Proceedings of Digital Frontiers in Buildings and Infrastructure International Conference Series

DFBI 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

Comparison of Data Labelling Techniques for Automating Postcode Extraction in NLP-Supported Early-Stage Building Design

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

Data labelling is crucial for the success of Natural Language Processing (NLP) models, as the quality of labelled data directly affects model accuracy and performance. In early-stage construction design, automating the data extraction of textual data is essential for integrating physical and digital workflows. However, data labelling presents significant challenges, requiring careful trade-offs between time, cost, and accuracy to meet project-specific needs. This paper compares three primary data labelling techniques for postcode extraction from project documents: manual, rule-based, and hybrid machine learning approaches. A review of the seminal literature reveals that manual labelling delivers high accuracy and quality but is labour-intensive and better suited for small datasets or creating gold standards. Rule-based techniques, such as regular expressions (Regex), automate labelling for structured data using predefined patterns, offering efficiency but requiring domain expertise. Machine learning-driven methods, like Named Entity Recognition (NER), enable scalability for large datasets but often demand task-specific fine-tuning. Due to suboptimal NER performance in initial testing, a hybrid approach combining Regex with NER was developed and implemented using Google Colab. Through empirical evaluation of postcode extraction from construction project documents, the rule-based approach achieved 96.7% accuracy when compared against manual labelling as the gold standard, while the hybrid machine learning approach achieved 98% accuracy. This paper provides a comparative framework to guide practitioners in selecting the most appropriate data labelling technique based on their specific needs, balancing accuracy, efficiency, and scalability to optimise workflows and enhance automation in early-stage building design.

Keywords: Natural Language Processing (NLP); Data Labelling; Rule-Based Models; Early-stage Building Design; Postcode Extraction

Cite this article:
Lastname, F. M., & Lastname, F. M. (Year). Title of the article. Title of the Journal, Volume 2025, Page range: 266-276

Highlights

- Rule-based regex extraction achieved 96.7% accuracy for UK postcode identification in construction emails.
- Hybrid approaches combining regex with NER delivered optimal 98% accuracy with minimal processing overhead.
- Hierarchical extraction strategies significantly outperformed single-method approaches for construction data.

1 Introduction and Background of Study

The Architecture, Engineering, Construction, and Operations (AECO) industry has long been characterised as document-centric (Opitz et al., 2014; Rezgui & Zarli, 2006), 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 (Opitz et al., 2014; Zhu et al., 2001). This document-centric nature results in vast amounts of unstructured textual data being produced and shared through natural language (Wang et al., 2012), creating significant challenges for digital management and information processing.

Studies have shown that over 80% of data in the construction industry exists in unstructured formats (Wu et al., 2022), making it difficult to extract and utilise valuable information efficiently. Natural Language Processing (NLP), defined as a set of techniques that help machines understand human languages, has emerged as a promising approach to address these challenges. NLP can transform unstructured textual information into structured data, facilitating improved information management in construction projects (Di Giuda et al., 2020). As suggested by the European Union, pairing BIM with digitalisation technologies like NLP can help realise the full potential of digital transition in the construction sector (Locatelli et al., 2021).

In early-stage building design, automating the extraction of textual data is essential for integrating physical and digital workflows. This process enables the extraction of valuable information from design briefs, client requirements, regulations, and project communications, facilitating informed decision-making and enhancing design quality. However, the success of NLP models heavily depends on the quality of data labelling, which directly affects model accuracy and performance.

This paper compares three primary data labelling techniques specifically for postcode extraction from construction project documents: manual labelling, rule-based techniques using regular expressions (Regex), and a hybrid machine learning approach combining Named Entity Recognition (NER) with rule-based methods. Postcodes represent a structured yet variable form of data commonly found in construction project documentation, making them an ideal candidate for comparing the effectiveness of different labelling approaches. The extraction of postcodes from project communications is particularly valuable as it provides location-specific information that can inform early-stage design decisions, site logistics planning, and regulatory compliance verification.

1.1 NLP in Building Design and Construction

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 (Yan et al., 2020; Darko et al., 2020).

Many researchers have attempted to analyse construction documents automatically, with early studies focusing on classifying or clustering construction documents for efficient management (Caldas et al., 2002). However, as construction projects have become larger and more complex, these document-level analyses have proven insufficient (Moon et al., 2018). Recent advances in data storage, computer processing, and deep learning methodologies have enabled NLP in the construction industry to enter a new phase (Jallan et al., 2019), allowing for more sophisticated analyses and applications.

Progressive advancements in text analytics and NLP have driven progress in construction-related studies, enabling various construction domains to achieve a degree of automation in recent years (Ding et al., 2022). 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 (Shamshiri et al., 2024).

1.2 Data Labelling Techniques

Data labelling is a critical step in developing effective NLP models, as it directly impacts model performance and accuracy. In the context of construction, three main approaches to data labelling have been utilised: manual labelling, rule-based techniques, and machine learning-driven methods. Each approach has distinct characteristics, advantages, and limitations that make it suitable for different types of NLP tasks.

1.2.1 Manual Data Labelling

Manual data labelling involves human annotators assigning predefined categories or tags to text elements, creating a gold standard dataset for training and evaluating NLP models. While this approach typically yields high-quality labelled data, it is labour-intensive and time-consuming, making it impractical for large datasets (Shamshiri et al., 2024).

In construction projects, experienced engineers possess domain knowledge crucial for accurate labelling. Their expertise allows them to interpret ambiguous or context-dependent information accurately. However, the process requires significant human resources. Moon et al. (2021) reported that six construction practitioners were involved in manually assigning word labels to 4,659 construction specification sentences for training an NER model. Each practitioner read every sentence and assigned appropriate categories to each word, with every labelled sentence cross-checked to ensure consistency.

The development of trustworthy labelled datasets requires following common principles for open data, such as the FAIR principles (Findable, Accessible, Interoperable, and Reusable) (Wilkinson et al., 2016). To ensure quality, multiple domain experts should annotate the dataset in parallel, with conflicts resolved by finding agreement among all annotators (Fuchs, 2021).

While manual labelling provides the highest accuracy and quality control, it is generally better suited for smaller datasets or for creating gold standards against which automated methods can be evaluated.

1.2.2 Rule-Based Data Labelling

Rule-based data labelling techniques automate the labelling process using predefined patterns, rules, and heuristics to identify and categorise text elements. These techniques, commonly implemented using regular expressions (Regex), are particularly effective for structured data with consistent patterns (Shamshiri et al., 2024).

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 (Locatelli et al., 2021).

In construction applications, rule-based approaches typically involve developing rules at different levels of complexity: simple rules based on basic features such as part-of-speech (POS) tags and phrase structure grammar (PSG) tags (Zhang & El-Gohary, 2016); tuple-patterns formed by several tokens or n-grams, such as subject-verb-object (SVO) tuples (Al Qady & Kandil, 2010); and complex rules created by combining simple rules with logical operators (AND, NOT, OR) (Xu & Cai, 2021).

Many rule-based approaches also use gazetteer lists—lists of fixed terms used to extract specific information types (Caldas et al., 2002; Ding et al., 2022; Fuchs, 2021). These lists are well-suited for extracting information types with little variation, such as negations, quantity units, and comparative relations.

Rule-based approaches in construction have shown promise in specific domains, as seen in Zhang and El-Gohary's extraction rules using phrase structure grammar (2016), Salama and El-Gohary's automated compliance checking framework (2011), and Lee et al.'s extraction model for poisonous clauses in international contracts (2019). Despite achieving high precision, these methods typically suffer from low recall and limited generalisation beyond their development corpus (Shamshiri et al., 2024).

1.2.3 Machine Learning-Driven Data Labelling

Machine learning-driven data labelling approaches use algorithms to learn patterns from data and automatically assign labels, reducing the need for manual rule crafting. Named Entity Recognition (NER), a subfield of machine learning-based information extraction, labels each word with predefined, informative categories such as names, locations, and objects (McCallum & Li, 2003).

The machine learning approach to NER has gained popularity in recent construction studies due to its robustness, expandability, and reduced resource requirements for model training compared to rule-based methods (Moon et al., 2020). Unlike rule-based approaches that rely on manually crafted patterns, machine learning models automatically identify usage patterns of words within text and acquire semantic information based on learning algorithms (Ratinov & Roth, 2009).

Two main types of NER approaches exist in the context of construction: Syntactic NER, which uses syntactic information and performs well with small, clean datasets because the syntactic expressions needed to determine word categories can be easily extracted (Newman et al., 2006). Semantic NER relies on semantic information and is known for its robustness and expandability compared to syntactic approaches (Ratinov & Roth, 2009).

Recent advances in deep learning have significantly enhanced NER performance. The Bidirectional Long Short-Term Memory (Bi-LSTM) architecture, often combined with Conditional Random Fields (CRF), has emerged as one of the most powerful architectures for NER tasks (Moon et al., 2020; Moon et al., 2021). Bi-LSTM contains two LSTM layers (forward and backward) that capture relationships between words in both directions and over long distances, allowing for more context-aware labelling (Moon et al., 2021).

The introduction of pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers) in 2018 has further revolutionised NLP tasks. BERT and other Transformer-based models learn contextual relations between words and can tackle a broad set of NLP tasks successfully (Devlin et al., 2019). These models address the challenge of limited training data through transfer learning, where models pre-trained on large corpora are fine-tuned for specific tasks (Nguyen et al., 2020).

While machine learning approaches offer scalability and adaptability, they have limitations. Their main drawback is reliance on manual feature engineering, which is time-consuming and restricts model generalisation to specific datasets (Shamshiri et al., 2024), though deep learning models attempt to address this through automatic feature extraction. These approaches typically require large training datasets, with studies showing poor performance from insufficient data (Fuchs, 2021). Many machine learning models suffer from "black-box" characteristics that hinder interpretability (Paliwal & Kumar, 2011), unlike the transparent nature of rule-based systems.

2 Methodology

This study employed a systematic approach to compare three data labelling techniques for postcode extraction from construction project documents and email communications. The methodology was designed to provide a comprehensive evaluation of manual labelling, rule-based extraction using regular expressions, and a hybrid machine learning approach combining NER with rule-based methods.

2.1 Dataset and Data Preparation

The dataset consisted of construction project email communications and documents. These contained various types of postcodes relevant to early-stage building design. A total of 278 emails were collected and processed from real construction projects, including design briefs, client communications, contractor correspondence, and regulatory submissions. Each email contained structured information, including sender details, subject lines, email bodies, and attached documents such as PDFs, Word documents, and Excel files. Of these, 103 emails were used for postcode labelling and analysis. The postcodes were categorised into two main types: project-related postcodes and administrative postcodes.

Data preparation involved several preprocessing steps. These were implemented in Google Colab environment. The email files were organised into individual folders, with each email folder containing the original .msg file and any extracted attachments. PDF attachments were automatically extracted from email messages and converted to text using Optical Character Recognition (OCR) to capture both textual content and information embedded in images. Excel files were processed to extract textual content from all sheets and columns.

The dataset preparation process involved mounting Google Drive storage, installing required Python libraries, including spaCy for NLP processing, extract-msg for email parsing, PyMuPDF for PDF processing, pytesseract for OCR functionality, and pandas for data manipulation. The text extraction process converted all textual content to lowercase to ensure consistent processing across different extraction methods.

2.2 Implementation of Extraction Techniques

2.2.1 Manual Labelling

Manual labelling was performed to create a gold standard dataset against which automated techniques could be evaluated. The author and a colleague with construction industry experience manually reviewed each email communication and associated document to identify and extract relevant postcodes. This process proved highly time-consuming, requiring careful examination of each communication to distinguish between project-related postcodes and administrative postcodes from client offices or company addresses.

Through this systematic manual review, several key insights emerged about the location and context of project-relevant postcodes within construction communications. The process revealed specific patterns in how and where project postcodes typically appeared, which varied significantly from standard administrative postcodes. These observations became crucial for understanding the contextual clues that could differentiate between relevant project information and routine company correspondence.

The annotation was conducted twice by both annotators to ensure consistency, with the primary author performing a final verification check on all labelled data. This iterative manual analysis process informed the development of the hierarchical extraction strategies and contextual scoring mechanisms later implemented in the automated approaches.

2.2.2 Rule-Based Approach (Regex)

The rule-based approach was developed based on patterns identified during the manual labelling process. It utilised regular expressions to extract UK postcodes that follow standardised formats. The primary regex pattern implemented was: `\b[A-Z]{1,2}[0-9R][0-9A-Z]? ?[0-9][A-Z]{2}\b`. This pattern captures the standard format of UK postcodes, consisting of an outward code (1-2 letters followed by 1-2 digits) and an inward code (a digit followed by 2 letters), separated by an optional space.

The extraction process implemented a hierarchical approach to maximise accuracy and relevance. The system first attempted to extract postcodes from email subjects, as these often contained project-specific information. If no postcode was found in the subject, the system then examined the first three lines of email bodies, which typically contain the most relevant project information. Additional patterns were implemented to extract postcodes following the word "postcode" in various formats.

When no postcodes were found through direct pattern matching, the system searched converted PDF text files that had been processed through OCR and Excel files. Finally, the system implemented signature detection to extract postcodes from email content while avoiding footer information that might contain irrelevant company addresses. Signature patterns included common email closings such as "Best regards," "Kind regards," "Thanks," and company-specific terms.

A smart postcode selection function was implemented to identify the most contextually relevant postcode when multiple postcodes were present. This function scored postcodes based on surrounding keywords that indicated project relevance, such as "project," "site," "address," "location," "property," "building," "client," "premises," "survey," and "works." The function penalised postcodes appearing near the company footer information and prioritised postcodes appearing earlier in documents.

2.2.3 Hybrid Machine Learning Approach

Initially, a pure Named Entity Recognition approach was attempted using spaCy's pre-trained English language model. However, this approach proved insufficient due to several limitations: the pre-trained model was not optimised for construction domain terminology, postcodes do not fit standard named entity categories, and the model struggled to distinguish between project-relevant postcodes and administrative addresses within construction communications.

Consequently, a hybrid approach was developed that combined the pattern recognition capabilities of regular expressions with the contextual understanding of Named Entity Recognition. A custom spaCy pipeline was configured with entity ruler patterns to recognise postcode entities using both full postcode patterns and first token patterns for partial matches.

The NER component utilised the pre-trained English language model "en_core_web_sm" enhanced with custom entity patterns. Two primary entity labels were defined: "POSTCODE" for complete postcodes and "POSTCODE_TOKEN" for first token matches. The entity ruler was positioned before the standard NER component in the processing pipeline to ensure custom patterns took precedence.

The hybrid extraction process implemented a fallback mechanism where regex extraction was attempted first, followed by NER-based extraction if regex methods failed to identify relevant postcodes. This approach leveraged the efficiency and precision of regex for well-formatted postcodes while utilising NER's contextual understanding for ambiguous or non-standard cases.

The NER extraction functions implemented both full postcode extraction and first token extraction capabilities. Full postcode extraction prioritised earlier occurrences in documents, while first token extraction focused on identifying partial postcodes that might be formatted differently. The system processed email subjects with the highest priority, followed by PDF/text files, email bodies before signatures, and Excel files with the lowest priority to avoid extracting irrelevant administrative postcodes.

The hybrid approach incorporated contextual analysis to distinguish between project-related postcodes and administrative information. The system analysed the surrounding text for project-relevant keywords and applied scoring mechanisms to select the most appropriate postcode when multiple candidates were identified. Signature detection and stopping point identification ensured that the extraction focused on relevant content while avoiding company footer information.

3 Results and Discussion

The comparative evaluation of the three data labelling techniques revealed significant differences in performance, efficiency, and practical applicability. This analysis focused on postcode extraction in construction project documents. The analysis was conducted using accuracy metrics comparing automated extraction results against the manually labelled gold standard dataset.

The rule-based approach using regular expressions achieved an accuracy of 96.7% when compared to manual labelling results. This demonstrates the effectiveness of regex patterns for identifying structured data elements like postcodes that follow standardised formats. The hybrid machine learning approach, combining NER with rule-based methods, achieved the highest accuracy at 98%, highlighting the value of combining multiple extraction techniques to handle variations in data formatting and contextual presentation.

Processing time analysis revealed substantial differences between approaches. Manual labelling required over 10 minutes per email for trained annotators to carefully review content and determine contextual relevance. The rule-based approach processed emails in under 1 second, while the hybrid approach required less than 3 seconds per email, including additional processing time for NER analysis.

Table 1 presents a performance comparison based on implementation results and observations from this study.

Table I. Comparative performance of manual, rule-based, and hybrid data labelling techniques for postcode extraction in construction project documents, evaluated using accuracy, processing time, scalability, and contextual understanding

Title 1 Criteria	Manual Labelling	Title 3 Regex	Hybrid
Accuracy (%)	100 (Gold Standard)	96.7	98
Processing Time per Email	>10 minutes	<1 second	<3 seconds
Scalability	Poor	Excellent	Excellent
Consistency	Variable (Annotator dependent)	High	High
Setup Complexity	Low	Medium	High
Contextual Understanding	Excellent	Limited	Good
Resource Requirements	High	Low	Low
Cost per Document	High	Very low	Low

Error analysis revealed that the 3.3% accuracy gap between rule-based and manual approaches primarily resulted from cases involving non-standard postcode formatting, ambiguous contextual placement, and OCR errors in PDF processing. The hybrid approach reduced this gap to 2% by successfully handling most non-standard formatting cases through its NER component.

The hybrid approach excelled in handling edge cases that challenged pure rule-based methods, including postcodes with non-standard formatting and cases where multiple postcodes required contextual disambiguation. The NER component's contextual understanding proved valuable for selecting the most relevant postcode when multiple candidates were present.

The scalability assessment revealed that automated approaches could handle unlimited document volumes without proportional increases in human resources, while manual labelling requirements scaled linearly with dataset size. This makes automated approaches essential for processing large construction project document collections. Additionally, automated approaches provided consistent results across similar document types, while manual labelling could exhibit variability between different annotators despite training and guidelines.

4 Conclusions

This study successfully compared three data labelling techniques for extracting postcodes from construction project email communications and associated documents, demonstrating the practical viability of automated information extraction approaches in early-stage building design. The hybrid machine learning approach achieved the highest accuracy at 98% compared to manual labelling, while the rule-based regex approach achieved 96.7% accuracy, both representing significant performance levels for practical applications.

The findings reveal that while manual labelling provides the gold standard for accuracy and contextual understanding, automated approaches can achieve near-equivalent performance with substantial improvements in efficiency and scalability. The hybrid approach proved most effective by combining the pattern recognition strengths of regular expressions with the contextual understanding capabilities of Named Entity Recognition, successfully handling variations in postcode formatting and document structure that challenged pure rule-based methods.

The 98% accuracy achieved by the hybrid approach demonstrates that sophisticated automated techniques can reliably extract structured information from unstructured construction communications. This performance level supports the integration of automated information extraction systems into early-stage building design workflows, enabling efficient processing of large document collections while maintaining high accuracy standards.

The approach demonstrated in this study could be applied in practice by integrating automated information extraction into early design workflows, enabling rapid processing of client requirements and site information. Furthermore, these techniques could enhance digital twin systems and BIM platforms by automatically populating location-based data from project communications, supporting more informed decision-making in construction project management.

The research contributes a practical framework for selecting appropriate data labelling techniques based on specific project requirements. For applications requiring the highest accuracy with unlimited human resources, manual labelling remains optimal. For high-volume processing with structured data formats, rule-based approaches offer excellent efficiency with good accuracy. For complex applications requiring both accuracy and flexibility, hybrid machine learning approaches provide superior performance, justifying their additional computational requirements.

Future work should explore several promising research directions to extend this study's findings. The application of these techniques to other technical design data commonly found in construction communications would broaden their practical utility. Investigation of transfer learning approaches and the development of domain-specific pre-trained language models for construction could further improve NER performance for construction-specific terminology and contexts, enhancing the contextual understanding capabilities of machine learning approaches. Additionally, integrating these extraction techniques into broader Building Information Modelling workflows could enhance digital twin development by efficiently extracting relevant information from project communications and populating BIM models with real-world project data.

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

Al Qady, M., & Kandil, A. (2010). Concept relation extraction from construction documents using natural language processing. Journal of Construction Engineering and Management, 136(3), 294-302.

Caldas, C. H., Soibelman, L., & Han, J. (2002). Automated classification of construction project documents. Journal of Computing in Civil Engineering, 16(4), 234-243.

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.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

Di Giuda, G. M., Locatelli, M., Schievano, M., Pellegrini, L., Pattini, G., Giana, P. E., & Seghezzi, E. (2020). Natural Language Processing for Information and Project Management. In Digital Transformation of the Design, Construction and Management Processes of the Built Environment (pp. 95-102). Springer International Publishing.

Ding, Y., Ma, J., & Luo, X. (2022). Applications of natural language processing in construction. Automation in Construction, 136, 104169.

Fuchs, S. (2021). Natural Language Processing for Building Code Interpretation: Systematic Literature Review Report. Technical Report, The University of Auckland.

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), 04520009.

Jallan, Y., Brogan, E., Ashuri, B., & Clevenger, C. M. (2019). 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, 11(2), 04519024.

Lee, J., Yi, J. S., & Son, J. (2019). Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based NLP. Journal of Computing in Civil Engineering, 33(2), 04019003.

Locatelli, M., Seghezzi, E., Pellegrini, L., Tagliabue, L. C., & Di Giuda, G. M. (2021). Exploring natural language processing in construction and integration with building information modeling: A scientometric analysis. Buildings, 11(12), 583.

McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003 (pp. 188-191).

Moon, S., Chung, S., & Chi, S. (2020). Bridge damage recognition from inspection reports using NER based on recurrent neural network with active learning. Journal of Performance of Constructed Facilities, 34(6), 04020119.

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.

Moon, S., Shin, Y., Hwang, B. G., & Chi, S. (2018). Document management system using text mining for information acquisition of international construction. KSCE Journal of Civil Engineering, 22(12), 4791-4798.

Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. In Proceedings of the International Conference on Intelligence and Security Informatics (pp. 93-104).

Nguyen, M. T., Phan, V. A., Linh, L. T., Son, N. H., Dung, L. T., Hirano, M., & Hotta, H. (2020). Transfer learning for information extraction with limited data. Communications in Computer and Information Science, 1215, 469-482.

Opitz, F., Windisch, R., & Scherer, R. J. (2014). Integration of document- and model-based building information for project management support. Procedia Engineering, 85, 403-411.

Paliwal, M., & Kumar, U. A. (2011). Assessing the contribution of variables in feed forward neural network. Applied Soft Computing, 11(4), 3690-3696.

Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning (pp. 147-155).

Rezgui, Y., & Zarli, A. (2006). Paving the way to the vision of digital construction: A strategic roadmap. Journal of Construction Engineering and Management, 132(7), 767-776.

Salama, D. M., & El-Gohary, N. M. (2011). Semantic modeling for automated compliance checking. In International Workshop on Computing in Civil Engineering 2011 (pp. 641-648).

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

Wang, C. C., Plume, J., & Jim, P. (2012). A review on document and information management in the construction industry: From paper-based documents to BIM-based approach. In Proceedings of the 2012 International Conference on Construction and Real Estate Management (pp. 369-373).

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3(1), 1-9.

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

Xu, X., & Cai, H. (2021). Ontology and rule-based natural language processing approach for interpreting textual regulations on underground utility infrastructure. Advanced Engineering Informatics, 48, 101288.

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.

Zhang, J., & El-Gohary, N. M. (2016). Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking. Journal of Computing in Civil Engineering, 30(2), 04015014.

Zhu, Y., Raja, R. A. I., & Cox, R. F. (2001). Web-based construction document processing via malleable frame. Journal of Computing in Civil Engineering, 15(3), 157-169.