Joint CSCE Construction Specialty & CRC Conference 2025
Conférence conjointe spécialisée en construction de la SCGC et CRC-2025

Montreal, Quebec
July 28-31, 2025 / *28-31 juillet 2025*

ADVANCEMENT OF CONSTRUCTION CONTRACT ANALYSIS WITH NATURAL LANGUAGE PROCESSING AND LARGE LANGUAGE MODELS: A LITERATURE REVIEW

Meta Soy[1]*, Tianru Zhao[1], Hosam Hegazy, Ph.D.[1], Jiansong Zhang, Ph.D.[1], and Emad Elwakil, Ph.D.[1]

[1] School of Construction Management Technology, Purdue University, USA

soym@purdue.edu, zhao1463@purdue.edu, hegazyh@purdue.edu,
zhan3062@purdue.edu, eelwakil@purdue.edu

ABSTRACT: Construction contracts are complex and contain lengthy textual information which is legally binding. To fully digest those contracts to extract key information is hard and time consuming and requires expertise to work on that. Natural language processing (NLP), due to its ability to understand human-like writing, were researched on in several ways to support the contract checking, validating, developing, or information extraction. With the recent booming use of Large Language Models (LLMs), a powerful subset of NLP, its application for textual information is further increasing. This study aims to present a systematic literature review conducted based on the use of NLP and LLMs in support of construction contract analysis. Filtering through 81 papers from the Scopus to 19, the review focuses on how NLP and LLM are used in the construction contract including the method, techniques, and results. A Systematic approach for applying NLP and LLMs in construction contract analysis is also presented. The review concludes that the use of NLP and LLMs in construction contract analysis could improve the accuracy of the document, minimize the potential misunderstanding between the stakeholders, reduce the contact reviewing effort, and fast track the development process.

KEYWORDS: Construction Contracts, Natural Language Processing, Large Language Model, NLP, LLM, Contract Analysis

1. INTRODUCTION AND BACKGROUND

In the construction industry, there is a large amount of textual information in digital form which is used in all the construction processes, ranging from contract, specification, documents for planning, procurement, method statement, and many others (Yan et al., 2022). Among those, the construction contract is the main legally binding document which contains all elements and aspects of the agreement between the parties involved in the project. Such crucial documents require experts to prepare, analyze, and fully understand them before any final agreement can be reached. Generally, whether following the format of AIA, Consensus, FIDIC, NEC, or others, the construction contract standards at its core are textually based, legally binding documents. It contains information to clarify the expectations of each party, illustrate risk management, indicate all aspects of legal protection, show the project management roadmap, and ensure monetary responsibility and security, among other details.

With the advancement of technology, Natural Language Processing (NLP), a field of computer science and artificial intelligence that can understand natural language such as human writing and speech, has been extensively researched to facilitate the analysis of construction documents in several ways. Some of those examples include extracting concepts and its relations for contract review and management (Al Qady and Kadil, 2010), detecting poisonous clauses (Lee et al. 2019), classifying the contact requirements (Hassan and Le, 2020), and so on. In the review done by Hussain et al. (2024), many methods and techniques of NLP were employed to support various textural processing in construction management. While the benefits of NLP in supporting textual documentation are clear, new capacities of NLP in the form of large language models (LLMs) have emerged. LLMs gained extensive popularity after the public release of ChatGPT by OpenAI at the end of 2022 (OpenAI, 2022). Large Language models (LLMs) are deep learning neural network architectures which were trained on billions of parameters and required massive amounts of textual input for training. LLMs such as the GPT series by OpenAI, Gemini by Google, Llama by Meta, Sonnet by Claude, and others can understand long and complex human-like text with remarkable accuracy, even within the context of their trained data. This leads to the booming use of LLMs in various sectors.

Seeing the surge in popularity of both NLPs and LLMs and their contributions, they are started being used in the construction sector. An initial review of NLP and LLM usage was performed based on the Scopus database. With a focus on construction contracts, the literature of related papers was digested, considering current search trends and the application of NLP and LLMs, in the context of contract analysis. So, this paper aims to cover the application of NLP and LLM in the construction contract from 2000 to 2025.

2. METHODOLOGY

The Prisma Criteria for systematic review are followed for the processes as shown in Figure 1. The papers are selected based on their related construction contents and from the construction-oriented database. Those selected papers are summarized, analyzed, in the use of NLP and LLM in their contract context application.

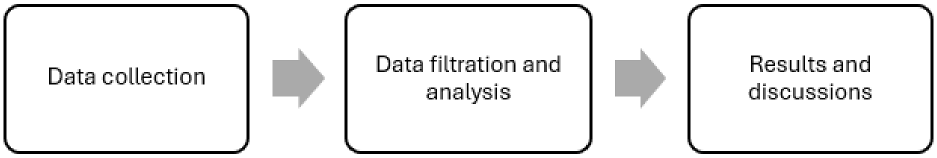


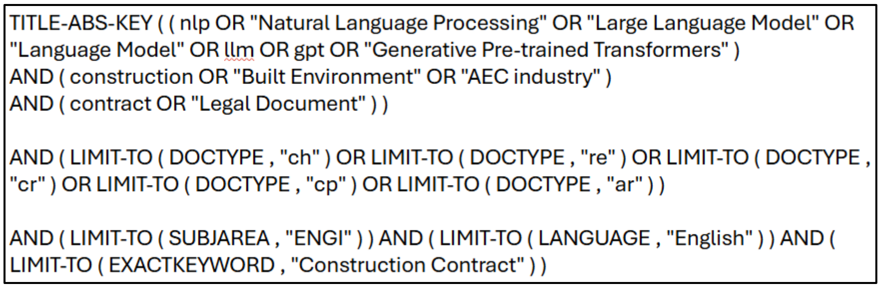
Figure 1: Research methodology

2.1 Data collection

The data was collected from Scopus, a peer-reviewed and indexed database. Firstly, the search criteria were defined based on the search term, and the keywords are shown in Table 1. The search criteria were placed in the advanced search mode with “AND” and “OR” operators. Figure 2 illustrates one of the advanced search combinations.

Table 1: Search keyword

LLM related	Main Key Word	Industry Related
NLP	Contract	Construction
Natural Language Processing	Legal document	Built Environment
LLM		AEC Industry
GPT		
Large Language Model		
Language Model		
Generative Pre-trained Transformers		



```
TITLE-ABS-KEY ( ( nlp OR "Natural Language Processing" OR "Large Language Model" OR
"Language Model" OR llm OR gpt OR "Generative Pre-trained Transformers" )
AND ( construction OR "Built Environment" OR "AEC industry" )
AND ( contract OR "Legal Document" ) )

AND ( LIMIT-TO ( DOCTYPE , "ch" ) OR LIMIT-TO ( DOCTYPE , "re" ) OR LIMIT-TO ( DOCTYPE ,
"cr" ) OR LIMIT-TO ( DOCTYPE , "cp" ) OR LIMIT-TO ( DOCTYPE , "ar" ) )

AND ( LIMIT-TO ( SUBJAREA , "ENGI" ) ) AND ( LIMIT-TO ( LANGUAGE , "English" ) ) AND (
LIMIT-TO ( EXACTKEYWORD , "Construction Contract" ) )
```

Figure 2: Exact advanced search options

2.2 Data filtration and analysis

The initial search results with selected search terms or search term combinations resulted in numerous papers. Therefore, specific filtering criteria are applied. The first one is to select only in English publications and published in the related Engineering domain journal or conferences in the last decade. After applying the first filtration, as shown in Figure 1, the number of papers was reduced from 81 to 60. The second filtration focuses on papers with extracted keywords such as "Contracts", “Construction Contract", "Construction Contracts”, “Condition Of Contract”, “Conditions Of Contract”, “Construction Contract Risk”, “Contracts Provisions”, “Contract Preparation”, “Contract Management”, “Contract Complexity”, “Contract Risk”, “Contractual Documents”, “Contractual Function”, “Contractual Functions”, “Exculpatory Clauses”, “Engineering Contracts”, “Contract Clause”, “Contract Conditions”, “Contract Document”, “Project Contract”, “Contract Administration”. The term “smart contract” is not considered because it refers to blockchain technology, which is not the focus of this paper. After this 2nd filtration, 27 papers are obtained. The 3rd filtration focuses on the papers that study or emphasize contract analysis. There are 19 papers in the final selection.

3. RESULTS AND DISCUSSIONS

After filtering the papers from the selected database, the papers are reviewed based on the specific focus, such as the aim, the method used, the application of the visual programming language, and the architectural components that the authors are experimenting with or selecting for their studies, as shown in Table 2.

The reviewed papers show different aspects of applying NLP (rule-based, machine learning-based, and deep learning-based) to automate information extraction from the contract. These studies offer a range of insights into the effectiveness and challenges associated with NLP in identifying missing information in the contract, predicting risk-handling actions, and calculating semantic similarity, as shown in Figure 3 and Figure 4.

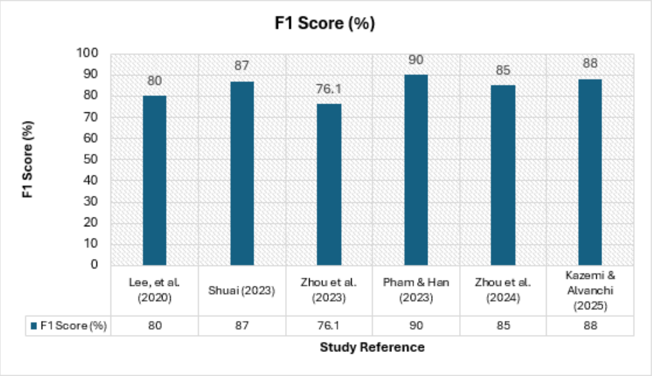


Figure 3: F1 score value according to the recent publication

Table 2: Sample Literature Summary

Study	Application Area	Methodology	Data Source	Contribution
Al Qady & Kandil, 2010	Concept relation extraction in construction contracts	Natural Language Processing (NLP), automated relation extraction	Construction contract documents	Utilized NLP for automatic concept relation extraction, improving document categorization and retrieval
Hassan & Le, 2020	Requirements identification in contracts	Machine Learning (ML), NLP	Construction contract documents	Developed NLP and ML-based framework for automatic classification of contractual requirements
Jafari et al., 2021	Generating reporting requirement extraction	NLP, Simulation, Cost Estimation	Construction contracts	Applied NLP and simulation to predict time and cost of contractual reporting requirements
Lee et al., 2020	Risk identification in FIDIC contract cases	Rule-based NLP, risk detection	FIDIC contract cases	Implemented rule-based NLP for detecting missing contractor-friendly clauses
Padhy et al., 2021	Identification of exculpatory clauses in contracts	NLP-based classification	Standard contract conditions	Developed an NLP model to detect risk-shifting clauses in contracts
Xue et al., 2022	Contract summarization	Deep Learning (DL), Text Summarization	Construction contracts	Tested deep learning models for summarizing lengthy construction contracts
Candaş & Tokdemir, 2022a	Identification of vague clauses in FIDIC Silver Book contracts	NLP, Machine Learning	FIDIC Silver Book contracts	Applied NLP and ML to detect ambiguous contractual clauses, improving clarity and dispute prevention
Candaş & Tokdemir, 2022b	Review coordination in construction contracts	Multi-label Text Classification	Construction contracts	Developed an NLP and ML-based system for multi-label classification of contract clauses by department relevance
Shuai, 2023	Risk identification in contract modifications	Rationale-augmented NLP, BERT	Construction contracts	Developed a rationale-augmented NLP framework using BERT for detecting unilateral contractual changes
Hassan et al., 2023	Digitalization of scope of work requirements	NLP framework	Scope of work obligations	Developed an NLP framework for extracting and structuring scope of work obligations in contracts
Fu et al., 2023	Machine-coding construction contract functions	DeBERTa-based NLP model	Construction contracts	Developed a DeBERTa-based NLP model for measuring contract complexity based on control, coordination, and adaptation functions
Pham & Han, 2023	Prediction of risk-handling actions in contracts	NLP, ML, DL, Text Classification	FIDIC project construction contracts	"Classified contract clauses into risk identification, risk allocation, and risk response categories by applying a multitask classification model integrating shared layers and task"
Zhou et al., 2023	Contract missing clauses Detection	NLP, DL, Text Understanding, Text Classification	Public legal document websites	Applied DL and NLP models for understanding and classifying contract to avoid missing clauses

Gao et al., 2024	Analysis of long construction contracts	Large Language Models (LLMs), text segmentation, intelligent Q&A	FIDIC contracts	Developed an LLM-based approach with two-stage segmentation and intelligent Q&A to enhance readability and efficiency
Wong et al., 2024	Risk identification in construction contracts	Knowledge-augmented LLMs	Construction contract risk databases	Utilized knowledge-augmented LLMs to improve contract risk identification without fine-tuning
Zhou et al., 2024	Identification of risks in construction contract clauses	NLP, Ontology	FIDIC, AIA, NEC	Provided theoretical support for the semantic analysis and pragmatic analysis studies of construction contract clauses
Kazemi & Alvanchi, 2025	Risk detection in contracts written in complex script systems	BERT-based NLP, deep neural networks	Farsi construction contracts	Implemented BERT-based NLP for detecting risky contract statements in Farsi contracts with high accuracy
Dikmen et al., 2025	Construction contract risk and responsibility assessment	NLP, ML, Text Classification	FIDIC templates, actual contracts	Applied NLP and ML models for classifying contract clauses into risk and responsibility categories, improving bid preparation and risk management
Kim et al., 2025	Inherent risks identification in a contract document	NLP, BERT	FIDIC	Using NLP and BERT to develop a method for generating automatically risk sentence identification rules to identify risks from the contract document.

The studies cover diverse categories of contract analysis, from identifying the requirements (Hassan et al., 2020; Jafari et al., 2021,Hassan et al., 2023) to finding the missing clause in the contract comparing to the standard contract (Zhou et al., 2023,) or missing contractor's friendly clause (Lee et al., 2020) and identifying the potential risks (Wong et al., 2024, Zhou et al. 2024, Kazemi & Alvanchi 2025, Dikmen et al. 2025, Kim et al. 2025). Besides, the abnormal case of the contract clause such as vague (Candaş & Tokdemir, 2022a) or risk-shifting (Padhy et al., 2021) were also investigated. Beyond the strong capability of contract summarization (Xue et al., 2022, Gao et al., 2024), NLP based framework was also used in prediction of risk mitigation action (Pham & Ham, 2023), linking risks to relevant responsible department (Candaş & Tokdemir, 2022b) or classifying the complexity of contracts (Fu et al., 2023). This indicates that NLP coupling with ML and DL and other supported technologies can be applied to different contract categories, which is flexible and could be widely used.

From the literatures, the NLP is seen playing a crucial role in those studies. NLP was used along side various technologies such as the rule-based algorithm (Hassan et al., 2020; Lee et al., 2020; Shuai et al., 2023), machine learning-based algorithm (Shuai et al., 2023) and deep learning-based ones (Kazemi & Alvanchi 2025). Ontology (Zhou et al., 2024) was also playing an important part of those application. Bidirectional Encoder Representations for Transformers (BERT), a deep learning model, is used extensive in supporting the application development (Fu et al., 2023, Shuai, 2023, Wong et al., 2024, Kim et al., 2025, Kazemi & Alvanchi, 2025). LLM such as Llama model and GPT model was only seen using with the research of Gao et al. (2024). While many researchers consider BERT as a deep learning model, the model itself can be a consider a type of LLM. The versatility of NLP and LLM, especially the use of BERT, allows for their seamless integration with other technologies, depending on the objectives and challenges of the specific application.

The research methods are varied for the studies; at the testing stage, manual extraction (Padhy et al., 2021) or classification (Candaş et al., 2022a) are applied to identify the clauses and sentences, while some of the papers directly train the model using the prepared dataset (Hassan et al., 2020; Kazemi et al., 2025). However, for the validation stage, most of the studies finished the tasks manually (Hassan et al., 2020; Lee et al., 2020; Padhy et al., 2021; Candaş et al., 2022; Wong et al., 2024), some of them leverage the model per se (Shuai et al., 2023; Zhou et al.,2024; Pham et al., 2023).

The reviewed papers indicate a strong consensus on the effectiveness of NLP and LLM in extracting target information, predicting and handling potential risks, and identifying missing information from the contract. NLP and LLM save the time of going through the contract manually for the professionals and, at the same time, detect the risk and missing information, which enables the construction projects to develop more efficiently and accurately, as shown in Figure 3 and Figure 4.

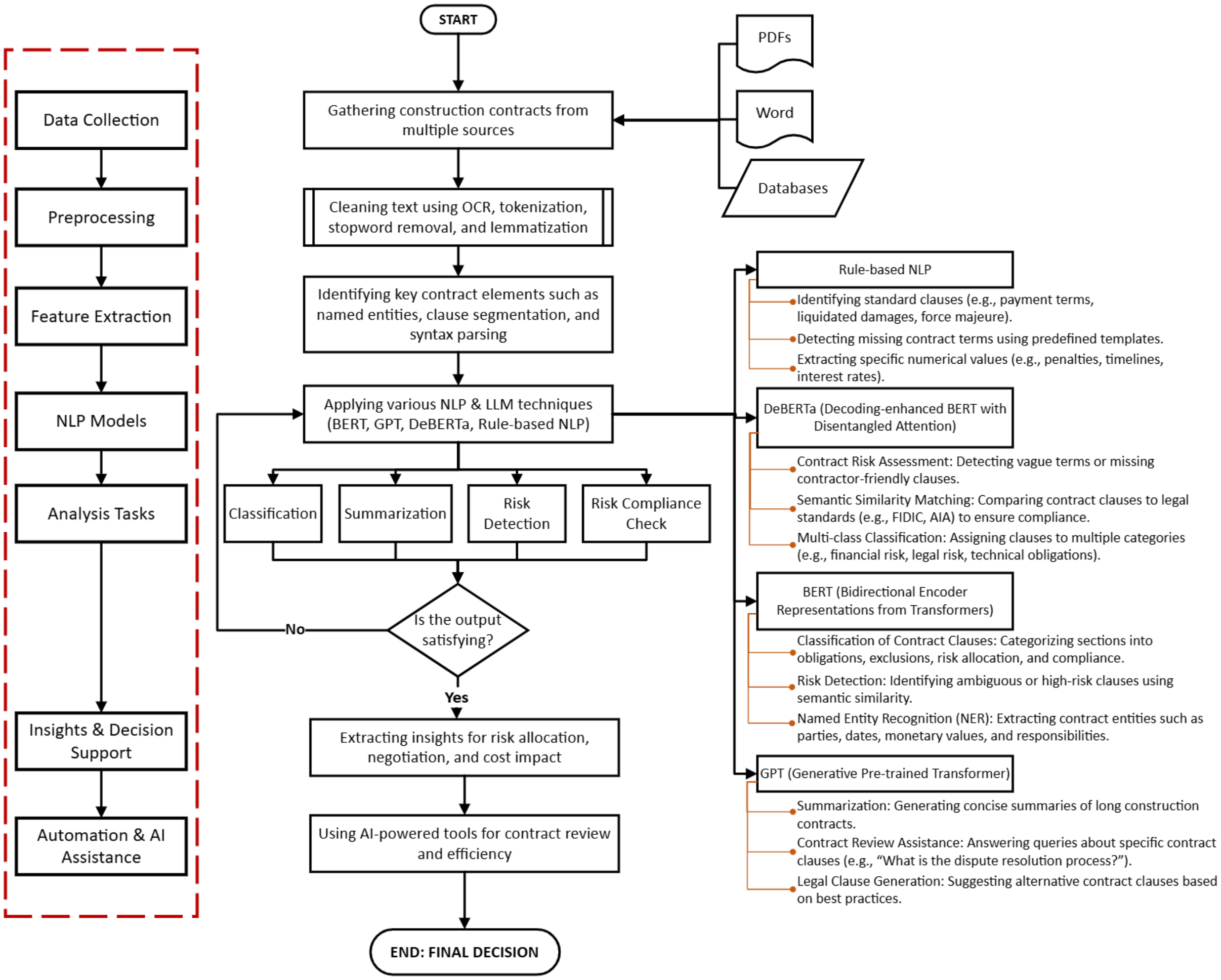

Data Collection
Preprocessing
Feature Extraction
NLP Models
Analysis Tasks
Insights & Decision Support
Automation & AI Assistance
START
Gathering construction contracts from multiple sources
PDFs
Word
Databases
Cleaning text using OCR, tokenization, stopword removal, and lemmatization
Identifying key contract elements such as named entities, clause segmentation, and syntax parsing
Applying various NLP & LLM techniques (BERT, GPT, DeBERTa, Rule-based NLP)
Classification
Summarization
Risk Detection
Risk Compliance Check
No
Is the output satisfying?
Yes
Extracting insights for risk allocation, negotiation, and cost impact
Using AI-powered tools for contract review and efficiency
END: FINAL DECISION
Rule-based NLP
Identifying standard clauses (e.g., payment terms, liquidated damages, force majeure).
Detecting missing contract terms using predefined templates.
Extracting specific numerical values (e.g., penalties, timelines, interest rates).
DeBERTa (Decoding-enhanced BERT with Disentangled Attention)
Contract Risk Assessment: Detecting vague terms or missing contractor-friendly clauses.
Semantic Similarity Matching: Comparing contract clauses to legal standards (e.g., FIDIC, AIA) to ensure compliance.
Multi-class Classification: Assigning clauses to multiple categories (e.g., financial risk, legal risk, technical obligations).
BERT (Bidirectional Encoder Representations from Transformers)
Classification of Contract Clauses: Categorizing sections into obligations, exclusions, risk allocation, and compliance.
Risk Detection: Identifying ambiguous or high-risk clauses using semantic similarity.
Named Entity Recognition (NER): Extracting contract entities such as parties, dates, monetary values, and responsibilities.
GPT (Generative Pre-trained Transformer)
Summarization: Generating concise summaries of long construction contracts.
Contract Review Assistance: Answering queries about specific contract clauses (e.g., "What is the dispute resolution process?").
Legal Clause Generation: Suggesting alternative contract clauses based on best practices.

Figure 4: Systematic approach for applying NLP and LLMs in construction contract analysis

4. CONCLUSION

For a long time, the project managers must read and check contracts manually, which is time-consuming and error prone. However, with the rise of NLP and LLM, these two models can be applied to skim over the contract, find if there are any discrepancies or risks that need to be handled, and evaluate the semantic similarity with other comparable documents. This paper presented a review on applying NLP and LLM technologies in facilitating contract management processes, highlighting how these tools can enhance project managers' efficiency, accuracy, and decision-making. The findings underscore the potential of NLP and LLM to revolutionize contract management in the construction industry and beyond. As these technologies evolve, their integration into contract analysis systems could pave the way for even more sophisticated applications, such as real-time risk monitoring, predictive data analytics, and dynamic contract generation. Ultimately, this paper contributes to the growing body of research demonstrating AI-driven tools' value in addressing complex, real-world challenges.

5. 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. 10.1061/(ASCE)CO.1943-7862.0000131

Candaş, A. B., & Tokdemir, O. B. (2022a). Automated Identification of Vagueness in the *FIDIC Silver Book* Conditions of Contract. *Journal of Construction Engineering and Management*, *148*(4). 10.1061/(ASCE)CO.1943-7862.0002254

Candaş, A. B., & Tokdemir, O. B. (2022b). Automating Coordination Efforts for Reviewing Construction Contracts with Multilabel Text Classification. *Journal of Construction Engineering and Management*, *148*(6). 10.1061/(ASCE)CO.1943-7862.0002275

Dikmen, I., Eken, G., Erol, H. & Birgonul, M.T. (2025). Automated construction contract analysis for risk and responsibility assessment using natural language processing and machine learning. Computers in Industry. 166. 10.1016/j.compind.2025.104251

Fu, Y., Xu, C., Zhang, L. & Chen, Y. (2023). Control, coordination, and adaptation functions in construction contracts: A machine-coding model. *Automation in Construction*. 152. 10.1016/j.autcon.2023.104890

Gao, Y., Gan, Y., Chen, Y., & Chen, Y. (2024). Application of large language models to intelligently analyze long construction contract texts. *Construction Management and Economics*, 1–17. 10.1080/01446193.2024.2415676

Hassan, F. ul, & 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).10.1061/(ASCE)LA.1943-4170.0000379

Hassan, F. ul, Le, T. & Le, C. (2023). Automated Approach for Digitalizing Scope of Work Requirements to Support Contract Management. *Journal of Construction Engineering and Management*, *149*(4). 10.1061/JCEMD4.COENG-1252

Hussain, F., Mehta, S., Soy, M., Zhang, J. (2024). Natural Language Processing for Construction Management: A Literature Review. Proc., Construction Research Congress 2024: Advanced Technologies, Automation, and Computer Applications in Construction, ASCE, Reston, VA, 607 - 618.

Jafari, P., Hattab, M. Al, Mohamed, E. & AbouRizk, S. (2021). Automated Extraction and Time-Cost Prediction of Contractual Reporting Requirements in Construction Using Natural Language Processing and Simulation. Applied Sciences. 11(13), 6188, 10.3390/app11136188

Kazemi, M. H., & Alvanchi, A. (2025). Application of NLP-based models in automated detection of risky contract statements written in complex script system. *Expert Systems with Applications*, *259*, 125296. 10.1016/j.eswa.2024.125296

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*(3). 10.1061/(ASCE)CP.1943-5487.0000807

Lee, J., Ham, Y., Yi, J.-S., & Son, J. (2020). Effective Risk Positioning through Automated Identification of Missing Contract Conditions from the Contractor’s Perspective Based on FIDIC Contract Cases. *Journal of Management in Engineering*, *36*(3). 10.1061/(ASCE)ME.1943-5479.0000757

OpenAI. (2024). Introducing ChatGPT. <https://openai.com/index/chatgpt/> (Dec. 7, 2024).

Padhy, J., Jagannathan, M., & Kumar Delhi, V. S. (2021). Application of Natural Language Processing to Automatically Identify Exculpatory Clauses in Construction Contracts. *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, *13*(4). 10.1061/(ASCE)LA.1943-4170.0000505

Pham, H. T. T. L., & Han, S. (2023). Natural Language Processing with Multitask Classification for Semantic Prediction of Risk-Handling Actions in Construction Contracts. Journal of Computing in Civil Engineering, 37(6). 10.1061/JCCEE5.CPENG-5218

Shuai, B. (2023). A rationale-augmented NLP framework to identify unilateral contractual change risk for construction projects. *Computers in Industry*, *149*, 103940. 10.1016/j.compind.2023.103940

Wong, S., Zheng, C., Su, X., & Tang, Y. (2024). Construction contract risk identification based on knowledge-augmented language models. *Computers in Industry*, *157–158*, 104082. 10.1016/j.compind.2024.104082

Xue, X., Hou, Y. & Zhang, J. (2022). Automated Construction Contract Summarization Using Natural Language Processing and Deep Learning. *In Proc. 39th International Symposium on Automation and Robotics in Construction*. 459-466. 10.22260/ISARC2022/0063

Yan, H., Ma, M., Wu, Y., Fan, H., & Dong, C. (2022). Overview and analysis of the text mining applications in the construction industry. *Heliyon*. 10.1016/j.heliyon.2022.e12088

Zhou, H., Gao, B., Tang, S., Li, B., & Wang, S. (2023). Intelligent detection on construction project contract missing clauses based on deep learning and NLP. *Engineering, Construction and Architectural Management*. 10.1108/ECAM-02-2023-0172

Zhou, H., Zhou, L., Gao, B., Huang, W., Huang, W., Zuo, J., & Zhao, X. (2024). Intelligent identification of risks in construction contract clauses based on semantic reasoning. *Engineering, Construction and Architectural Management*. 10.1108/ECAM-05-2023-0527