SASBE 2025 aims to encourage the international exchange of innovative ideas between researchers from academia and industry. In addition to knowledge dissemination, the conference offers a valuable platform for professional networking, particularly benefiting university professors, graduate students, and postdoctoral researchers.

Research Article

Enhancing Hybrid NLP Approaches with the Strategic Shift Toward retrieval-augmented generation for Design Phase Communication Documents Processing

Ghazal Salimi[1], Farzad Rahimian[1], Alessandro Di Stefano[2], Edlira Vakaj[3]

[1] School of Computing, Engineering & Digital Technologies, Teesside University, UK
[2] Department of Computing & Games, Teesside University, UK
[3] The Faculty of Computing, Engineering and The Built Environment, Birmingham City University, UK
Correspondence: g.salimi@tees.ac.uk

Abstract

While hybrid Natural Language Processing (NLP) approaches that combine rule-based methods with machine learning have shown promise for processing construction documents, their effectiveness for extracting technical data from variable client communications remains limited. This paper presents a systematic evaluation of traditional hybrid NLP methods for information extraction from construction communication documents, documenting both achievements and fundamental limitations that drive the need for more advanced approaches. Building upon a dataset of 103 manually annotated client emails containing seven key technical features (postcodes, U-values, materials, deck types, falls), we implemented and evaluated hybrid approaches combining regular expression (Regex) pattern matching with Named Entity Recognition (NER). While achieving high accuracy for structured attributes like UK postcodes (98%), the hybrid Regex+NER approach plateaued at approximately 60% accuracy for complex, contextually variable attributes such as U-values and material specifications. Through systematic analysis, we identify fundamental limitations of traditional hybrid approaches when applied to the linguistic variability inherent in professional client communications. These limitations include brittleness when facing inconsistent terminology, data-hungry requirements for domain-specific training, and the inability to handle the contextual complexity of technical communications in early-stage design phases. Based on these empirical findings, we present the strategic rationale for moving toward Retrieval-Augmented Generation (RAG) approaches, which address the identified limitations through contextual understanding, reduced data requirements, and improved handling of document-length variability. The research contributes quantitative baseline performance data for traditional methods and a methodological framework for evaluating NLP approaches in construction communication processing, providing researchers with empirical evidence about the practical boundaries of traditional hybrid methods.

Keywords: Hybrid NLP; Construction Communication; Information Extraction; RAG; Early-Stage Design.

Highlights

- Hybrid NLP methods achieved 98% accuracy for postcodes but only 60% for complex technical specifications in construction emails.
- Three core limitations identified: inconsistent terminology, poor contextual understanding, and excessive training data requirements.
- Provides baseline performance data and evaluation framework supporting the strategic shift toward RAG approaches for construction communication processing.

1 Introduction and Background of Study

The Architecture, Engineering, Construction, and Operations (AECO) industry has long been characterised as document-centric, where documents serve as interfaces for accessing and navigating collections of information. Despite the increasing adoption of Building Information Modelling (BIM), the information flow in construction projects remains heavily reliant on document exchange. This document-centric nature results in vast amounts of unstructured textual data being produced and shared through natural language, creating significant challenges for digital management and information processing.

It is commonly known that the built environment has a big influence on the environment all over the world (UNSTATS, 2010). More than 40% of global energy consumption and greenhouse gas emissions are attributable to the building and construction sector (IEA, 2019). In addition, they pollute the air, water, and soil, use a lot of resources, and produce a lot of waste (Sharma et al., 2010; Cabeza et al., 2014). The earliest phases of design offer the most promise for reducing carbon. To effectively reduce embodied carbon in buildings, researchers and designers are responsible for developing and applying early-stage design methods aimed at lowering embodied carbon (WGBC, 2019). The use of technologies in the early stages of design greatly aids in reducing the carbon footprint in many ways. In the design, engineering and construction sectors, advanced technology integration has attracted a lot of attention. Numerous studies have concentrated on process optimisation and productivity enhancement through automation and data-driven methodologies (Farooqui et al., 2020; Cerquiteli et al., 2021). Making these crucial design choices, however, requires quick access to reliable technical data, which is usually concealed in unstructured formats such as client communications, project specifications, and design notes (Czvetko et al., 2022).

Smart construction is heavily dependent on data to bridge the physical and virtual worlds. This data encompasses both structured forms (e.g., sensor data) and unstructured forms (e.g., images and text). While existing smart construction systems, such as BIM and GIS, concentrate on structured data stored in table databases, over 80% of construction data is unstructured, primarily in text form (Wu et al., 2021). Text data is also easier and more cost-effective to collect compared to images and audio (Gharehchopogh & Khalifelu, 2011). Therefore, it is essential and economical to automatically and intelligently extract and interpret text data.

One promising strategy to deal with these issues is Natural Language Processing (NLP), which is a collection of methods that assist computers in comprehending human languages (Olivetti, et al., 2020; Shah, 2020). Better information management in building projects is made possible by NLP's ability to convert unstructured textual data into structured data. The full potential of the digital transition in the construction industry can be realised by combining BIM with digitalisation technologies like natural language processing, as recommended by the European Union. Automating the extraction of text data is essential for linking digital and physical work in early building design. This process helps extract valuable information from design briefs, client requirements, regulations, and project communications, enabling better decision-making and higher design quality.

Working with client communications in early design stages is particularly challenging. These emails and messages contain crucial technical details that influence design decisions - such as thermal performance requirements, material preferences, structural specifications, and building regulations. The difficulty is that these communications are typically informal, with people using varied terminology to describe the same concepts. While automated extraction has proven highly effective

for structured attributes with consistent patterns (Salimi et al., 2025), complex technical specifications present greater challenges. This inconsistency makes it very difficult to extract technical information automatically using traditional methods. Moreover, manually reviewing these documents is extremely time-consuming and slows the entire design process. When important information is overlooked or misinterpreted due to rushed email reviews, it can result in poor design decisions. One promising method for automated data extraction from unstructured text and technical drawings is natural language processing. In recent years, NLP has been applied in construction contexts for tasks such as filtering information from reports, classifying technical drawings, developing expert systems, and checking compliance against contracts and standards (Yan et al., 2020; Darko et al., 2020). The applications of NLP in construction can be categorised into four main scenarios: filtering information to extract key data from noisy texts, organising documents by automatically grouping them based on different backgrounds, developing expert systems that integrate expert knowledge, and automated compliance checking.

Several studies have investigated the use of NLP for information extraction and data analysis in the construction industry. For example, Roca et al. (2021) employed NLP to evaluate the environmental impacts of construction materials, while Zhao et al. (2019) applied NLP and machine learning to environmental assessment reports to identify critical sustainability issues. In the context of construction-specific applications, Koc and Sariyildiz (2018) developed an NLP-based system to extract building material data from construction documents, while Wang et al. (2021) used NLP to analyse social media data related to construction projects. Recent work has also explored automated extraction from technical drawings and diverse file formats (Salimi et al., 2024). Many researchers have attempted to analyse construction documents automatically, with early studies focusing on classifying or clustering construction documents for efficient management (Koc & Sariyildiz' 2018; Wang et al., 2021). However, as construction projects have become larger and more complex, these document-level analyses have proven insufficient.

NLP in the construction sector has entered a new phase thanks to recent developments in data storage, computer processing, and deep learning techniques, which enable more complex analyses and applications. Recent years have seen a degree of automation in a number of construction-related domains thanks to developments in text analytics and natural language processing. The evolution of NLP applications in construction research shows significant growth starting from 2013, coinciding with the introduction of the Word2Vec word embedding method, and continuing through advancements like BERT in 2018 (Iranmanesh et al., 2025).

Despite these advances, significant gaps remain in understanding the practical effectiveness and limitations of hybrid NLP approaches when applied to the specific challenge of client communication processing. The early-stage design phase of construction places significant emphasis on design tools, while insufficient attention is directed towards the time required for collecting technical data prior to initiating the design process. The majority of studies only show NLP techniques on small datasets or in controlled settings, paying little attention to the contextual complexity and linguistic variability present in casual professional communications (Rogers et al., 2022; Zulkarnain & Putri, 2021; Hung et al., 2021). Furthermore, there is a lack of empirical data regarding the performance limits of traditional hybrid approaches and the factors that contribute to these limitations.

Despite advancements in NLP applications for construction and design engineering, several gaps remain in the literature. Traditional NLP approaches in construction have typically relied on rule-based

methods and supervised learning paradigms. Rule-based approaches using regex patterns and keyword matching have been effective for structured data extraction but struggle with the linguistic variability inherent in client communications, where the same technical information can be expressed in multiple ways. The rule-based approach evolved during the 1970s to 2010 period and represents one of the earliest methods in NLP evolution. These systems are based on complex sets of manually written rules, offering high interpretability but limited flexibility when dealing with noisy or ambiguous text data. Supervised learning methods require extensive labelled datasets that are expensive to create and maintain, particularly for domain-specific construction terminology.

Large language model (LLM) advancements in recent years have opened up new avenues for overcoming the drawbacks of conventional NLP techniques. A paradigm shift from supervised learning to pre-training followed by fine-tuning has resulted from the remarkable success of pre-trained language models in natural language processing. Compared to conventional methods, these models have a number of benefits, such as contextual comprehension, few-shot learning, and the capacity to handle longer, more complicated documents (Iranmanesh et al., 2025). Retrieval-Augmented Generation (RAG) has emerged as a particularly promising solution to LLM limitations. RAG combines the contextual understanding capabilities of LLMs with external knowledge retrieval, grounding model outputs in verifiable sources (Wang et al., 2023; Lee et al., 2024). Kumarasinghe and Kirikova (2025) demonstrated the value of RAG for attribute extraction from unstructured documents in knowledge graph construction, showing improved scalability and precision (Lee et al., 2024). Parallel studies in BIM have highlighted how large language models can enable natural language interfaces for BIM and enhance compliance checking and workflow optimisation (Kumarasinghe & Kirikova, 2025; Iranmanesh et al., 2025; Wang et al., 2023). Collectively, these developments signal a shift towards hybrid frameworks that integrate retrieval, generation, and semantic validation, aligning closely with the challenges faced in automating technical data extraction from client communications in construction.

The literature review reveals a significant gap in systematic evaluation of traditional hybrid NLP approaches when applied to variable client communications in construction contexts. There is not enough empirical data regarding the performance limits of conventional approaches when dealing with the linguistic variability and contextual complexity of informal professional communications, despite the fact that previous research has shown success with structured documents and controlled datasets. Furthermore, rather than methodically recording limitations and their root causes, the majority of current studies concentrate on showcasing improvements. This gap in understanding constrains both method development and practical implementation decisions, as practitioners lack evidence-based guidance about when traditional approaches are sufficient versus when more advanced methods are necessary.

This research addresses these gaps through a systematic evaluation of hybrid NLP approaches for technical data extraction from construction client communications. Building on a dataset of 103 manually annotated client emails containing seven key technical features, we implement and rigorously evaluate hybrid approaches combining regular expression pattern matching with Named Entity Recognition. Our investigation reveals both the capabilities and fundamental limitations of these traditional methods, providing empirical evidence about their performance boundaries when applied to variable professional communications.

The primary contributions of this work are threefold. First, we provide quantitative baseline performance data for hybrid NLP approaches applied to construction client communications, establishing benchmarks for future method comparisons. Second, we systematically document the limitations of traditional hybrid approaches, identifying specific factors that constrain their effectiveness for variable textual data. Third, we present a methodological framework for evaluating NLP approaches in construction communication processing, contributing to more rigorous assessment practices in construction informatics research.

Our results show that hybrid methods work very well for simple, structured data like UK postcodes (98% accuracy). However, they struggle with complex information that depends on context, like materials, where accuracy drops to around 60%. When we looked at the errors, we found three main problems: the methods break down when people use different words for the same thing, they can't understand how context changes meaning, and they need lots of specialised training data to work properly.

These empirical findings have important implications for both researchers and practitioners in construction informatics. For researchers, our work provides evidence-based understanding of where traditional methods fail and why more advanced approaches are necessary. For practitioners, our analysis offers realistic expectations about the capabilities and limitations of current hybrid NLP technologies, informing decisions about technology adoption and implementation strategies.

Based on our systematic evaluation, we argue that the identified limitations necessitate a strategic shift toward more advanced approaches such as Retrieval-Augmented Generation (RAG), which can address the contextual understanding gaps and data requirements that constrain traditional methods (Kumarasinghe & Kirikova, 2025). While implementation of such advanced approaches represents the next phase of our research, the empirical foundation established in this work provides essential baseline data and methodological insights for evaluating future developments.

2 Methodology

2.1 Dataset Development

The study utilised client emails collected through an industry partnership, providing access to authentic construction communication data. From an initial collection of 278 client emails, a rigorous selection and annotation process was implemented to create a high-quality dataset suitable for systematic evaluation. Seven key technical features were identified for extraction based on their frequent occurrence in construction communications and their importance for early-stage design processes: sender, subject, postcode, U-value, material, deck type, and fall. These features represent a range of extraction challenges, from highly structured attributes (postcodes) to contextually variable technical specifications (U-values, materials).

Manual annotation was performed on all 278 emails, with subsequent quality control processes removing duplicates and irrelevant entries. This refinement process resulted in a final dataset of 103 high-quality annotated examples, each containing verified labels for the seven target features. The dataset provides ground truth data for systematic evaluation of extraction accuracy across different attribute types and complexity levels. Preprocessing steps included tokenisation, lemmatisation, and OCR-based text extraction from email attachments using tools such as PDFMiner, PyMuPDF, and EasyOCR (Salimi et al., 2024). This preprocessing pipeline created a clean, structured dataset suitable

for experimental evaluation while maintaining the linguistic variability inherent in authentic client communications.

2.2 Hybrid NLP Implementation

The experimental approach focused on the systematic evaluation of hybrid methods combining rule-based and machine learning techniques. The implementation progressed through multiple stages to establish comprehensive baseline performance data, building on established preprocessing pipelines for construction documents (Salimi et al., 2024). Initial experiments focused on regex-based extraction, particularly for attributes with clear structural patterns. UK postcodes served as the primary test case for regex effectiveness, as they follow well-defined formatting rules (Salimi et al., 2025). However, when applied to more variable attributes such as U-values, materials, deck types, and falls, regex approaches struggled due to the inconsistent ways such information is expressed in client communications.

To address the limitations of pure rule-based approaches, Named Entity Recognition was introduced as a complementary method. NER models offered greater flexibility in identifying entities in free text and proved particularly effective for recognising attributes like U-values and materials that appear in variable linguistic contexts. A hybrid pipeline combining regex and NER was then implemented to leverage the strengths of both approaches. This architecture used regex for highly structured attributes where pattern matching was reliable, while employing NER for contextually variable attributes requiring semantic understanding.

2.3 Evaluation Framework

Performance evaluation focused on accuracy measurement across the seven target features, with particular attention to understanding how accuracy varied based on attribute complexity and linguistic variability. The evaluation methodology established quantitative baselines for comparing different extraction approaches and identifying performance limitations.

3 Results and Discussion

3.1 Performance Analysis

The evaluation of hybrid NLP methods shows a clear split in performance depending on what type of information we're trying to extract. The hybrid approach combining regex and Named Entity Recognition worked very differently across our seven target features - achieving 98% accuracy for UK postcodes but only around 60% for more complex attributes like U-values and materials.

Figure 1. Workflow of Hybrid Regex+NER Approach for Technical Data Extraction

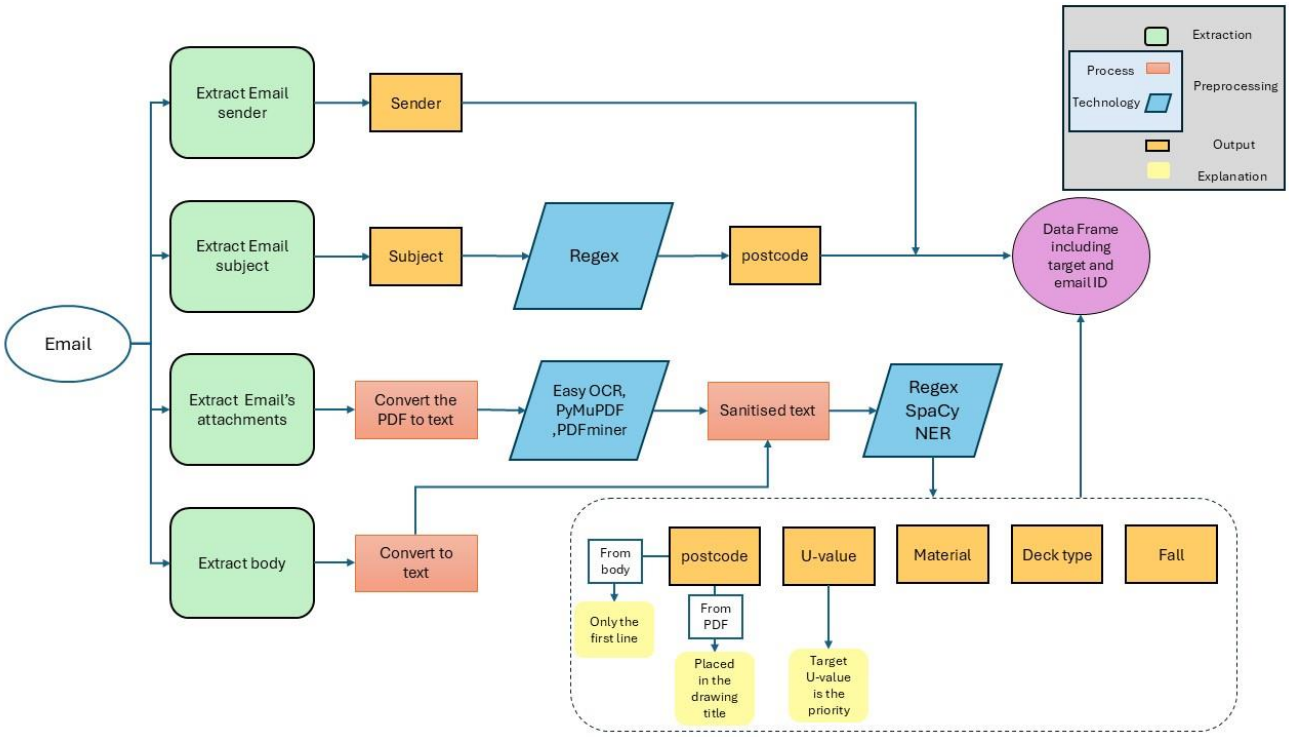


Figure 1 shows how the hybrid system works. The system handles emails through different paths: it directly extracts sender and subject information, uses regex patterns for postcodes, and applies a more complex process for technical details that involves converting PDFs, cleaning up text, and using both regex and NER together.

The 98% accuracy for postcodes makes sense because they follow clear rules (like "TS1 3BA" or "SW1A 1AA"). This confirms that regex works well for structured data with consistent patterns. However, the drop to 60% accuracy for complex technical information reveals serious problems. These attributes appear in many different ways - a U-value might be written as "U-value of 0.15 W/m^2K," "thermal transmittance: 0.15," or just "meeting Part L requirements." Each needs a different extraction logic.

3.2 Why Hybrid Approaches Struggle

Our analysis identified three main problems with traditional hybrid methods.

The first issue is inconsistent terminology. People use different words for the same thing. Roofing materials might be called "single-ply membrane," "TPO roofing," "thermoplastic membrane," or just "membrane system." Each variation needs to be explicitly programmed into regex patterns or included in NER training examples. As new terms appear, the system requires constant updates to keep working properly.

The second problem is missing context. Many technical details aren't stated directly but implied through context. A client might write "we need to exceed current Building Regulations" instead of giving exact U-values, or "similar to Project Alpha specifications" instead of listing materials. Regex can't interpret these references at all, and NER models struggle with understanding connections

across longer documents. This is especially problematic in early design when clients often reference previous discussions or standards without repeating all the details.

The third limitation is the need for lots of training data. The NER part of our hybrid system needs extensive domain-specific training to work with construction terminology. Our 103 annotated emails were enough for testing, but nowhere near enough for building a robust system. Creating quality training data is expensive and time-consuming - it requires real domain experts, not just anyone. This creates a major barrier, especially for smaller construction firms.

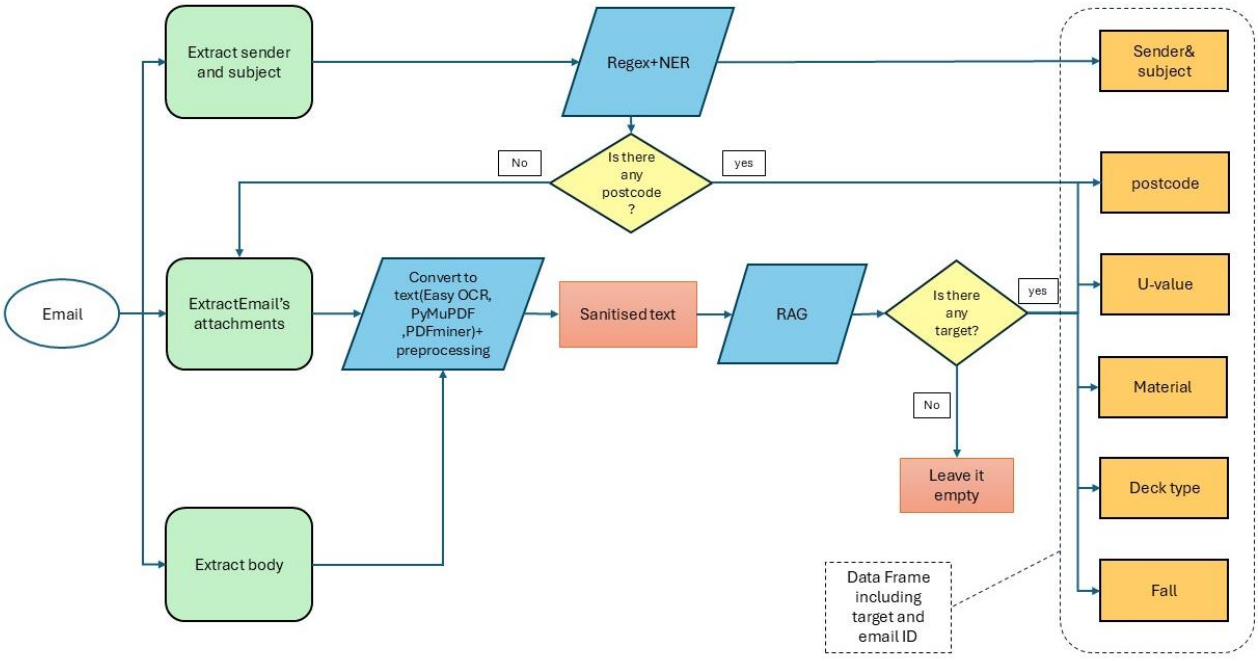
3.3 What This Means Practically

These findings have important implications. For researchers, they show that small improvements to traditional methods won't solve the problem - we need fundamentally different approaches. For practitioners, hybrid methods work fine for structured data like postcodes, but if you need to extract complex technical specifications automatically, traditional methods simply won't reach acceptable accuracy levels.

3.4 Moving Towards RAG Approaches

Based on these limitations, we propose shifting to Retrieval-Augmented Generation (RAG) approaches. **Figure 2** shows the RAG pipeline architecture, which works very differently from the traditional hybrid approach.

Figure 2. RAG-Based Approach for Construction Communication Processing



Instead of just pattern matching and entity recognition, RAG combines retrieval (finding relevant information) with generation (producing structured outputs using large language models). This addresses our three main problems. First, it handles varied terminology because pre-trained language models already understand that different terms can mean the same thing, without needing every variant programmed in. Second, it provides better context understanding since large language models

can interpret implied information and cross-references across entire documents. Third, it needs less training data because RAG systems can work effectively with minimal fine-tuning, utilising knowledge already built into the pre-trained models.

The shift from Figure 1 to Figure 2 represents a fundamental change in approach - from trying to list all possible patterns to using contextual understanding and knowledge retrieval.

3.5 Research Contributions

This research provides a systematic framework for evaluating NLP approaches in construction contexts. The progression from building a dataset, implementing hybrid methods, measuring performance, and analysing failures creates a template that other researchers can follow. Importantly, we document not just what works but where and why methods fail - crucial information for making real progress.

4 Conclusions

This study systematically evaluates hybrid NLP approaches for extracting technical data from construction client communications. Using 103 manually annotated emails with seven key features, we show that hybrid Regex+NER methods achieve excellent accuracy (98%) for structured attributes but only 60% for complex technical specifications - too low for practical use in early-stage design.

Our key contributions include baseline performance data for comparing future methods, documentation of three fundamental limitations (terminology issues, context problems, and data requirements), and an evaluation framework for construction NLP research. These findings support moving towards RAG approaches that address these limitations through better contextual understanding and lower data requirements.

Several limitations merit acknowledgement. Our dataset of 103 emails, whilst authentic, is relatively small and may not represent all construction sectors. We focused mainly on accuracy without analysing other important factors like cost-benefit trade-offs. Future work should implement and test RAG approaches to see if they deliver on their promise. Researchers should also explore combining RAG with rule-based methods for structured attributes and address practical challenges like integrating with BIM workflows and designing user interfaces for verifying automated extractions.

This work helps construction informatics by clearly showing what current NLP technologies can and cannot do. As the construction industry goes digital, automated processing of unstructured communications will be essential - and this research provides the foundation for developing technologies that actually work in practice.

Acknowledgements
The study was conducted at Teesside University in collaboration with TaperedPlus company and received financial support from the company as a studentship.

Conflicts of Interest
The authors declare no conflict of interest.

References

Cabeza, L., Rincon, L., Vilarino, V., Perez, G., & Castell, A. (2014). Life cycle assessment (LCA) and life cycle energy analysis (LCEA) of buildings and the buildings sector: A review. Renewable and Sustainable Energy Reviews, 29, 394–416. doi: 10.1016/j.rser.2013.08.037

Cerquitelli, T., Pagliari, D. J., Calimera, A., Bottaccioli, L., Patti, E., Acquaviva, A., & Poncino, M. (2021). Manufacturing as a data-driven practice: Methodologies, technologies, and tools. Proceedings of the IEEE, 109(4), 399–422. doi: 10.1109/JPROC.2021.3056006

Czvetkó, T., Kummer, A., Ruppert, T., & Abonyi, J. (2022). Data-driven business process management-based development of Industry 4.0 solutions. CIRP Journal of Manufacturing Science and Technology, 36, 117–132. doi: 10.1016/j.cirpj.2021.12.002

Darko, A., Chan, A. P., Adabre, M. A., Edwards, D. J., Hosseini, M. R., & Ameyaw, E. E. (2020). Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities. Automation in Construction, 112, 103081. doi: 10.1016/j.autcon.2020.103081

Farooqui, A., Bengtsson, K., Falkman, P., & Fabian, M. (2020). Towards data-driven approaches in manufacturing: An architecture to collect sequences of operations. International Journal of Production Research, 58(16), 4947–4963. doi: 10.1080/00207543.2020.1735660

Gharehchopogh, F. S., & Khalifelu, Z. A. (2011). Analysis and evaluation of unstructured data: Text mining versus natural language processing. In 2011 5th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1–4). IEEE. doi: 10.1109/ICAICT.2011.6111017

Huang, J., Johanes, M., Kim, F. C., Doumpioti, C., & Holz, G.-C. (2021). On GANs, NLP and architecture: Combining human and machine intelligences for the generation and evaluation of meaningful designs. Technology|Architecture + Design, 5(2), 207–224. doi: 10.1080/24751448.2021.1967060

IEA. (2019). Global status report for buildings and construction 2019 – towards a zero-emissions, efficient and resilient buildings and construction sector. Retrieved from https://www.iea.org/reports/global-status-report-for-buildings-and-construction-2019

Iranmanesh, S., Saadany, H., & Vakaj, E. (2025). LLM-assisted Graph-RAG information extraction from IFC data. In 2025 European Conference on Computing in Construction.

Koc, M., & Sariyildiz, S. (2018). An NLP-based building material data extraction system for construction documents. Automation in Construction, 86, 124–137. doi: 10.1016/j.autcon.2017.11.018

Kumarasinghe, A., & Kirikova, M. (2025). Automated data science project knowledge graph construction using a document retrieval augmented generation pipeline. Riga Technical University. Posted: May 6, 2025.

Lee, S., Ahn, S., Kim, D., & Kim, D. (2024). Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety management knowledge retrieval. Automation in Construction, 168, 105846. doi: 10.1016/j.autcon.2024.105846

Olivetti, E. A., Cole, J. M., Kim, E., Kononova, O., Ceder, G., Han, T. Y.-J., & Hiszpanski, A. M. (2020). Data-driven materials research enabled by natural language processing and information extraction. Applied Physics Reviews, 7(4), 041317. doi: 10.1063/5.0021106

Roca, X., Casals, M., & Roca, F. (2021). A new approach to the environmental impact analysis of building materials using NLP techniques. Sustainability, 13(10), 5758.

Rogers, A., Gardner, M., & Augenstein, I. (2022). QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. ACM Computing Surveys. doi: 10.1145/3560260

Salimi, G., Rahimian, F., Di Stefano, A., & Vakaj, E. (2025). Comparison of data labelling techniques for automating postcode extraction in NLP-supported early-stage building design. In Proceedings of Digital Frontiers in Buildings and Infrastructure International Conference Series (DFBI 2025).

Salimi, G., Rahimian, F., Okonta, E. D., Oliver, S., Di Stefano, A., Vakaj, E., & Lane, N. (2024). Enhancing construction design efficiency: An approach to data extraction with natural language processing for technical drawing. Paper presented at [Conference Name].

Shah, V. (2020). Advancements in deep learning for natural language processing in software applications. International Journal of Computer Science and Technology, 4(3), 45–56. Retrieved from https://ijcst.com.pk/index.php/IJCST/article/view/373

Sharma, A., Saxena, A., Sethi, M., Shree, V., & Varun. (2011). Life cycle assessment of buildings: A review. Renewable and Sustainable Energy Reviews, 15(1), 871–875. doi: 10.1016/j.rser.2010.09.008

UNSTATS. (2010). Greenhouse gas emissions by sector (absolute values). Retrieved from https://unstats.un.org/unsd/environment/air_greenhouse_emissions_by_sector.htm

Wang, H., Li, J., Wu, H., Hovy, E., & Sun, Y. (2023). Pre-trained language models and their applications. Engineering, 25, 51–65. doi: 10.1016/j.eng.2022.04.024

Wang, Y., Li, Q., Li, X., Li, Z., & Li, J. (2021). Using natural language processing for social media data analysis of construction projects. Journal of Management in Engineering, 37(2), 04020027. doi: 10.1061/(ASCE)ME.1943-5479.0000927

World Green Building Council & Ramboll. (2019). Bringing embodied carbon upfront: Coordinated action for the building and construction sector to tackle embodied carbon. World Green Building Council.

Wu, C., Wu, P., Wang, J., Jiang, R., Chen, M., & Wang, X. (2021). Ontological knowledge base for concrete bridge rehabilitation project management. Automation in Construction, 121, 103428. doi: 10.1016/j.autcon.2020.103428

Yan, H., Yang, N., Peng, Y., & Ren, Y. (2020). Data mining in the construction industry: Present status, opportunities, and future trends. Automation in Construction, 119, 103331. doi: 10.1016/j.autcon.2020.103331

Zhao, X., Huang, Q., Xue, X., Wang, Q., & He, G. (2019). An approach for environmental impact assessment based on natural language processing and machine learning. Journal of Cleaner Production, 228, 1570–1579. doi: 10.1016/j.jclepro.2019.04.343

Zulkarnain, & Putri, T. D. (2021). Intelligent transportation systems (ITS): A systematic review using a natural language processing (NLP) approach. Heliyon, 7(12), e08615. doi: 10.1016/j.heliyon.2021.e08615