Unlocking Insights: Analysing Construction Issues in Request for Information (RFI) Documents with Text Mining and Visualisation

Muneeb Afzal, Johnny Kwok Wai Wong, and Alireza Ahmadian Fard Fini

***Abstract*— Request for Information (RFI) is an essential communication and decision support tool that assists project teams in identifying and resolving construction queries. RFI occurrences are common throughout the project lifecycle and predominantly comprise issues related to conflicting drawings and specifications, unclear requirements, vague contract documents or unexpected site conditions that inhibit project progress. Initiating RFIs or drafting their responses requires project resources, and failure to address them timely leads to poor project performance, decreasing chances of project success. RFIs are typically unstructured textual documents, and their manual content analysis for knowledge extraction is arduous and time-consuming. Previous research has successfully harnessed the potentials of Natural Language Processing (NLP) and text mining to process unstructured text documents and extract useful information. While NLP and text mining approaches have been applied in different domains, their application in construction industry is limited, particularly for analysing RFI datasets. Hence, the present research study analyses RFIs and their query subjects through unsupervised learning approach. Key contributions include the implementation of Latent Dirichlet Allocation (LDA), a text-mining algorithm to identify predominant topics and themes to classify the issues discussed within RFIs. This analysis successfully identified and highlighted issues related to structural discrepancies, construction coordination, building fixtures, building specifications and construction drawings as prominent issues mentioned in the project RFIs. As exploratory research, the findings of this study aim to enhance understanding of RFI issues and inspire future investigations that can delve deeper into specific aspects of RFI review process and motivate future studies, which may lead to dissecting RFI issues through specific issues.**

I. Introduction

The construction industry is complex [1], characterised by intense operations and has a high interdependency among project teams [2]. Effective stakeholder communication is crucial in steering construction projects in the right direction [3]. In this regard, request for information serves as a communication channel and facilitates identifying and resolving construction queries [4]. RFIs are lodged when a contractor or a subcontractor encounters a project issue requiring resolution or further clarification. These RFIs are then escalated to the project consultant or owner for advice. RFI responses tend to drive on-site work or control design decisions. However, if the responses are not satisfactory or the query has discrepancies, another cycle of RFI is initiated [6].

Despite its significance, RFI as a communication channel is often viewed negatively by the body of knowledge and is considered a "necessary evil" for the project [6]. It is largely because of the resources consumed in initiating an RFI and then reviewing and establishing responses. Rectifying an error within existing construction documents is time-consuming and can result in additional costs and delays to a construction project [7]. Delays in responding to RFI can make the schedule critical and negatively impact the on-site progress [8]. Additionally, if there is a delay in responding to RFI or in case a response is not provided at all, it can create frustration and a sense of distrust [9] between project team members. Both academia and industry are utilising limited capabilities to expedite the overall RFI process.

While the use of common data environments (Aconex and Procore) [10] has streamlined the RFI process by enabling online communication and ensuring tracked information, it is important to note that there is still a room for improvement in leveraging emerging technologies and data-driven analytics to enhance the RFI process even further and increase the overall project efficiency. Similarly, in the academic realm, researchers have traditionally relied on manually reviewing the content of the RFIs [11] to extract patterns, insights and lessons learnt. However, their approach can be time-consuming, impractical, and prone to errors.

To fill this gap, this research proposes a novel approach to address the challenges of RFI review process by implementing natural language processing and text mining techniques to extract patterns and themes related to issues mentioned in RFIs. To achieve this, the authors applied unsupervised learning - Latent Dirichlet Allocation algorithm to the RFI dataset. This research also presents visualisations to clearly and effectively represent the patterns and insights extracted from the RFI dataset. The results of this study have significant implications for the construction industry, as project teams can gain valuable insights into the root causes, visualise the problem and accordingly take proactive steps to prevent similar issues from arising in future. Additionally, this research could inform the development of more sophisticated RFI management tools to optimise the review process and improve communication among stakeholders in the construction industry.

Corresponding author: Muneeb Afzal
The authors are with the School of Design, Architecture and Building, University of Technology, Sydney, Australia
PO Box 123 Broadway NSW 2007 Australia
(muneeb.afzal@student.uts.edu.au)

II. Background and Research Objectives:

RFI documents are critical for construction projects, as they bridge the gap between contractors and designers, allowing teams to seek clarification and advice related to project problems or construction documents [12]. The timely response to RFI queries is essential for maintaining project momentum, as delays in drafting responses can significantly impact project schedules [8]. While industry solutions such as Aconex, Procore, and Autodesk 360 have facilitated tracking, monitoring, and communication of project concerns, there is little evidence to suggest that these platforms can significantly reduce the number of RFIs or their review period. Furthermore, these platforms may present challenges such as data loss and legal issues.

Traditionally, manual data mining or content analysis has been the approach for extracting patterns and lessons learned from RFIs [13]. Researchers have developed classification frameworks to determine the reasons behind RFIs, enabling deeper analysis of issues, such as why they occurred and how they could have been resolved [13]. However, manual review of RFIs poses several challenges, such as being time-consuming and labour-intensive, which can be impractical on an already under-resourced construction site. Additionally, the learnings can only be implemented once the project is completed.

To overcome these difficulties, researchers have been using advanced methods of Natural Language Processing and text mining to extract valuable information from unstructured documents [14]. Early efforts in this field were made by Caldas and Soibelman [15], who devised an automated hierarchical document classification system to manage construction documents. Further, Tixier et al. [16] created a content-analysis tool that can identify the reasons and consequences of injury reports, while Zhang and Ashuri [18] used data mining techniques to extract valuable insights from building information model (BIM) log files that measure design productivity. In another study, Mahfouz et al. [19] examined hidden legal knowledge in differing site condition (DSC) litigation cases. Recently Lee and Yi [20], employed topic modelling approach for predicting uncertainty through the RFIs. However, their research only deals with the RFIs related to the pre-bidding stage. Usually, in construction projects, the number of RFIs increases as the project progresses. These efforts mark significant progress towards integrating data-driven approaches to improve traditional construction processes.

To the best of the author's knowledge, no study has disassembled the RFI based on the issues through autonomous approaches of natural language processing. This study aims to address this research gap by implementing an unsupervised learning technique LDA to understand and visualise the insights from the RFIs. Contrary to the traditional content analysis approach, this approach won't require professionals to filter out each RFI one by one according to their cause. Visualisations generated through these algorithms can help develop a clear understanding of the prevailing issues recorded in the RFIs, and accordingly, preventive measures can be taken to avoid them.

III. Dataset Description and Characteristics:

The dataset comprises of RFIs obtained from construction projects in Middle Eastern countries. In total, 1800 RFI documents were collected from past construction projects, providing valuable insights into the RFI lexicon and overall request for information process applicable in the region. Despite the extensive information available within this dataset, there were a few limitations concerning the available dataset. For example, RFIs were gathered from construction projects of differing natures (infrastructure highways and metro rails), residential and buildings. Hence the findings may not be generalised to a specific type of construction project.

Additionally, the dataset may be biased towards certain regions of Middle East, which could limit the generalisability of the findings to other locations. Furthermore, there may be some variations within the RFI lexicon between different types of projects (infrastructure, residential etc.), and the dataset may not fully capture these differences. Lastly, some RFIs were also limited regarding the details provided, which may impact the ability to conclude the data. Therefore, improving the dataset by augmenting it with additional RFIs from various countries and diverse project types may be necessary.

IV. Data Pre-processing:

To derive useful patterns from the RFI dataset, this research study will employ an unsupervised learning approach by implementing Latent Dirichlet allocation, a natural language processing algorithm used for topic modelling. Prior to feature selection, the study implemented several data pre-processing techniques, which are as follows:

- Lowercasing: The complete text dataset was transformed into lowercase form with lowercasing. This is to ensure that words with similar meanings but different capitalisations such as “Design” and “design” are treated as one term.
- Punctuation removal: With punctuation removal, all punctuation marks from the dataset, for example, commas, periods, exclamation marks and question marks, were removed, as they do not contribute to the text semantics and often become a source to noise in the dataset.
- Stop words removal: For this study, all stop words pertaining to English language, such as “the", “and”, "a", "in", and "to" were removed.
- Numeric digit removal: Within an RFI query, it is common to encounter large amounts of numeric data such as reference numbers and drawings numbers. While

these numbers are helpful in RFI tracking, it was decided to remove them for the present study and make dataset more accessible for analysis.

- Tokenisation: Tokenisation is an essential step of NLP and text mining, and it breaks down an RFI query into individual words or other meaningful units called tokens. This step ensures that each RFI statement from the dataset is analysed and processed individually.
- Lemmatisation: Lemmatisation refers to reducing the words to their root form; for example, experiencing and connectivity are shortened to experience and connect, respectively. By lemmatising RFIs, it is possible to identify recurring issues and themes across multiple RFIs.

These pre-processing steps ensure that meaningful information and patterns are extracted from the RFI dataset and that the irrelevant or confounding information is minimised.

V. Insights into Prominent Topics and Keywords:

A. Unsupervised Approach Implementing Latent Dirichlet Allocation

RFIs are unstructured text documents. For processing documents like these, both supervised and unsupervised learning techniques can be applied. Unsupervised learning discovers patterns and relationships from text documents without the need of labelled datasets. For the present research study 1800 requests for information were analysed through the lens of unsupervised learning and later on clustered with regards to the predominant topics identified. For analysis the research study implemented a statistical modelling technique called Latent Dirichlet Allocation, which is an application of natural language processing. LDA views each RFI document as a combination of different topics. The goal of this research was to implement topic modelling through LDA on novel dataset of RFI through unsupervised approach. 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. 1) 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. 2 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. Table 3 provides the top 20 keywords, the topic number and suggested topic/theme. 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 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 (Requests for Information), and ensuring the specifications meet regulatory requirements.

• Topic 9 - Building Maintenance and Renovation: Once a building is constructed, keeping it in good condition is essential. 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.

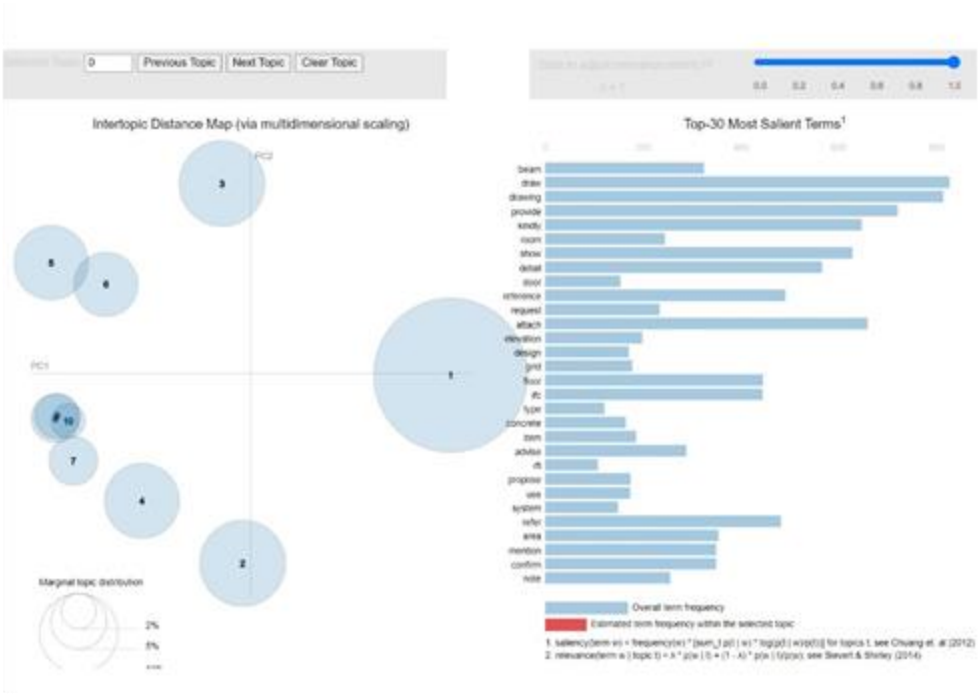
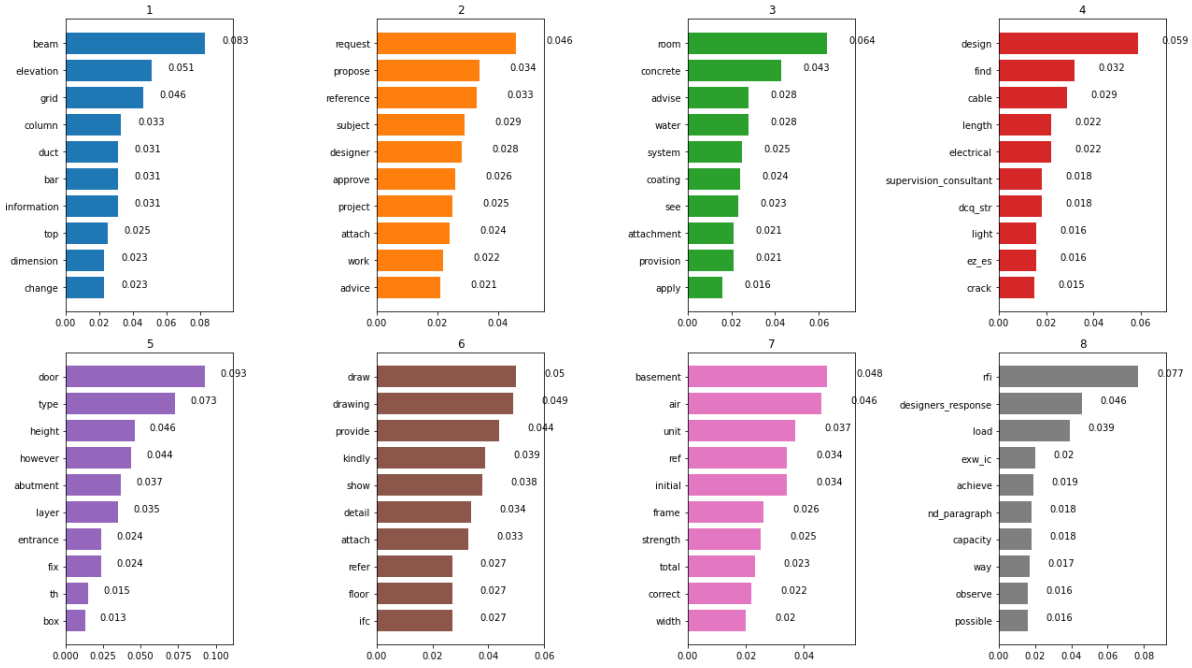


Figure 1. Representation of topics identified by LDA on 1800 construction requests for information



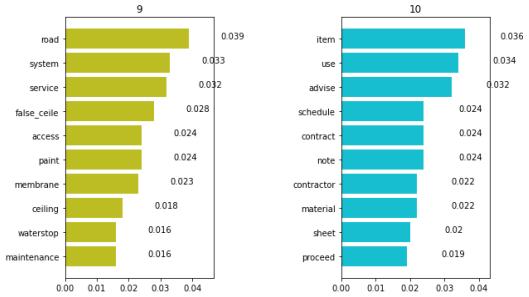


Figure 2. Results of the topic modeling of the RFIs

Table 1. Top 20 keywords within each topic and proposed theme/topic

Topic No:	Top 20 Keywords	Suggested Theme/Topic
1	beam, elevation, grid, column, duct, bar, information, top, dimension, change, steel, roof, shaft, response, equipment, run, main, part, tower, case.	Structural Discrepancies
2	request, propose, reference, subject, designer, approve, project, attach, work, advice, shop, due, tile, approval, review, structure, site, date, joint, opening.	Construction Approval
3	room, concrete, advise, water, system, coating, see, attachment, provision, apply, base, drainage, clash, supply, tank, power, attached, high, well, force, kind.	Coordinating Construction Systems
4	design, find, cable, length, electrical, supervision_consultant, dcq_str, light, ez_es, crack, construction_joint, require, space, receive, waterproof, meter, avoid, subcontractor, pole, purpose.	Electrical Installation
5	door, type, height, however, abutment, layer, entrance, fix, th, box, cabinet, second, water_proofe, go, comply, attachment_point, gate, hold, unit_located, contrast.	Building Fixtures
6	draw, drawing, provide, kindly, show, detail, attach, refer, floor, ifc, area, mention, confirm, section, reference, level, clarify, require, wall, location.	Construction Drawings
7	basement, air, unit, ref, initial, frame, strength, total, correct, width, projection, post_tensioning, rebar, retail, inner, vent_city, conflict, characteristic, outer, channel	Structural Stability
8	rfi, designers_response, load, exw_ic, achieve, nd_paragraph, capacity, way, observe, possible, base, calculation, enclose, confirmation, exceed, screed, guarantee, apparent, value, observation, rating, smdb, threshold, workshop, fact, rpj_ri, almost, traffic, connect, chw	Building Specifications
9	road, system, service, false_ceiling, access, paint, membrane, ceiling, waterstop, maintenance, rack, general_note, chamber, appear, application, datum, experience, internal, mep_service, switch.	Building Maintenance and Renovation
10	item, use, advise, schedule, contract, note, contractor, material, sheet, proceed, panel, requirement, follow, make, specification, require, available, cover, order, glass, document.	Procurement Management

B. Topic Clustering and Visualisation

1) Topic Clustering

We 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. 3) developed through the TSNE algorithm, we were able to identify 10 topics and their respective clusters within the corpus of RFI documents.

One notable finding was that Topic 6 (brown), which pertains to "Drawing Specifications", 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.

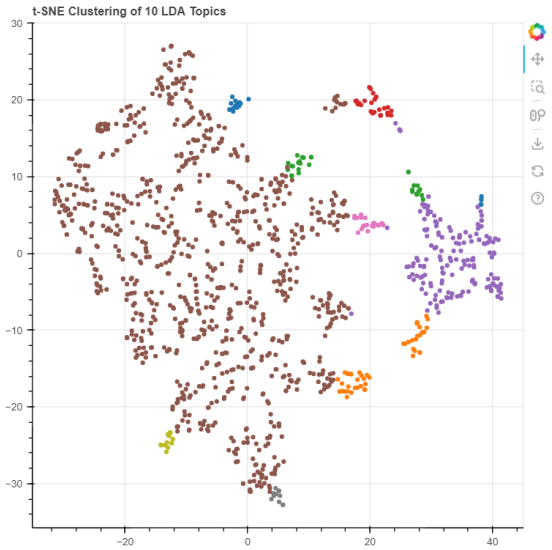


Figure 3. Topic clusters using TSNE algorithm.

2) Word Clouds for Construction Insights

Word clouds have become a popular tool for identifying 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, we utilised a word cloud (Fig. 4) 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.

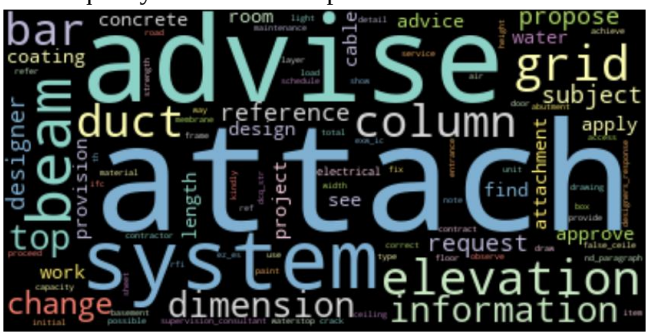


Figure 4. Developed word cloud from RFI dataset.

VI. Conclusion

In construction projects, requests for information play a vital role in enabling effective communication and informed decision-making. However, manually analysing RFIs for knowledge extraction is time-consuming and prone to errors due to the unstructured nature of these requests. Moreover, the construction industry has yet to fully leverage the potential of data analytics to support RFI processing. To address these challenges, our research proposes a novel approach employing natural language processing and text mining techniques to analyse RFI datasets. The study successfully identified and visualised the prevalent topics and themes discussed in RFIs, enhancing our understanding of RFI issues and providing inspiration for future research. The findings of this study have significant implications for the construction industry, as project teams can gain valuable insights into the root causes of problems and take proactive measures to prevent similar issues from arising in the future. Therefore, we strongly recommend developing more sophisticated RFI management tools that utilise data-driven approaches to support responding to RFIs, expedite the overall review process, and enable project teams to focus on more strategic tasks.

References

[1] S. Bertelsen, "*Complexity-Construction in a New Perspective*", 2003 IGLC-11, Blacksburg, Virginia.

[2] B. Shirazi, D.A. Langford, and S.M. Rowlinson, *"Organizational structures in the construction industry"*. Construction Management & Economics, 1996, 14(3), pp.199-212.

[3] O.A. Ejohwomu, O.S. Oshodi, and K.C. Lam, "*Nigeria's construction industry: Barriers to effective communication*", Engineering, Construction and Architectural Management, 2017, 24(4), pp.652-667.

[4] M. Abdel-Monem, and T. Hegazy, "*Enhancing construction as-built documentation using interactive voice response"*, J. Constr. Eng. Manag. 139 (7) (2013) 895–898. https://10.1061/(ASCE)CO.1943-707 7862.0000648

[5] L. Liao, E. A. L. Teo, R. Chang, and L. Li, *"Investigating Critical Non-Value Adding Activities and Their Resulting Wastes in BIM-*

Based Project Delivery", Sustainability 12 (1) (2020) 355, 719 https://doi: 10.3390/su12010355.

[6] A. Aibinu, S. Carter, V. Francis and P. Vaz-Serra, "*Request for information frequency and their turnaround time in construction projects"*, Built Environment Project and Asset Management 10 (1) 727 (2019) 1-15, https://doi.org/10.1108/BEPAM-10-2018-0130

[7] P.E.D. Love, J. Zhou, C.P. Sing, J.T. Kim, "*Assessing the impact of RFIs in electrical and instrumentation engineering contracts"*, J. Eng. Des. 25 (2014) 4–6, https://doi.org/10.1080/09544828.2014.935305

[8] D. Kelly and B. Ilozor, "*Performance outcome assessment of the integrated project delivery (IPD) method for commercial construction projects in USA"*, International Journal of Construction 722 Management 1–9 (2020), https://doi: 10.1080/15623599.2020.1827340.

[9] M. Philips-Ryder, J. Zuo, and X. H. Jin, "*Evaluating Document Quality in Construction Projects –Subcontractors' Perspective*", International Journal of Construction Management 13 (3) (2012) 77–806 94, https://doi: 10.1080/15623599.2013.10773217.

[10] M. Das, X. Tao, J.C.P Cheng, "*A Secure and Distributed Construction Document Management System Using Blockchain"*. In: Toledo Santos, E., Scheer, S. (eds) Proceedings of the 18th International 876 Conference on Computing in Civil and Building Engineering. ICCCBE 2020. Lecture Notes in 877 Civil Engineering, 98 (2020). https://doi.org/10.1007/978-3-030-51295-8_59

[11] J. J. Kim, A. L. Petrov, J. Lim, and S. Kim, "*Comparing Cost Performance of Project Delivery Methods Using Quantifiable RFIs": Cases in California Heavy Civil Construction Projects*, International Journal of Civil Engineering 20 (3) 2021 323–335, https://doi: 10.1007/s40999-021-00658-0

[12] MCGRAW-HILL (2008), "*Building information modeling (BIM), transforming design and construction to achieve greater industry productivity"*, SmartMarket Report: Design and Construction 704 Intelligence. McGraw Hill Construction (2008)

[13] A. S. Bhat, E.A. Poirier, & S.S. French, "*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 (2017), 779 955–964.

[14] Y. Jallan, E. Brogan, B. Ashuri, and C.M. Clevenger,. "*Application of natural language processing and text mining to identify patterns in construction-defect litigation cases"*, Journal of legal affairs and dispute resolution in engineering and construction, 2019, 11(4)

[15] C.H. Caldas, and L. Soibelman, "*Automating hierarchical document classification for construction management information systems"*, Autom. Constr. 2003, 12 (4): 395–406. https://doi.org/10.1016/S0926-5805 (03)00004-9.

[16] A.J.P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, "Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports." Autom. Constr. 2015 62 (Feb): 45–56. https://doi.org/10.1016/j.autcon.2015.11.001.

[17] L. Zhang, and B. Ashuri, "BIM log mining: Discovering social networks." Autom. Constr. 2018 91 (Jul): 31–43. https://doi.org/10.1016 /j.autcon.2018.03.009

[18] T. Mahfouz, A. Kandil, and S. Davlyatov. "Identification of latent legal knowledge in differing site condition (DSC) litigations." J. Autom.Constr, 2018 94 (Oct): 104–111. https://doi.org/10.1016/j.autcon.2018.06.011.

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