM-based request for information management.

Although BIM has drastically improved the landscape of AEC industry in terms of design and execution, research has pointed to discrepancies and inefficiencies related to issue handling and RFI management. At the award of contract stage, the general contractor seeks clarifications from consultants to finalise their BIM models (Liao et al., 2020). Eventually, the contractor must re-create the MEP design models due to insufficient documents from designers. During the design phase, due to poor BIM collaboration among different services, design issues are not resolved until the execution, resulting in a large number of RFIs, and a potential project delay (Liao et al., 2020). Poor BIM management practice creates excessive paperwork for all the parties involved in the project.

Similarly, architects and engineers take long to respond to contractors' RFIs as their design models do not directly guide site work (Liao et al., 2020), causing delay risk. With such a scenario at hand, the speciality contractors are also left with inconsistent and uncoordinated models and drawings and unresolved clashes to carry out execution activities to manage the tight construction schedule models (Liao et al., 2020), starting the vicious cycle of RFIs. Furthermore, trade contractors rarely incorporate BIM (Lam, 2014); this practice leads to executing construction activities that are unplanned and inconsistent with the design models of the in-house BIM. Moreover, clashes detection during execution leads to abortive works, resultantly requiring extra time and person-hours to re-design, re-do and rebuild the construction work (Liao et al., 2020). These issues can be streamlined by developing a BIM execution plan (BEP) that properly narrates how federated models will be updated to showcase the final design intent. Further, BEP should include the frequency and extent to which change orders and RFI responses affect design intent and should be incorporated into the design authoring models.

Despite the overall decrease in RFIs due to improved visualisation and experiential experience, the nature of the RFIs generated as a result of poor BIM coordination practices and BIM-based design deficiencies remains unclear, primarily due to the absence of research on this aspect (Liao et al., 2020). There is also a relationship between BIM professionals' skill level in design modelling and on-site project productivity based on rework and RFIs (Fan et al., 2014). If the

BIM modeller does not incorporate the required level of detail for the speciality trade, then BIM will not serve as an accurate virtual representation of the on-site components. Hence, inexperienced BIM modellers become the primary reason for BIM failure, which increases the amount of rework in the field and the number of related RFIs (Fan et al., 2014). Moreover, with BIM in practice, not all the stakeholders have the same level of information, especially trade contractors. This gap in knowledge between stakeholders becomes the source of errors during design and construction, leading to the vicious cycle of RFIs. To assuage these negative implications, BEP should narrate a framework acting on which team members determine whether an RFI response disturbs design intent and a mechanism to track any changes in design authoring models that arose due to these RFI responses. Considering the limitations of BIM for RFI management, both industry and academia have explored other technologies to streamline the process, which are discussed in the following section.

2.4.4 Other digital tools and platforms to aid request for information management

The construction industry has observed a subtle shift in managing RFIs from paper/email-based channels to online. Table 2.3 points out the technologies mentioned in the literature to aid different aspects of the RFI processes. Currently, several electronic data management systems (EDMS) are available that allow construction stakeholders to manage their RFI documents (Kähkönen and Rannisto, 2015). EDMS are Web-based platforms that enable tracking changes and traceability (Pradeep et al., 2021) for RFIs. This makes their usage and applicability more efficient than traditional paper/email-based RFI communication, lacking data traceability. With the advent of BIM in the construction industry, more sophisticated data management provisions are now available. A common data environment (CDE) is one such platform that facilitates model-based project management (Pradeep et al., 2021). CDE platforms that support RFI administration and handling include Aconex (Das et al., 2020; Aibinu et al., 2019), Autodesk 360 (Das et al., 2020) and Procore. CDE provides a collaborative platform for project stakeholders to exchange information through cloud technology. Both EDMS and CDE have helped the construction industry handle RFIs; however, they present a few challenges. Some of these challenges include; complex data architecture, storage problems, interoperability issues within EDMS (Kähkönen and Rannisto, 2015) and data loss and legal issues within CDEs (Pradeep et al., 2021). Academic researchers are now embracing emerging technologies to develop transparent and secure file handling platforms to deal with these challenges. Studies by Das et al. (2020) and Pradeep et al. (2021) provide a decent start towards developing

Blockchain-backed prototypes to develop better RFI management systems. However, these studies still require industry validation to substantiate the results claimed in the studies.

Table 2-3. Technologies related to RFI management.

Technology category	Technology/Software	Functionality to aid RFI management	Citation
Common data environment	Aconex, Autodesk 360	RFI administration/management/handling	Das et al., 2020 (✓)
	Aconex		Aibinu et al., 2019
Emerging technologies	Blockchain	RFI administration/management/handling	Das et al., 2020
		Information exchange and data security	Pradeep et al., 2021 (✓)
ERP System	RFI Automation tool	RFI administration/management/handling	Ekstrom and Bjornsson, 2005
Immersive technologies	Mixed Reality	Design coordination and collaboration	Alizadehsalehi et al., 2019 (✓)
Other ICT provisions	Interactive Voice Response, Email	Communication	Abdel-Monem and Hegazy et al., 2013
Software solution	Unity 3D, 3DS Max	Design Visualisation	Castronovo et al., 2018 (✓)
	Unity 3D	Design coordination and collaboration	Alizadehsalehi et al., 2019 (✓)
	CMiC	RFI administration/management/handling	Sawyer, 2007
Text mining	Metadata modelling	Content analysis	Zhu et al., 2007; Mao et al., 2007
	Topic modelling	Risk assessment	Lee and Yi, 2017
	QCA software	Content analysis	Soh et al., 2020
Web-browser interface/web-based platform	Electronic Data Management System (EDMS)	RFI administration/management/handling	Kähkönen and Rannisto, 2015 (✓); Papajohn et al., 2018

(✓): BIM-enabled

Since RFIs comprise unstructured text, different researchers have explored the text mining approach to decode the content of the RFIs. In this regard, Zhu et al. (2007) and Mao et al. (2007) developed metadata models to facilitate the processing of the RFIs. Lee and Yi (2017) recently used text mining through the topic modelling approach and categorised RFIs based on their risks. Researchers have also used immersive technologies to improve various aspects of design, such as design visualisation, coordination and collaboration to proactively detect the issues, which may later become part of the RFI process. Alizadehsalehi et al. (2019) focused on mixed reality. They integrated it with BIM to enhance stakeholders' coordination which indirectly leads to timely resolution of issues, ultimately reducing the negative implications of the RFI process. Most of the approaches mentioned in the literature had an indirect impact on the RFIs and limited to no research aims to reduce the review time of RFIs using technologies.

2.4.5 Classification methods within the literature

RFI has been criticised for the lack of structure (Mao et al., 2007). Unstructured information presented in RFI can be related to any issue, depending on the clarification by the contractor. Different researchers have characterised the reasons or causes of RFIs to analyse their content. Through this analysis, industry stakeholders can delve into important questions such as: why the issue existed, and how it could have been addressed? (Bhat et al., 2017).

Based on these explorations, best practices and strategies can also be devised to reduce the turnaround time and improve the overall RFI process (Kim et al., 2021). Considering the anatomy of an RFI, contractors or owners do not classify the RFIs based on the issue/reason codes (Kim et al., 2021); hence researchers have developed different classification methodologies to categorise RFIs and analyse their content. Figure 2.11 provides an alluvial diagram representing different classification frameworks and the sub-classifications incorporated in the literature. Past research has classified RFIs based on the property code, discipline code, work-element code, category/type of RFI and reason/issue behind RFI. The individual node height is an indicator of frequency. Within these categories, classification based on the category/type of RFI, and reason/issue has been the predominant categories. Each researcher has categorised the RFIs based on their own experience and knowledge (Kim et al., 2021). Hence, it can be observed that few RFI sub-classifications belong to two different categories. For example, design clarification is termed as a type category by Bhat et al. (2017) and, at the same time, an issue category by Hanna et al. (2012).

Human driven codification highlights multiple sub-classification categories, such as alternative design solutions, approvals, confirmation, conflict, construction coordination, design coordination, design modification, value engineering, constructability issues, differing site conditions, omission, divergence and design clarification, as predominant and recurring issues within the RFIs, indicating their significance in construction projects. Interestingly, few researchers have incorporated multiple classification methods to analyse the RFI documents. For example, Filho et al. (2016a, 2016b) devised a two-stage RFI classification methodology. Firstly, the RFI was classified based on correction, omission, verification and divergence categories. Then, the authors reposed a coding process to organise the RFIs into issues, such as, conflicts, poor alignment, impracticable ceiling height, level differences and absence of structural elements as the predominant problems. Similarly, Bhat et al. (2017) followed a three-step codification process to annotate and analyse the RFIs. These classifications can be valuable to the project team as they enhance communication, navigate the project team and align the project towards intended and desired results (Brazee, 2014).

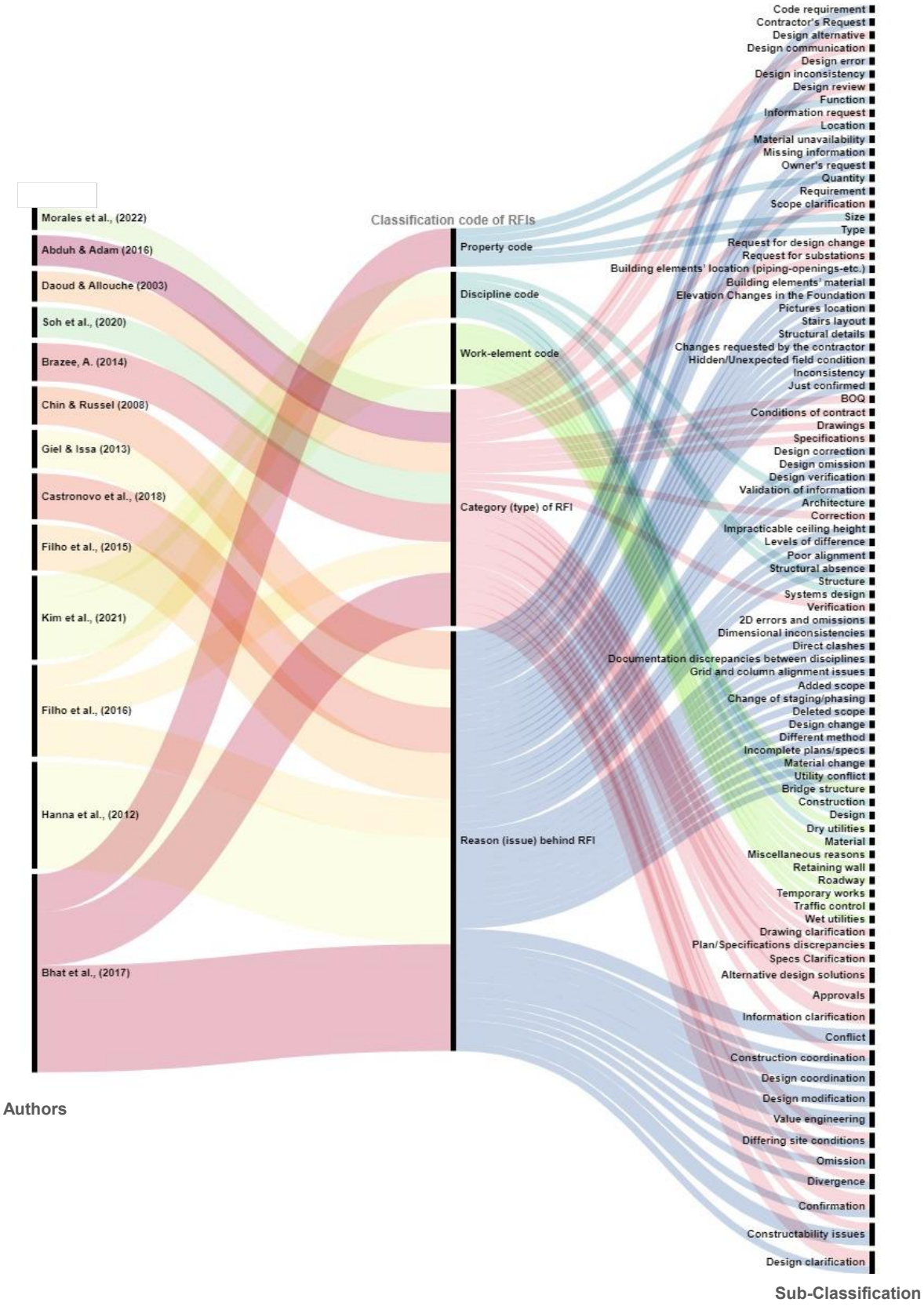


Figure 2-11. RFI classifications mentioned in the literature.

2.5 A way forward on RFI management

With the viewpoints taken from the above themes, Figure 2.12 illustrates the strategies and best practices for the better overall management of the RFI process that can potentially reduce the number of subsequent RFIs. These best practices and strategies can be divided into the following six groups:

i) *Effective communication:* Effective communication is vital to handling the RFI process effectively. Clear expectations of the project outcomes should be defined at earlier stages by the owner (Liao et al., 2020). Otherwise, unexpected scope deviations at later stages delivered through RFIs and other project documents make difficult for contractors to execute and lead to cost implications and schedule delays. Hence, project program meetings should be conducted to share information and foster meaningful and open exchanges between the teams (Hanna et al., 2012). Particularly, these project program meetings should be integrated with the RFI management system so that strategies are promptly discussed towards swiftly closing active RFIs.

ii) *Standardised RFI process:* RFIs should be prepared in a standardised format with a proper mentioning of elements, such as; number and date, the identity of the author, a reference to the design or specification document, suggested solution, response time, a tracking number and an indication to schedule and cost impacts (Hanna et al., 2012). Hanna et al., (2012) further add; that an RFI should seek only one clarification, ideally with a probable solution to decrease the response time. Batching RFIs is considered malpractice (Andrews, 2005). From consultant's/owner's end, the response should be timely and guide towards resolving the problem. In case of a delay, the submitter should be notified. Furthermore, the RFI logs should be updated consistently, and in case of any change, they should be communicated to all the concerned parties (Andrews, 2005). A change order request should be initiated if an RFI changes the scope of work or a change in contract conditions.

iii) *BIM-based measures:* For the projects incorporating BIM, to manage any RFIs that are generated due to the non-standardised BIM-design coordination, a contractual framework should be established that promotes RFI-based collaboration (Piroozfar et al., 2019; American Institute of Architects, California Council [AIACC], 2014), especially for speciality designers (Morales et al, 2022). This minimises errors (Liao et al., 2020), raising fewer RFIs. BIM implementation plan should incorporate a digitised RFI management system that enforces all the stakeholders' participation from the earlier design stages

(Alreshidi et al., 2017). The lack of skilled BIM labour (in speciality contractors) is an issue that hampers desirable BIM implementation. RFIs that are caused due to such modelling mistakes can be managed by frequent BIM-QA/QC and code compliance checks (Donato et al., 2017). Further, the client body can arrange special training to upskill the team. This way, additional costs, duplicate efforts, and RFIs generated due to poor BIM modelling practices can be minimised.

iv) *Realistic timelines for design:* Insufficient durations should not be imposed upon designers to develop and specifications, and tender documents and review design alternatives. This leads to design documents remaining often ambiguous, incomplete, and containing myriad of unresolved issues within disciplines (Daoud and Allouche, 2003). An accurate schedule ensures the timely completion of design works. On the contrary, an unrealistic schedule may backfire on the design team and the quality of design packages. The time saved in the design phase could accrue several times during bidding and execution (Daoud and Allouche, 2003). Incorporating lessons learnt from findings of previous projects can provide an understanding of the mistakes that can be avoided when designing for future projects.

v) *Constructability exercises:* The design firms should actively engage in constructability exercises to avoid consecutive interruptions to the work progress. Sorting errors initially reduces the number of RFIs; otherwise, the same omissions exacerbate the morale of workforce, creating frustration, demotivation, and low productivity. Jarkas et al., (2015) and Papajohn & El Asmar (2021) suggest that the appointment of a high calibre project manager at earlier stages can ensure; a) scrutiny of contract documents (Hanna et al., 2012; Daoud and Allouche, 2003), b) detection of pitfalls, and missing details, and c) eventually prepare an in-depth RFIs to avoid downstream risks.

vi) *Leveraging RFI process:* Although literature labels the RFI process as undesirable (Aibinu et al., 2019), the process can be leveraged to minimise project risks. This can be achieved if the proper criterion for a justified RFI is established. For example, RFIs that can be answered by field staff, loosely worded RFIs, or unrelated RFIs can be detected early and sent back to the initiator or discouraged (Papajohn & El Asmar, 2021). RFI submittal should not be batched. Papajohn & El Asmar (2021) adds that RFI logs should not be looked at as a reason to initiate project claims. In case of intricate design issues, a project meeting should be held to resolve them instead of abusing RFIs to document claims. As per Andrews

(2005), contractor and client/designer should mutually agree on the type and scope of issues translating to an RFI.

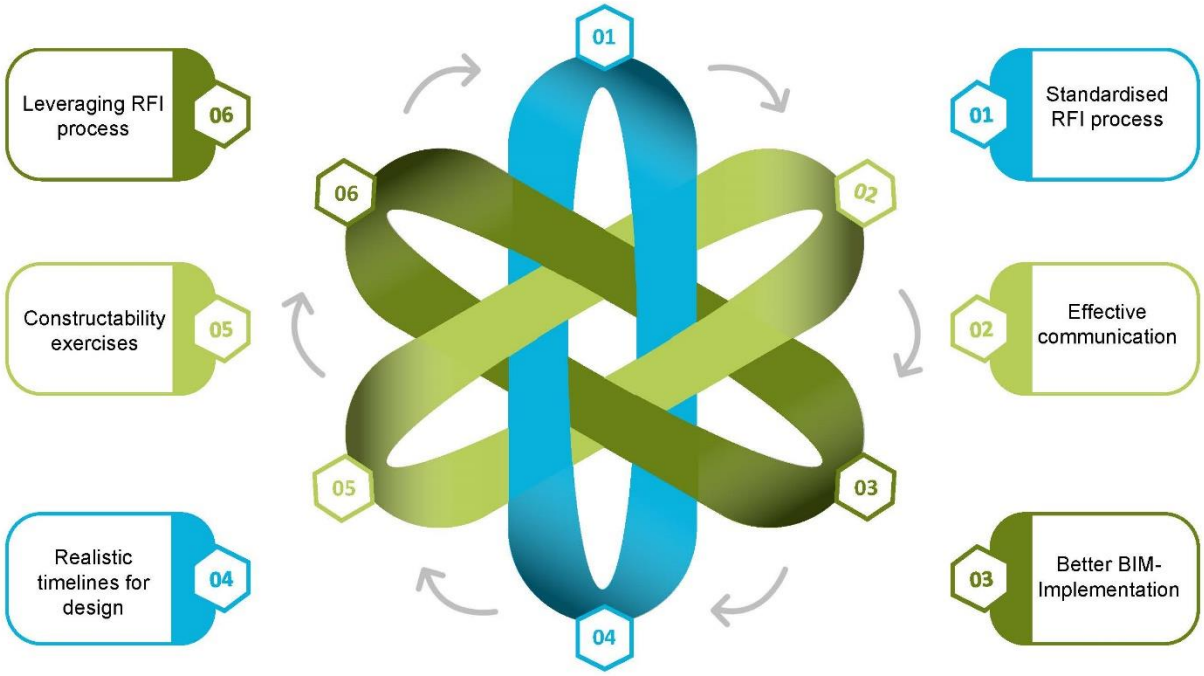


Figure 2-12. Best practices for a way forward on RFI management

2.6 Research gaps

This review identifies gaps in RFI management research and proposes further investigation to provide solutions and enhance understanding.

2.6.1 Improved information exchange platforms

In this review study, several shortcomings of existing RFI management platforms were identified, including:

- complex folder structure;
- slow and ineffective system usage;
- combination of documents and digital data; and
- different end-users and project needs.

This problem becomes more complex in BIM-enabled projects. BIM standardisation and software interoperability remain the challenges encountered by industry and academic. Many speciality contractors and suppliers do not have appropriate BIM suites and skilled labour to operate efficiently. Furthermore, the specific RFIs arising from BIM-related issues have not been extensively explored. The optimal data exchange will happen only if the design team, the

prime contractor, all subcontractors and suppliers/vendors have full access to the BIM documents and use the same BIM software packages along with efficient platforms that streamline RFI sharing, tracking and monitoring

2.6.2 Automated approaches to determine request for information priority

Streamlining RFIs requires prompt responses from consultants/owners to avoid negative impacts on the project schedule (Papajohn and El Asmar, 2021). RFIs should indicate the date and priority of the requested clarification, as prioritisation can aid in timely responses and avoiding delays (American Institute of Steel Construction, 2016). A priority provided by the contractor will be backed by approximate estimations. Predictive data analytics can assist in determining RFI priority, allowing for allocation of resources and timely closure of urgent RFIs, thus avoiding disruption to the project's critical path.

2.6.3 Analysing request for information content through text mining

Analysing RFI content is essential for learning from past experiences and enhancing design practices for future projects. To achieve this, advanced techniques such as text mining and natural language processing must be used to comprehend the unstructured content within RFIs. Natural language processing is a realm within artificial intelligence that aids in dissecting, comprehending, and manipulating human languages, encompassing unstructured text documents as well as human speech (Hassan and Le, 2021). Leveraging these approaches will provide valuable insights from RFIs, enabling the identification of various risks. Consequently, this will enhance the handling of conventional RFI documents and reduce the turnaround time of RFIs. However, the reason this type of research has not been extensively explored yet is because the adoption of advanced text mining approaches in the construction field is progressing slowly. Extracting knowledge from construction documents using text mining remains a challenging task. Additionally, the sensitivity of RFI documents makes it even more difficult to access relevant documents, as they often reveal project mistakes and are not readily available.

2.6.4 Critical risk factors identification

This review developed a preliminary CLD to map different challenges associated with the RFI process. However, it serves as a qualitative conceptualisation due to the lack of relationships mentioned in the literature. Future research may incorporate industry feedback to develop a systems dynamic (SD) model to establish the missing links in the literature between different variables through quantitative analysis. A robust SD model will serve both body of knowledge

and industry professionals, a holistic understanding of the causality relationships within the RFI process.

2.7 Comparative analysis and novel insights: situating the review within the construction innovation landscape

The present study aligns with literature reviews on NLP in construction, revealing limited research on leveraging NLP and text mining to extract valuable insights from construction documents (Wu et al., 2022). Additionally, the scarcity of research on applying text mining techniques to construction documents underscores data sharing challenges due to sensitive project information (Baek et al., 2021; Yan et al., 2020). Notably, this literature review advances construction innovation knowledge by raising awareness of ongoing trends and emphasising the need for text mining and NLP techniques to extract valuable information from unstructured construction documents.

Distinctively, previous research lacks a comprehensive examination of the evolution from traditional construction processes to advanced technologies, such as BIM, while advocating for analytics integration. For instance, a specific article by Sompolgrunk et al. (2021) focuses on BIM's impact on the RFI process, while research by Rowlinson (2017) and Karasu et al. (2022) delve into BIM implementation within IPD settings. Similarly, Ali et al. (2022) attempt to address construction claims but limit their suggestions to BIM and other digital technologies, overlooking the transformative potential of advanced analytics in text mining research within the construction domain. Unlike the saturated body of knowledge surrounding BIM-related research (Abdal Noor and Yi, 2018), this study provides a fresh perspective.

A pioneering aspect of this research is its comprehensive examination of the entire RFI lifecycle, from traditional practices to BIM-enabled processes and beyond. Such a holistic approach makes it a trailblazer in the construction innovation landscape. The present literature review focuses on a critical construction communication process, thoroughly dissecting its limitations, technological advancements and future trajectory for further enhancement. Emphasising the adoption of recommended strategies, the potential benefits are substantial, ranging from driving advanced practices to optimising project resources and enhancing construction business profitability. This systematic research approach and extensive coverage of RFI processes has the potential to lead and serve as valuable roadmap for other construction specific documents or processes. By adapting the methodology and insights derived from this research, other construction-related documents, such as change orders, contract documents,

specifications, claims and request for proposals, can be subject to similar analyses. The integration of text mining and NLP methods, as briefed in this study, can reveal actionable insights from construction documents, leading to improved decision-making, streamlined processes and desired project outcomes.

In addition, this research recognises the transformative potential of intelligent large language models such as ChatGPT, that can streamline knowledge extraction from construction documents. Embracing these technologies can revolutionise the construction industry, offering efficient data analysis and improved decision-making. In conclusion, the adoption of cutting-edge technologies offers transformative possibilities for the industry, leading to increased efficiency, better project outcomes and heightened competitiveness. As the construction landscape evolves, embracing advanced text mining and NLP will play a pivotal role in shaping the future of construction innovation.

2.8 Summary of the chapter

This literature review serves as a comprehensive review of existing studies on RFI process in construction industry. A number of themes in RFI management literature were identified and discussed in this study, including:

- risk mapping;
- influence of different project delivery methods;
- BIM and RFI management;
- other digital tools and platforms to aid RFI management; and
- classification methods within the literature.

It is revealed that the RFI processes directly impact project interface, success and productivity. RFI frequency and response time have been regarded as essential metrics to compare different project settings. In this review study, it also found that no project delivery method can be deemed most suitable for minimising the risks due to the RFIs.

Overall, this literature review serves as a comprehensive reference for academicians and industry experts interested in the RFI processes in the construction sector. For academia, this study provides a holistic understanding of the RFI process by discussing its repercussions in both traditional and BIM-enabled lifecycles. The CLD diagram was developed to illustrate the causal relationships between different project parameters that influence the RFI process. Further, this study summarises all the classification frameworks developed by different

researchers to decode the unstructured content of the RFIs. Based on the discovered themes and critical literature analysis, this study presents the best practices and strategies. This review extends the body of knowledge and may help world of practice in:

- reducing the occurrence of RFIs; and
- in case of such events, recommend measures to reduce the RFI review period.

While existing research has made efforts to categorize and classify RFIs to uncover their underlying causes, such methods often rely on manual processes, making them time-consuming and inefficient. Addressing the research gaps identified in the literature review, this thesis focuses on one main gap-text mining techniques to analyse RFI content in a more automated and scalable manner. The research questions outlined in Chapter 1, Section 1.3, are closely aligned with this gap. In response, Chapter 3 presents the research methodology, employing a design science research framework to systematically investigate the problem. Chapters 4 and 5 detail the development and evaluation of NLP models designed for information extraction and text classification from RFIs. These models aim to generate actionable insights and enhance RFI handling processes, ultimately contributing to the resolution of the research questions set forth in Section 1.3.

Chapter 3: Research methodology

This chapter outlines the methodological framework adopted in this thesis, selecting the design science research framework. This framework guides the research, addressing the research questions through its principles. The subsequent sections detail how the design science research framework is applied to employ natural language processing for text classification and information extraction from RFIs. The chapter concludes with a visual representation of the components of the design science research framework and their relation to the various chapters in this thesis.

3.1 Determining the methodological framework

Opting an appropriate methodological framework depends on the specifics of the research problem and the research objectives (Yin, 2017). These factors emphasise the need to tackle a practical issue through the development of a viable solution. Accordingly, the design-oriented framework incorporates understanding the practical problem, creating and refining a solution, assessing its effectiveness, and disseminating the findings (Hevner et al., 2004). Overall design-oriented framework can be categorised into action, constructive, and design science research (DSR) methodologies.

The constructive research approach utilizes a question-driven design that creates a logical framework linking empirical data to the original research questions of the study, ultimately leading to well-founded conclusions (Oyegoke, 2011). The primary goal of constructive research is to develop innovative solutions to theoretical and practical challenges. These solutions are often generated through managerial problem-solving techniques, which include the creation of models, diagrams, and strategic plans. Adhering to a design-oriented framework, this approach seeks to enhance management science research by formulating theoretical constructs (Lukka, 2003). The constructive research approach emphasizes the design of constructs and the development of practical solutions (Oyegoke, 2011). Market research is utilised as a method for testing and evaluating these proposed solutions (Piirainen & Gonzalez, 2013). Several studies have effectively utilized the constructive research approach, including Kasanen et al. (1993), who developed a model for capital budgeting in organizations, and Lindholm (2008), who established a corporate real estate management framework utilising this methodology.

Action research is recognized as a research method that incorporates a design-oriented framework aimed at developing solutions to address significant and practical challenges (Järvinen, 2007). As noted by Baskerville and Myers (2004), this approach requires the researcher to be actively engaged within the social system, thus providing a contextual backdrop for problem-solving. Action research leverages a design-oriented framework to confront issues and create solutions that are relevant to the complex interactions within organizational settings and operational processes (Avison et al., 2001). It emphasizes the construction and testing of theories while addressing urgent practical problems in real-world environments (Azhar et al., 2010). In this methodology, the researcher investigates the current conditions of a specific problem area, identifies relevant issues, actively participates in implementing changes to improve the situation, evaluates the effects of these changes, and reflects on the entire process and its outcomes to generate new insights (Naoum, 2001; Baskerville, 1999). Action research is a well-established methodology that has found successful application across diverse fields, including IT (Lindgren et al., 2004), education (McTaggart, 1991), management (Susman & Evered, 1978), and healthcare (Meyer, 2003).

3.1.1 Design science research

Design science research (DSR), a prevalent problem solving and design-oriented research framework, is aimed at advancing human knowledge through understanding practical problems and devising solutions by developing innovative artefacts (Vom Brocke et al., 2020; Gregor and Hevner, 2013). DSR strives to expand the foundations of both science and technology by developing solutions that address specific problems and improve the environments in which they are applied, while also contributing to the growth of design knowledge (Vom Brocke et al., 2020). DSR differs from the constructive and action research approaches in several key aspects (Ekankaye, 2022). As noted by Iivari and Venable (2009), DSR does not always require active collaboration between researchers and practitioners, which is typically essential in the other two methodologies. This distinction arises from DSR's goal of developing a broad framework for design-focused research (Baskerville, 2008), making it applicable to a wider range of contexts beyond individual organizations. Consequently, DSR can utilise alternative forms of testing and validation beyond traditional market testing (Piirainen & Gonzalez, 2013).

The DSR which is essentially a problem-solving approach has its roots from engineering, sciences (Simon, 1996) and IT disciplines (Hevner et al., 2004). Over the past two decades, this research methodology has sparked significant interest in enhancing organizational

capabilities (Vom Brocke et al., 2020; Watson et al., 2010). According to Gregor and Hevner (2013), the goal of DSR is to expand organisational potential by developing innovative artifacts, which may include methods, constructs, models, or practical applications. DSR seeks to develop understanding of how products can be created or organized, typically through human intervention, to meet specific objectives (Vom Brocke et al., 2020). It functions as a fundamental research paradigm in multiple disciplines, including architecture, economics, business, information technology, and engineering, where it is applied to create innovative solutions for relevant design challenges (Vom Brocke et al., 2020). Previous studies have utilised the DSR process to develop various innovations. For instance, Ekankaye et al., (2024) utilised DSR principles and developed a deep learning model integrated with computer vision for construction progress monitoring for indoor construction projects. Pradeep et al. (2020), developed a blockchain-based prototype using DSR to promote the information exchange within construction projects. The prototype was designed to integrate blockchain technology into design management processes—such as design review, coordination, and RFIs and Bodenbenner et al. (2013) utilised DSR to design an efficient electricity demand response system.

3.1.2 Mapping the research process with design science research framework

A structured process for implementing the DSR framework was introduced by Vaishnavi and Kuechler (2004), which was later refined into a standardized methodological framework by Peffers et al. (2007). The present research also utilised DSR framework, with the goal at analysing the problem and exploring different courses of action, thereby enhancing knowledge and contributing to the theoretical and practical realm. This commonly adopted framework is illustrated and mapped onto the research thesis through Figure 3.1. The DSR framework utilized in this study draws inspiration from the fields of construction informatics and information systems, drawing specifically from studies by Ekankaye (2022); Pradeep et al., (2021); Peffers et al. (2007), and Vaishnavi and Kuechler (2004). The following steps detail how design science research is applied in this study for automating text classification and information extraction from construction RFIs using natural language processing.

3.1.2.1 DSR Step 1: Problem formulation

The first phase of the DSR framework focuses on defining the problem. During this phase, the focus is on formulating the problem, highlighting its importance and uncovering its root causes.

For this research, problem formulation began with a systematic literature review of the existing RFI process within the construction sector. To justify the problem's significance, it must be situated within practical contexts, and its underlying causes must be understood (Johannesson & Perjons, 2014). Thus, the literature review (chapter 2) provided a comprehensive understanding of the RFI process by examining the predominant themes in RFI-related research. One critical theme was the risks associated with the RFI process. Often labelled a "necessary evil" in the literature, it was crucial to comprehend the consequences of inadequate management of this process. The evaluation was conducted using the principles of systems thinking to comprehensively understand the risks associated with the RFI process and their impact on construction projects. Following this theme was the discussion on another critical aspect: existing solutions and practices within both research and industry aimed at streamlining the RFI process. A detailed feature review of predominant industry solutions was conducted to identify discrepancies and gaps. This thorough review provided a deeper understanding of the missing elements, paving the way for devising and developing a solution that advances academic research and serves as a viable tool for industry practitioners: this thematic analysis and feature review of existing solutions identified gaps in the current approaches.

3.1.2.2 DSR Step 2: Establishing requirements

The next stage within this framework aims to develop a tangible product to address issues by outlining the requirements. This phase typically involves envisioning a solution and delineating development requirements (Ekanayake 2022; Pradeep et al., 2021; Johannesson & Perjons, 2014). Building on the gaps identified in the literature (chapter 2), the proposed solution seeks to enhance existing common data environment platforms, suggesting the addition of analytics and insights to help decision-makers swiftly review and resolve the RFIs. This enhancement can be achieved through automated text classification and information extraction using natural language processing techniques, given the inherently unstructured nature of RFI queries and responses. This approach aims to expedite RFI processing by developing advanced models that leverage state-of-the-art NLP techniques to extract actionable insights from RFIs.

3.1.2.3 DSR Step 3: Solution development

The third step of this research framework emphasises on creating a solution that effectively tackles the identified problem (Ekanayake 2022; Pradeep et al., 2021; Johannesson & Perjons, 2014; Vaishnavi & Kuechler, 2004). In this step, the technical and performance features of the

proposed solution are formulated to align with the established requirements (Gregor & Hevner, 2013). This research (chapter 4 and 5) uses NLP techniques and machine learning algorithms to develop novel models that facilitate automated information extraction from RFIs. This solution addresses the deficiencies in existing CDEs and provides a pathway for researchers to leverage techniques for efficient text classification and information extraction from RFIs. Optimisation of the models is achieved through various feature extraction techniques, hyperparameter tuning, and maintaining low training and validation loss, resulting in high overall accuracy for classifying RFIs or extracting key entities from them. Different algorithms are employed based on the task at hand. For example, the next chapter compares deep learning-based RNNs with traditional machine learning models for phase-wise separation of RFIs. This chapter also explores leveraging topic modelling application of NLP through the implementation of LDA algorithm. In the fifth chapter, two distinct models are developed. The first classifies RFIs based on the predominant issue using a CNN algorithm. The second part of this chapter utilises BERT, BiLSTM, and their ensemble versions (BERT-CRF and BiLSTM-CRF) to extract key entities from RFIs.

3.1.2.4 DSR Step 4: Solution testing

During the fourth step of the DSR framework, the developed solution is tested to determine how effectively it solves the problem and adheres to the requirements. Depending on performance requirements, testing can involve a demonstration or the deployment of the solution. At a minimum, the solution must prove its effectiveness in at least one real-world case (Piirainen & Gonzalez, 2013). The methodology advances to concluding phase only when the solution is considered viable and meets the set requirements (Hevner et al., 2004). For the phase-wise separation model (model 1) presented in chapter 4, the research employs a train-test split method and compares the performance of the best-performing machine learning model against human participants in coding RFIs. For models 2 and 3 (chapter 5), this thesis again applies the train-test split method to test RFIs collected from real construction projects, thereby evaluating the information extraction and text classification capabilities of the developed NLP-driven models.

3.1.2.5 DSR Step 5: Research communication

After completion of testing and evaluation of the developed solution against the established requirements, the final stage of this framework comprises the conclusion phase, wherein the

results are presented to the intended stakeholders. This research concludes with a guide or roadmap (chapter 6) for industry practitioners on integrating the models within the traditional paper/email-based RFI process and the more advanced CDE-based RFI management. Additionally, for the research community, the findings of this research will be disseminated through this thesis, along with journal publications and conference presentations. The DSR framework, which encompasses articulating the problem statement, determining solution requirements, developing the solution, testing it, and sharing the findings, corresponds to the process of automating text classification and information extraction from construction RFI documents using natural language processing techniques. The next two chapters provide details on data collection, NLP techniques and algorithmic architecture, performance evaluation of techniques, testing and validation and results of the developed models.

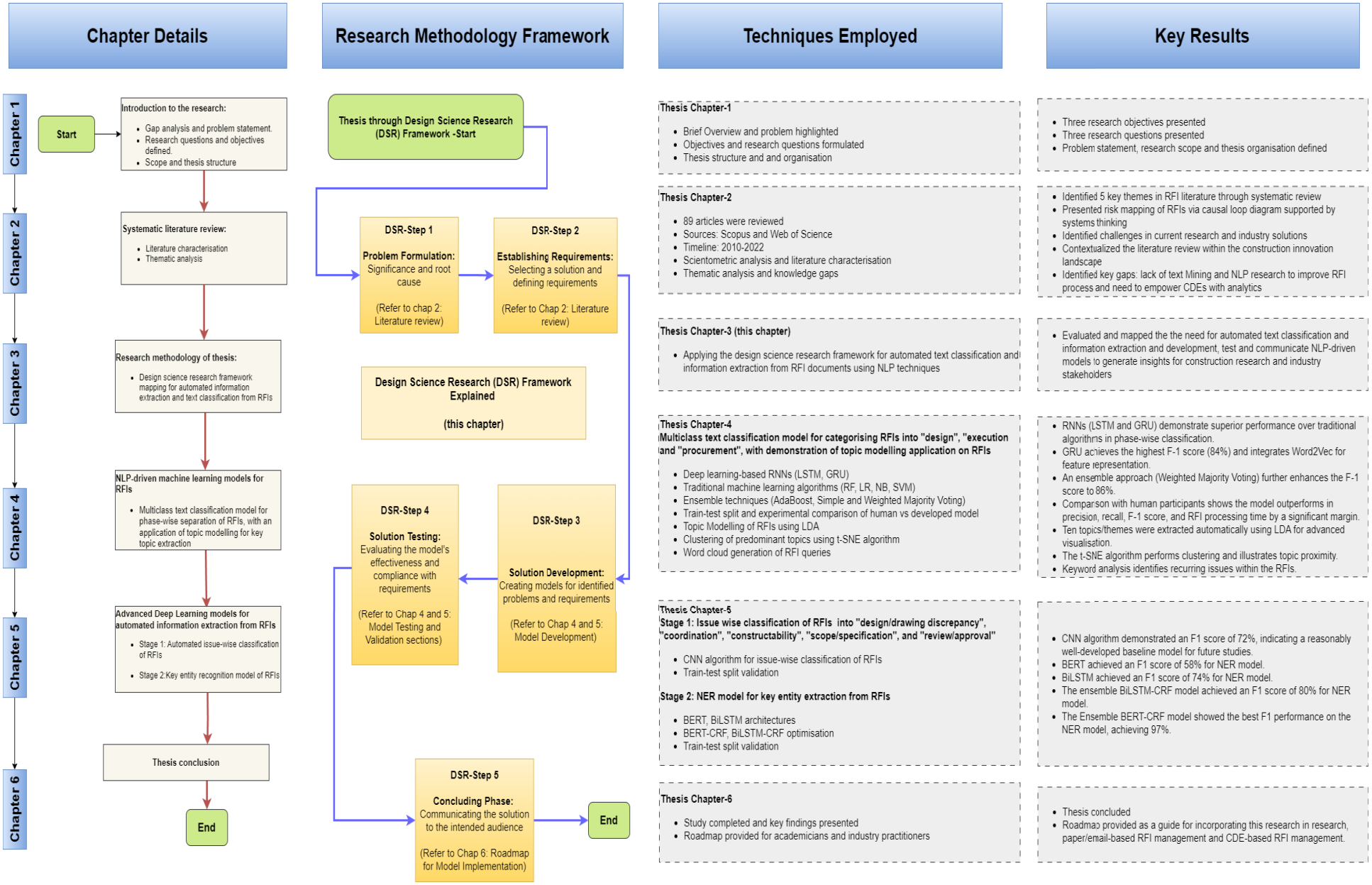


Figure 3-1. Research methodology mapped with design science research framework.

Chapter 4: Automated phase-wise separation for advanced RFI management through natural language processing

Chapter 4 delineates the application of supervised machine learning algorithms and NLP techniques for efficient text classification from RFIs. The current RFI process lacks insights and analytics that can enhance the handling and management of RFIs by generating actionable insights. While researchers have analysed RFI content to extract insights and patterns, the process has been manual, laborious and impractical for resource-constrained projects. To address this issue, this study proposes a multiclass text classification model that categorises RFIs by project phase. This model aims to improve information flow, identify trends and patterns within the RFIs, and expedite the RFI review process. The model utilises text mining and natural language processing techniques and compares two deep learning based recurrent neural network (RNN) architectures and four traditional machine learning algorithms along with ensemble learning methods. The developed model was validated through an experiment in which it surpassed human participants in categorising and reviewing RFIs based on the project phase. Additionally, the chapter applies topic modelling using Latent Dirichlet Allocation (LDA), a text-mining algorithm, to identify predominant topics and themes within RFIs. This analysis successfully uncovered key issues related to structural discrepancies, construction coordination, building fixtures, specifications, and construction drawings as the most prominent concerns discussed in RFIs. As exploratory research, the findings of this study enhance our understanding of RFI-related issues and lay the groundwork for future investigations that could further dissect specific aspects of the RFI review process, encouraging more detailed studies into particular problem areas.

4.1 Background:

The request for information is an important communication channel and a critical project control tool (Hughes et al., 2013) employed by stakeholders in the architecture, engineering, and construction (AEC) industry to seek clarifications related to project documents. Information requests may cover construction plans, drawings, specifications, and agreements (Abdel-Monem and Hegazy, 2013). Raising project concerns, seeking timely responses, and resolving the issues efficiently through RFIs, may steer the project towards the right direction, increasing the chance of project success (Morales et al. 2022). The construction project knowledge base has described RFI as a 'necessary evil' (Aibinu et al., 2019) because there are always project documents that will still be inaccurate, or ambiguous that require timely

clarification (Jarkas and Bitar, 2012). Failing to address these inaccuracies results in downstream rework or activities that may get on the critical path leading to the risks of schedule delay (Kelly and Ilozor, 2020) and cost overruns (Love et al., 2014). Therefore, there has been a consistent emphasis on reducing the review period and minimising the number of RFIs to mitigate their negative impact (Afzal et al., 2024).

The RFI process has evolved alongside in parallel with the advancements in the construction industry's technological landscape. Initially handled through email or paper-based formats, RFIs are now exchanged using more sophisticated and organized common data environment (CDE) platforms. While these advancements have enhanced the tracking, handling, and management of RFIs, there has not been much effort in generating actionable insights automatically from within RFIs. Previously researchers have manually analysed the RFI content to articulate important questions such as, i) why the issue emerged? and ii) how they could have been efficiently resolved? (Bhat et al. 2017). In this regard, many researchers have exploited manual content analysis to extract non-trivial knowledge, but little has been done to automate the content analysis process. While this approach proves effective in understanding the underlying issues when dealing with a limited number of documents, manually reviewing hundreds of documents remains time-consuming and labour-intensive (Fang et al., 2020). To improve the efficiency of content analysis, a few researchers have devised various classification schemes for RFI categorisation (Kim et al., 2021; Bhat et al., 2017). Such valuable endeavours are mainly focused on documenting insights and lessons learned to facilitate informed decision-making for future projects. However, employing manual content analysis remains still an impractical and inefficient approach for ongoing construction projects. Therefore, there is a need for investigating an automated system that can quickly classify RFIs, generate valuable insights and potentially enhance the existing RFI process.

To bridge this gap, this study develops a domain-specific multiclass text classification model that effectively and efficiently classifies RFIs into major construction project phases; “design”, “execution”, and “procurement”. This automated classification scheme provides a comprehensive view of RFI distribution across project phases, enabling several actionable insights. First, it aids RFI manager in automated screening and routing RFI directly to the concerned party. Then the trend analysis reveals patterns in RFI submissions over time, highlighting peak periods and recurring issues specific to project phases. By automatically identifying common issues within each phase, such as architectural discrepancies in design RFIs, teams can implement targeted improvements to reduce future occurrences. Comparing

response times and resolution rates across phases informs performance benchmarking and process optimisation efforts (Seyis and Özkan, 2024). Predictive analytics based on historical RFI data help anticipate future volumes and types of RFIs, supporting proactive planning and risk management. These insights collectively enhance decision-making, improve project efficiency, and mitigate costs and schedule impacts associated with RFIs in construction projects. The study employs the capabilities of natural language processing (NLP), deep learning (DL) based recurrent neural network (RNN) and traditional machine learning (ML) approaches for automating classification of RFIs obtained from construction projects. Significant contributions of this work include:

- The present research aims to introduce an NLP-driven RFI workflow that will significantly enhance the RFI process both within traditional and CDE-based settings. To achieve this goal, deep learning-based RNNs are compared with traditional machine learning algorithms and thoroughly trained and tested on a dataset of 2,273 RFIs.
- This innovative approach equips project supervisors with the ability to identify and comprehend the issues encountered during a specific project phase. Furthermore, design teams will be empowered to enhance their design practices and construction documents by gaining insights into the nature of the issues within each project phase.
- This study also aims to investigate the influence of different text features and ensemble techniques, which are not commonly used in the construction domain (Zhang et al., 2018). This research aims to contribute to the advancement of knowledge and practices within the construction industry.

This chapter is organized into seven sections. First section provides an overview of manual content analysis practices, followed by a detailed review of features offered by prevalent common data environment solutions available in the market. The second section reviews NLP-driven text classification. The third section, proposed model, details the model for phase-wise separation of RFIs, including NLP pipeline, and methodology. The results and discussion section presents the outcomes of the proposed model and discusses the approaches incorporated. In the experimental evaluation section, algorithmic assessments are validated by comparing performance with domain-specific experts. The following section applies topic modelling to uncover common themes and predominant topics from the RFIs. Finally, the conclusion section summarizes the study's findings and implications.

4.2 Research gap:

To expedite RFI processing, both industry and research entities have implemented various measures. In the industry, various electronic data management systems and common data environment platforms, like Autodesk BIM 360 (now known as Autodesk Construction Cloud) (Jaskula et al., 2023), Aconex (Pradeep et al., 2021), and Procore (Jaskula et al., 2023), are available to support and streamline the RFI process. Table 4.1 provides a comprehensive review of the features offered by the commonly employed CDEs. It's important to note that these CDEs support numerous features and various types of documentation beyond RFIs. However, this study specifically focuses on reviewing features relevant to RFI handling. Although these environments have improved the overall RFI processing and communication, there is no evidence that their deployment on a construction project reduces RFI frequency. Besides, implementing them presents challenges such as data loss and storage problems, complex data architecture, legal challenges, and interoperability issues (Pradeep et al., 2021). From the tabular review, it is evident that these platforms are not fully utilising the vast amounts of data they generate. Integrating AI within these platforms is a research direction still in its infancy (Jaskula et al., 2024). Zawada et al. (2024) emphasise the need to analyse and interpret BIM-based CDE datasets to extract valuable information that aids decision-making and enhances project efficiency. With the goal to leverage construction RFI dataset, this study develops automated text classification model, which can subsequently guide project decision-making and be utilised in future projects.

Compared with industry practices, the body of knowledge has relied on manual text classification to extract patterns, identify errors and issues, and uncover valuable insights and lessons learned from construction RFIs (Bhat et al., 2017). Accordingly, researchers have codified RFIs into distinct categories based on their subjective assessments. Figure 4.1 presents a summary of classification strategies recorded recently in the literature. RFIs have been previously classified based on various criteria, including discipline code, property code, work-element code, issue behind the RFI, and category/type of RFI (Afzal et al., 2024). These classification codes have likely been selected to facilitate a more in-depth analysis of the factors that lead to the generation of RFIs. While manual content analysis can offer value in studying RFI patterns, it also presents various challenges, as outlined below:

- Large volume of data: Extracting insights from large datasets can be resource-intensive (Grimmer and Stewart, 2013), posing challenges for the construction sector dealing with limited resources.

- Subjectivity: Human-driven codification or analysis can be subject to the biases and interpretations of the individuals involved (Downe-Wamboldt, 1992). Without established guidelines or standards, tasks like codifying RFIs can lead to inconsistent results.
- Inaccuracies: Processing textual documents through non-automated means can be prone to errors (Rosenberg et al., 1990) and inconsistencies, especially when handling significant volumes of data.
- Limited scalability and scope: Manual content analysis is typically not scalable (Groen et al., 2018), as it relies on human labour and cannot be easily automated. This can make it difficult to analyse large datasets promptly, resulting in delays in extracting useful information. As a result, manual content analysis becomes impractical for construction projects.

Considering the limitations of human-driven codification, recent research has increasingly focused on leveraging advanced NLP and ML techniques for automated text classification of RFIs. For example, Lee and Yi (2017) utilized various machine learning algorithms, including artificial neural networks (ANN), support vector machines (SVM), k-nearest neighbours (KNN), and naïve Bayes (NB), to conduct pre-bid risk classification of RFIs. They also employed the Latent Dirichlet Allocation (LDA) algorithm for topic modelling of the RFI corpus. Shrestha et al. (2023) classified pre-bid RFIs based on their criticality, applying ANN for this purpose. While these initiatives are at the forefront of analysing and extracting information from RFIs, they do not directly impact the execution of projects. Furthermore, all of these studies focus on the pre-bid phase, whereas RFIs from the delivery phase, entail most of the project issues, drive project communication and contribute towards project success. Furthermore, there is a dearth of studies applying NLP and text mining techniques to construction documentation due to a lack of available datasets (Baek et al., 2021), and this scarcity extends to construction RFIs as well. Hence this research aims to extend the features of existing CDEs for generating more actionable insights in RFI management, and to convert the practice of human-driven RFI codification into an automated.

Table 4-1. Feature review of predominant CDEs utilised by industry stakeholders.

Sr. No.	Feature	Description	CDEs		
			Autodesk BIM-360 | Autodesk Construction Cloud	Procore	Oracle Aconex
1	RFI Creation	Ability to create and submit RFIs within the platform.	✔	✔	✔

2	RFI Tracking	Tools to monitor the status and progress of RFIs from submission to resolution.	✔	✔	✔
3	RFI Assignment	Ability to assign RFIs to specific team members or groups.	✔	✔	✔
4	RFI Response Management	Mechanisms to manage and document responses to RFIs.	✔	✔	✔
5	RFI Notification	Automated notifications and alerts for new RFIs, responses, and status changes.	✔	✔	✔
6	RFI Attachments	Capability to attach documents, drawings, and other files to RFIs.	✔	✔	✔
7	RFI History and Audit Trail	Detailed log of all actions taken on RFIs, including submissions, responses, and status changes.	✔	✔	✔
8	RFI Codes	Issues mentioned in the RFIs (For example, phase, code compliance, specifications etc)	✔	X	✔
9	RFI Analytics	Actionable insights from the unstructured content of RFIs	X*	X	X
10	RFI Workflow Customization	Ability to customize RFI workflows to fit specific project needs.	✔	X	✔
11	BIM-enabled	Integration with Building Information Modelling (BIM)	✔	X	✔
12	RFI Security and Permissions	Control over who can view, create, respond to, and manage RFIs, ensuring sensitive information is protected.	✔	✔	✔
13	Mobile Accessibility	Access to RFI features through mobile apps, allowing team members to manage RFIs on the go.	✔	✔	✔
14	Collaborative RFI Review	Collaborative tools to allow multiple stakeholders to review and comment on RFIs simultaneously.	✔	✔	✔
15		Source of Feature Review	Autodesk Construction Cloud Software	Procore Learning Platform Provided by Procore	Interactive Demo Provided by Oracle Aconex

* Insights can be extracted, but user needs to input all the information. Currently there is no mechanism in Autodesk Construction Cloud to generate insights from inputted unstructured text.

This chapter involves leveraging NLP, deep learning-based recurrent neural networks (RNNs), traditional machine learning algorithms, various feature extraction and ensemble techniques, advancing beyond previous studies that utilised text mining methods. Our model will

automatically categorize RFIs into three project phases: design, execution, and procurement. This approach has been adapted from Kim et al. (2021), who manually classified RFIs based on design, craft/construction, and materials.

Resolving an issue within a particular project phase is crucial; otherwise, it may later translate into another issue/RFI emerging in another project phase. For example, a design issue left unresolved during the design phase has the potential to emerge as a craftsmanship or constructability challenges. Alternatively, it may surface as an out-of-sequence activity during execution phase, thereby impeding project progress. Similarly, imprecise and miscalculated procurement estimations can easily translate into on-site fabrication, assembly, and misalignment errors, leading to variations and change orders. Through phase-wise separation, relevant stakeholders can precisely identify issues within each project phase, enabling them to curb them within the same phase by taking mitigating measures proactively. Additionally, this approach can generate insights and lessons learned from an ongoing project that can later be utilised to minimise the discrepancies in contract documents of future projects. Designers can analyse and incorporate all the designing/detailing mistakes they made in previous projects, thereby improving their skills and expertise. With this classification, contractors or execution teams may know the exact solution of a problem already encountered in previous projects.

4.2.1 NLP-driven text classification with ML and DL algorithms

RFIs are unstructured documents in textual format, and it is necessary to convert them to machine-readable format for further processing. Natural language processing is a field of artificial intelligence that helps analyse, understand, and process human languages, including unstructured text documents and human speech (Hassan and Le, 2021). According to Hassan and Le (2020), NLP is equipped with an efficient mechanism to process unstructured documents with semantic and lexical ambiguities into structured (normalised) data to drive rule-based and machine-learning approaches for text classification. Text classification is either single-labelled or multi-labelled (Cai et al., 2020). Single-labelled text classification has only one label or class assigned to the text statement. In contrast, for multi-labelled text, there is a possibility of two or more labels or classes for a given text statement (Cai et al., 2020). Within single-labelled text classification, if there are two distinct classes in the dataset, then it is a binary text classification. When there are more than two categories or classes, the text classification becomes a multi-class text classification. The present research study categorises the RFI documents into three distinct categories; design, execution, or procurement, which

makes the model under consideration a multi-class text classification model. Building upon this classification approach, this study demonstrates text classification of RFI documents based on supervised machine and deep learning algorithms.

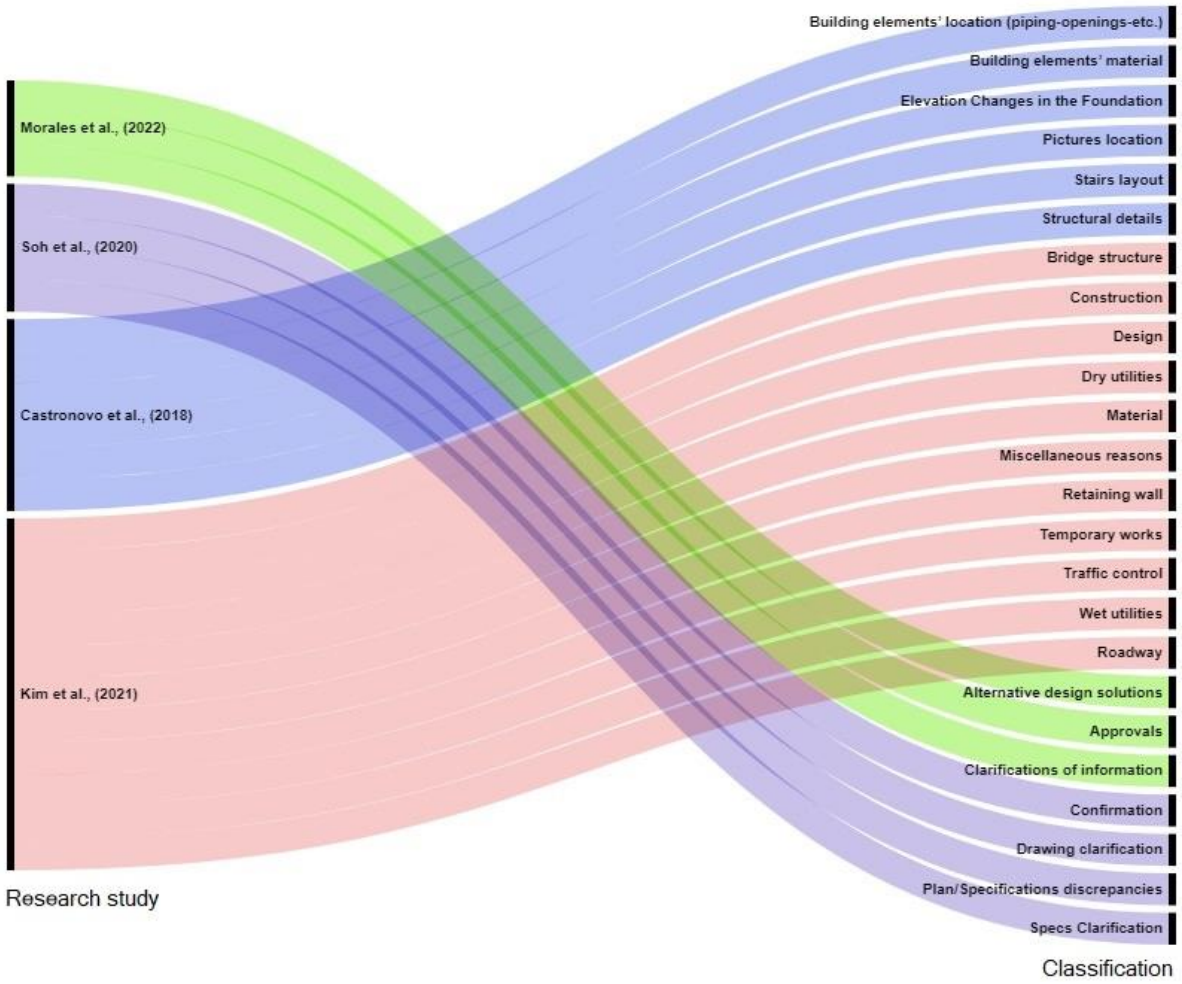


Figure 4-1. Recent human driven RFI classifications recorded in the literature.

The use of supervised learning techniques in text classification is a well-established approach, in which labelled data is used to train the model. Compared to unsupervised learning, supervised learning relies on human codification or labelling of data, making it more efficient (Iqbal et al., 2019). Table 4.2 presents the detailed characteristics and functions of the employed traditional machine learning and deep learning approaches for the phase-wise classification of RFIs. For this research study, traditional machine learning algorithms, including random forest (RF), logistic regression (LR), naïve Bayes (NB), and support vector machine (SVM), were utilised, along with deep learning-based recurrent neural networks (RNNs) such as long short-term memory (LSTM) and gated recurrent unit (GRU).

Traditional ML algorithms and RNN architectures have been selected based on their strengths and applicability to the specific text classification problem. Each machine learning approach offers distinct advantages in handling different aspects of the data. For example, RF, a decision tree-based algorithm, is capable of providing lower bias and variance. It mitigates overfitting issues by employing random split selection (Khalef et al., 2021). In contrast, LR exhibits great advantages over complex models. It does not assume normality in data and is robust, flexible, and easy to interpret (Pohar et al., 2004). Further, SVM is selected for its capacity to deliver superior performance in text classification (Abu Sheika and Inkpen 2010). It creates a hyperplane that maximizes the gap between positive and negative training data (Joachims, 1998). NB is preferred simple, easy to implement and draws better accuracy in large datasets (Vidhya and Aghila, 2010). In short, traditional algorithms are simpler and computationally efficient however may not be as effective as deep learning algorithms in certain scenarios (Sewak et al., 2018). Hence, for a thorough comparison and to determine the optimal approach for testing on the RFI dataset, deep learning based neural network architectures were also employed. Specifically, within the realm of deep learning, this study implemented recurrent neural networks; as GRU and LSTM. They are designed to address the challenges of processing sequential data (Ahmadzadeh et al., 2021). GRU and LSTM excel at modelling long-range dependencies and mitigating the vanishing gradient problem in RNNs (Noh, 2021), making them suitable choices for accurately capturing the sequential information present in textual documents.

The implementation of these algorithms was carried out using Python, as the primary programming environment. A wide range of open-source libraries was utilized to support various stages of the machine learning pipeline, including Scikit-learn, TensorFlow, and Gensim. Additional libraries such as NLTK, Pandas, NumPy, and Matplotlib supported pre-processing, data handling, and visualization tasks.

Table 4-2. Comparison of characteristics: supervised machine learning algorithms and deep learning based neural networks

Machine Learning Algorithm	**Technical Description for Text Classification**	**Advantages**	**Disadvantages**
Random Forest (RF)	RF constructs classification trees based on different features and aggregates their outputs (Wang, 2012) to classify documents.	1. Handles high-dimensional data effectively. 2. Reduces overfitting by averaging predictions from multiple trees.	1. Requires careful parameter tuning. 2. Less interpretable compared to linear models.

Logistic Regression (LR)	Logistic regression, a probabilistic classifier, relates categorical dependent variables to independent variables, providing a probability distribution over class labels for each test sample (Hosmer and Lemeshow, 2000).	1. Simplicity in implementation. 2. Efficient in forecasting categorical outcomes.	1. Assumes a linear relationship between features and target, which may not hold in complex text data 2. Considered sensitive to outliers.
Support Vector Machine (SVM)	In text classification, SVM converts text features into a high-dimensional feature space (Shin et al., 2000) to categorise them.	1. Effective in high-dimensional spaces. 2. Capable of dealing with both linear and non-linear data.	1. Demands optimisation of hyperparameters. 2. Can be time-consuming for large datasets.
Naïve Bayes (NB)	NB calculates the likelihood of a document belonging to a specific class based on its features (Hassan and Le, 2021).	1. Fast training and prediction. 2. Requires a relatively small amount of training data.	1. Exhibits limitations with highly correlated features.
Gated Recurrent Unit (GRU)	GRU employs gating mechanisms to control the flow of information. It is equipped to understand better connections between words or phrases that are in distance to each other (Zulqarnain et al., 2019).	1. Computationally efficient and less prone to overfitting 2. Handles long-term dependencies while mitigating the vanishing gradient problem.	1. May not capture as complex relationships as LSTM. 2. Limited memory compared to LSTM.
Long Short-Term Memory (LSTM)	LSTM is a type of RNN with memory cells and gating mechanisms to learn and remember information over long pieces of text. It can also handle a problem where the network either forgets or gets too overwhelmed with information.	1. Captures long-term dependencies more effectively than traditional RNNs. 2. Mitigates vanishing/exploding gradient problem.	1. Computationally expensive compared to simpler RNN architectures. 2. May overfit on smaller datasets if not properly regularised.

4.3 Proposed model for phase-wise separation of RFIs

Key literature distinguishes project lifecycle into various stages such as feasibility, design, execution (Alexander et al., 2019), procurement and operations and management (Liu et al., 2017). RFI dissemination starts at pre-bidding, however, for this research, the RFI dataset obtained did not have any RFI from bidding or pre-bidding; hence, all the gathered RFIs belong to design, execution or procurement. Another assumption made for this research is that each RFI statement is single disciplinary, meaning that they belong to only one of the three project phases: design, execution, and procurement. The current research endeavour employed various NLP techniques, such as text pre-processing and feature extraction methods, so that RFI queries could be utilised as input for machine learning algorithms. Six commonly used machine learning approaches including the traditional algorithms and neural networks were implemented to develop the RFI classification model. Subsequently, these models were subjected to ensemble learning techniques, with the aim to observe improvements in the overall

classification performance of the RFI classification model. Fig. 4.2 shows the workflow of the classification model leading to development of best performing model. Additionally, an experimental evaluation was conducted to compare the performance metrics of the best-performing model with human participants, leading towards a more informed AI-

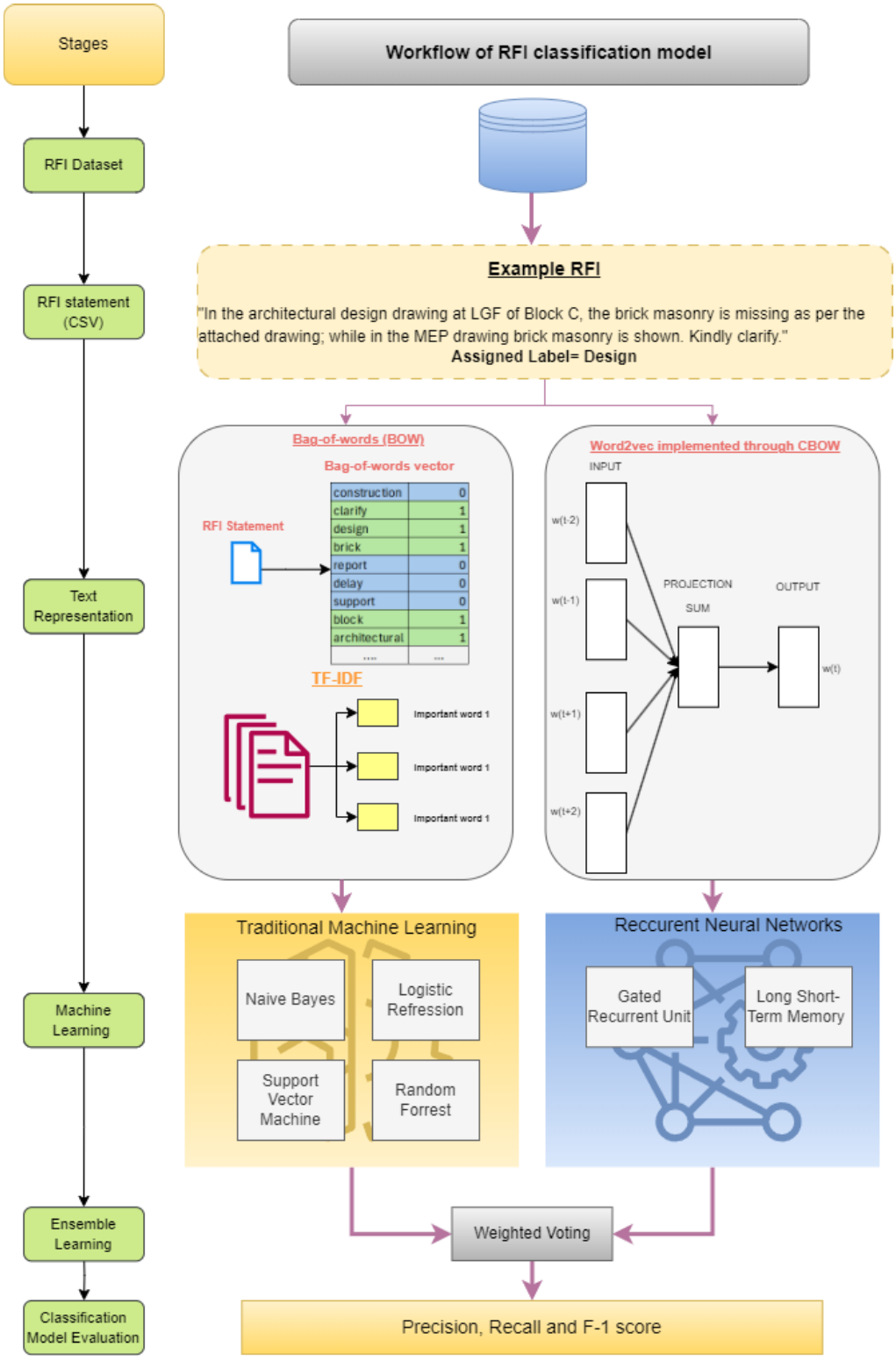


Figure 4-2. Workflow of the multiclass text classification model using machine and deep learning.

driven workflow to improve information flow, and review process, while promoting knowledge mining for future projects. This research advances the RFI management process by directly improving and impacting it, providing a mechanism that can be utilised by researchers to extract information in an automated manner and a tangible solution that can be employed by industry stakeholders for efficient decision making. The following sections discuss into the proposed methodology in detail.

4.3.1 Dataset and text pre-processing

For this research, a labelled dataset of 2,273 requests for information statements from eight different construction projects executed in the Middle East (Qatar and United Arab Emirates) was curated. These statements were manually assigned a project phase based on the challenge they were presenting. During design phase, many decisions are made related to development and coordination of plans, drawings, specifications, and layout of the project. Hence, RFIs that sought these clarifications or confirmations were labelled as 'design'. Execution phase is when the construction of the project is carried out. RFIs concerning unexpected site conditions, site preparation, workmanship, and any design change directing site work, were assigned 'execution' category. Procurement phase involves acquiring materials, equipment and services required for the project. Hence, RFIs highlighting issues like rectifying a mistake in bill of quantity, negotiating prices and delivery schedules or any other problems that disrupt on-site supply chain of materials were labelled as 'procurement'. The final dataset for this study included 1,147 RFI statements related to design, 599 pertaining to execution and 527 categorised as procurement. In the next phase, text pre-processing techniques were applied to convert the RFI queries to machine-readable format. The steps for text processing as described in detail below:

- Lowercase and punctuation removal: Initially, the dataset was transformed by converting all text to lowercase, reducing word variation, while also removing punctuation marks that do not contribute to text semantics.
- Stop words removal: In the model development, all the English-language stop words (such as "the", "etc", and "for") were removed to ensure the reduction in noise and improve the models' computational efficiency.
- Tokenisation: Tokenization involves splitting a character sequence, comprising words, punctuations, or symbols, into tokens, as demonstrated by the segmentation of the RFI

statement "water lifting chamber detail, and IFC drawings are missing" into individual tokens ['water', 'lifting', 'chamber', 'detail', ',', ‘and’, 'IFC', 'drawings', 'are', 'missing', '.'].

- Lemmatisation: Lemmatisation is a technique applied to all the words in corpus to reduce them to their base forms. For instance, "inspecting", "inspected", and "inspection" were reduced to "inspect".
- Parts-of-speech (POS) Tagging: In this study, POS tagging was employed to discern the grammatical structure and word roles within the corpus, identifying verbs like "refer" and "clarify" tagged as VB, and nouns like "drawing" and "offset" tagged as NN.

4.3.2 Text representation

For enabling ML algorithms to decode unstructured text, conversion of textual data into vectors representing the most relevant features is standard practice. In this context, a feature refers to an attribute or word within a textual statement that embodies its semantic meaning (Hassan and Le, 2020). This study implemented three distinct approaches to construct representation. By utilising these three approaches, this study aimed to create effective representation vectors for machine learning algorithms, enabling them to analyse and interpret unstructured textual data. These feature extraction methods are discussed below:

- **Bag-of-Words (BoW) Model**: BoW model is a widely utilised NLP technique to represent textual statements as numerical value features (Baker et al., 2020). In the context of text classification, a Bag of Words model records the frequency of each bag, created for each word or type, without taking into consideration the grammatical structure or sequence of the words (Qader et al., 2019). The BoW model operates by considering all the statements in the corpus, transforming them into numeric vectors. In this process, each element in the vector corresponds to a word within the corpus, effectively capturing the textual information in a numerical representation. The BoW model assumes that the frequency of occurrence of words in the text document carries important information about the document's content (Kontoghiorghes and Colubi, 2023).
- **Term frequency – invert document frequency (TF-IDF):** TF-IDF is a statistical technique capable of evaluating the importance of words within a document (Zhou, 2022). It is a feature weighting and representation method that offers improvements over BoW method. Unlike BoW, TF-IDF accounts for the importance of words in a document relative to their occurrence across the entire corpus. This is achieved by using term frequency (TF) parameter, which assigns the weights to individual words based within a document, and

inverse document frequency (IDF) parameter, which quantifies the rarity of features across the corpus (Hassan and Le, 2021). Thus, this weighting approach helps identify words with higher informational value and distinctiveness within a text document. Eq. (4.1) was employed in this study to calculate the TF-IDF scores.

$$\mathrm{TF}-\mathrm{IDF}=\frac{\mathrm{n}_t}{\mathrm{N}}\times\left(1+\log\frac{\mathrm{F}}{\mathrm{F}_t}\right) \qquad (4.1)$$

Where n_t = number of occurrences of a term *t* in a document; N= total number of words in the document; F=total number of documents; and F_t = number of documents that include the term *t*.

- **Word2Vec Model:** Besides the BoW and TF-IDF methods, this research study also employed the Word2Vec model for feature representation for recurrent neural networks. The Word2Vec model, based on artificial neural networks, generates multidimensional vectors that can capture the semantic meaning of each unique word in the entire corpus (Kim and Chi, 2019). Unlike the BoW representation, where a single vector represents each statement, the Word2Vec representation assigns a vector to each word (Fahad and Le, 2020). As a result, a text statement is transformed into an array of word vectors, effectively capturing its meaning as input for machine learning algorithms. In this study, the Continuous Bag-of-Words (CBOW) algorithm was utilised for Word2Vec representation, wherein the prediction of the current word is based on the input of surrounding word vectors.

4.3.3 Model training with traditional ML and RNNs

After the feature extraction of the RFI statements, a diverse set of four supervised machine learning algorithms, namely naive Bayes, support vector machine, logistic regression, and random forest, were implemented to develop an RFI classification model. It is to be noted that the performance of these traditional ML algorithms is mainly dependent on dataset quality and problem domain (Salama and El-Gohary, 2013; Nigam et al., 1999). For traditional machine learning algorithms, this research study considered a k-fold cross-validation approach to minimise biased estimates (Hassan and Le, 2021). In this method, the given dataset is divided into k equal-sized subsets.

Consequently, the model is iteratively constructed k times, wherein one subset is designated as the test set, while the remaining subsets are combined to form the training set (Jayashree and

Srikanta, 2011). This continues until each k subset has been consecutively employed as the test set over k iterations. The final assessment of the model's performance is evaluated by calculating the average of the performance metrics obtained from all k subsets. Considering its proven efficacy in prior investigations (Salama and El-Gohary, 2013), a k-value of 10 was judiciously chosen for this study. In the context of each k-fold iteration, the dataset consisting of 2,273 RFI statements was randomly split into two separate parts. A total of 2,045 RFI statements (90%) were utilised for training purposes, whereas the remaining 227 RFIs (10%) were set aside for testing of the developed models. This criterion was adapted from previously published research study (Hassan and Le, 2021), enabling the models to learn from a substantial amount of data.

For deep learning-based approach, the methodology utilised the abilities of recurrent neural networks. Specifically, two types of RNN architectures, LSTM and GRU, were implemented and compared with other traditional algorithms. These RNN architectures were chosen due to their capability of capturing sequential information and handling long-range dependencies within the input data (Birnbaum et al., 2019). The training process of the RNN models began with the preparation of the input data. The RFIs, represented as sequences of words, were encoded as word embedding vectors. Unlike traditional algorithms, no stop words removal or feature selection was applied to retain the order of the words in the input. The RFIs were converted into a matrix format, where each row represented a word embedding vector and each column represented a specific feature. The matrix had a predefined maximum length, and if an RFI was shorter than this maximum length, the remaining rows were filled with zero vectors.

Once the input data was prepared, the training architecture (Fig. 4.3) of the RNN models was implemented. The LSTM and GRU architectures followed a similar structure. The input matrix was fed into the RNN layer, which processed the words of the RFI sequentially. The RNN layer updated its hidden state at each step, considering both the current word embedding vector and the accumulated knowledge from previous hidden states (Lai et al., 2015). This allowed the model to capture the contextual information, and long-term dependencies present in the RFIs. The output of the RNN layer was a sequence of hidden states, which represented the encoded information of the input (Hassan and Le, 2021). To predict the category of an RFI, the final hidden state from the RNN layer was passed through a softmax function (Yoon et al., 2018), which produced a probability distribution over different category. The category with the highest probability was selected as the predicted label for the RFI.

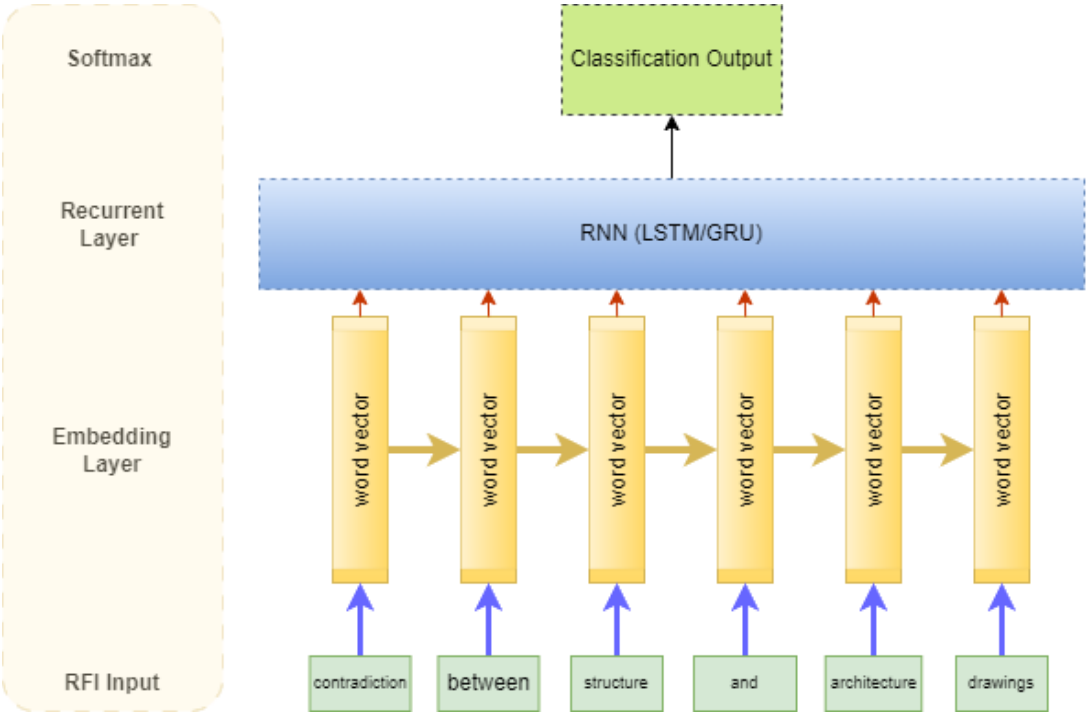


Figure 4-3. Architecture of RNN model for text classification, adapted from Hassan and Le, (2021)

4.3.4 Ensemble models

The present study further examined the impact of two ensemble learning techniques: boosting and voting, to enhance the final model's robustness (Bühlmann, 2012). Boosting is an ensemble technique combining weak models' learnings to develop a strong learner. The approach relies on iterative training, whereby each new model is provided with the data and allowed to make misclassifications. However, the sequential iterations leverage the misclassifications performed by the former, enabling them to make better predictions. Through this iterative process, it is anticipated that each subsequent weak learner will demonstrate incremental improvements in effectively handling challenging instances compared to their predecessors (Freund, 1995).

In particular, this study employed Adaptive Boosting (AdaBoost) technique which improves the boosting algorithm by focusing on difficult patterns (Onan et al., 2016). AdaBoost can adjust the weights assigned to training instances based on the prior classification errors. Each weak model is trained on a modified version of training data, increasing the misclassification weights (Sun et al., 2016). This way subsequent model pays more attention to the previously misclassified samples, learning to become a more efficient ensemble model.

Furthermore, the study also employed the voting algorithm to synthesise the outcomes of the traditional ML algorithms and RNNs. In this regard, two voting techniques; simple majority

voting and weighted majority voting, were utilised. In simple voting, all votes from all the classifiers carry equal weight (Hassan and Le, 2021). Conversely, in weighted voting, weights are assigned to the votes depending on the F-scores of the algorithm. By assigning weights considering the F-scores, it is ensured that the significance of predictions made by more competent classifiers is heightened, thereby enhancing the classification performance of the voting algorithm (Bouziane et al., 2011).

4.3.5 Classification performance evaluation

To measure the performance of the developed models, the present research study incorporated different metrics such as precision, recall, accuracy, and F-1 score (Priyadarshini et al., 2022). These metrics are calculated through equations (4.2)-(4.5). The formulae for these metrics indicate four important factors: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Precision represents the percentage of instances correctly labelled as design out of all instances the model identified as design. Recall indicates the percentage of correctly labelled design RFIs compared to all RFIs that are design related. Since precision and recall are noted as dual performance metrics (Buckland and Gey 1994), F-1 score is considered as a combined factor for both recall and precision to evaluate the effectiveness of the model (Abu Sheikha and Inkpen 2010). Lastly, accuracy represents the overall correctness of the model, across all the classes, including both positive and negative instances.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{4.2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4.3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{4.4}$$

$$\text{FI} - \text{score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}\times\text{Recall}} \tag{4.5}$$

4.4 Results and discussion

4.4.1 Classification performance of traditional machine learning approach

In order to develop a high-performing RFI classification model, traditional machine learning algorithms (LR, SVM, RF, and NB) were evaluated using pre-processed RFI query statements. Table 4.3 compares the performances of these algorithms on three class labels (design,

execution, and procurement) using two feature extraction techniques (Bag-of-Words and TF-IDF). Both feature extraction techniques almost yielded similar results. Table 4.3 shows in both techniques, logistic regression delivered promising results with the F-1 scores of 77% on design, 57% on execution, 68% on procurement for BoW and 76% on design, 57% on execution, and 69% on Procurement for TF-IDF method. SVM exhibited comparable performance on par with logistic regression. RF demonstrated strengths in capturing positive instances, particularly in the design class (recall = 95% and 96%, for Bow and TF-IDF respectively, but delivered poor results for execution class. The performance of NB algorithm was suboptimal, with the lowest F-1 score of 69% for design, 52% for execution and 62% for procurement on BoW method. The choice of feature representation technique did not consistently favour one over the other and the findings underscore the importance of considering class-specific characteristics and algorithmic suitability in multi-class classification tasks, as highlighted by variations in performance across different algorithms and class labels. While LR and SVM exhibited robustness, RF and NB faced challenges in specific class contexts. Poor performance on the 'execution' class may be attributed to class imbalance, semantic overlap with other classes, and higher linguistic variability, making it difficult for traditional models to classify accurately. These insights advocate for a context-dependent approach when designing machine learning models for multi-class classification scenarios. Moving forward, the overall performance metrics of the traditional machine learning algorithms were compared with sophisticated recurrent neural network approaches in the next step.

Table 4-3. Performance of machine learning algorithms using different feature representation techniques.

		Bag-of-Words			TF-IDF		
Traditional Machine Learning algorithms	Class label	Precision (%)	Recall (%)	F-1 Score (%)	Precision (%)	Recall (%)	F-1 Score (%)
LR	Design	71	85	77	71	84	76
	Execution	65	53	57	60	56	57
	Procurement	74	65	68	77	64	69
SVM	Design	69	83	75	69	86	76
	Execution	61	49	54	64	53	58
	Procurement	73	59	65	78	64	70
RF	Design	66	95	76	66	96	76
	Execution	71	36	46	72	39	49
	Procurement	77	55	63	83	51	62
NB	Design	68	73	69	64	81	70
	Execution	55	51	52	61	43	49
	Procurement	61	64	62	66	56	60

4.4.2 Evaluation of machine learning approach versus deep learning algorithms

With the detailed class-wise evaluation, the performance of traditional classification algorithms was compared with that of recurrent neural networks. Table 4.4, presents the overall precision, recall and F-1 scores for traditional machine learning algorithms and RNNs. In this regard, feature extraction techniques such as BoW, and TF-IDF were utilised for traditional machine learning algorithms, while Word2Vec was applied to establish word embeddings for RNNs. As shown in Table 4.4, both RNNs (LSTM and GRU) outperformed traditional algorithms in terms of their F-scores. The GRU model achieved high precision (86%), recall (85%), and F-1 score (84%), while the LSTM model showed slightly lower scores (82% precision, 82% recall, and 81% F-1 score). Among traditional algorithms, LR and SVM had highest F-scores. Additionally, the TF-IDF feature extraction delivered better results than the BoW method. Both LR and SVM attained 68% in F-score when passed through TF-IDF. Same algorithms achieved F-scores of 67% and 65% through BoW method, respectively. The high performance of RNNs can be attributed to using Word2Vec-based word embeddings. These embeddings can capture the dependencies and context between words, which is crucial for accurately classifying unstructured documents like RFIs (Hassan and Le, 2021).

Table 4-4. Comparison of performance for machine learning and deep learning algorithms

	Bag-of-words			TF-IDF			Word2Vec		
Machine Learning algorithms	Precision (%)	Recall (%)	F-1 score (%)	Precision (%)	Recall (%)	F-1 score (%)	Precision (%)	Recall (%)	F-1 score (%)
LR	70	68	67	69	68	68	–	–	–
SVM	68	64	65	70	68	68	–	–	–
RF	71	62	61	74	62	62	–	–	–
NB	61	63	61	64	60	60	–	–	–
RNN (GRU)	–	–	–	–	–	–	86	85	84
RNN (LSTM)	–	–	–	–	–	–	82	82	81

4.4.3 Performance evaluation of ensemble classifiers

This research further explored the performance of two ensemble techniques: voting (simple majority and weighted majority) and boosting (specifically Ada Boosting). Table 4.5 illustrates the performance of these ensemble classifiers in comparison with the best-performing traditional machine learning algorithm (LR) and recurrent neural network (GRU). Here AdaBoost, in combination with LR (best performing traditional algorithm), aims to improve

the previously achieved performance of LR, however, both F-score and recall decreased when AdaBoost combined with LR model. One possible explanation for this can be the ensemble model becoming more complex, eventually overfitting the training data. Furthermore, it is noteworthy that GRU attained better results than AdaBoost ensemble and simple majority voting methods. The weighted majority voting model outperformed the RNN models by achieving the highest precision, recall and F-score of 89%, 85% and 86%, respectively. The high performance exhibited by weighted majority voting model can be attributed to the combination of diverse models (Onan et al., 2016).

Table 4-5. Performance comparison of the ensemble models and best individual models

Machine learning algorithms	Precision (%)	Recall (%)	F-1 Score (%)
Weighted majority voting	89	85	86
Simple majority voting	78	73	74
AdaBoost + LR	72	59	59
LR	68	68	68
RNN-GRU	86	85	84

4.5 Experimental evaluation

4.5.1 RFI annotation exercise

An experiment was conducted to assess the performance of the developed model through an RFI annotation exercise. To conduct this evaluation, a labelling exercise comprising 40 RFI statements was designed in Qualtrics® platform from the RFIs obtained from an ongoing project. As part of the evaluation, the model's output was benchmarked against human-driven classification, wherein participants manually assigned each RFI to its respective project phase based on their domain expertise and contextual understanding of construction workflows. This RFI annotation exercise—referred to as human-driven classification—served as the baseline for comparison, allowing for an assessment of the model's accuracy, consistency, and efficiency in replicating expert judgement. The initial labelling of the RFI statements was carried out by one of the authors, followed by a review and validation process conducted by industry professionals with relevant domain expertise. No discrepancies were identified during the review, thereby affirming the accuracy and reliability of the assigned labels. Within this dataset, 21 RFIs belonged to the design phase, while a total of 10 and 9 RFIs belonged to execution and procurement phases, respectively. This distribution was intentionally chosen to

ensure adequate coverage of RFIs across different project phases, allowing for a comprehensive assessment of the model's performance throughout the project lifecycle.

An introductory section outlining the study objectives and providing information about the different project phases was established to guide participants in accurately annotating the RFI statements. Subsequently, three participants with diverse experience and expertise in AEC research and industry were chosen to participate in the evaluation. Three experienced professionals were sufficient to ensure expert, consistent annotations without introducing variability. Similarly, 40 RFIs were selected to provide a representative yet manageable dataset across key project phases, allowing meaningful model comparison while avoiding participant fatigue. This RFI annotation exercise was forwarded to the participants through email, and the time taken, and the accurate answers provided by the participants were recorded by the Qualtrics® automatically. The annotation performance of the participants was measured and subsequently compared to the machine learning-based classification model.

4.5.2 Results of experimental study

The metrics employed in this experimental study included precision, recall, accuracy, and processing time. Among these parameters, processing time and recall hold utmost importance as they are critical in ensuring accurate and timely annotation of RFIs. The ultimate goal of this pilot study is to alleviate industry professionals from the arduous task of manual content analysis by achieving a recall value of 100% (Hassan and Le, 2021) and completing the annotation process within the shortest possible timeframe. For this experiment best-performing ensemble model within the previously developed models, i.e., weighted majority voting, was compared against the performance of human participants. The results of the conducted annotation experiment are presented in Fig. 4.4. On an unseen dataset, the weighted majority voting model demonstrated an effective 80% recall rate establishing itself as a highly promising foundation for the automated phase-wise separation of RFIs.

Conversely, the participants achieved a relatively lower mean recall of 54%. Remarkably, the model showcased superior performance in terms of accuracy and precision, attaining scores of 83% and 85%, respectively. In comparison, manual content analysis yielded 55.8% and 59% mean accuracy and precision scores, respectively. The precision for human-driven classification was calculated for each of the three participants, and the overall mean precision was derived by averaging their individual scores across all phases. Furthermore, the time taken to classify the RFIs was a critical factor examined in both experimental settings. The ensemble

model surpassed the human participants by completing the automated classification within 47.2 seconds. In contrast, human participants took an average of 13 minutes to complete the task. The results of this experiment lead to the development of a robust solution that addresses the limitations of manual content analysis of RFIs, discussed earlier. The results demonstrate the efficiency of the machine learning model in handling large volumes of data, presenting minimal bias and fewer errors. Additionally, the model exhibits enhanced scalability and a broader scope of implementation. This validation resulted in the development of an improved RFI management workflow and a graphical interface designed for stakeholders, which will be detailed in the subsequent section.

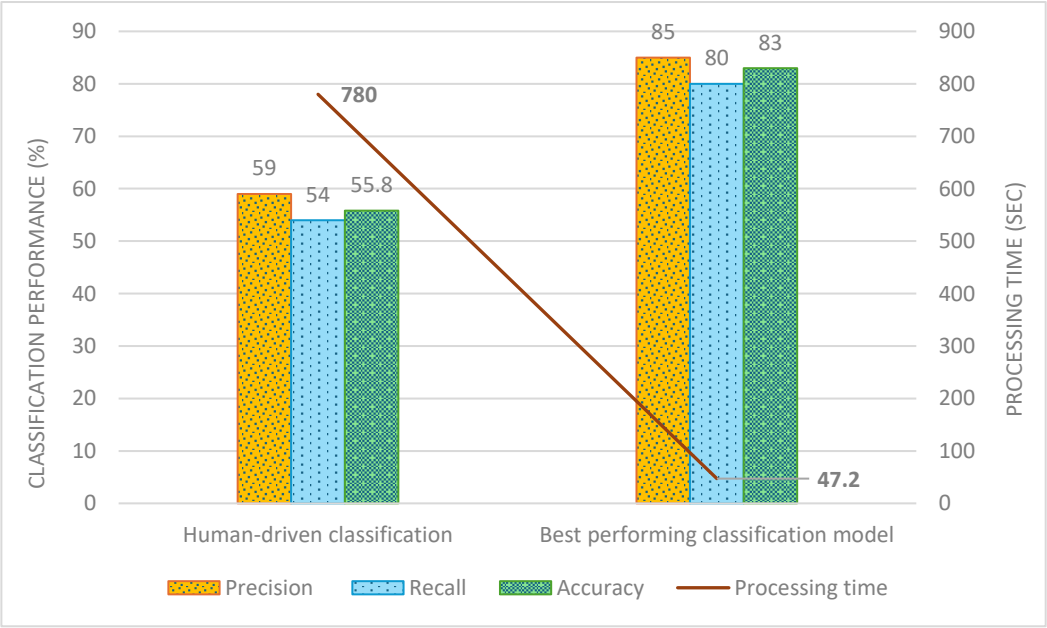


Figure 4-4. Performance comparison between human-driven and best performing RFI classification model.

4.6 Analysing construction issues in RFI documents with text mining and visualisation

This section discusses an approach to discover and visualise to extract patterns and themes related to issues mentioned in RFIs. To achieve this, unsupervised learning based-Latent Dirichlet Allocation algorithm was applied to the RFI dataset. The same RFIs collected for phase-wise separation were also utilized for this NLP application. Pre-processing steps, including lowercasing, punctuation removal, stop word elimination, numeric digit removal, tokenization, and lemmatization, were carried out to ensure that meaningful information and patterns were extracted, while minimizing irrelevant or confounding data.

4.6.1 Insights into prominent topics and keywords through topic modelling

LDA algorithm views each RFI document as a combination of different topics. The present research implements topic modelling through LDA on novel dataset of RFI through unsupervised approach. Topic modelling, an NLP application, helps identify clusters or groups of similar words within the given corpus. This implementation was performed with the aid of Python-programming language, with the help of different libraries such as Gensim, Natural Language Toolkit (NLTK) and pyLDAvis (Fig. 4.5) for the visualisation of the results. Through different iterations the final number of the topics chosen for this research was 10. Each topic produced was a combination of different keywords from each RFI document. Fig. 4.6 represents the top 10 keywords within each topic with their weightage. Usually the overarching theme/topic of the words is the representation of top 5-10 words within each category. It must be noted that the suggested topic is provided with context of the RFIs. A detailed description of the topics and their semantic representation is provided below:

- Topic 1 - Structural Discrepancies: This topic concerns structural discrepancies that can arise during construction, including issues with beams, elevations, grids, columns, ducts, and steel. It is necessary for the contractors to rapidly identify these problems and address them promptly to ensure that the building's structure is sound.
- Topic 2 - Construction Approval: Obtaining approval for a construction project can be a lengthy and complicated process. This topic focuses on the steps involved in getting approval, including submitting requests, and proposals, and working with designers and other stakeholders to gain approval concerning any project activity or scope. It also involves managing the work process of construction projects, ensuring that everything runs smoothly and is completed on time.
- Topic 3 - Coordinating Construction Systems: To ensure that a building functions correctly, it's essential to coordinate various construction systems, such as water, power, and drainage. This topic covers the provision and coordination of these systems, including managing rooms, concrete, water, and other essential components.
- Topic 4 - Electrical Installation: This topic focuses on electrical installations, including designing them, choosing the suitable cable, determining cable length, and ensuring that the installation functions as intended. Contractors must also consider lighting, space, and waterproofing to ensure that electrical installations meet the building's needs.
- Topic 5 - Building Fixtures: Building fixtures, such as doors, cabinets, and gates, play a crucial role in making a building functional and attractive. This topic covers different types of fixtures and their compliance with regulations such as height, layer, and others.

- Topic 6 - Construction Drawings: Accurate construction drawings are essential to ensure a building is constructed as intended. This topic focuses on creating and managing construction drawings, including floor plans, areas, and sections.
- Topic 7 - Structural Stability: This topic focuses on ensuring a building is structurally stable, including dealing with issues such as the basement, frame, and rebar. It also involves managing buildings' ventilation and air conditioning systems, such as air vents and channels.
- Topic 8 - Building Specifications: Accurate building specifications are crucial to ensure a building is safe and meets the requirements. This topic focuses on load capacity, RFIs and ensuring the specifications meet regulatory requirements.
- Topic 9 - Building Maintenance and Renovation: This topic focuses on maintaining and renovating buildings, including managing systems such as water stops, chambers, and switches.
- Topic 10 - Procurement Management: This topic focuses on managing procurement processes, including contracts, materials, and schedules. It also involves managing requirements, orders, and documentation to ensure the procurement process runs smoothly.

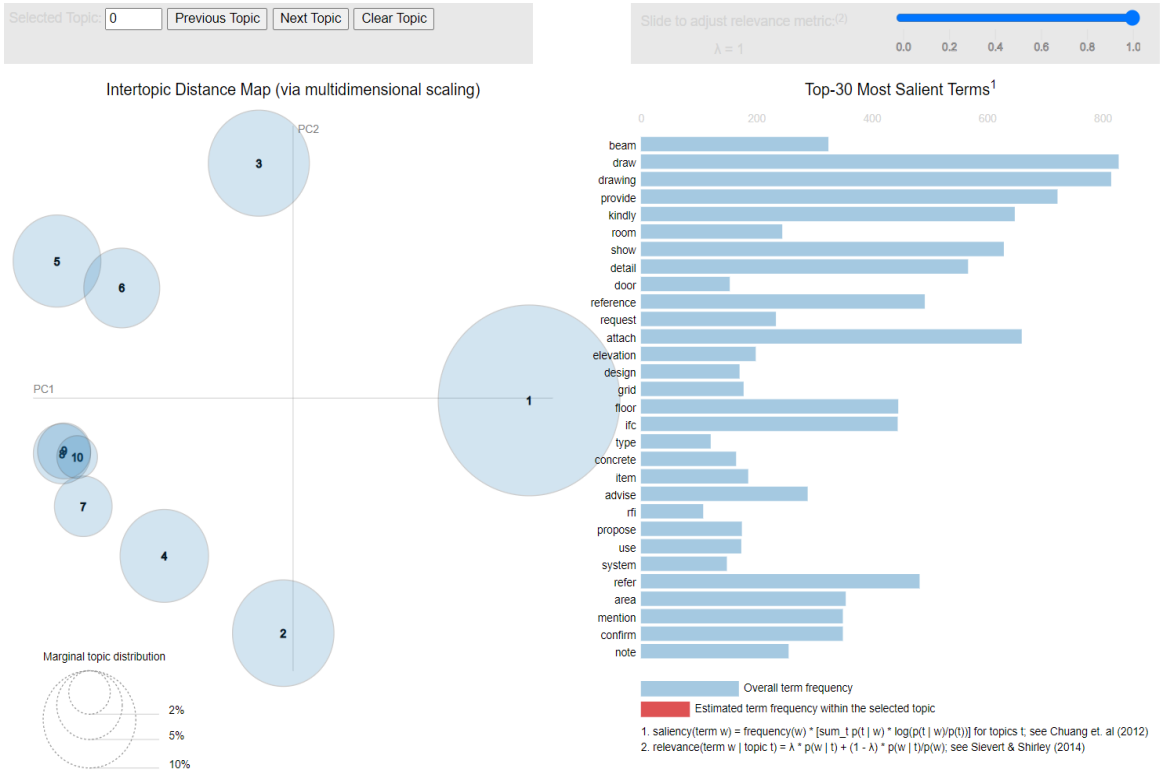


Figure 4-5. Representation of topics identified by LDA on construction RFIs.

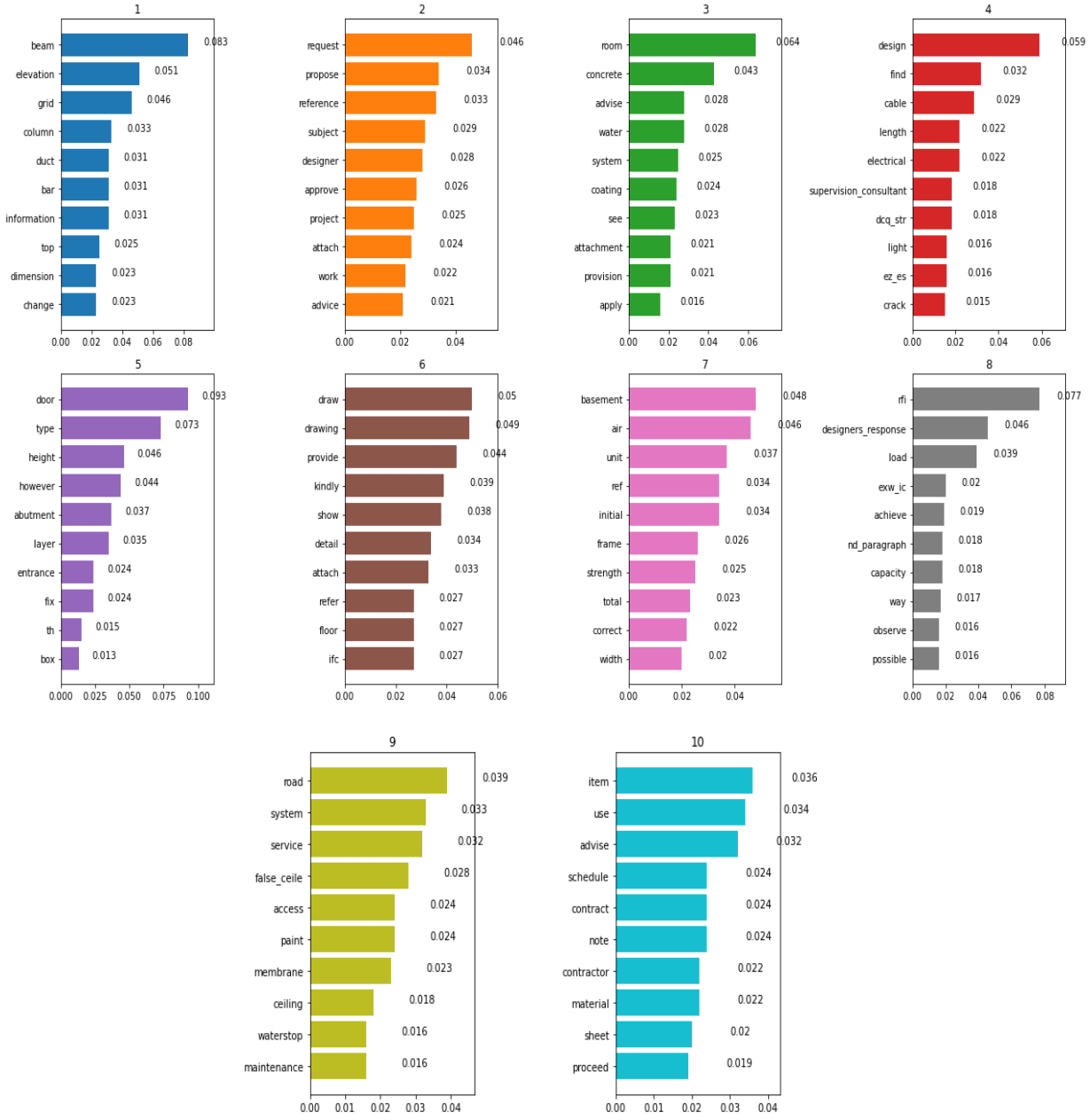


Figure 4-6. Results of the topic modelling of the RFIs

4.6.2 Topic clustering and word cloud visualisation

Subsequently this research implemented TSNE (t-Distributed Stochastic Neighbor Embedding) algorithm, for visualising clusters of 10-topics derived from within the corpus of RFI documents. This technique works by reducing the dimensionality of high-dimensional data, allowing for more effective visualisation of the underlying patterns and structures. Based on the scattered plot (Fig. 4.7) developed through the TSNE algorithm, 10 topics were identified and their respective clusters within the corpus of RFI documents.

One notable finding was that Topic 6 (brown), which pertains to "Construction Drawings", was spread all over the plot, indicating that it shared some overlap with other topics such as Topic

2 - "Construction Approval", Topic 7 – "Structural Stability", and Topic 9 – "Building Maintenance and Renovation". This insight could potentially suggest that drawing specifications are a critical component that overlaps with multiple aspects of construction projects, including approval processes, structural stability, and maintenance and renovation.

In addition to the previous observation, it is worth noting that topic 2 (Construction Approval), topic 3 (Coordinating Construction Systems), and topic 5 (Building Fixtures) appear to be clustered closely together in the scatter plot. This suggests that these topics may share similar underlying patterns and relationships within the RFI corpus. Identifying these clusters and relationships can provide valuable insights into potential areas for improvement in the construction process, such as streamlining construction approval procedures or optimising the coordination of construction systems and building fixtures. It must be noted here, that the TSNE algorithm is a non-linear dimensionality reduction technique which means that it may not maintain all the information present in the original high dimension data. While it has efficiently identified the clusters and patterns from the RFI corpus, some details may be lost in dimensionality reduction process.

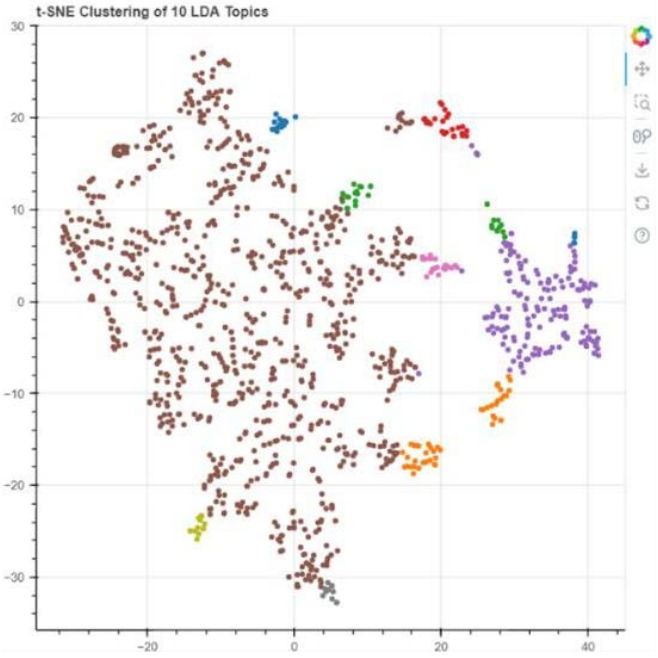


Figure 4-7. Topic clusters using TSNE algorithm.

Next, word cloud was generated to identify key themes and topics in a text corpus. By displaying frequently occurring words in larger font sizes, word clouds allow for a quick and easy analysis of the most important keywords in a dataset. In this study, a word cloud (Fig. 4.8) is utilised to gain important insights into a construction project based on the keywords of the

requests for information. The word cloud shows that the words "attach" and "advise" are the most common in the RFI keywords, highlighting the importance of providing additional attachments and seeking advice when making decisions. The presence of words such as "elevation," "dimension," "grid," "reference," and "design" suggests discrepancies in the construction drawings. Additionally, the appearance of words like "beam," "bar," "coating," "column", "concreting," "electrical," "duct," "cable," and "water" indicates discussions around various building components. Resolving issues related to these components is critical as failure to do so can have a detrimental effect on the project. By analysing this information, construction teams can identify problematic areas of their project and improve the quality of their construction drawings to minimise RFIs. This insight can also be used to inform future projects and improve the overall quality of construction processes.

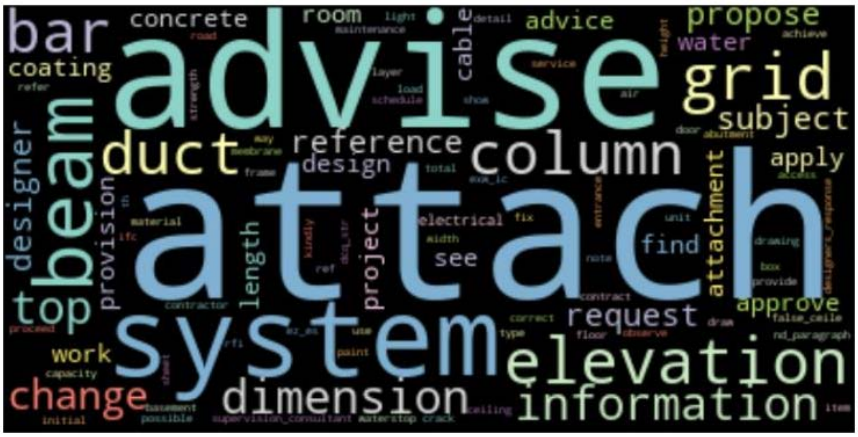


Figure 4-8. Developed word cloud from RFI dataset.

4.7 Summary of the chapter

This study examined the abilities of supervised machine learning and NLP techniques to develop an automated model for efficient phase-wise classification of RFIs. The developed model serves as pilot research for reliable identification of issues and assigning them to their corresponding design, execution, and procurement phases.

The training of these models was conducted on a novel dataset of 2,273 RFI documents that were collected from construction projects. Four traditional machine learning algorithms (LR, NB, RF, and SVM) were implemented to determine their performance on this classification task. Both BoW and TF-IDF were techniques utilised for feature representation, and both showed almost slightly better results in terms of performance metrics. Furthermore, recurrent neural networks GRU and LSTM were also tested, and both showed superior results on performance metrics as compared to the traditional machine learning algorithms. This study

also explored ensemble models, including voting, and boosting. Compared with the best-performing machine learning model (GRU), weighted majority voting ensemble model displayed superior performance. The performance of this ensemble model was compared with that of human participants in an RFI annotation experiment. The model demonstrated superior results in terms of recall, precision, F score, and processing time. Finally, the research concluded by successfully identifying and visualising the prevalent topics and themes discussed in RFIs, through the application of topic modelling, enhancing our understanding of RFI issues and providing inspiration for future research.

While this research explored two distinct applications of NLP—topic modelling and text classification—using traditional machine learning and deep learning algorithms, there remains a need for more advanced NLP techniques to extract deeper and more critical insights from RFIs. This sets the stage for the next chapter, which focuses on developing domain-specific Named Entity Recognition (NER) and issue classification models. The models in chapter 5 aim to identify specific RFI issues and extract key entities using state-of-the-art NLP methods, capturing nuances that may be overlooked by conventional approaches.

Chapter 5: A two-step deep learning-driven NLP pipeline for efficient information extraction from construction RFIs

This chapter presents the two-step approach adopted for developing an advanced model for information extraction from RFIs. Building on the previous chapter, this chapter first leverages a supervised deep learning model based on a convolutional neural network (CNN) for issue-wise classification of RFIs. Next, the bidirectional long short-term memory (BiLSTM) and bidirectional encoder representations from transformers (BERT) models are utilised to develop a named entity recognition (NER) model. These models are then enhanced through an ensemble method by integrating a conditional random field (CRF) layer. The developed models hold promise for empowering project stakeholders to mitigate the impact of RFIs by facilitating prompt responses and transforming RFIs into actionable insights. These insights can guide decisions to ensure project success. Moreover, the knowledge gained can inform the development of future projects, enhancing both design and documentation quality.

5.1 Introduction:

Construction documentation, including drawings, contracts, specifications, and schedules, may often neglect details essential for project execution. This can lead to ambiguities arising during the project lifecycle (Panahi et al., 2023), necessitating additional input and clarification from the design team. Accordingly, a request for information document is employed as a means for the stakeholders to convey these concerns and seek resolutions (Abdel-Monem and Hegazy, 2013). RFI process aims to resolve issues as efficiently as possible and ensure that the project can proceed with minimal interruption. However, it is essential to recognise that drafting a suitable response to RFI consumes resources, including time and cost (Love et al., 2014). The consequences of not responding to RFIs promptly include delayed execution, overburdened design teams, and impeded information flow (Afzal et al., 2024). Due to such factors, the RFI process has often been deemed a necessary evil (Aibinu et al., 2019). Hence, the literature suggests that appropriate measures should be implemented to ensure that RFIs serve as communication channels rather than becoming sources of out-of-sequence work (Ibrahim et al., 2020), which can ultimately become part of the project's critical path (Kelly and Ilozor, 2020).

Accordingly, different technologies have been employed to address RFI challenges and enhance the efficiency in processing RFIs. Building Information Modelling (BIM) has offered improved workflow associated with RFIs. BIM facilitates error reduction in drawings by

providing visualisation capabilities that aid in resolving model misalignments (Sompolgrunk et al., 2021), leading to improved issue management. However, adopting BIM introduces its challenges, such as modelling errors, software interoperability issues, uncoordinated models, and unresolved clashes, thus initiating a new cycle of RFIs (Afzal et al., 2024). Comparably, common data environments like Aconex, Autodesk Construction Cloud (Das et al., 2020) and Procore (Sandoval et al., 2023) are systems employed in the construction industry for transmitting RFIs between the sender (contractor) and the recipient (client/consultant/designer). The integration of digital platforms has transformed the process of RFI exchange, shifting it from a traditional email or paper-based method to a centralised channel that enables central tracking and management of RFIs (Pradeep et al., 2021). However, these data platforms are associated with potential risks, including data loss, legal issues, and complicated data architecture (Afzal et al., 2024). Further, these BIM-enabled CDEs are not fully utilising the vast amounts of data they generate. Zawada et al. (2024) emphasise the importance of analysing and interpreting BIM-based CDE datasets to extract valuable information that enhances decision-making and improves project efficiency.

Similarly, in construction research, various mechanisms have been developed to understand the complexities of the RFI process, all aimed at improving it. Prior research can be divided into human-driven RFI codification and automated RFI assessments. In the former, researchers have performed manual content analysis to extract critical questions, such as the underlying issue causing an RFI and potential solutions (Bhat et al., 2017). These classifications are based on RFI category/type, property code, discipline code, work-element code, and the reason behind the RFI, with every researcher categorising RFIs according to their expertise. This manual method is inefficient and impractical for the resource-constrained construction industry. The latter method utilises automated approaches driven by natural language processing (NLP) and machine learning (ML). For example, Lee and Yi (2017) utilised various ML algorithms, including support vector machine, artificial neural network (ANN), naïve Bayes, and k-nearest neighbours, to conduct pre-bid risk classification of the RFIs. Additionally, they employed the Latent Dirichlet Allocation (LDA) algorithm for topic modelling of the RFI corpus. Similarly, Afzal et al. (2023) utilised LDA and t-distributed stochastic neighbour embedding algorithms to classify and cluster predominant topics and themes within RFIs. Shrestha et al. (2023) classified pre-bid RFIs based on impactful RFIs and those of a less critical nature. They applied ANN for this purpose. Another research study (Panahi et al., 2023) incorporated computer vision and NLP to extract symbols and texts from

drawings, performed an optimised search through the database of RFIs from previous projects, summarised the results using Chat-GPT and returned the RFIs most related to the under-review drawing sheets.

All the initiatives mentioned above are at the forefront of analysing and extracting information from RFIs, but they do not effectively streamline construction projects. There is a critical need to generate insights from existing CDE platforms currently limited in this capacity. Additionally, a segment of research continues to rely on manual content analysis and information extraction from RFIs. Moreover, most of the studies leveraging state-of-the-art technologies focus on the pre-bid phase, yet RFIs during the delivery phase can significantly impact project outcomes. Furthermore, there is an overall shortage of studies due to a lack of datasets that apply advanced NLP methods (Baek et al., 2021) to study RFIs and create models that can either reduce the review period of RFIs or help stakeholders manage them more efficiently. No studies have yet applied advanced deep learning methods, particularly state-of-the-art transformers, to enhance information extraction from the inherently unstructured RFI datasets. Therefore, this study seeks to extract more significant information from RFIs, including the coded issues, and identify critical entities to improve RFI review process efficiency. Significant contributions of this work include:

- A two-step approach is introduced; in first step, a convolutional neural network (CNN) is utilised for multiclass text classification of issues from RFIs. In the subsequent step, bidirectional long short-term memory (BiLSTM) and bidirectional encoder representations from transformers (BERT) are employed to develop domain-specific named entity recognition (NER) models for RFIs, enabling the extraction of key entities. Applying these NLP techniques to the RFI corpus for issue extraction and NER development presents an innovative approach yet to be explored in the current RFI-related literature. Therefore, the developed models will be the baseline for future development and research endeavours.
- RFIs are categorised based on the issues they address, drawing upon prior literature where manual content analysis was utilised. The current study identifies if the RFI presents any issues: coordination, constructability, review/approval, scope/specification clarification, and design/drawing discrepancy. In the following phase, the study also pinpoint key entities, such as problematic components, locations, and drawing references from within the RFIs. Stakeholders integrating these NLP-driven models into their RFI processes can

swiftly assess the magnitude of the issue and access pertinent information to expedite the review and closure of the RFI.

- Different deep learning-based methods are compared and evaluated on a novel dataset of RFIs from the delivery phase. Subsequently, NER approaches were enhanced by adding a layer of conditional random field (CRF) into the architecture, creating an ensemble model to achieve improved performance. The ensemble model combinations of BERT-CRF achieve a recall of 96%, a precision of 99%, and an F-1 score of 97%. Similarly, the BiLSTM-CRF combination scores 89% precision, 76% recall, and an F-1 score of 80%. These ensemble approaches, which are not extensively utilised within the construction informatics body of knowledge, are explored in this study.

In summary, focusing on the construction RFI process, this research aims to propose an advanced two-step approach driven by a deep learning-based NLP pipeline for automated and efficient information extraction from RFIs. Combining deep learning algorithms for automated issue identification and key entity extraction from RFI documents aims to improve information flow and overall tracking and management of RFIs. Furthermore, this research intends to guide construction teams in analysing RFI content quickly and making effective decisions. The complete research process for this study is illustrated in Fig 5.1. The organisation of this chapter is as follows: Section 2 presents theoretical background, focusing on challenges related to RFI management, and the use of NLP applications of text classification and NER for construction sector. Section 3 presents the CNN model developed for issue-wise classification of RFIs. Section 4 introduces the NER model designed for extracting key entities, providing a detailed discussion of the architecture of the models and their performance. In Section 5, the article outlines the study's contributions and presents a potential example of how the developed models can be applied in practical settings. Finally, section 6 presents the chapter’s conclusions, highlighting its limitations and outlining potential avenues for future research.

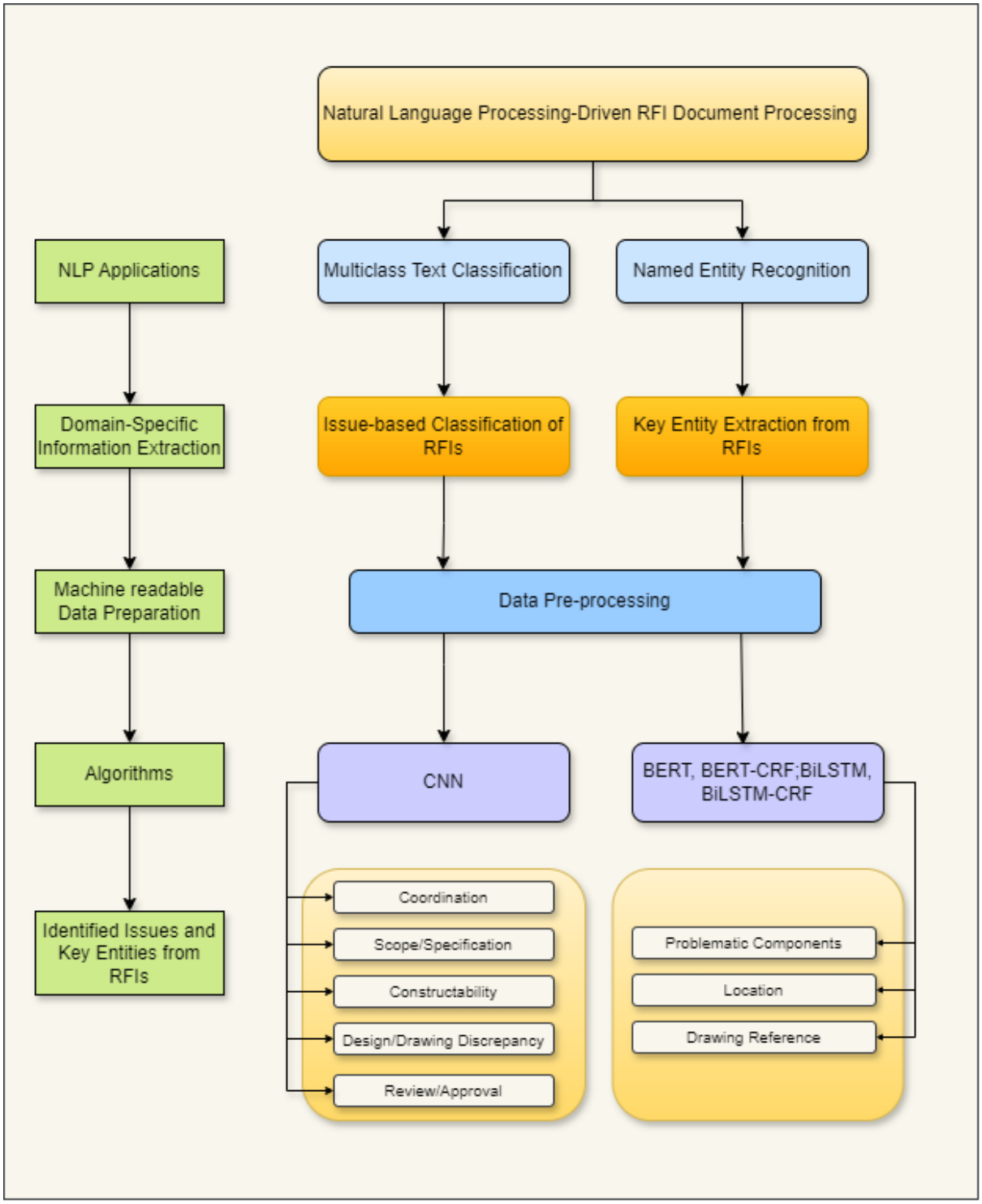


Figure 5-1. Overview of the research.

5.2 Research background

5.2.1 Theoretical background

RFIs are essential queries related to design issues or site execution challenges. Failure to respond timely can lead to implications, including trust issues, site delays, budget overruns,

and litigation (Afzal et al., 2024). Projects often encounter numerous RFIs, occasionally concerning minor details, which complicates the timely provision of responses. As contractors tend to use the RFI process for nearly all project communications, inefficient handling of RFIs may lead to unanswered or delayed RFIs and, thus, can potentially make a case for future delay claims. Subsequently, parties involved may use such records to justify contract increase amounts or time (Hughes et al., 2013).

Project delays and potential claims may result when RFIs involve activities on the critical path (Papajohn and El Asmar, 2021), mainly if responses are not provided promptly (Papajohn et al., 2018). Moreover, delays in RFI responses critically impact labour productivity (Jarkas et al., 2014) and project communication (Papajohn et al., 2018). Furthermore, uncertainties, risks, and project delays related to RFIs can significantly affect the cost performance of both the contractor and the client (Liao et al., 2020). Contractors incur costs in drafting RFIs and compiling supporting documentation, while clients or consultants face expenses in preparing RFI responses (Papajohn and El Asmar, 2021). In addition to financial costs, the process also entails a significant time commitment for both parties. Contractors and subcontractors struggle to identify informational errors across various drawings, a process that is both time-consuming and prone to mistakes. Consultants often attribute delays in addressing RFIs to inadequate fees and a lack of skilled workforce (Philips-Ryder et al., 2013). Reducing and effectively managing RFIs can improve communication and foster trust between project stakeholders (Zuppa et al., 2016). To achieve this, it is essential to implement measures that streamline the RFI process by extracting actionable insights, enabling stakeholders to review and respond more efficiently.

5.2.2 Natural language processing for construction documentation

NLP is a subset of AI (Candaş and Tokdemir, 2022) that enables machines to comprehend and analyse text language or speech-based datasets, typically containing unstructured information. This research will focus on reviewing two distinct applications for efficiently extracting insights from unstructured RFIs:

- Text classification involves assigning predefined categories to text statements based on their content. This application of NLP has been utilised in various studies within construction documentation. For instance, researchers have applied text classification techniques to extract requirements and non-requirements (Hassan and Le, 2020) and perform phase-wise separation (Hassan and Le, 2021) from construction contract

documents. Another important application involves classifying construction site accidents for improving construction safety analysis (Cheng et al., 2020).

- Named entity recognition is an NLP application that identifies and categorises named entities in text statements into predefined categories (Zhong and Goodfellow, 2024). Few examples related to NER in construction research include automated specification review (Moon et al., 2021), rule-based extraction of electrical and plumbing information (Wu et al., 2022), and extraction of regulatory requirements (Zhang and El-Gohary, 2021).

Both text classification and NER can be considered broader applications of information extraction from textual data. Information extraction focuses on automatically retrieving and restructuring information from textual documents for analysis (Candaş and Tokdemir, 2022). These documents can vary in structure, from unstructured to semi-structured, necessitating techniques for converting them into a format suitable for machine interpretation. The subsequent sections explore different research endeavours related to both NLP applications within the construction sector.

5.2.3 Previous studies on text classification in the construction sector

Text classification involves defining the category or categories to which a text statement or a document belongs (Manning et al. 2009). A category typically represents a class; for instance, a text about clashes in BIM models can be labelled as "coordination" to indicate a design coordination issue. Text classification methods include manual, rule-based, and machine-learning approaches (Salama and El-Gohary, 2013). In rule-based approach, handcrafting is utilised which requires domain expertise for manually creating classification rules that specify the criteria for labelling a text segment and associating it with a particular category (Salama and El-Gohary, 2013). While manual text classification generally achieves high precision and recall, it demands considerable labour for development and maintenance (Manning et al., 2009). ML-based text classification is categorised into supervised and unsupervised learning (Ayyasamy et al. 2010). Unsupervised learning operates without human guidance (Russell and Norvig 2010). Conversely, supervised machine learning involves human input, where a series of labelled text datasets is provided. This dataset is split into a training dataset, which trains the model with the rules for accurate labelling, and a testing dataset, for evaluating the model's performance.

Text classification has seen numerous applications related to construction documentation. For instance, in construction safety, ML algorithms can streamline processing of extensive safety documents like reports, accident records, and work plans, extracting valuable insights and knowledge. This approach helps managers identify safety risks more effectively and develop robust accident prevention strategies and safety decisions. For example, Tixier et al. (2016) developed an automated system of ML algorithms like random forest and gradient boosting to assess construction worker injury risks. Similarly, Zhang et al. (2018) also employed ML to build an ensemble model, optimised with sequential quadratic programming, to classify construction accident causes.

Traditional machine learning algorithms like support vector machine (SVM) and random forest excel in text classification tasks, but manual feature extraction and cleaning can be inefficient, especially for complex, large-scale tasks. Deep learning offers functionality and flexibility (Luo et al., 2023) that outperforms traditional methods. With complex multilayer neural network architectures, deep learning simulates the human brain's operation, automatically learning features and uncovering hidden relationships in data. Deep learning adapts more efficiently to construction applications than traditional methods requiring manual pre-training. Despite significant progress in applying deep learning to text-mining tasks in construction documentation (Luo et al., 2023), a notable gap remains in addressing RFIs. Current deep-learning approaches often fall short of effectively classifying and managing the diverse and complex issues present in RFIs. This gap is critical because RFIs involve many queries and issues requiring precise and context-aware classification to improve handling and processing. Therefore, this study will use a deep learning-based CNN algorithm to address this gap by more effectively classifying critical issues within RFIs.

5.2.4 Previous studies on construction domain-specific NER

NER serves as a pragmatic means for extracting valuable information from unstructured text data. Its primary objective is identifying and classifying terms or phrases into predefined entity types. Two fundamental techniques are employed for entity extraction: rule-based and ML-based approaches (Baek et al., 2023). Rule-based NER functions by identifying entities based on predefined patterns in text. This approach necessitates linguistic and domain-specific knowledge to create semantic and syntactic rules manually. Previous research has relied on manually created rules for task like analysing regulations and safety incident reports (Zhong et al. 2012). However, these approaches are time-consuming and pose restricted adaptability (Zhou et al., 2022; Zhang and El-Gohary 2013). Accordingly, diverse automated and semi-

automated techniques (Zhang and El-Gohary 2015) have emerged for various construction documentation applications, integrating ontology and rule-based information extraction (Ren and Zhang, 2021). While these efforts may have effectively automated information extraction, the manual construction of rules and ontological models remains necessary, often resulting in limited reusability (Baek et al., 2023).

Recent research has employed machine learning algorithms capable of automatically extracting complex features to overcome these limitations. These algorithms are also more robust in inputting data variations than rule-based approaches. Machine learning-based NER models extract entities by learning from labelled datasets; however, feature generation is essential for computers to process text data effectively. Unlike rule-based methods, machine-learning approaches are adept at handling variations as they autonomously discern and learn intricate patterns in textual data (Wu et al., 2022). With the significant advancements in the NLP field, current NER models designed for general entity recognition have reached performance levels that closely approximate human capabilities (Papers with Code 2022). However, developing domain-specific NER models remains challenging due to distinct lexicons in typical documents and inadequate training datasets (Lison et al., 2020).

This has led researchers to leverage deep learning. Recently, there has been a significant increase in the adoption of deep learning techniques for NER applications. For example, Zhong et al. (2020) proposed a hybrid model combining BiLSTM and CRF to recognise named entities and their relationships, enabling automated extraction of qualitative construction constraints. Liu et al. (2023) developed a contrastive learning-based framework to accurately extract entities and relationships from safety documents, reducing error propagation with limited training data. Furthermore, according to Devlin et al. (2018), the introduction of transformer-based pre-trained language models like BERT has significantly improved the performance of NLP models. Leveraging the transformer architecture, these models grasp the contextual meaning of individual tokens by considering both right and left contexts (Vaswani et al. 2017). Recent NER models in construction have utilised transformer-based pre-trained models to extract information on building defects (Jeon et al. 2022), inspection data (Li et al. 2021), and semantic elements for building compliance checks (Zheng et al. 2022). Accordingly, the research aimed to develop a novel NER model tailored explicitly for RFIs, an application yet to be fully explored. While there has been research developing NER models and applying deep

learning to construction documentation, studies focusing on RFIs are still non existing. Hence this work intends to address this gap and enhance the body of knowledge in this field.

5.3 Deep learning-based issue classification model

5.3.1 Data preparation for issue classification

To identify the issues outlined in the RFIs, datasets were obtained from projects completed within the last five years in the Middle East, specifically in Qatar and the United Arab Emirates. These datasets were sourced from companies specialising in the region's construction projects. The datasets were classified into five distinct categories, which were informed by existing literature. The process of RFI codification has been conducted by previous researchers, and the labels for the issues were obtained from their work.

Using the classifications in the literature and with the aid of industry professionals who possess domain expertise in addressing the RFIs, all RFI statements were initially categorised based on the issues they presented, and subsequently, a model was developed to classify these RFIs. Accordingly, RFI statements were classified into five broad categories: coordination, constructability, design/drawing discrepancy, review/approval, and scope/specification clarifications. This broad categorization enabled a more streamlined and impactful classification approach, effectively accommodating the dataset limitations. For instance, “design” and “drawing discrepancy” often share common causes, such as ambiguities in design intent or missing details, and typically require similar resolutions, like clarification from the design team. Similarly, "review/approval" and "scope/specification" clarifications frequently overlap in their nature and resolution paths, making it more efficient to handle them as combined categories.

This simplified set of categories enables RFI managers to more effectively prioritize, route, and resolve issues, highlighting the most common and high-impact RFI types. By focusing on these five consolidated categories, which play a critical role in project success, RFI managers are better equipped to respond promptly and accurately. Some issues may be resolved with minimal intervention, such as quick clarification, while others require in-depth discussions among experts to address more complex challenges. Automated detection of these categories can further support RFI managers by enabling timely, appropriate actions, particularly for issues requiring coordination across multiple disciplines.

Fig. 5.2 depicts the categorical distribution across all classes, revealing an uneven distribution. Evidently, the dataset lacks samples from certain categories and contains an insufficient number of samples for specific categories (e.g., review/approval = 190 entries). Therefore, during the dataset-splitting process, the samples in each category were divided according to a specific proportion (Liu and Yang, 2023). This helps the model learn the data features effectively. Subsequently, CNN was employed to construct a classification model and conduct supervised multiclass text classification for the RFI statements.

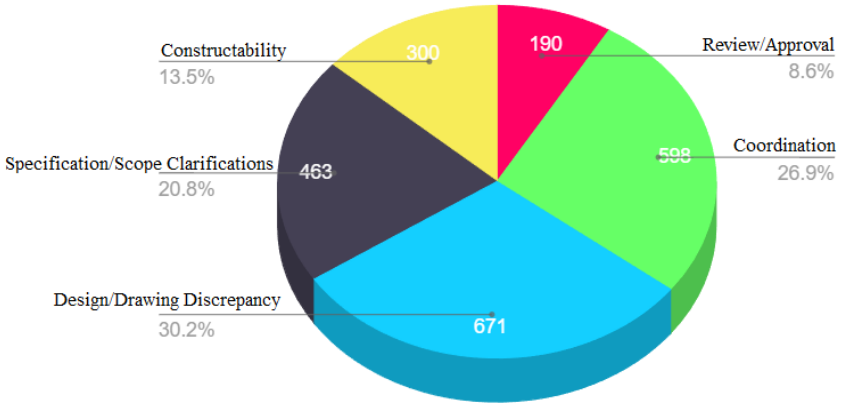


Figure 5-2. Categorical distribution of issues within RFI dataset.

5.3.2 CNN architecture and optimisation for RFI issue classification

CNNs are documented as highly effective deep neural networks and have been widely used across various AI domains such as computer vision, NLP and image processing (Fang et al., 2020). Advancements in word embeddings, vector space models, and the feature learning capabilities of CNN have led to excellent results in text classification, demonstrating its effectiveness not only in computer vision but also in capturing essential text features through its multilayer architecture, resulting in higher classification accuracy (Luo et al., 2023). A typical CNN (Fig. 5.3) includes an input layer, convolutional and pooling layers, a fully connected layer, and an output layer module (Zhu and Chen, 2020). Luo et al. (2023) delineate the architecture of CNN, where the convolutional and pooling layers collaboratively constitute a convolution group. This group conducts feature extraction from local to global levels in a layer-wise manner. Subsequently, the fully connected layer produces a vectorised representation of the extracted features. This feature vector is subsequently fed into the softmax output layer for text classification based on the RFI statement issues.

For this research a CNN model was developed following hyperparameter tuning to address the issue classification with training accuracy as the primary objective. Various combinations of

hyperparameters were tested to find the configuration that yielded the best performance. The final parameters were: convolution kernel size of 5, 128 filters, an embedding dimension of 50, 128 neurons in the fully connected layer, and a learning rate of $1 \times 10^{\wedge -3}$. For this research, the RFI statements were partitioned into two distinct groups: a training set and a testing set, with a split ratio of 90:10, consistent with previous research in construction documentation (Hassan and Le, 2021). The model was trained for 20 epochs with a batch size of 16, using the Adam optimiser.

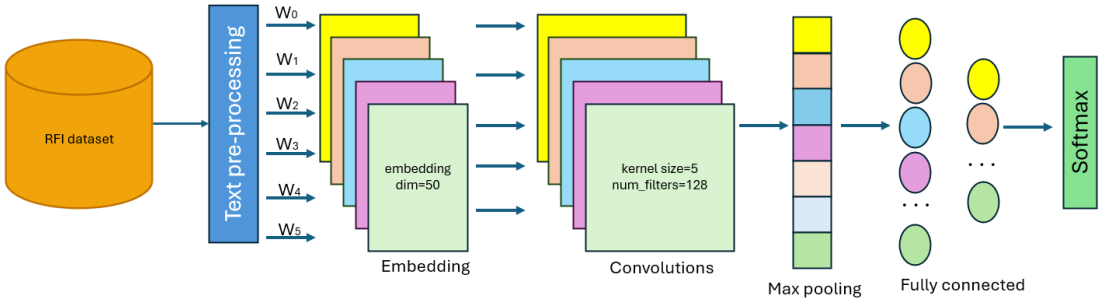


Figure 5-3. CNN architecture for issue-wise classification of RFIs modified from Liu and Yang, (2023).

5.3.3 Performance evaluation and classification results

After the model is developed, its classification performance is assessed using F1 score, precision, and recall, as the primary evaluation metrics (Romijnders et al., 2021), as specified in equations (5.1)–(5.3). In the equations, TP represents true positives, FP represents false positives, and FN represents false negatives, indicating the prediction results for each category.

$$Precision = \frac{TP}{TP + FP} \qquad (5.1)$$

$$Recall = \frac{TP}{TP + FN} \qquad (5.2)$$

$$FI - score = \frac{2 \times Precision \times Recall}{Precision \times Recall} \qquad (5.3)$$

In addition to accuracy, both the macro-average and weighted average were computed to assess the model's test results, as described through Equations (5.4) and (5.5), respectively. Here, the indicator represents the set of performance metrics (precision, recall, and F1-score), and support means the number of samples in each respective category (Liu and Yang, 2023).

$$macro\ avg = \frac{\sum_{i=1}^{num\_classes} indicator_i}{num\_classes} \qquad (5.4)$$

$$weighted\ avg = \frac{\sum_{i=1}^{num\_classes} indicator_i \times support_i}{\sum_{i=1}^{num\_classes} support_i} \qquad (5.5)$$

Table 5-1. Classification results of CNN for issue-wise classification of RFIs.

Issue in RFI	Precision	Recall	F1-score
Constructability	0.70	0.83	0.76
Coordination	0.68	0.76	0.72
Design/Drawing Discrepancy	0.79	0.73	0.76
Review/Approval	0.71	0.69	0.70
Scope/Specification	0.70	0.61	0.65
Accuracy	0.72	0.72	0.72
Macro avg	0.72	0.72	0.72
Weighted avg	0.73	0.72	0.72

Applying the above equations, it was determined that each category's average predicted F-1 value exceeds 70%, with the highest at 76% and the lowest at 65%. The performance assessment of the CNN algorithm, illustrated in Table 5.1, highlights notable variations in precision scores across different categories. Particularly significant is the higher precision observed for categories such as "design/drawing discrepancy" (79%) and "review/approval" (71%), indicative of the model's robust accuracy in identifying issues within these categories. Interestingly, both "constructability" and "scope/specification" achieved identical precision scores of 70%. Conversely, the precision score for "coordination" (68%) is relatively lower, suggesting a heightened occurrence of false positives in predictions related to this category.

Similarly, recall scores also fluctuate across categories, with categories like "constructability" (83%) exhibiting higher recall rates compared to the category with the lowest score "scope/specification" (61%). This indicates the model's ability to effectively capture a more significant proportion of actual issues within specific categories while potentially missing some in others. The model's performance is further highlighted by its F1 scores, frequently employed as combined recall and precision metrics to evaluate model effectiveness. Categories with higher F1 scores, including "constructability", with a score of 76%, show a satisfactory balance between precision and recall, indicating overall solid performance in issue classification. On the other hand, categories with lower F1 scores, such as "scope/specification" with a score of 65%, may indicate areas where the model's performance could be enhanced through additional

data. The model generally exhibits reasonable performance across all categories, with a weighted average F1 score of 72%. The F1 score, which ranges from 0 to 1, provides a balanced measure of precision and recall, with higher values indicating better overall performance. The achieved score of 0.72 is considered satisfactory for the scope of this study. Further improvements could be achieved by refining the model and expanding the training dataset.

5.4 NER model development for construction RFIs

When an RFI is received, it is processed by the RFI manager, who assigns it to the responsible person or party within the team. In instances where the RFI is deemed trivial and the manager possesses the requisite knowledge to provide an immediate resolution, they may respond directly. In this regard, three critical entities underpin the effective management of RFIs: the problematic component (e.g., walls, floors, ceilings), the location (e.g., academic block, canteen, ground level), and the drawing reference embedded within the RFI (e.g., DWG-DIC-HB-05-WS-02 or A-0001). These entities were identified through consultations with industry experts who engage with RFIs regularly. The extraction of these key entities facilitates the RFI manager or project manager pinpoint the specific building component experiencing an issue, its precise location, and the relevant drawings that must be consulted to resolve the problem instantly. This systematic approach prevents a single RFI from cascading into multiple subsequent RFIs by ensuring meticulous attention during the design and execution phases.

5.4.1 Data preparation for NER model

Table 5.2 mentions the extracted tags along with their corresponding information types. The RFIs were initially collected in PDF format, which made their content unsuitable for direct computer analysis. Consequently, each PDF file was first converted to TXT format. While open-sourced platforms were initially utilised for this process automatically, it was found that this method compromised the essential content of the RFIs. Subsequently, the decision was made to manually produce the text format of the RFIs, ensuring the preservation of crucial information for further NER labelling. The text files were then converted to JSON format, facilitating NER labelling for the problematic components (PRO), location (LOC), and respective drawing reference (DWG). This labelling was performed using an open-source platform online through NER Annotator for SpaCy. The same dataset used for issue classification was also employed for NER model development, with the same train test split.

Table 5-2. Entity categories.

Entity name	Annotation tag	Type
Problematic Component	PRO	Text
Location	LOC	Text
Drawing Reference	DWG	Mixed (number and text)

5.4.2 Bidirectional long short-term memory

Traditional neural networks cannot remember previous input states, prompting researchers to explore RNNs, which have demonstrated efficiency in capturing long-term dependencies (Adil et al., 2021). However, traditional RNNs are prone to vanishing or exploding gradients during backpropagation, resulting in network instability. Hochreiter and Schmidhuber (1997) introduced LSTM (Fig. 5.4a) networks to address these issues by incorporating a cell state mechanism, enabling the preservation of temporal information over extended sequences. Unlike traditional RNNs, LSTM networks include essential components such as forget gates, input gates, and output gates, facilitating the controlled information flow within the network architecture (Adil et al., 2021). BiLSTM (Fig 5.4b), which extends the traditional LSTM architecture (Guo et al., 2015), stands for bidirectional long short-term memory. It can capture input characteristics from both past (via the forward pass state) and future (via the backward pass state) contexts over a specified duration (Liu and Yang, 2023). The BiLSTM approach differs from unidirectional LSTM such that it processes the same input in both forward and backward directions. This bidirectional processing preserves additional contextual information, significantly enhancing the model's performance (Adil et al., 2021). Tao and Liu (2018) detail the procedure of forward pass state through the following equations:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \quad (5.6)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (5.7)$$

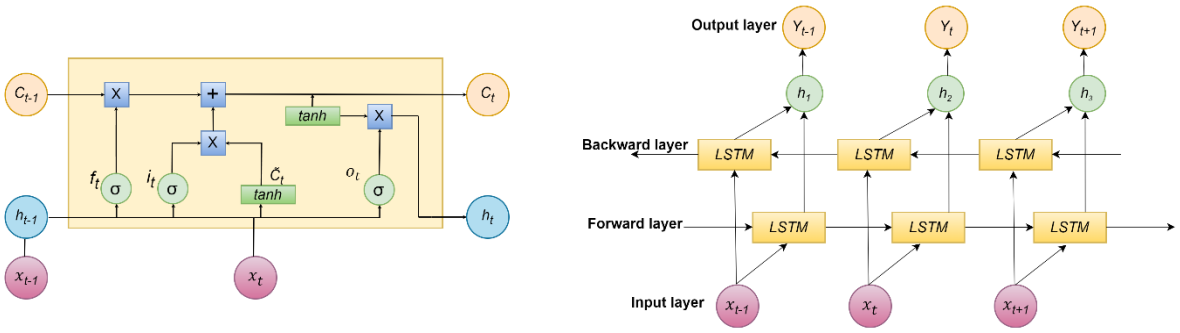
$$C_t = tanh(W_c[h_{t-1}, x_t] + b_c) \quad (5.8)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (5.9)$$

$$C_t = f_t \times C_{t-1} + i_t \times C_t \quad (5.10)$$

$$h_t = o_t \times tanh(C_t) \qquad (5.11)$$

The LSTM network operates on a sequence of inputs, transforming them into hidden states. The forget gate (f_t), in equation (5.6), decides which information to discard based on both the previous hidden state (h_{t-1}) and the current state (x_t) (Wang et al., 2021). Equation (5.7) describes the input gate, which decides what information should be retained in the cell state. It consists of two layers: the first layer, composed of sigmoid activations, determines the updated value, while the second layer, depicted as tanh in equation (5.8), generates a set of candidate values (C_t) that may augment the cell state (Gunter and Önder, 2016; Yang et al., 2014). These layers combine to update the cell state. The output gate (o_t), as detailed in equation (5.9), controls the output information. Equation (5.10) shows the updated cell state, calculated by summing the previous cell state scaled by the forget gate and the candidate cell state scaled by the input gate (Hochreiter and Schmidhuber, 1997). The cell state is then passed through the tanh function (Equation (11)) and multiplied by the sigmoid output. Throughout Equations (5.6) – (5.11), σ and tanh are activation functions defining neuron outputs in the network, while W and b represent weight matrices and bias vectors, respectively (Adil et al., 2021).



(a) LSTM cell

(b) BiLSTM network

Figure 5-4. Illustrative block of LSTM cell and its variant BiLSTM, adapted from Wadawadagi and Pagi, (2020)

5.4.3 Bidirectional encoder representations from transformers

Pre-training models such as BERT (Fig. 5.5) have demonstrated cutting-edge performance across a range of NLP tasks (Wang et al., 2021). They leverage semantic knowledge gained from extensive unlabelled corpora during pre-training, which is then transferred to downstream tasks, enhancing performance (Liu and Yang, 2023). This approach aligns with transfer learning, where a pre-trained model is adapted for a new task, saving time and computational resources while enhancing learning accuracy (Prottasha et al., 2022). As a result, BERT's pre-

training model was chosen to develop a NER model for RFI documents. This decision was based on two key reasons. First, BERT has consistently achieved impressive performance across various NLP tasks, including sentence classification and question answering (Nguyen et al., 2020). Second, a pre-trained model is particularly advantageous for this situation, given the limited number of RFI documents from various construction projects for training and testing.

This aligns with the overarching challenge in NLP research within the construction sector, where access to extensive textual datasets remains a persistent constraint (Chung et al., 2023). As a pragmatic choice, leveraging existing models rather than training an entire network from scratch is often more feasible (Nguyen et al., 2020). This research utilises BERT for transfer learning to fine-tune the model on domain-specific RFIs to address this imperative. BERT's input representation encompasses the integration of three key embeddings (Liu and Yang, 2023), including:

(1) Token Embeddings: represent fixed-dimensional vectors and depict each word or subword token in the input text.
(2) Position Embeddings: enable BERT to understand the relative positions of tokens within the input sequence, allowing it to capture contextual relationships effectively.
(3) Segment Embeddings: distinguish between tokens belonging to different segments by assigning distinct vectors to each segment. This allows BERT to differentiate between tokens from different sentences and incorporate information from each segment separately.

According to (Liu and Yang, 2023), BERT's encoder and decoder modules use a multilayer bidirectional transformer architecture for fine-tuning. This approach allows each word to be contextually encoded with all other words in the sentence, leading to a more balanced representation. Each encoder module includes a multi-head self-attention mechanism with input vector matrices.

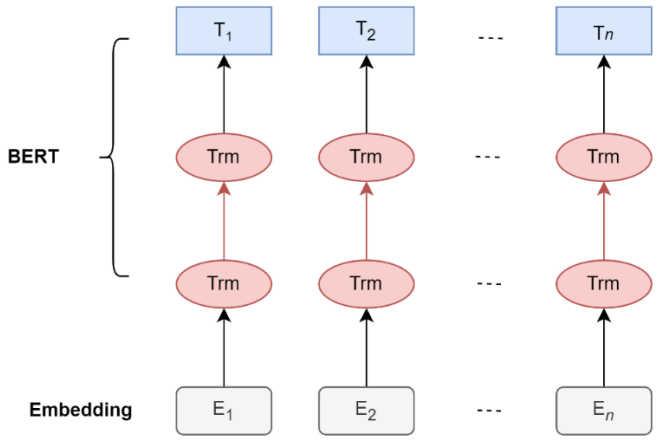


Figure 5-5. Illustrative block of BERT pre-trained language model, adapted from Liu and Yang, (2023)

5.4.4 Conditional random field-based ensemble model.

The main challenges in developing named entity recognition for RFI documents are as follows:

- Complex structure: RFI documents contain various construction-related details and queries, including specifications, material confirmations, and design clarifications. This diversity in content, combined with the detailed and technical nature of the requests, makes it difficult to process these texts intelligently.
- Scarcity of corpus: There is a significant shortage of appropriate corpora for this task. Construction project details and RFI responses are often confidential, making it difficult to acquire large quantities of relevant data necessary for training robust NER models.
- Context dependency and specialised terminology: RFIs in the construction industry feature complex structures and exhibit strong contextual semantic relevance. Additionally, the field includes many specialised terminologies, further complicating the NER task.

To address these challenges, a conditional random field (CRF) layer is applied to optimise the developed NER models. Further stacking CRFs on BiLSTM (Ju et al., 2018; Lample et al., 2016) and BERT (Li et al., 2024) has shown promising outcomes for NER. The ensemble approach leverages CRF's robust feature-matching capabilities, which have demonstrated success in different sectors, including network security (Ma et al., 2021) and medicine (Kang et al., 2021). By incorporating CRF into the models, the research aims to enhance text embedding quality and improve overall model training performance. Nguyen et al. (2020) provides a comprehensive explanation of CRFs (Fig. 5.6). CRFs are designed to predict label sequences globally for given input sequences. For a given input sequence $X = (x_1, x_2, \ldots, x_n)$, CRFs learn to predict each x_i by maximising the logarithmic probability during training.

Specifically, for a sentence sequence $X = (x_1, x_2, \ldots, x_n)$, and its corresponding sequence of labels $Y = (y_1, y_2, \ldots, y_n)$, CRFs compute the probability of Y conditioned on X as $P(Y|X)$. When predicting y_i, CRFs consider the current input x_i (the current word) and the previous states of the preceding words to estimate $P(y_i = 1|X)$. Using CRFs for prediction can be seen as a sequence labelling task. For an input document consisting of segments, CRFs classify each segment as either an extracted value or a tag.

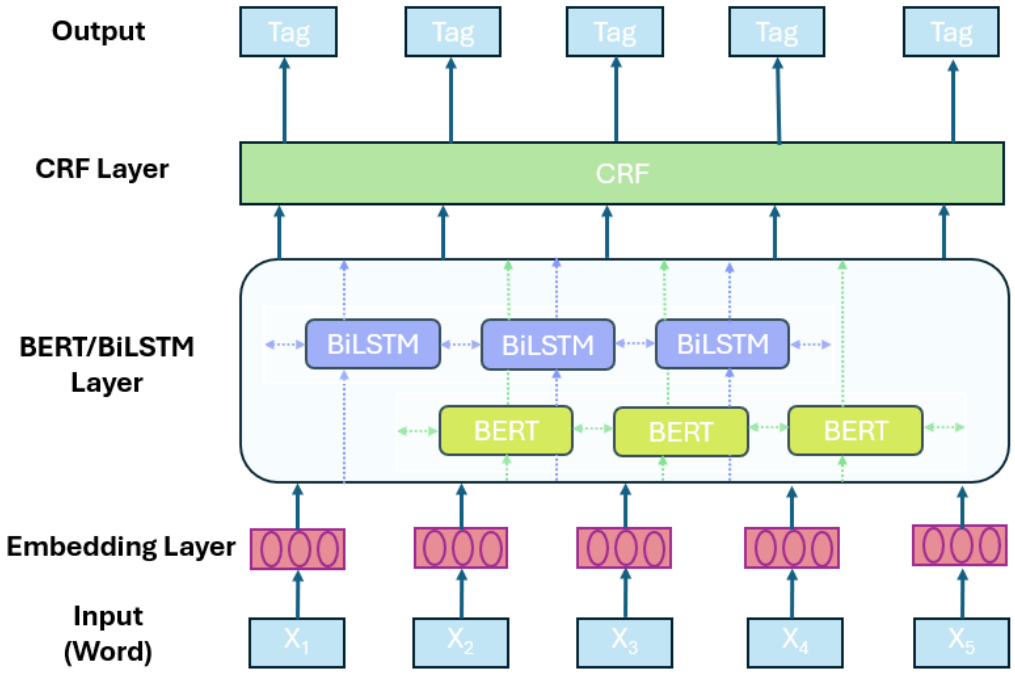


Figure 5-6. Illustrative block of BERT/BiLSTM-CRF ensemble model for NER of RFIs, adapted from Liu and Yang, (2023)

5.5 NER model evaluation

Based on the above architectures and ensemble settings, an NER model is proposed utilising various approaches for the pre-processed RFI documents. Specifically, 90% of the annotated RFI dataset was allocated for training, while the remaining 10% was reserved for testing to evaluate the generalisability and performance of the developed NER models. Table 5.3 below shows the evaluation of performance metrics, including precision, recall, and F1-score for the developed models. The performance metrics for the NER models BERT, BERT-CRF, BiLSTM, and BiLSTM-CRF reveal significant differences in their precision, recall, and F1 scores across all three entities. The comparison of the BERT, BERT-CRF, BiLSTM, and BiLSTM-CRF models revealed notable differences in their performance across NER categories, with significant improvements observed upon adding the CRF layer. The BERT

model alone yielded moderate results, achieving its highest F1-score of 0.65 for the "problematic component" category while performing poorly on the "location" category with an F1-score of only 0.38. BERT's overall macro and weighted averages were 0.56 and 0.58, respectively. However, incorporating the CRF layer into the BERT model (BERT-CRF) substantially enhanced performance across all metrics. For instance, the F1 score for the "problematic component" jumped to 0.99, and the "location" category improved to an F1 score of 0.96. The macro and weighted averages reached 0.97, highlighting the significant boost in precision, recall, and F1 scores across all categories because of the integration of the CRF layer.

Similarly, the BiLSTM model showed reasonable performance, achieving its best F1 score of 0.80 in the "problematic component" category, yet underperformed in the "drawing reference" category with an F1-score of 0.37. The macro and weighted averages for BiLSTM were 0.63 and 0.74, respectively. Adding the CRF layer to the BiLSTM model (BiLSTM-CRF) resulted in a marked performance improvement. The F1-score for "problematic component" increased to 0.86, and while the "drawing reference" category still had a relatively low F1-score of 0.38, overall metrics improved. The macro and weighted averages improved to 0.68 and 0.80, respectively. It can be assumed that the CRF layer enhanced model performance by providing better sequence labelling capabilities, as it considers the context of the entire sequence, reducing labelling errors that occur when considering tokens in isolation. This contextual understanding is crucial for NER tasks, where the label of one token often depends on neighbouring tokens. The BERT-CRF model, which benefits from both BERT's robust contextual embeddings and CRF's sequence-level optimisation, demonstrated substantial improvements in precision, recall, and F1 scores compared to BERT alone. Similarly, the BiLSTM-CRF model exhibited enhanced performance metrics compared to the BiLSTM model. However, the improvement was less pronounced than in the BERT-CRF case, likely due to BERT's superior contextual embedding capabilities over BiLSTM. In summary, adding the CRF layer significantly boosted the performance of both BERT and BiLSTM models, as evidenced by improved precision, recall, and F1 scores across all entity categories, underscoring the effectiveness of combining contextual embeddings with sequence-level optimisation for NER tasks.

Table 5-3. NER performance.

Model	NER Category	Precision	Recall	F1-Score
BERT	Problematic Component	0.62	0.68	0.65

	Location	0.54	0.29	0.38
	Drawing Reference	0.57	0.73	0.64
	Macro Avg	0.58	0.57	0.56
	Weighted Avg	0.59	0.59	0.58
BERT-CRF	Problematic Component	0.99	0.98	0.99
	Location	0.98	0.94	0.96
	Drawing Reference	0.98	0.95	0.97
	Macro Avg	0.99	0.96	0.97
	Weighted Avg	0.99	0.96	0.97
BiLSTM	Problematic Component	0.90	0.72	0.80
	Location	0.76	0.71	0.74
	Drawing Reference	0.25	0.70	0.37
	Macro Avg	0.64	0.71	0.63
	Weighted Avg	0.80	0.71	0.74
BiLSTM-CRF	Problematic Component	0.98	0.77	0.86
	Location	0.87	0.72	0.79
	Drawing Reference	0.24	0.89	0.38
	Macro Avg	0.70	0.79	0.68
	Weighted Avg	0.89	0.76	0.80

5.6 Theoretical contributions and practical implications of developed models

This study guides the research towards harnessing advanced NLP techniques and algorithms for analysing RFIs, previously reliant on manual categorisation and information extraction (Morales et al., 2022; Soh et al., 2020; Kim et al., 2021). Although NLP has made strides in IT, its application in construction documentation is domain-specific (Chung et al., 2023). By leveraging deep learning-driven models such as CNN for text classification and BiLSTM and BERT for NER, this research advances the field, paving the way for more sophisticated methods tailored to the construction domain and focusing on novel RFI corpora. Future work will build upon these advancements to refine and enhance RFI analysis.

One of the significant practical contributions of this study lies in the transition from manual RFI content analysis to automated analysis, facilitating stakeholders in making informed decisions. Fig. 5.7 illustrates how NLP-driven models can be implemented in construction settings, enabling RFI managers to streamline the RFI process by directing RFIs to the appropriate individuals and analysing trends to address recurring issues pre-emptively. Integrating automated models into existing RFI workflows, whether via email exchanges or common data environments, offers numerous benefits for practitioners. These text analyses empower RFI managers, project managers, and teams to identify problematic areas, expedite reviews, and proactively mitigate issues, ultimately improving project efficiency. Furthermore, existing CDEs and BIM-based platforms generate substantial data, yet they often underutilise it for efficient project management. The NLP pipeline proposed in this chapter exemplifies how the data already generated and stored within existing CDE platforms can be leveraged to derive actionable insights. Such utilisation of vast data can inform future projects, facilitating the creation of design plans and specifications that integrate lessons learned and proactively address potential issues.

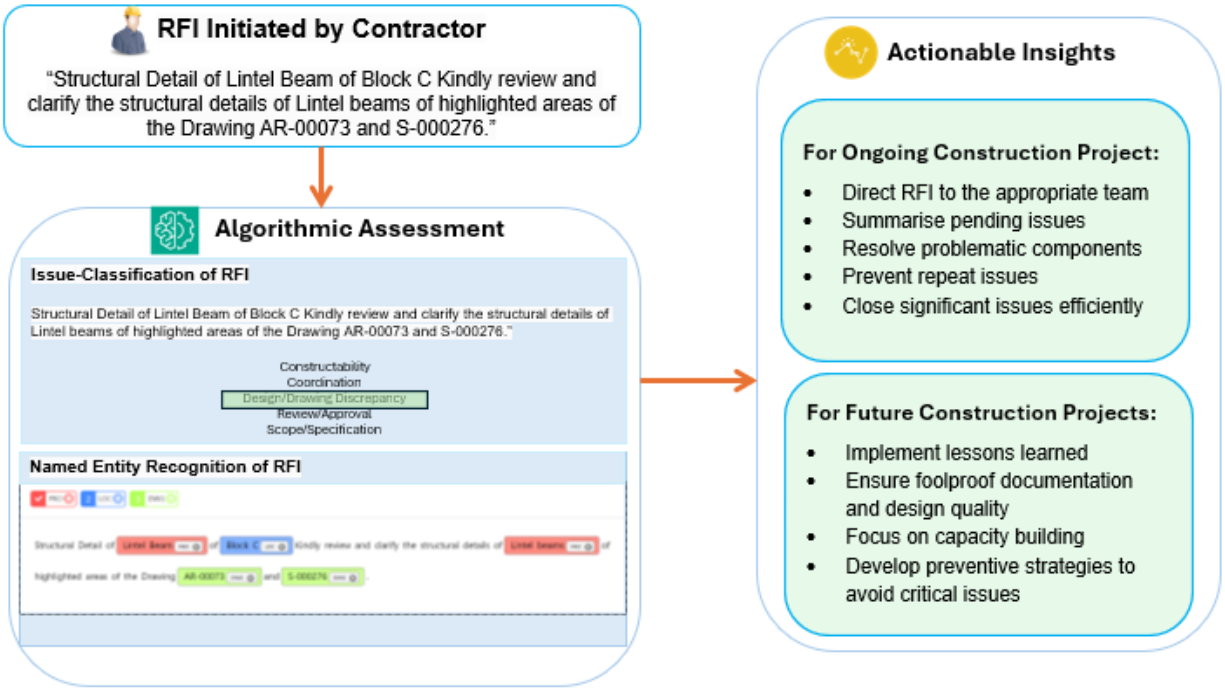


Figure 5-7. Potential usage of the developed text classification and NER model for RFI management.

5.7 Summary of the chapter:

This research focused on analysing RFI documents, classifying the issues within them, and identifying key entities essential for rapid issue resolution. For this study, RFIs were collected

from construction projects, and in the initial phase, a CNN-based text classification model was developed to classify the issues presented in the RFIs. These issues were categorised into constructability, coordination, design/drawing discrepancy, review/approval, and scope/specification. The CNN model performed well across classification performance metrics. In the subsequent phase, novel NER models were developed using BERT and BiLSTM architectures to extract entities such as problematic components, locations, and drawing references.

Integrating a CRF layer significantly enhanced these models' performance, with the BERT-CRF ensemble showing marked improvements in entity extraction accuracy and the BiLSTM-CRF model outperforming the BiLSTM baseline. These enhancements are due to the effective combination of contextual embeddings and sequence modelling facilitated by the CRF layer.

In terms of limitations, it is acknowledged that no research endeavour is flawless or fully exhaustive, and this study can benefit from further refinements and enhancements. One key challenge was a class imbalance across categories and entities in the NLP model. Future research should focus on incorporating larger datasets to improve the training, testing, and validation of these models. Ideally, fine-grained evaluation in the CNN-driven text classification model would be preferable; however, this requires a more comprehensive dataset for each class. Additionally, the current approach merges a few issues to create broader categories, enabling project teams to understand overarching issues. With richer data spanning a wider range of project RFIs, the model can be further refined to identify specific issues within RFIs. Data limitations thus pose a primary constraint on improving and refining results.

It must also be noted that RFIs are inherently varied based on project-specific conditions. Due to differences in construction methods, materials, and strategies, RFIs from infrastructure projects use different jargon than those from residential buildings. Consequently, a one-size-fits-all approach is inadequate. The developed model is currently suitable for the project settings it was designed for. However, to adapt it to different project settings, the algorithms must be trained on larger datasets to handle RFIs from diverse project environments, including pre-bidding and delivery stages.

Future investigations are required to automatically extract more pertinent features from RFIs to improve efficiency in addressing RFIs based on their severity, potentially leading to earlier issue resolution. This process may involve integrating various state-of-the-art large language

models with a user-friendly graphical interface that stakeholders in their live projects can utilise.

Chapter 6: Conclusion and recommendations

This chapter concludes the thesis by presenting the key findings and providing a detailed roadmap on integrating the developed models for efficient text classification and information extraction during the RFI processes. Each research objective is addressed, highlighting the novelty, value, and implications of the research. Additionally, this chapter offers a comprehensive roadmap, including a conceptual representation of a dashboard, demonstrating how the developed models can provide data-driven insights for industry practitioners.

6.1 Review of research objectives

As highlighted in the introduction chapter, this research aimed to achieve three key objectives. Table 6.1 outlines how these objectives have been met, indicating the specific chapters where they are addressed and summarising the related key findings. The present research recognised the limited applications of NLP techniques for extracting information from unstructured construction documents, including RFIs. Extracting insights from construction RFIs is crucial, as failing to process RFIs in a prompt manner, negatively impacts projects. Both direct and indirect issues arise from the emergence of RFIs and their delayed processing. Although the construction industry has evolved from email/paper-based RFI exchanges to more streamlined CDE-based RFI tracking and management, these platforms primarily serve as systematic exchange mechanisms and do not leverage the big data they generate. Therefore, a text-mining driven solution is required to support both paper/email-based and CDE-based RFI management mechanisms. Limited studies have applied machine learning algorithms to RFI datasets. Researchers have often relied on manual content analysis to understand RFIs and extract patterns, an impractical approach for construction projects due to its time-consuming and laborious nature. Furthermore, NLP-driven RFI solutions are rare, with only a few studies exploring NLP techniques to extract information from RFI content. Most of these studies use RFI corpora from the bidding or pre-bidding phases, limiting their applicability to the project delivery phase. Additionally, the application of advanced deep learning-based RNNs and transformers for extracting insights from RFI corpora is not explored within the existing body of knowledge. Therefore, there is a pressing need of models that incorporate NLP advancements to generate actionable insights from RFI datasets, which can be leveraged for project success.

This research aimed at developing advanced models that integrate state-of-the-art NLP and deep learning approaches for efficient text classification and information extraction from RFIs.

These automated applications can be utilised in ongoing and future projects. The thesis achieves this broader aim through three research objectives: (1) investigating the existing RFI process, along with the repercussions of delayed RFI resolution and existing solutions for streamlining RFI handling, (2) developing a domain-specific text classification model for automated phase-wise classification of RFIs and using topic modelling for advanced visualisation, and (3) developing a domain-specific issue classification model identifying major issues from RFIs and an NER model for efficient key entity extraction from RFIs. Given the nature of the problem and the overall aim of the thesis—to address existing limitations in RFI handling within both traditional and advanced construction setups—the design science research framework was chosen as the underpinning methodological framework. Design science research involves understanding a practical problem, designing and developing a solution to address the problem, testing the solution, and communicating the results. The framework was integrated into the research methodology chapter to achieve the study's objectives. The subsequent sections present outcomes corresponding to each research objective.

Table 6-1. Research objectives, achievement criteria, location in thesis, and key findings.

Objective No.	Objective Details	How is it addressed?	Where is it addressed ?	Key findings
1	Synthesise and summarise the challenges in the RFI process, the repercussions of delayed RFI resolutions, and the solutions proposed by academia and industry to streamline the RFI process.	Through a thorough literature review encompassing 89 systematically retrieved research articles.	Chapter 2	• Delayed processing of RFIs and frequent emergence of RFIs were identified as key factors jeopardizing overall project success. • Review of existing RFI handling solutions highlighted the need for advanced NLP-driven methods. • Advanced NLP methods can generate actionable insights from unstructured RFIs and improve project outcomes.

2	Developing a text classification model to categorise RFIs into construction phases for efficient routing and applying topic modelling to cluster RFIs and uncover themes.	This objective had two parts. The first part of this objective leveraged supervised learning approaches for phase-wise categorisation of RFIs. The developed model classifies the RFIs into predominant phases "design", "execution" and "procurement". In the second part, topic modelling, algorithms such as LDA and t-SNE, were applied to extract and cluster key topics and themes from the RFIs.	Chapter 4	• A domain-specific multiclass text classification model was developed using RNNs, demonstrating impressive performance. • Performance was further enhanced with ensemble techniques, with the weighted majority model achieving an F1 score of 86%. • The LDA algorithm was used to identify and visualize predominant topics in RFIs. • Advanced visualizations were developed for stakeholders to facilitate a deeper understanding of project RFIs.
3	Developing an issue classification model and an NER model to extract key entities from RFIs for automated information extraction.	To achieve this objective, a two-stage approach was adopted. The first performed domain specific issue-wise classification into "coordination", "constructability", "design/drawing discrepancy", "review/approval", and "scope/specification". In the next stage an NER model was developed utilising BiLSTM, BERT and the performance of this model was further evaluated by the integration of CRF layer into the above-mentioned architectures. Key entities extracting using NER were "problematic components", "location", and "drawing reference".	Chapter 5	• The CNN-based issue classification model was developed, achieving an overall F-1 score of 72%. • A novel NER model was developed to extract key entities. Both BERT (F1 score of 58%) and BiLSTM (F1 score of 74%) models saw significant improvement with the integration of a CRF layer. The BERT-CRF model achieved an F1 score of 97%, while the BiLSTM-CRF model achieved 80%.

6.1.1 Objective 1: Synthesising and assimilating RFI literature to investigate shortcomings in the RFI process, the repercussions of delayed RFI resolution, and

solutions proposed by both academic research and practical applications to streamline RFI process.

This objective was accomplished through a systematic review of the literature. Predominant themes were extracted from this review to comprehensively understand the state-of-the-art in the RFI process within the construction sector. This research objective aligns with the initial steps of the design science research framework, which emphasise the development of a solution to the identified problem. The literature review, coupled with an analysis of risks inherent in the RFI process and a review of existing solutions, contributes to a thorough understanding of the problem and the establishment of necessary requirements. Chapter 2 of the thesis effectively fulfils this research objective.

- **Existing RFI practices and challenges within**

RFIs are considered a necessary evil in the construction sector, and delays in their processing exacerbate project difficulties. If these issues are not addressed promptly, they can lead to downstream risks such as design omissions, complex designs, uncoordinated drawings, inconsistent project specifications, and field execution problems. The emergence of new issues often triggers another cycle of RFIs, causing recurrent disruptions that lower project productivity, extend project timelines, and increase construction costs. Activities mentioned in the RFIs can lead to project delays and potential claims if they become part of the critical path, often due to untimely responses. Additionally, delays in RFI responses critically affect project communication and labour productivity. Risks and uncertainties associated with RFIs, and schedule delays can significantly impact project cost performance for both clients and contractors. Contractors incur expenses drafting RFIs and gathering supporting documentation, while clients or consultants accrue costs in preparing RFI responses. It is also challenging, time-consuming, and error-prone for contractors or subcontractors to detect informational errors from multiple drawings. This research elucidates the challenges and risks associated with the RFI process using systems thinking approach, detailed in Section 2.4.1 of this thesis. Due to the challenges associated with the emergence and timely resolution of RFIs, both researchers and industry professionals have developed solutions.

- **Limitations of the existing RFI industry and academic solutions**

BIM facilitates error reduction in drawings by providing visualisation capabilities that aid in resolving model misalignments. It also mitigates 2D errors and omissions, minimises discrepancies in grid and column alignments, and decreases direct clashes. However, adopting BIM introduces its challenges, such as modelling errors, software interoperability issues, uncoordinated models, and unresolved clashes. These issues can lead to another cycle of RFIs, which emerge due to BIM implementation itself.

Therefore, a trade-off exists between the RFIs stemming from BIM adoption and those arising from its non-implementation. Comparably, common data environments like Aconex, Autodesk Construction Cloud and Procore are platforms employed in the construction industry for transmitting RFIs between the sender and the responder. The integration of digital platforms has unquestionably transformed the process of RFI exchange, shifting it from a traditional email or paper-based method to a centralised, streamlined channel that enables effective management, organisation, and tracking of RFIs. Section 2.4.4 details the technologies supporting RFI management, while section 4.2 further examines the features of the predominant CDEs in detail. There are potential risks associated with using these data platforms, including data loss, legal issues, and complicated data architecture. Furthermore, these BIM-enabled CDEs are currently underutilizing the vast amounts of data they generate. This limitation hinders their potential to streamline the RFI process effectively. Automated extraction of information from unstructured RFI queries can empower decision-makers to resolve RFIs swiftly during projects. Moreover, these insights can be leveraged post-project completion to integrate lessons learned into future projects.

- **Need for NLP-driven solutions to extract insights from unstructured RFI statements**

The challenges and limitations of existing solutions highlight the importance of analysing RFI content to learn from past experiences and enhance design practices in future projects. Advanced techniques like text mining and natural language processing are essential to understand the unstructured content within RFIs. NLP, a branch of artificial intelligence, helps in analysing, understanding, and manipulating human languages, including unstructured text documents and speech. Utilising these approaches provides valuable insights from RFIs, allowing the identification of various risks and improving the handling of traditional RFI documents, thereby reducing turnaround times. However, the slow adoption of advanced text mining techniques in the construction field has limited this direction of research. Extracting insights from construction documents using text mining is inherently challenging, exacerbated

by the unstructured nature of RFI queries and responses. Recognising the limitations of current methods and the necessity of embracing advanced NLP techniques, this study takes a significant step forward by developing NLP models for efficient text classification and information extraction from RFIs. This advancement is accomplished through objectives 2 and 3 of this research.

6.1.2 Objective 2: Developing a domain-specific text classification model to categorise RFIs into construction phases for efficient routing and automated classification, supplemented by topic modelling to cluster RFIs and uncover predominant themes.

The second objective of this research is to guide through the stages of model development, aligning with the design science research framework, following the articulation of our problem in the literature review. This objective necessitates the development of machine learning model for phase-wise text classification of RFIs and then uncovering predominant themes from within them using topic modelling approach. The detailed accomplishment of this objective is outlined in the following subsections.

- **Supervised traditional machine learning algorithms vs deep learning-based RNN:**

The research presented in Chapter 4 compared two deep learning-based recurrent neural networks with four traditional machine learning algorithms. The RNN-based models include GRU and LSTM. In contrast, the traditional machine learning algorithms include RF, LR, SVM and NB (Section 4.2.1). The logic behind this comparison was determining which supervised algorithms work best for the phase-wise classification of RFIs. The algorithms were trained on a training dataset to perform multiclass text classification, categorising RFIs into their phases: design, execution, and procurement. Precision, recall, and F1 scores were used as the classification performance metrics to measure the effectiveness of the algorithms. When tested on an unseen dataset, the deep learning algorithms outperformed the traditional machine learning algorithms across all criteria, with the GRU performing (86% precision, 85% recall, and 84% F-1 score). This superior performance can be attributed to their advanced capability to capture complex patterns and dependencies in the data, highlighting the potential of deep learning models for RFI classification.

- **Comparison between different feature extraction techniques:**

In Chapter 4 of the research thesis, in Section 4.4.1 and 4.4.2 performance comparisons were conducted on different feature extraction techniques. These techniques are critical as they assess the effectiveness of converting unstructured RFIs into machine-readable formats for algorithmic training and testing. Bag-of-words and TF-IDF methods were utilised for traditional machine learning algorithms, while the Word2Vec method was employed for the RNNs. The Word2Vec method demonstrated the best performance in comparison with TF-IDF. This can be attributed to its ability to capture the semantic relationships between words, resulting in more meaningful feature representations for the classification tasks. In traditional machine learning algorithms, TF-IDF performed better than the BoW method. This can be attributed to its ability to weight terms based on their importance in an RFI corpus, thereby highlighting more distinctive features for classification tasks.

- **Performance enhancement using ensemble techniques:**

This study also examined the effectiveness of ensemble techniques—voting (simple and weighted majority) and boosting (AdaBoost)—compared to LR and GRU. AdaBoost, combined with LR, aimed to enhance LR's performance but decreased F-score and recall, possibly due to overfitting. GRU outperformed AdaBoost and simple majority voting. The weighted majority voting achieved the highest precision (89%), recall (85%), and F-score (86%), credited to leveraging diverse model combination.

- **Validation of the models:**

In line with the design science research framework, once a tangible solution is developed it needs to be tested. The validation mechanism adopted for addressing this question aligns with existing NLP studies that employ the train-test split method. From the total RFI dataset (2227 RFIs), 2045 RFIs were allocated for training and 227 RFIs for testing. After efficient hyperparameter tuning, this study achieved desirable scores and proceeded with further experimental validation (Section 4.5) for the practical implementation of the model. For instance, in a practical evaluation, RFIs are managed by an RFI manager who reviews and forwards them to the relevant teams. Accordingly, the experimental assessment involved classifying an unseen set of 40 RFIs by three subject matter experts, whose performance was then compared with the best-performing machine learning model. The machine learning model consistently outperformed human experts across all metrics: achieving 85% precision, 80% recall, and an 83% F-1 score, with a processing time of 47.2 seconds. In comparison, human

experts achieved 59% precision, 54% recall, and a 55.8% F-1 score, requiring significantly more time at 780 seconds.

- **Advanced visualisation using topic modelling:**

In Chapter 4, topic modelling was employed to uncover key topics and themes from RFIs. First, pre-processing of the RFI corpus was conducted using NLP techniques and then unsupervised learning model based on LDA algorithm was iteratively trained to identify ten optimal topics, each associated with construction-related keywords. Visualisations were generated using Gensim, NLTK, and pyLDAvis libraries. Additionally, the t-SNE algorithm visualised topic proximity, offering valuable insights into RFI issues. These visualisations provide industry stakeholders with a clear understanding of prevalent RFI topics, inspiring future research endeavours.

6.1.3 Objective 3: Developing a domain-specific issue classification model, followed by creating an NER model to efficiently extract key entities from RFIs for automated information extraction.

The third objective also involved developing two distinct NLP models: an issue classification model from RFIs, and a named entity recognition model for extracting key entities from RFIs. As part of the model development process, this phase required iterative stages of model creation and testing using real construction project data.

- **CNN-driven issue classification of RFIs:**

In Chapter 5, advanced information extraction techniques were integrated, focusing on a two-stage approach to develop models for extracting key issues from RFIs. The first model employed a deep learning-based CNN algorithm (Section 5.3 - model architecture) to identify "coordination", "constructability", "design/drawing discrepancies", "scope/specification clarification", and "review/approval"—critical issues typically central to an RFI. Hyperparameter tuning enhanced the CNN's performance, which was evaluated through metrics like precision, recall, and F1 score, achieving an overall accuracy of 72%. Addressing class imbalance within the RFIs involved applying various techniques to mitigate overfitting, culminating in the optimal model. This model establishes a foundational benchmark for future studies examining RFI issue extraction methodologies.

- **NER model development using BERT and BiLSTM:**

In the next phase of this chapter, a novel model was developed utilising a BiLSTM-based RNN architecture and BERT to extract key entities from RFIs. This marks the first application of an NER model explicitly tailored for RFIs. The identified key entities include "problematic components" (e.g., walls, floors, tiling), "location" (e.g., academic block, canteen), and "drawing reference" (e.g., DWG-001-0009-000A, DWG-00001). These entities are crucial as they provide decision-makers with essential information at a glance, facilitating quicker RFI resolution and information flow. The data underwent pre-processing and was transformed into a machine-readable format using distinct text representation techniques for both BiLSTM and BERT. Subsequently, the models were trained, tested, and evaluated on precision, recall, and F1 score metrics. Comparing BiLSTM and BERT, the BiLSTM model demonstrated superior performance in terms of F1 score.

- **Improved NER model through CRF layer integration:**

To enhance the developed models further, this research integrated a CRF layer into the BERT and BiLSTM architectures (Section 5.5), resulting in BERT-CRF and BiLSTM-CRF configurations, respectively. This integration notably boosted the overall performance scores of both models. The BERT-CRF achieved an impressive F1 score of 97%, while the BiLSTM-CRF achieved 80%. These improvements can be attributed to the CRF layer's ability to model sequential dependencies and enhance the models' accuracy in entity recognition tasks, ensuring more precise identification and labelling of key entities within RFIs.

- **Validation of the developed models:**

The validation of the models was conducted using the train-test split method, where 90% of the RFI dataset was allocated for training the models, and the remaining 10% was reserved for testing. The same strategy was applied to both models developed to address this question: the issue classification model using CNN and BERT/BiLSTM-based NER models for key entity extraction. This method is widely recognised and aligns with best practices in studies that employ text mining methods for analysing unstructured documents within the domain's body of knowledge. By employing this approach, the study ensures robustness and reliability in evaluating the performance and generalisability of the developed models for RFI analysis, aligning with the established design science research framework.

6.2 Original contributions and significance of research

This research delivers original contributions that hold significance for both the academic community and industry stakeholders. The novelty of this study lies in its domain-specific application of deep learning and natural language processing to automate the classification and information extraction from construction request for information documents, which has been largely overlooked in existing literature.

Three distinct models were developed: a phase-wise separation model, an issue classification model, and a named entity recognition model. Although these models build upon existing deep learning architectures such as CNN, BiLSTM, and BERT, their tailored application to RFIs—along with the creation of a novel real-world dataset—constitutes a unique and meaningful advancement. These models set a new benchmark for baseline performance in this domain, offering foundational resources and methodological direction for future research in construction informatics.

The study also contributes methodologically by evaluating various feature representation and ensemble learning techniques, which remain underexplored in construction NLP research. Furthermore, it advances theoretical understanding by uncovering patterns in the RFI process, elucidating the nature and gravity of construction-related issues, and proposing an intelligent NLP-based framework for automating traditionally manual processes. Collectively, these contributions address a critical gap in the literature and provide significant opportunities for further development and academic inquiry.

6.3 Implications of the thesis

This research study provides substantial contributions from both scientific and applied standpoints. The following subsections discuss the theoretical and practical contributions of this research thesis.

6.3.1 Theoretical implications

This study provides a pioneering and comprehensive investigation into the potential of deep learning algorithms and various NLP techniques for automated text classification and information extraction from RFIs. While NLP has made significant strides in IT, its application in construction documentation remains domain-specific and relatively uncommon. By

leveraging deep learning-based RNNs and CNNs for text classification, and BiLSTM and BERT for NER, this research advances the field and paves the way for more sophisticated methods tailored to the construction domain. While these models utilize established algorithms, their application in this context leads to the development of novel models that serve as baseline references for future research. All models are built on a novel corpus of RFIs from real construction projects, specifically focusing on RFIs from the delivery phase. Previous studies have largely overlooked this phase, making this study a crucial contribution by addressing this gap. Future research can build upon this foundation by utilizing the dataset, strategies, and algorithms proposed herein to enhance this work and extend its applicability to other construction-related datasets. Lastly, this study explores different feature representation techniques to drive machine learning and deep learning algorithms on an RFI corpus. Furthermore, ensemble techniques have not been widely utilised in construction-related research; this study also contributes by evaluating different ensemble approaches in the RFI dataset.

Lastly, the developed models significantly advance the existing body of knowledge by proposing a pilot model that represents a substantial leap forward from manual content analysis of RFIs to a more unbiased, faster, accurate, advanced, and automated approach. Another theoretical benefit of this research is that it provides a deep understanding of the existing RFI process, including the associated challenges, risks, and gravity of issues, as well as the solutions available within academic and industry settings. Researchers can use this study to gain insights into the RFI process and follow the directions suggested to further streamline the RFI process.

6.3.2 Practical implications

The developed models in this research offer unique value to industry stakeholders. For example, phase-wise separation allows project teams to confine issues within a specific phase, ensuring they do not re-emerge in subsequent phases. This phase-wise identification can also aid RFI managers in automatically routing RFIs to the appropriate departments, such as design, execution, or procurement. Further for ongoing projects, stakeholders can actively determine the sources of RFIs and adjust their design, execution, and procurement methods to reduce the number of RFIs. The issue identification model, which employs CNN, helps project teams identify the origins of their issues mentioned in the RFIs. This capability facilitates knowledge acquisition, and the application of lessons learned, enabling informed decision-making. Action measures can be devised based on the criticality or risk posed by specific issues. For example,

coordination issues often necessitate coordination meetings. Automatically identifying these issues helps stakeholders address them promptly by implementing the action plans suggested by the RFI manager. This proactive approach ensures earlier collaboration and resolution. Next, the NER model identifies problematic components, locations, and drawing references, providing critical information that helps teams understand the source of issues and locate relevant drawings, with all the necessary details to the issues. By addressing issues at their root, teams can prevent new cycles of RFIs from starting.

Project managers can use this data to develop executive summaries and receive daily updates on unresolved issues from each phase. Acting on this information enables them to take mitigating measures before issues become part of the critical path. This knowledge can also be used to improve the quality of future documents, such as drawings, specifications, and contract documents, leading to a reduction in the number of RFIs. For practitioners, integrating automated models into existing RFI workflows, whether via email exchanges or common data environments, offers numerous benefits. These text analyses empower RFI managers, project managers, and teams to identify problematic areas, expedite reviews, and proactively mitigate issues, ultimately improving project efficiency. Furthermore, existing CDEs and BIM-based platforms generate substantial amounts of data, yet they often underutilize this data for efficient project management. The NLP pipeline proposed in this paper demonstrates how data already generated and stored within existing CDE platforms can be leveraged to derive actionable insights. Utilizing this vast data can inform future projects, facilitating the creation of design plans and specifications that incorporate lessons learned and proactively address potential issues. The most significant advantage of these developments is their potential integration into existing CDE platforms through plugins. This integration would allow the utilization of big data, enabling the generation of valuable insights for project teams. Lastly, this research provides a detailed roadmap that can be utilized by researchers and industry practitioners on how the developed models can be integrated into construction settings. This roadmap is supported by the proposed dashboard discussed in subsequent section, which can be incorporated into both traditional RFI management systems and more modern BIM/CDE-enabled RFI management systems.

6.4 Roadmap for utilising developed models in traditional and CDE-driven RFI management

The current RFI process is predominantly managed through common data environment platforms. In regions where technological adoption is limited, RFIs are often exchanged in a simpler manner, mostly through email exchanges. This section serves as a blueprint and guide for unifying these models into a central dashboard. This dashboard can be integrated within both traditional construction management settings and more sophisticated environments utilizing BIM platforms and CDEs. The graphical representation of the proposed dashboard is presented though Fig. 6.1. It must be noted that this dashboard serves as a visual representation, and its actual construction is designated as the scope for future research.

6.4.1 Situating the model within email-based and CDE-cased RFI exchange

For traditional setups where RFIs are exchanged through email, often due to factors like lack of software, awareness, or affordability, the developed dashboard can be implemented as is. An RFI manager, for instance, can install the system as a standalone software product and use it in conjunction with email exchanges. With features such as email embedding for receiving and tracking RFIs, this model can function as a lightweight CDE. Once integrated into the system, both contractors and responders/consultants can utilize it to benefit from its analytical capabilities. For construction environments where advanced technologies are utilised the system can be converted into a plugin to be embedded into the CDE. For example, Autodesk

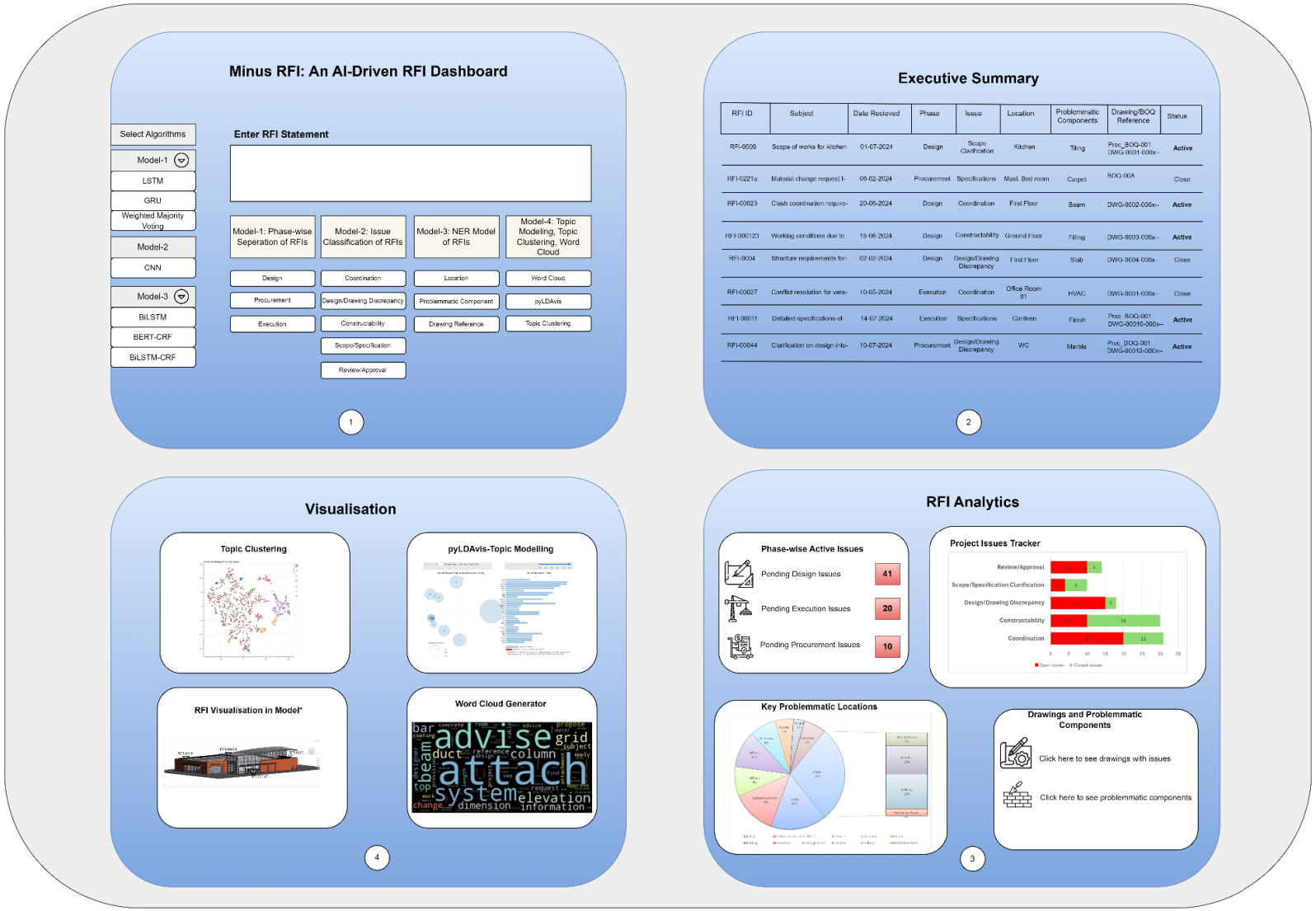


Figure 6-1. Proposed end product for academic and industry stakeholders.

Construction Cloud facilitates API integrations through their Partner Program. Similar approaches can be adopted by other platforms to enhance RFI handling and tracking. The following sub-section will guide the development of the system.

6.4.2 Step-by-step actions for centralised model development:

Step 1: Model integration

1) **Model deployment**: Containerize NLP models (Docker).
2) **Model API development**: Implement RESTful APIs (FastAPI/Flask).
3) **Model orchestration**: Manage with Kubernetes.

Step 2: CDE development

1) **Backend development**: API endpoints, data storage integration.
2) **Frontend development**: User-friendly dashboard, visualization components.

Step 3: Dashboard and analytics

1) **Analytics integration**: Topic modelling, issue tracker, NER, phase-wise classification
2) **Dashboard features**: Filters, search, drill-down capabilities, close/active RFIs, BIM integration.

Step 4: Testing and validation

1) **Unit resting**: Component and API testing.
2) **Integration testing**: Backend, NLP models, frontend interaction.
3) **User acceptance testing**: Validate functionality and usability.

Step 5: Deployment and monitoring

1) **Deployment**: Cloud platform deployment.
2) **Monitoring and maintenance**: Performance tracking, regular updates.

Step 6: Documentation and training

1) **Documentation**: System, APIs, user guides.

2) **Training**: User training sessions.

Step 7: Continuous improvement

1) **Feedback Loop**: Gather user feedback.

2) **Iterative Development**: CI/CD pipelines for updates.

6.5 Limitations of the study and suggestions for future research:

The current thesis highlights the limitations of the study derived from the answers to each research question.

1 **Preliminary CLD for RFI process challenges:** This review paper introduces a preliminary CLD to map various challenges associated with the RFI process. Due to limited documented relationships in the literature, this remains a qualitative conceptualisation. Future research should integrate industry feedback to develop a system dynamics model, quantitatively establishing the missing links between different variables. A robust system dynamics model will provide both the academic community and industry professionals with a comprehensive understanding of the causal relationships of challenges/risks within the RFI process.

2 **Leveraging advancements in NLP:** While the current thesis employs effective techniques, recent advancements in the NLP space, particularly with the emergence of large language models, offer new opportunities. LLMs can be used to develop more effective text mining models with relatively less training data and improved accuracy. The current models serve as baseline models; future research can build on these to extract the same or other key information from RFIs with higher accuracy. Additionally, this research involved manual conversion of text to a machine-readable format. Future studies could employ state-of-the-art Optical Character Recognition (OCR) algorithms to automate the extraction of information from RFIs, followed by further pre-processing to prepare the data for algorithms.

3 **Development of a unified RFI dashboard and a plugin:** Another limitation of this study is the lack of a visually appealing dashboard or plugin that can be utilised by both traditional paper-based RFI processing setups and more advanced CDE-based setups. Future research can develop an interface using Python libraries to exhibit all the algorithms, integrating and presenting their results in a user-friendly manner for those without machine learning expertise. This dashboard could be used as-is in email-based RFI exchange setups and integrated as a plugin or API with CDE platforms that enable RFI handling and management.

4 **Enhanced RFI analysis:** As discussed in Chapter 2, this thesis focuses on incorporating and extracting specific information or entities from the unstructured content of RFIs. Future research could train algorithms and develop models to estimate critical factors such as the criticality of RFIs and their impact on schedule and cost. While these assessments are currently available, they often rely on the requestor's intuition. Additionally, future research could target key entities such as actions for the reviewer, specifications/standards, and required information.

5 **Training datasets:** This thesis marks the beginning of a new research direction by integrating text mining into RFI document analysis. It has developed several NLP models for extracting information from project RFIs, achieving high scores in its initial efforts. However, further advancements are needed. These models should be trained on large datasets encompassing various project types such as infrastructure, residential, commercial, and industrial, ensuring broad applicability across different project scenarios. Moreover, the models should be tested in real-time settings within both BIM and non-BIM/CDE environments to assess their effectiveness in assisting RFI managers or project teams in promptly responding to and resolving RFIs.

References

Abdal Noor, B. and Yi, S. (2018), Review of BIM literature in construction industry and transportation: meta-analysis*, Construction Innovation*, Vol. 18 No. 4, pp. 433-452.

Abdel-Monem, M., & Hegazy, T. (2013). Enhancing Construction As-Built Documentation Using Interactive Voice Response. *Journal of Construction Engineering and Management*, 139(7), 895–898. https://doi.org/10.1061/(ASCE)co.1943-7862.0000648.

Abu Sheikha, F., & Inkpen, D. (2010). Automatic classification of documents by formality. *In Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 2010*. New York: IEEE.

Adil, M., Wu, J. Z., Chakrabortty, R. K., Alahmadi, A., Ansari, M. F., & Ryan, M. J. (2021). Attention-based STL-BiLSTM network to forecast tourist arrival. *Processes*, 9(10), 1759.

Afzal, M., Wong, J. K. W., & Ahmadian Fard Fini, A. (2024). Towards digital approach for managing request for information (RFI) in construction projects: a literature review. *Construction Innovation.*

Afzal, M., Wong, J. K. W., & Fini, A. A. F. (2023, August). Unlocking Insights: Analysing Construction Issues in Request for Information (RFI) Documents with Text Mining and Visualisation. In *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)* (pp. 1-6). IEEE.

Ahmadzadeh, E., Kim, H., Jeong, O., Kim, N., & Moon, I. (2022). A deep bidirectional LSTM-GRU network model for automated ciphertext classification. *IEEE Access*, 10, 3228-3237.

AIB (2005). *Getting it right the first time: A plan to improve standards in project documentation within the building and construction industry*, Australian Institute of Building, Canberra, ACT

Aibinu, S., Carter, S., Francis, V., & Vaz-Serra, P. (2019). Request for information frequency and their turnaround time in construction projects. *Built Environment Project and Asset Management*, 10(1), 1-15.

Alexander, J., Ackermann, F., & Love, P. E. (2019). Taking a holistic exploration of the project life cycle in public–private partnerships. *Project Management Journal*, 50(6), 673-685.

Ali, B., Aibinu, A.A. and Paton-Cole, V. (2022), Closing the information gaps: a systematic review of research on delay and disruption claims, *Construction Innovation.*

Ali, N., Chen, S. S., Srikonda, R., & Hu, H. (2014). Development of concrete bridge data schema for interoperability. *Transportation Research Record*, 2406(1), 87-97. https://doi.org/10. 3141/2406-10.

Alizadehsalehi, S., Hadavi, A., & Huang, J. C. (2019, June). BIM/MR-Lean construction project delivery management system. In *2019 IEEE Technology & Engineering Management Conference (TEMSCON)* (pp. 1-6). IEEE. https://doi: 10.1109/TEMSCON.2019.8813574.

Alreshidi, E., Mourshed, M., & Rezgui, Y. (2017). Factors for effective BIM governance. *Journal of Building Engineering*, 10, 89-101. doi: 10.1016/j.jobe.2017.02.006.

Alves, T. D. C., Pestana, A. C. V., Gilbert, E., & Hamzeh, F. (2016). Lean principles for the management of construction submittals and RFIs. *Journal of Professional Issues in Engineering Education and Practice*, 142(4), 05016004. https://doi: 10.1061/(ASCE)EI.1943-5541.0000285.

American Institute of Steel Construction. (2016). *Code of standard practice for steel buildings and bridges*. American Institute of Steel Construction.

American Institute of Architects. (2007). *Integrated project delivery: A guide*. http://www.aia.org/ipdg

American Institute of Architects, California Council (AIACC). (2014). *Integrated project delivery: An updated working definition.*

Andrews, W. (2005). RFI recommendations. *Modern Steel Construction*, 46(10), pp. 37-41

Avison, D., Baskerville, R., & Myers, M. (2001). Controlling action research projects. *Information Technology & People*. 14(1), 28–45. https://doi.org/10.1108/09593840110384762

Ayyasamy, R. K., Tahayna, B., Alhashmi, S., Eugene, S., and Egerton, S.(2010). Mining Wikipedia knowledge to improve document indexing and classification. *Proc., 10th Int. Conf. on Info. Science, Signal Processing and their Applications (ISSPA*), Vol. 10, IEEE, Washington, DC, 806–809.

Azhar, S., Ahmad, I., & Sein, M. K. (2010). Action research as a proactive research method for construction engineering and management. *Journal of construction engineering and management*, 136(1), 87-98.

Bademosi, F. M., & Issa, R. R. (2022). Automation and Robotics Technologies Deployment Trends in Construction. *Automation and Robotics in the Architecture, Engineering, and Construction Industry*, 1-30.

Baek, S., Han, S. H., & Jung, W. (2023). Automated identification of active players for international construction market entry using natural language processing. *Journal of Management in Engineering*, 39(5), 04023025.

Baek, S., Jung, W., & Han, S. H. (2021). A critical review of text-based research in construction: Data source, analysis method, and implications. *Automation in Construction*, 132, 103915. https://doi.org/10.1016/j.autcon.2021.103915.

Baker, H., Hallowell, M. R., & Tixier, A. J. P. (2020). Automatically learning construction injury precursors from text. *Automation in Construction*, 118.

Barlish, K., & Sullivan, K. (2012). How to measure the benefits of BIM—A case study approach. *Automation in construction*, 24, 149-159., http://dx.doi.org/10.1016/j.autcon.2012.02.008.

Baskerville, R. (1999). Investigating information systems with action research. *Communications of the association for information systems*, *2*(1), 19.

Baskerville, R., & Myers, M. D. (2004). Special issue on action research in information systems: Making IS research relevant to practice: Foreword. *MIS Quarterly*, 329-335.

Baskerville, R. (2008). What design science is not. *European Journal of Information Systems*, 17(5), 441-443. https://doi.org/10.1057/ejis.2008.45

Bhat, A. S., Poirier, E. A., & French, S. S. (2017). Investigating the potential of BIM to address project delivery issues. *6th CSCE-CRC International Construction Specialty Conference 2017* - Held as Part of the Canadian Society for Civil Engineering Annual Conference and General Meeting 2017, 2, 955-964.

Bilbo, D., Bigelow, B., Escamilla, E., & Lockwood, C. (2015). Comparison of construction manager at risk and integrated project delivery performance on healthcare projects: A comparative case study. *International Journal of Construction Education and Research*, 11(1), 40-53. https://doi: 10.1080/15578771.2013.872734.

Birnbaum, S., Kuleshov, V., Enam, Z., Koh, P. W. W., & Ermon, S. (2019). Temporal FiLM: Capturing Long-Range Sequence Dependencies with Feature-Wise Modulations. *In Advances in Neural Information Processing Systems*, 32.

Bodenbenner, P., Feuerriegel, S., & Neumann, D. (2013). Design science in practice: designing an electricity demand response system. In *Design Science at the Intersection of Physical and Virtual Design: 8th International Conference, DESRIST 2013, Helsinki, Finland, June 11-12, 2013. Proceedings 8* (pp. 293-307). Springer Berlin Heidelberg.

Bouziane, H., Messabih, B., & Chouarfia, A. (2011). Profiles and majority voting-based ensemble method for protein secondary structure prediction. *Evolutionary Bioinformatics*, 7, 171-189.

Brandon, P., & Lu, S. L. (2008). *Clients Driving Innovation*. Willy-Blackwell Publishing Ltd.

Brazee, A. (2014). The Anatomy of a Request for Information (RFI). http://blog.procore.com/blog/bid/371063/The-Anatomy-of-a-Request-For-Information-RF

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45(1), 12–19.

Bühlmann, P. (2012). Bagging, boosting and ensemble methods. *Handbook of computational statistics: Concepts and methods*, 985-1022.

Cai, L., Song, Y., Liu, T., & Zhang, K. (2020). A hybrid BERT model that incorporates label semantics via adjustive attention for multi-label text classification. *IEEE Access*, 8, 152183-152192.

Candaş, A. B., & Tokdemir, O. B. (2022). Automating coordination efforts for reviewing construction contracts with multilabel text classification. *Journal of Construction Engineering and Management,* 148(6), 04022027.

Castronovo, F., Awad, B., & Akhavian, R. (2018). Implementation of virtual design reviews in the generation of as-built information. *Construction Research Congress 2018* (pp. 285-294).

Chan, A.P.C., Chan, D.W.M. and Yeung, J.F.Y. (2009), Overview of the application of 'fuzzy techniques' in construction management research*, J. Constr. Eng. Manage*., Vol. 135 No. 11, pp. 1241-1252, doi: 10.1061/(ASCE)CO.1943-7862.0000099.

Cheng, M. Y., Kusoemo, D., & Gosno, R. A. (2020). Text mining-based construction site accident classification using hybrid supervised machine learning. *Automation in Construction*, 118, 103265.

Chin, C-S and Russell, J S (2008) Predicting the expected service level and the realistic lead time of RFI process using binary logistic regression In: *Dainty, A (Ed) Procs 24th Annual ARCOM Conference*, 1-3 September 2008, Cardiff, UK, Association of Researchers in Construction Management, 739-748.

Chung, S., Moon, S., Kim, J., Kim, J., Lim, S., & Chi, S. (2023). Comparing natural language processing (NLP) applications in construction and computer science using preferred reporting items for systematic reviews (PRISMA). *Automation in Construction*, 154, 105020.

Daoud, O. E., & Allouche, E. N. (2003). Bid queries as a gauge for quality control of documents. *Proceedings of the Canadian Society for Civil* Engineering, Moncton, NB, Canada, 4-7.

Das, M., Tao, X. and Cheng, J.C.P. (2020), A secure and distributed construction document management system using blockchain, in Toledo Santos, E., Scheer, S. (Eds), *Proceedings of the 18th International Conference on Computing in Civil and Building Engineering. ICCCBE 2020. Lecture Notes in Civil Engineering*, Vol. 98, pp. 652-659.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding.* Preprint, submitted October 11, 2018. http://arxiv.org/abs/1810.04805

Donato, V., Lo Turco, M., & Bocconcino, M. M. (2017). BIM-QA/QC in the architectural design process. *Architectural Engineering and Design Management*, 14(3), 239-254. https://doi: 10.1080/17452007.2017.1370995.

Downe-Wamboldt, B. (1992). Content analysis: method, applications, and issues. *Health Care for Women International,* 13(3), 313-321.

Ekanayake, B. J. (2022). *Improving automation in computer vision-based indoor construction progress monitoring: A deep learning approach* (Doctoral dissertation, University of Technology Sydney). ProQuest Dissertations Publishing.

Ekanayake, B., Wong, J. K. W., Fini, A. A. F., Smith, P., & Thengane, V. (2024). Deep learning-based computer vision in project management: Automating indoor construction progress monitoring. *Project Leadership and Society*, 5, 100149.

Ekström, M. A., & Björnsson, H. C. (2005). Valuing flexibility in architecture, engineering, and construction information technology investments. *Journal of construction engineering and management*, 131(4), 431-438.

El Asmar, M., Hanna, A. S., & Loh, W. Y. (2013). Quantifying performance for the integrated project delivery system as compared to established delivery systems. *Journal of construction engineering and management*, 139(11), 04013012.

Ellis, G. Construction RFI: A Comprehensive Guide [And Template] - Digital Builder. [online] Digital Builder (2022). Available at: <https://constructionblog.autodesk.com/construction-rfi-guide-template/> [Accessed 1 March 2022].

Fan, S. L., Skibniewski, M. J., & Hung, T. W. (2014). Effects of building information modeling during construction. *Journal of Applied Science and Engineering*, 17(2), 157-166.

Fang, W., L. Ding, P. E. Love, and C. Zhou. (2020). Computer vision ap-plications in construction safety assurance. *Autom. Constr*.110 (Feb):103013.https://doi.org/10.1016/j.autcon.2019.103013

Fang, W., Luo, H., Xu, S., Love, P. E., Lu, Z., & Ye, C. (2020). Automated text classification of near misses from safety reports: An improved deep learning approach. *Advanced Engineering Informatics*, 44, 101060.

Filho, J.B.P.D., Angelim, B.M., Guedes, J.P., Silveira, S.S. and Neto, J.P.B. (2016a), Constructability analysis of architecture-structure interface based on BIM, IGLC 2016 – *24th Annual Conference of the International Group for Lean Construction*, pp. 73-82.

Filho, J.B.P.D., Angelim, B.M., Guedes, J.P., De Castro, M.A.F. and Neto, J.D.P.B. (2016b),Virtual design and construction of plumbing systems, *Open Engineering*, Vol. 6 No. 1, pp. 730-736

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and computation*, 121(2), 256-285.

Gajendran, T., and Brewer, G. (2012). Collaboration in public sector projects: unearthing the contextual challenges posed in project environments. *Engineering Project Organization Journal*, 2(3), 112-126.

Gao, X., & Pishdad-Bozorgi, P. (2019). BIM-enabled facilities operation and maintenance: A review. *Advanced engineering informatics*, 39, 227-247.

Giel, B. K., & Issa, R. R. (2013). Return on investment analysis of using building information modeling in construction. *Journal of computing in civil engineering*, 27(5), 511-521. https://doi: 10.1061/(asce)cp.1943-5487.0000164.

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS quarterly*, 337-355.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267-297.

Groen, E. C., Schowalter, J., Kopczynska, S., Polst, S., & Alvani, S. (2018). Is there Really a Need for Using NLP to Elicit Requirements? A Benchmarking Study to Assess Scalability of Manual Analysis. In *REFSQ Workshops.*

Gunter, U.; Önder, I. (2016). Forecasting city arrivals with Google Analytics. *Ann. Tour. Res*. 2016, 61, 199 212.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., S. Lew, M., (2015). Deep learning for visual understanding: A review. *Neurocomputing* 187, 27–48. http://dx.doi.org/10. 1016/j.neucom.2015.09.116.

Hanna, A. S. (2016). Benchmark performance metrics for integrated project delivery. *Journal of Construction Engineering and Management*, 142(9), 04016040. http://dx.doi.org/10.1061/(ASCE)CO.1943- 7862.0001151

Hanna, A. S., E. J. Tadt, and G. C. Whited. (2012). Request for information: Benchmarks and metrics for major highway projects. *J. Constr. Eng. Manage*. 138 (12): 1347–1352. https://doi.org/10.1061/(ASCE)CO .1943-7862.0000554.

Hasan, A., Baroudi, B., Elmualim, A. and Rameezdeen, R. (2018), Factors affecting construction productivity: a 30-year systematic review, *Engineering, Construction and Architectural Management*, Vol. 25 No. 7.

Hassan, F. U., & Le, T. (2020). Automated requirements identification from construction contract documents using natural language processing. *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, 12(2).

Hassan, F. U., & Le, T. (2021). Computer-assisted separation of design-build contract requirements to support subcontract drafting. *Automation in Construction*, 122.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 75-105.

Higgins, Jr., D. , Fryer, S. , Stratton, R. , Simpson, D. & Reginato, J. (2012). Using the Forward Thinking Index to Reduce Delays Related to Request for Information Process, *20th Annual Conference of the International Group for Lean Construction*.

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

Hooper, B. and Haris, M. (2010). *2020 Vision* , Royal Institution of Chartered Surveyors, London (2010).

Hosmer, D. W., & Lemeshow, S. (2000). Applied logistic regression. *In Applied Logistic Regression* (pp. 118-128). https://doi.org/10.1080/00401706.1992.10485291

Hughes, N., Wells, M., Nutter, C. L., and Zach, J.G. (2013). *Impact & Control of RFIs on Construction Projects*, NAVIGANT, Chicago, IL, https://www.cmaanet.org/sites/default/files/resource/Impact%20%26%20Control%20of%20RFIs%20on%20Construction%20Projects.pdf

Ibrahim, M.W., Hanna, A.S., Russell, J.S., Abotaleb, I.S. and El-Adaway, I.H. (2020), Comprehensive analysis of factors associated with out-of-sequence construction, *Journal of Management in Engineering*, Vol. 36 No. 4.

Iivari, J., & Venable, J. R. (2009). Action research and design science research- Seemingly similar but decisively dissimilar [Paper presentation]. *17th European Conference on Information Systems (ECIS)*, Verona, Italy.

Iqbal, M., Karim, A., & Kamiran, F. (2019). Balancing prediction errors for robust sentiment classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(3), 1-21. https://doi. org/10.1145/332879

Issa, R. R., Fukai, D., & Danso-Amoako, M. O. (2003). Evaluation of computer anatomic modelling for analysing pre-construction problems. In *Construction Research Congress: Wind of Change: Integration and Innovation* (pp. 1-8).

Jarkas, A. M., & Bitar, C. G. (2012). Factors affecting construction labour productivity in Kuwait. *Journal of Construction Engineering and Management*, 138, 811–820.

Jarkas, A.M., Radosavljevic, M. and Wuyi, L. (2014), Prominent demotivational factors influencing the productivity of construction project managers in Qatar, *International Journal of Productivity and Performance Management*, Vol. 63 No. 8, pp. 1070-1090.

Jarkas, A. M. (2015). Factors influencing labour productivity in Bahrain's construction industry. *International journal of construction management*, 15(1), 94-108.

Jarkas, A. M., Al Balushi, R. A., & Raveendranath, P. K. (2015). Determinants of construction labour productivity in Oman. *International Journal of Construction Management*, 15(4), 332-344. https://doi: 10.1080/15623599.2015.1094849.

Järvinen, P. (2007). Action research is similar to design science. *Quality & Quantity*, 41(1), 37-54. https://doi.org/10.1007/s11135-005-5427-1

Jaskula, K., Papadonikolaki, E. and Rovas, D. (2023), Comparison of current common data environment tools in the construction industry, *2023 European Conference on Computing in Construction, presented at the 2023 European Conference on Computing in Construction*, doi: 10.35490/EC3.2023.315.

Jaskula, K., Kifokeris, D., Papadonikolaki, E., & Rovas, D. (2024). Common data environments in construction: state-of-the-art and challenges for practical implementation. *Construction Innovation.*

Jayashree, R., & Srikanta, M. K. (2011). An analysis of sentence-level text classification for the Kannada language. In *Proceedings of the 2011 International Conference of Soft Computing and Pattern Recognition, SoCPaR 2011* (pp. 147-151).

Jeon, K., G. Lee, S. Yang, and H. D. Jeong. (2022). Named entity recognition of building construction defect information from text with linguistic noise. *Autom. Constr*. 143 (Nov): 104543. https://doi.org/10 .1016/j.autcon.2022.104543

Joachims, T. (1998). Text categorization with SVM: Learning with many relevant features. In *European Conf. on Machine Learning*, 137–142. Berlin: Springer.

Johannesson, P., & Perjons, E. (2014). *An introduction to design science*. Springer.

Ju, M., Miwa, M., Ananiadou, S. (2018). A neural layered model for nested named entity recognition. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, Volume 1 (Long Papers). ACL, pp. 1446–1459.

Kähkönen, K., & Rannisto, J. (2015). Understanding fundamental and practical ingredients of construction project data management. *Construction Innovation*, 15(1), 7-23. https://doi: 10.1108/CI-04-2014-0026

Kang, T., Perotte, A., Tang, Y., Ta, C., Weng, C. (2021). UMLS-based data augmentation for natural language processing of clinical research literature. *J. Amer. Med. Inform. Assoc*. 28 (4), 812–823.

Karasu, T., Aaltonen, K. and Haapasalo, H. (2022), The interplay of IPD and BIM: a systematic literature review*, Construction Innovation*, Vol. 23 No. 3, doi: 10.1108/CI-07-2021-0134.

Kasanen, E., Lukka, K., & Siitonen, A. (1993). The constructive approach in management accounting research. *Journal of management accounting research*, 5(1), 243-264.

Kelly, D., & Ilozor, B. (2020). Performance outcome assessment of the integrated project delivery (IPD) method for commercial construction projects in USA. *International Journal of Construction Management*, 0(0), 1–9. https://doi.org/10.1080/15623599.2020.1827340

Khalef, R., & El-adaway, I. H. (2021). Automated identification of substantial changes in construction projects of airport improvement program: Machine learning and natural language processing comparative analysis. *Journal of management in engineering*, 37(6), 04021062.

Kim, T., & Chi, S. (2019). Accident case retrieval and analyses: Using natural language processing in the construction industry. *Journal of Construction Engineering and Management*, 145(3), 04019004.

Kim, J. J., Petrov, A. L., Lim, J., & Kim, S. (2021). Comparing cost performance of project delivery methods using quantifiable RFIs: cases in California heavy civil construction projects. *International Journal of Civil Engineering*, 20(3), 323-335.

Kontoghiorghes, L., & Colubi, A. (2023). New metrics and tests for subject prevalence in documents based on topic modelling. *International Journal of Approximate Reasoning*, 157, 49-69.

Krippendorff, K. (2013), *Content Analysis: An Introduction to Its Methodology*, 3rd ed. SAGE, Los Angeles

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Lam, S.W. (2014). The Singapore BIM Roadmap. In *Proceedings of the Government BIM Symposium 2014, Singapore*, 13 October 2014; Available online: http://bimsg.org/wp-content/uploads/2014/10/BIM-SYMPOSIUM_MR- LAM-SIEW-WAH_Oct-13-v6.pdf

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C. (2016). Neural architectures for named entity recognition. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. ACL, pp. 260–270.

Lee, P. C., & Su, H. N. (2010). Investigating the structure of regional innovation system research through keyword co-occurrence and social network analysis. *Innovation*, 12(1), 26-40. https://doi.org/10.5172/impp.12.1.26.

Lee, J., & Yi, J. S. (2017). Predicting project's uncertainty risk in the bidding process by integrating unstructured text data and structured numerical data using text mining. *Applied Sciences*, 7(11), 1141.

Li, J., & Kassem, M. (2021). Applications of distributed ledger technology (DLT) and Blockchain-enabled smart contracts in construction. *Automation in construction*, 132, 103955. https://doi.org/10.1016/j.autcon.2021.103955.

Li, R., T. Mo, J. Yang, D. Li, S. Jiang, and D. Wang. (2021). Bridge inspection named entity recognition via BERT and lexicon augmented machine reading comprehension neural model. *Adv. Eng. Inform.* 50 (Oct): 101416. https://doi.org/10.1016/j.aei.2021.101416.

Li, J., Shi, Y., & Li, S. (2024). Analysis of Beijing Traffic Violations Based on the BERT-CRF Model. *Promet – Traffic & Transportation,* 36(2), 279–293.

Liao, H., Tang, M., Luo, L., Li, C., Chiclana, F., & Zeng, X. J. (2018). A bibliometric analysis and visualization of medical big data research. *Sustainability*, 10(1), 166. https://doi.org/10.3390/su10010166

Liao, L., Teo, E.A.L., Chang, R. and Li, L. (2020), Investigating critical non-value adding activities and their resulting wastes in BIM-Based project delivery, *Sustainability*, Vol. 12 No. 1, p. 355.

Lindgren, R., Henfridsson, O., and Schultze, U. (2004). Design principles for competence management systems: A synthesis of an action research study. *MIS Q*., 283, 435–472.

Lindholm, A. L. (2008). A constructive study on creating core business relevant CREM strategy and performance measures. *Facilities*. 26(7), 343–358. https://doi.org/10.1108/02632770810877976

Liu, C., & Yang, S. (2023). A text mining-based approach for understanding Chinese railway incidents caused by electromagnetic interference. *Engineering Applications of Artificial Intelligence*, 117, 105598.

Liu, J., H. Luo, W. Fang, and P. E. Love. (2023). Contrastive learning framework for safety information extraction in construction of paper. *Adv. Eng. Inform*. 58 (Oct): 102194. https://doi.org/10.1016/j.aei.2023 .102194.

Liu, J., Love, P. E., Sing, M. C., Smith, J., & Matthews, J. (2017). PPP social infrastructure procurement: Examining the feasibility of a lifecycle performance measurement framework. *Journal of Infrastructure Systems*, 23(3), 04016041.

Loosemore, M. (2014). Improving construction productivity: a subcontractor's perspective. *Engineering, construction and architectural management*, 21(3), 245-260.

Love, P. E. D., Zhou, J., Sing, C. P., & Kim, J. T. (2014). Assessing the impact of RFIs in electrical and instrumentation engineering contracts. *Journal of Engineering Design*, 25, 4-6.

Lukka, K. (2003). The constructive research approach. In L. Ojala & O.-P, Hilmola (Eds.), *Case study research in logistics*. Publications of the Turku School of Economics and Business Administration, Series B, 1(2003) (pp. 83-101). Turku School of Economics and Business Administration.

Luo, X., Li, X., Song, X., & Liu, Q. (2023). Convolutional Neural Network Algorithm–Based Novel Automatic Text Classification Framework for Construction Accident Reports. *Journal of Construction Engineering and Management*, 149(12), 04023128.

Ma, P., Jiang, B., Lu, Z., Li, N., Jiang, Z., (2021). Cybersecurity named entity recognition using bidirectional long short-term memory with conditional random fields. *Tsinghua Sci. Technol*. 26 (3), 259–265.

Manning, C. D., C. D. Manning, and H. Schütze. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Manning, C., Raghavan, P., and Shutze, H. (2009). *An introduction to information retrieval*, Cambridge University Press, Cambridge, U.K.

Mao, W., Zhu, Y., & Ahmad, I. (2007). Applying metadata models to unstructured content of construction documents: A view-based approach. *Automation in construction*, 16(2), 242-252.

McGraw-Hill (2008), Building information modeling (BIM), transforming design and construction to achieve greater industry productivity, *Smart Market Report: Design and Construction Intelligence*. McGraw Hill Construction (2008)

McTaggart, R. (1991). Principles for participatory action research. *Adult Educ*. Q., 413, 168–187.

Maryland DOT (MDOT) (2013), State highway administration, Maryland department of transportation design-build manual

Merigó, J. M., Mas-Tur, A., Roig-Tierno, N., & Ribeiro-Soriano, D. (2015). A bibliometric overview of the Journal of Business Research between 1973 and 2014. *Journal of Business Research*, 68(12), 2645-2653.

Meyer, J. (2003). Questioning design and method: Exploring the value of action research in relation to R&D in primary care. *Prim. Health Care Res. Dev*., 42, 99–108.

Andi, & Minato, T. A. (2003). Design documents quality in the Japanese construction industry: factors influencing and impacts on construction process. *International Journal of Project Management*, 21(7), 537-546

Moon, S., Lee, G., Chi, S., & Oh, H. (2021). Automated construction specification review with named entity recognition using natural language processing. *Journal of Construction Engineering and Management*, 147(1), 04020147.

Morales, F., Herrera, R. F., Rivera, F. M.-L., Atencio, E., & Nuñez, M. (2022). Potential Application of BIM in RFI in Building Projects. *Buildings*, 12(2), 145.

Musarat, M. A., Hameed, N., Altaf, M., Alaloul, W. S., Al Salaheen, M., & Alawag, A. M. (2021, December). Digital Transformation of the Construction Industry: A Review. In *2021 International Conference on Decision Aid Sciences and Application (DASA)* (pp. 897-902). IEEE.

Naoum, S. G. (2001). *Dissertation research and writing for construction students*, Butterworth-Heinemann, Stoneham, Mass.

Nasrallah, W. F., & Bou-Matar, R. (2008). Exponential, gamma, and power law distributions in information flow on a construction site. *Journal of construction engineering and management*, 134(6), 442-450.

Nguyen, M. T., Le, D. T., & Le, L. (2020). Transformers-based information extraction with limited data for domain-specific business documents. *Engineering Applications of Artificial Intelligence*, 97, 104100.

Nigam, K., Lafferty, J., & McCallum, A. (1999). Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering* (pp. 61–67).

Noh, S. H. (2021). Analysis of gradient vanishing of RNNs and performance comparison. *Information*, 12(11), 442.

Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 57, 232-247.

Oyegoke, A. (2011). The constructive research approach in project management research. International *Journal of managing projects in business*, 4(4), 573-595.

Panahi, R., Kivlin, J. P., & Louis, J. (2023). Request for Information (RFI) Recommender System for Pre-Construction Design Review Application Using Natural Language Processing, Chat-GPT, and Computer Vision. In *Computing in Civil Engineering 2023* (pp. 159-166).

Papajohn, D., Alleman, D., Asmar, M. E., & Molenaar, K. (2018). Exploring potential delays associated with requests for information in CM/GC highway construction. In *Construction Research Congress 2018* (pp. 640-649).

Papajohn, D. and El Asmar, M. (2021), Impact of alternative delivery on the response time of requests for information for highway projects, *Journal of Management in Engineering*, Vol. 37 No. 1.

Papers with Code. (2022). Named entity recognition on CoNLL 2003 (English). https://paperswithcode.com /sota/named-entity-recognition-ner-on-conll-2003.

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

Philips-Ryder, M., Zuo, J. and Jin, X.H. (2013), Evaluating document quality in construction projects – subcontractors' perspective, *International Journal of Construction Management*, Vol. 13 No. 3, pp. 77-94.

Piirainen, K. A., & Gonzalez, R. A. (2013). Seeking constructive synergy: Design science and the constructive research approach. In *Design Science at the Intersection of Physical and Virtual Design: 8th International Conference, DESRIST 2013, Helsinki, Finland, June 11-12, 2013. Proceedings 8* (pp. 59-72). Springer Berlin Heidelberg.

Piroozfar, P., Farr, E. R., Zadeh, A. H., Inacio, S. T., Kilgallon, S., & Jin, R. (2019). Facilitating building information modelling (BIM) using integrated project delivery (IPD): A UK perspective. *Journal of Building Engineering*, 26, 100907. https://doi: 10.1016/j.jobe.2019.100907

Pohar, M., M. Blas, and S. Turk. (2004). Comparison of logistic regression and linear discriminant analysis: A simulation study. *Metodoloski Zvezki* 1 (1): 143–161.

Pradeep, A. S. E., Yiu, T. W., Zou, Y., & Amor, R. (2021). Blockchain-aided information exchange records for design liability control and improved security. *Automation in construction*, 126, 103667.

Priyadarshini, I., Alkhayyat, A., Obaid, A. J., & Sharma, R. (2022). Water pollution reduction for sustainable urban development using machine learning techniques. *Cities*, 130, 103970.

Prottasha, NJ., Sami, AA., Kowsher, M., Murad, SA., Bairagi, AK., Masud, M., Baz, M. (2022). Transfer learning for sentiment analysis using BERT based supervised fine-tuning. *Sensors* 22 (11), 4157. http://dx.doi.org/10.3390/s22114157.

Psomas, L., & Alzraiee, H. (2020). A Technology Platform for a Successful Implementation of Integrated Project Delivery for Medium Size Projects. *Proceedings of the 37th International Symposium on Automation and Robotics in Construction, ISARC 2020: From Demonstration to Practical Use - To New Stage of Construction Robot, Isarc*, 449–456. https://doi.org/10.22260/isarc2020/0063

Qader, W. A., Ameen, M. M., & Ahmed, B. I. (2019, June). An overview of bag of words; importance, implementation, applications, and challenges. In *2019 international engineering conference (IEC)* (pp. 200-204). IEEE.

Ramos, D. Free Construction RFI Templates and Forms | Smartsheet. [online] Smartsheet (2022) Available at: <https://www.smartsheet.com/content/construction-rfi-templates> [Accessed 1 March 2022].

Ren, R., and J. Zhang. (2021). Semantic rule-based construction procedural information extraction to guide jobsite sensing and monitoring. *J. Comput. Civ*. Eng. 35 (6): 04021026. https://doi.org/10.1061 /(ASCE)CP.1943-5487.0000971.

Romijnders, R., Warmerdam, E., Hansen, C., Welzel, J., Schmidt, G., Maetzler, W. (2021). Validation of IMU-based gait event detection during curved walking and turning in older adults and Parkinson's Disease patients. *J. Neuroeng. Rehabil*. 18 (1), 28. http://dx.doi.org/10.1186/s12984-021-00828-0.

Rosenberg, S. D., Schnurr, P. P., & Oxman, T. E. (1990). Content analysis: A comparison of manual and computerized systems. *Journal of Personality Assessment*, 54(1-2), 298-310.

Rowlinson, S. (2017), Building information modelling, integrated project delivery and all that, *Construction Innovation*, Vol. 17 No. 1, pp. 45-49.

Russel, S., and Norvig, P. (2010). *Artificial intelligence: A modern approach*, 3rd Ed., Prentice Hall, New York

Salama, D. M., & El-Gohary, N. M. (2013). Semantic text classification for supporting automated compliance checking in construction. *Journal of Computing in Civil Engineering*, 30(1), 04014106.

Sandoval, I. S. M., Bernardo, J. P. A. O., Dionisio, S. R., & Aquino, A. H. (2023, October). Application of Information and Communication Technologies (ICT) to optimise the request for information (RFI) process in construction projects. In *2023 Congreso Internacional de Innovación y Tendencias en Ingeniería (CONIITI)* (pp. 1-6). IEEE.

Sawyer, T. (2007). Innovative tools help companies cut the data beast down to size: Vendors offer a range of new strategies to capture and leverage enterprise data. *Engineering New Record*, 26, 26-28.

Schuldt, S. J., Jagoda, J. A., Hoisington, A. J., & Delorit, J. D. (2021). A systematic review and analysis of the viability of 3D-printed construction in remote environments. *Automation in Construction*, 125, 103642. https://doi: 10.1016/j.autcon.2021.103642.4

Sewak, M., Sahay, S. K., & Rathore, H. (2018). Comparison of deep learning and the classical machine learning algorithm for the malware detection. In *2018 19th IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD)* (pp. 293-296). IEEE.

Seyis, S., & Özkan, S. (2024). Analyzing the added value of common data environments for organizational and project performance of BIM-based projects. *J. Inf. Technol. Constr*., 29, 247-263.

Shin, C. S., Kim, K. I., Park, M. H., & Kim, H. J. (2000). Support vector machine-based text detection in digital video. In *Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop* (Vol. 2, pp. 634-641). IEEE.

Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE Access*, 7, 53040.

Shrestha, R., Ko, T., & Lee, J. (2023) Natural Language Processing (NLP)-Driven Classification of Pre-Bid Request for Information (RFI). In *Computing in Civil Engineering 2023* (pp. 59-66).

Simon, H.A. (1996). *The Sciences of the Artificial*, MIT press. https://doi.org/ 10.1126/science.165.3896.886-a.

Simpeh, E. K., Ndihokubwayo, R. and Love, P. E. (2011). Field diagnosis of causes and effects of rework in higher education residential facilities. *Journal of Construction*, 4(1), 17-24.

Soh, M.F., Barbeau, D., Dore, S. and Forgues, D. (2020), Qualitative analysis of request for information to identify design flaws in steel construction projects, *Organization, Technology and Management in Construction*, Vol. 12 No. 1.

Sompolgrunk, A., Banihashemi, S. and Mohandes, S.R. (2021), Building information modelling (BIM) and the return on investment: a systematic analysis, *Construction Innovation*, Vol. 23 No. 1.

Sparksman, P. J. (2015). Quantifying the time and cost associated with the request for information (RFI) or technical query (TQ) process: A designer's perspective (Master's thesis, University of Southern Queensland).

Sun, B., Chen, S., Wang, J., & Chen, H. (2016). A robust multi-class AdaBoost algorithm for mislabelled noisy data. *Knowledge-Based Systems*, 102, 87-102.

Susman, G., and Evered, R. (1978). An assessment of the scientific merits of action research. *Adm. Sci.* Q., 234, 582–603.

Tao, F., Liu, G., (2018). Advanced LSTM: A study about better time dependency modelling in emotion recognition. In: *2018 IEEE International Conference on Acoustics, and Signal Processing (ICASSP), April* (2018) 25-20. Calgary, CANADA, pp. 2906–2910. http://dx.doi.org/10.1109/ICASSP.2018.8461750.

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Application of machine learning to construction injury prediction. *Automation in construction*, 69, 102-114.

Tribelsky, E. and Sacks, R. (2011). An empirical study of information flows in multidisciplinary civil engineering design teams using lean measures. *Architectural Engineering and Design Management*, 7(2), 85-101

Vaishnavi, V., & Kuechler, W. (2004). *Design research in information systems*. CRC Press.

Vaswani, A., A. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. (2017). Attention is all you need. In *Proc., Advances in Neural Information Processing Systems 2017. San Diego: Neural Information Processing Systems*. https://dl.acm.org/doi/abs/10 .5555/3295222.3295349

Vidhya, K. A., & Aghila, G. (2010). A Survey of Naïve Bayes Machine Learning approach in Text Document Classification. *International Journal of Computer Science and Information Security*, 7, 206–211.

Vom Brocke, J., Hevner, A., & Maedche, A. (2020). *Introduction to design science research*. Design science research. Cases, 1-13.

Wadawadagi, R., & Pagi, V. (2020). Sentiment analysis with deep neural networks: comparative study and performance assessment. *Artificial Intelligence Review*, 53(8), 6155-6195.

Wang, H. (2012). Pattern classification with random decision forest. *In 2012 International Conference on Industrial Control and Electronics Engineering* (pp. 128-130). IEEE.

Wang, C.-C.; Chien, C.-H.; Trappey, A. (2021). On the Application of ARIMA and LSTM to Predict Order Demand Based on Short Lead Time and On-Time Delivery Requirements. *Processes*, 9, 1157.

Wang, Y., Sun, Y., Ma, Z., Gao, L., Xu, Y. (2021). Novel tools for the management, representation, and exploitation of textual information. *Sci. Program*. 2021, 8812754.

Watson, R. T., Boudreau, M.-C., & Chen, A. J.W. (2010). Information systems and environmentally sustainable development: Energy Informatics and new directions for the IS community. *MIS Quarterly*, 34(1), 23–38.

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

Wu, L. T., Lin, J. R., Leng, S., Li, J. L., & Hu, Z. Z. (2022). Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web. *Automation in Construction*, 135, 104108.

Xu, Z., & Coors, V. (2012). Combining system dynamics model, GIS and 3D visualization in sustainability assessment of urban residential development. *Building and Environment*, 47, 272-287. https://doi.org/10.1016/j.buildenv.2011.07.012.

Yan, H., Yang, N., Peng, Y. and Ren, Y. (2020), Data mining in the construction industry: present status, opportunities, and future trends, *Automation in Construction*, Vol. 119, p. 103331.

Yang, Y.; Pan, B.; Song, H. (2014). Predicting Hotel Demand Using Destination Marketing Organization's Web Traffic Data. *J. Travel Res*. 2013, 53, 433–447.

Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Sage Publications.

Yoon, S., Byun, S., & Jung, K. (2018). Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop* (SLT), pp. 112-118. IEEE.

Zawada, K., Rybak-Niedziółka, K., Donderewicz, M., & Starzyk, A. (2024). Digitization of AEC Industries Based on BIM and 4.0 Technologies. *Buildings*, 14(5), 1350.

Zhang, F., Fleyeh, H., Wang, X., & Lu, M. (2018). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, 99, 238-248. https://doi.org/10.1016/j.autcon.2018.12.016.

Zhang, J., and N. M. El-Gohary. (2015). Automated information transformation for automated regulatory compliance checking in construction. *J. Comput. Civ. Eng*. 29 (4): B4015001. https://doi.org/10.1061 /(ASCE)CP.1943-5487.0000427.

Zhang, J., and N. M. El-Gohary. (2013). Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking. *J. Comput. Civ. Eng*. 30 (2): 04015014. https://doi .org/10.1061/(ASCE)CP.1943-5487.0000346

Zhang, R., & El-Gohary, N. (2021). A deep neural network-based method for deep information extraction using transfer learning strategies to support automated compliance checking. *Automation in Construction*, 132, 103834.

Zhang, Y., Liu, H., Kang, S. C., & Al-Hussein, M. (2020). Virtual reality applications for the built environment: Research trends and opportunities. *Automation in Construction*, 118, 103311. https://doi.org/10.1016/j.autcon.2020.103311.

Zheng, Z., X. Z. Lu, K. Y. Chen, Y. C. Zhou, and J. R. Lin. (2022). Pretrained domain-specific language model for natural language processing tasks in the AEC domain. *Comput. Ind*. 142 (Nov): 103733. https://doi.org/10.1016/j.compind.2022.103733

Zhong, B., Wu, H., Ding, L., Love, P. E., Li, H., Luo, H., & Jiao, L. (2019). Mapping computer vision research in construction: Developments, knowledge gaps and implications for research. *Automation in Construction*, 107, 102919.

Zhong, B. T., L. Y. Ding, H. B. Luo, Y. Zhou, Y. Z. Hu, and H. M. Hu. (2012). Ontology-based semantic modeling of regulation constraint for automated construction quality compliance checking. *Autom. Constr*. 28 (Dec): 58–70. https://doi.org/10.1016/j.autcon.2012.06.006.

Zhong, B., X. Xing, H. Luo, Q. Zhou, H. Li, T. Rose, and W. Fang. (2020). Deep learning-based extraction of construction procedural constraints from construction regulations. *Adv. Eng. Inform*. 43 (Jan): 101003. https://doi.org/10.1016/j.aei.2019.101003.

Zhong, Y., & Goodfellow, S. D. (2024). Domain-specific language models pre-trained on construction management systems corpora. Automation in Construction, 160, 105316.

Zhou, H. (2022). Research of Text Classification Based on TF-IDF and CNN-LSTM. *Journal of Physics: Conference Series*, 2171(1).

Zhou, Y.-C., Z. Zheng, J.-R. Lin, and X.-Z. Lu. (2022). Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking. *Comput. Ind*. 142 (Nov): 103746. https://doi.org/10.1016/j.compind.2022.103746.

Zhu, Y., and S. P. Chen. (2020). Text classification models with nearest neighbor attention and convolutional neural networks. [In Chinese.] *Small Microcomputer Syst*.41 (2): 375–380.https://doi.org/10.3969/j.issn.1000-1220.2020.02.025.

Zhu, Y., Mao, W., & Ahmad, I. (2007). Capturing implicit structures in unstructured content of construction documents. *Journal of computing in civil engineering*, 21(3), 220-227. https://doi:10.1061/(asce)0887-3801(2007)21:3(220).

Zulqarnain, M., Ghazali, R., Ghouse, M. G., & Mushtaq, M. F. (2019). Efficient processing of GRU based on word embedding for text classification. JOIV: *International Journal on Informatics Visualization*, 3(4), 377-383.

Zuppa, D., Olbina, S. and Issa, R. (2016), Perceptions of trust in the US construction industry, *Engineering, Construction and Architectural Management*, Vol. 23 No. 2, pp. 211-236, doi: 10.1108/ECAM-05-2015-0081.

Appendix Research ethics application approval

This research followed the research integrity standards and principles as outlined by the University of Technology Sydney in the “Research Ethics and Integrity Policy and the Research Management Policy”. The Ethics Application to the Human Research Ethics Committee (HREC) was approved under ETH23-8903. The confidentiality and anonymity of the collected RFIs was maintained to ensure ethically responsible research. The ethics approval email is presented as follows. Please note that the original thesis title of this study at the time of the ethics approval is shown below.

Your ethics application has been approved as low risk - ETH23-8903

From research.ethics@uts.edu.au <research.ethics@uts.edu.au>

To:
Johnny Wong <Johnny.Wong@uts.edu.au>;
Muneeb Afzal <Muneeb.Afzal@student.uts.edu.au>

Dear Applicant,

Re: ETH23-8903 - "Automated Knowledge Extraction from Request for Information (RFI) Documents Using Analytic Algorithms."

Your local research office has reviewed your application and agreed that it now meets the

requirements of the National Statement on Ethical Conduct in Human Research (2007) and has been approved on that basis. You are therefore authorised to commence activities as outlined in your application, subject to any conditions detailed in this document.

You are reminded that this letter constitutes ethics approval only. This research project must also be undertaken in accordance with all UTS policies and guidelines including the Research

Management Policy.

Your approval number is UTS HREC REF NO. ETH23-8903

Approval will be for a period of five (5) years from the date of this correspondence subject to the submission of annual progress reports.

The following standard conditions apply to your approval:

- Your approval number must be included in all participant material and advertisements. Any advertisements on Staff Connect without an approval number will be removed.

- The Principal Investigator will immediately report anything that might warrant review of ethical approval of the project to the Ethics Secretariat.

- The Principal Investigator will notify the Committee of any event that requires a modification to the protocol or other project documents, and submit any required

amendments prior to implementation. Instructions on how to submit an amendment application can be found here.

- The Principal Investigator will promptly report adverse events to the Ethics Secretariat. An adverse event is any event (anticipated or otherwise) that has a negative impact on participants, researchers or the reputation of the University. Adverse events can also include privacy breaches, loss of data and damage to property.

- The Principal Investigator will report to the UTS HREC or UTS MREC annually and notify the Committee when the project is completed at all sites. The Principal Investigator will notify the Committee of any plan to extend the duration of the project past the approval period listed above.

- The Principal Investigator will obtain any additional approvals or authorisations as required (e.g. from other ethics committees, collaborating institutions, supporting organisations).

- The Principal Investigator will notify the Committee of his or her inability to continue as Principal Investigator including the name of and contact information for a replacement.

This research must be undertaken in compliance with the Australian Code for the Responsible Conduct of Research and National Statement on Ethical Conduct in Human Research.

You should consider this your official letter of approval.

If you have any queries about this approval, or require any amendments to your approval in future, please do not hesitate to contact your local research office or the Ethics Secretariat.