sustainability

MDPI

Article

Transformer-Based NLP for Construction Contract Clause Classification: Implications for Sustainable Construction Project Governance

Anıl Demircan * and Latif Onur Uğur

Civil Engineering Department, Düzce University, 81620 Düzce, Türkiye; latifugur@duzce.edu.tr
* Correspondence: anildemircan@gmail.com or anildemircan@duzce.edu.tr

Abstract

Construction contracts are vital for governing responsibilities in large-scale infrastructure projects, but their increasing complexity often leads to interpretation difficulties, disputes, and delays. Despite advances in natural language processing (NLP), automated analysis of construction contract clauses remains limited in project management. This study proposes a text classification framework integrating transformer-based contextual embeddings (BERT, ALBERT, RoBERTa, and DistilBERT) with machine learning and deep learning models (RNN, GRU, and LSTM) to analyze FIDIC and JCT contract provisions. Two multi-class classification tasks were defined: Dataset 1 (DS1) focusing on obligations, operational actions, optional provisions, general statements, and Dataset 2 (DS2) covering cost, quality, and time dimensions. Experimental results show that deep learning models consistently outperform traditional machine learning algorithms. Specifically, LSTM combined with RoBERTa and DistilBERT achieved the highest accuracy levels of 98.06% and 98.33% for DS1. The framework may support transparent contract governance by enabling faster and more consistent identification of contractual clauses. From a sustainability perspective, the findings suggest potential process-level contributions to economic efficiency, administrative workload reduction, and decision-making support throughout the project lifecycle.

Keywords: construction contracts; natural language processing; text classification; transformer models; construction project management; contract governance; sustainable construction management

Academic Editor: Andreas Kanavos

Received: 30 April 2026
Revised: 7 June 2026
Accepted: 14 June 2026
Published: 3 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The construction industry plays a crucial role in global sustainable development due to its significant economic scale, extensive resource consumption, and long-term impacts on infrastructure systems and society. Beyond environmental considerations, sustainability in construction increasingly encompasses economic efficiency, institutional transparency, effective risk management, and the ability to support informed decision-making throughout the project lifecycle. Within this broader perspective, construction contract management represents a fundamental mechanism that influences project performance, stakeholder coordination, and the overall governance of construction projects.

Construction projects typically involve multiple stakeholders, complex technical requirements, and substantial financial investments. In such environments, contractual frameworks serve as the primary instruments that regulate relationships among project participants and define responsibilities, rights, and risk allocation mechanisms. Standardized construction contracts developed by organizations such as FIDIC (Fédération Internationale Des Ingénieurs-Conseils), JCT (Joint Contracts Tribunal), and NEC (New Engineering Contract) are widely used in both national and international projects. These contracts provide structured procedures for managing key project dimensions including cost control, quality assurance, time management, and dispute resolution. Consequently, effective contract management processes contribute not only to project efficiency but also to improved transparency, accountability, and sustainable governance in construction projects.

Despite their importance, construction contracts are often highly complex documents that contain numerous clauses describing obligations, procedural requirements, and risk allocation mechanisms. The legal and technical language used in such contracts can make interpretation difficult, particularly when large volumes of documents must be reviewed within limited timeframes. Manual contract analysis therefore becomes a labor-intensive process that is prone to inconsistencies and human error. In large-scale infrastructure projects where rapid identification of contractual obligations and risks is critical, traditional manual analysis approaches may prove insufficient.

In response to these challenges, recent research has increasingly explored the use of natural language processing (NLP) techniques to analyze textual information contained in construction documents. NLP methods have demonstrated significant potential for supporting automated contract interpretation, clause classification, and risk identification. For instance, in 2020, Hassan and Le [1] highlighted the limitations of manual contract analysis and proposed a machine learning-based framework for automatically identifying contractual requirements. Their study evaluated several algorithms, including Support Vector Machines (SVM), Logistic Regression, and Naive Bayes, and reported that the SVM model achieved the highest accuracy of 98.51% in identifying contractual requirements. Similarly, Zhou et al. [2]. introduced an intelligent framework based on deep learning and natural language processing (NLP) for detecting missing clauses in construction contracts. By transforming unstructured contract text into structured representations, the proposed model utilizes a multi-label classification mechanism to identify absent or incomplete contractual provisions. The study highlights the potential of data-driven approaches to enhance the accuracy and efficiency of contract analysis compared to traditional manual methods.

Nevertheless, a substantial body of NLP research continues to rely on traditional text representation techniques such as TF-IDF and Bag-of-Words, which model text as collections of independent terms and thus fail to capture semantic relationships between words. While such approaches are effective in identifying surface-level lexical patterns, their ability to represent contextual dependencies remains limited. In the construction domain, Deza et al. [3] examined the classification of bill of quantities (BoQ) item descriptions into International Construction Measurement Standard (ICMS) categories using machine learning and deep learning models. Their findings show that BoQ descriptions are typically short, technical, and largely descriptive, with discriminative information often embedded in local key features. Consequently, relatively simple feature-based or locally oriented models can achieve strong performance in BoQ-to-ICMS classification, while the additional benefit of recurrent contextual modeling may be limited for this specific type of short construction cost text.

More recently, transformer-based language models such as BERT and its variants have achieved substantial improvements in natural language processing tasks by

generating contextualized representations of textual data. These models are capable of capturing semantic dependencies across sentences and have demonstrated superior performance in tasks such as text classification, semantic analysis, and information extraction. Despite these advances, the application of transformer-based language models to construction contract analysis remains relatively limited. In particular, comprehensive studies comparing different transformer embeddings in combination with machine learning and deep learning architectures for contract clause classification are still scarce.

From a sustainability perspective, the absence of efficient tools for automated contract analysis may hinder proactive decision-making in construction project management. Ineffective contract interpretation can lead to delayed responses, contractual disputes, administrative inefficiencies, and suboptimal resource utilization, which ultimately affect project performance and long-term infrastructure sustainability. Therefore, the development of advanced analytical frameworks capable of supporting automated contract analysis represents an important step toward improving project governance and promoting sustainable construction management.

To address this need, this study aims to develop a comprehensive framework for the automated classification of construction contract clauses using transformer-based contextual embeddings integrated with machine learning (ML) and deep learning (DL) models. Contractual provisions extracted from the FIDIC Conditions of Contract for Design, Build and Operate and the JCT Design and Build Contract were analyzed through two multi-class classification tasks. The first task categorizes clauses according to obligation types (mandatory, required, optional, and general provisions), while the second focuses on key project performance dimensions (cost-, quality-, and time-related clauses). By systematically evaluating multiple transformer embeddings together with several ML and DL architectures, the study aims to identify high-performing model combinations capable of improving automated contract analysis.

The main contributions of this study can be summarized as follows:

- **Development of a structured dataset for construction contract clause analysis.**
  This study constructs a clause-level dataset derived from two widely used standard form contracts, namely the *FIDIC Conditions of Contract for Design, Build and Operate* and the *JCT Design and Build Contract*. The dataset enables systematic analysis of contractual provisions related to obligations and key project management dimensions.
- **Integration of transformer-based language models with machine learning and deep learning classifiers.**
  The study proposes a comprehensive NLP framework that combines multiple transformer-based contextual embeddings (BERT, ALBERT, RoBERTa, and DistilBERT) with both traditional machine learning algorithms and deep learning architectures for automated contract clause classification.
- **Comparative evaluation of multiple NLP model architectures for contract analysis.**
  By systematically comparing different transformer embeddings together with several machine learning and deep learning models, the study provides empirical insights into the effectiveness of advanced NLP techniques for analyzing complex contractual language.
- **Implications for sustainable construction contract management.**
  The proposed framework may support faster and more consistent identification of contractual obligations and project-related clauses. These process-level outputs may assist decision-making and contract governance and provide a basis for future empirical assessment of sustainability-related benefits in construction project management.

In the context of sustainable construction, effective contract management plays a crucial role in ensuring transparency, accountability, and efficient resource utilization. Poor interpretation of contractual clauses often leads to disputes, delays, and cost overruns, which negatively affect the economic and institutional dimensions of sustainability. Therefore, developing automated tools for contract analysis is not only a technical challenge but also a key requirement for improving sustainability in construction project governance. This also aligns with the growing need to integrate digital technologies into sustainable construction management systems.

The remainder of this paper is organized as follows. Section 2 presents a review of the existing literature on NLP-based construction contract analysis, transformer models, and recent developments in large language model applications. Section 3 describes the research objectives and the overall methodological framework of the study. Section 4 explains the dataset construction process, preprocessing procedures, text representation techniques, and the machine learning and deep learning models used in the experiments. Section 5 presents the experimental results and performance evaluation of the proposed models. Finally, Section 6 discusses the findings, limitations of the study, and directions for future research.

2. Literature Review

2.1. NLP Applications in Construction Document and Contract Analysis

The growing volume of digital construction documents has accelerated the use of natural language processing (NLP) techniques in construction management research. Early studies showed that unstructured textual data such as contract clauses, specifications, and bidding documents can be analyzed automatically to support information extraction and risk-related decision-making. For example, Lee and Yi [4] applied text mining methods to improve bidding risk prediction by integrating textual and numerical data, while Lee et al. [5] developed a rule-based NLP framework to extract problematic clauses from international construction contracts. These studies demonstrated the feasibility of using NLP to support contract review, although they largely relied on rule-based or shallow text analysis approaches.

As NLP applications in construction matured, machine learning models were increasingly adopted for clause classification and requirement identification. Hassan and Le [1] proposed an NLP-based framework for automatically identifying contractual requirements from construction contract documents and showed that Support Vector Machines outperformed other traditional classifiers. In a related study, Hassan et al. [6] developed a classification model to categorize design–build contract requirements into predefined project-phase groups, demonstrating that automated requirement classification can support subcontract scope preparation and project definition processes. Jafari et al. [7] further combined NLP, machine learning, and Monte Carlo simulation to extract contractual reporting requirements and estimate their time–cost implications, thereby linking automated text analysis to project management performance.

These studies collectively confirm that NLP techniques can improve the efficiency of contract-related information extraction in construction. However, much of the earlier literature still depended on traditional text representation methods and task-specific pipelines, limiting the ability of those models to capture the contextual nuances of legal and technical language.

2.2. Automated Classification of Contractual Risks, Obligations and Clauses

A substantial body of recent research has focused specifically on automated classification of contractual risks, obligations, vague terms, and missing clauses. Moon et

al. [8] developed a BERT-based model to identify contractual risk clauses in construction specifications and reported strong performance across seven risk categories, clearly showing the advantage of transformer-based contextual representations over conventional machine learning baselines. Candaş and Tokdemir [9] addressed the problem of vagueness in FIDIC contracts, proposing an NLP-based methodology to identify ambiguous contractual expressions that could lead to disputes. Similarly, Candaş [10] introduced a semantic analysis framework intended to reduce review time and improve consistency in construction text interpretation.

Another important stream of research concerns missing or insufficient clauses. Zhou et al. [2] combined BERT with MatchPyramid to detect missing clauses in construction contracts and demonstrated that deep learning models can support both multi-label clause classification and semantic similarity assessment in long contractual documents. Zhong et al. [11] examined dispute classification in construction contracts using deep learning-based text mining techniques, showing that such approaches can effectively support dispute analysis in construction contract management.

More recent studies have shifted toward risk-oriented clause analytics in standard-form or model contracts. Kazemi and Alvanchi [12] evaluated transformer and deep learning models for detecting risky contract statements in Persian construction contracts and found that BERT produced the best overall results, outperforming other deep architectures. Dikmen et al. [13] developed NLP and machine learning models for automated construction contract analysis aimed at classifying risk, responsibility, and entitlement clauses in FIDIC-based texts and real project contracts. Their findings showed that contextual models such as BERT outperform traditional feature-based methods, especially for complex contractual expressions.

Taken together, these studies show that contract-focused NLP research has moved from rule-based extraction toward data-driven classification of risk-related and obligation-related clauses. Nevertheless, most prior studies still concentrate on one narrowly defined task, such as risk detection, vagueness identification, or missing clause detection, rather than offering a broader multi-class framework for construction contract analysis.

2.3. Deep Learning, Transformer Models, and the Emerging Role of LLMs

The broader construction text mining literature also shows a clear transition from shallow representations to deep contextual language models. Gao et al. [14] proposed a multi-sentence context-aware risk identification model that combines RoBERTa with BiGRU and demonstrated that contextual multi-sentence modeling improves classification performance in risk-related text analysis. Lee and Lee [15] developed a BERT-based keyword-extraction engine combined with graph models to automatically extract predefined risk-assessment knowledge from unstructured construction-safety data and build an entity–relationship-based risk-assessment knowledge base. Deza et al. [3] also demonstrated that neural architectures can achieve strong performance in classifying construction cost documents, indicating that deep learning approaches are useful beyond legal texts and can support broader construction management tasks.

Since 2024, the literature has begun moving beyond conventional transformer classification toward large language model (LLM) applications. Wong et al. [16] presented an LLM-enhanced approach for construction contract risk identification based on domain knowledge, showing that LLMs can emulate expert-style review of real construction contracts when supported by contract-specific knowledge structures. In 2025, Zheng et al. [17] proposed a tuning-free framework that integrates LLMs with a nested contract knowledge graph using GraphRAG for contract risk identification, emphasizing both accuracy and interpretability in automated contract review. Zhang et al. [18] developed a

domain-adapted construction contract question answering model using supervised fine-tuning and reinforcement learning on a FIDIC-based QA dataset, indicating that LLM-style systems are increasingly being tailored for contract interpretation tasks. Xiao et al. [19] further proposed a generative knowledge-guided review system for construction documents, again showing that LLMs are becoming central to automated document review.

This shift becomes even more evident in 2026 studies. Tang et al. [20] examined NLP and large language models in construction contract management and explicitly identified a gap between current academic work and practical industry needs, arguing that adoption remains fragmented despite rapid advances. Likewise, Nasser et al. [21] used NLP, text mining, and deep learning to analyze construction contractual semantics and develop a taxonomy of contract selection factors, showing that semantic analysis can support contract selection and contractual decision-making. These studies are especially important because they indicate that the field is expanding from clause classification toward broader contract management intelligence, semantic structuring, and decision support.

Table 1 shows that the literature has evolved from rule-based and traditional machine learning approaches toward transformer-based and, more recently, LLM-based systems for construction contract analysis. However, the dominant focus of prior studies remains narrow and task-specific. A systematic framework comparing multiple transformer embeddings together with both classical machine learning and sequential deep learning models for multi-class construction contract clause classification is still limited. This gap provides the main motivation for the present study.

Table 1. Summary of Previous Studies on NLP-Based Construction Contract Analysis.

Study	Data Source	Method/Model	Research Task	Key Finding
Lee & Yi [4]	Pre-bid clarification documents (RFI text + numeric bid data)	Text mining + SVM (integrated model)	Bidding risk prediction	When textual and numerical data were used together, the model accuracy increased to 72.9%, compared to 52.1% when only numerical data were used.
Lee et al. [5]	International construction contracts	Rule-based NLP	Problematic clause extraction	Combined syntactic and semantic analysis for clause extraction.
Hassan & Le [1]	Construction contract documents	Machine learning (NB, SVM, LR, FFNN)	Requirement identification	SVM achieved the highest classification performance among the evaluated models
Hassan et al. [6]	Design–build contracts	NLP and machine learning (NB, SVM, LR, KNN, DT, FFNN)	Requirement classification	Logistic regression achieved the highest accuracy (~94%), and automated classification supports subcontract scope preparation
Jafari et al. [7]	Contract documents	NLP + ML + simulation	Reporting requirement extraction	Linked requirement extraction to time–cost estimation.
Moon et al. [8]	2807 clauses from 56 projects	BERT	Risk clause classification	BERT outperformed conventional models with high F1 performance.
Candaş & Tokdemir [9]	FIDIC clauses	NLP + ML	Vagueness detection	Identified ambiguous contractual expressions.
Deza et al. [3]	50,000+ BoQ items	CNN/RNN/TCN	Cost document classification	Deep models achieved strong classification performance.

Zhong et al. [11]	Construction dispute cases	Deep learning-based text mining	Dispute classification	Demonstrated effective classification of dispute types, highlighting the applicability of NLP in post-contract analysis
Zhou et al. [2]	1322 contracts	BERT + MatchPyramid	Missing clause detection	Demonstrated deep learning for clause omission analysis.
Gao et al. [14]	Construction risk texts	RoBERTa + BiGRU	Risk identification	Multi-sentence context improved performance.
Lee & Lee [15]	Construction safety data	BERT + graph model	Risk-assessment knowledge extraction	Developed an entity-relationship- based knowledge base for construction risk assessment.
Wong et al. [16]	Real construction contracts	LLM + domain knowledge	Contract risk identification	Showed that LLMs can emulate expert-style contract review.
Kazemi & Alvanchi [12]	Contract sentences	BERT, LSTM, CNN, etc.	Risky statement detection	BERT achieved the best overall performance.
Dikmen et al. [13]	FIDIC + project contracts	NLP + ML + BERT	Risk and responsibility classification	Contextual models outperformed traditional approaches.
Zheng et al. [17]	Construction contracts	LLM + Knowledge Graph (GraphRAG, NCKG)	Automated contract review and risk identification	Knowledge graph-enhanced LLMs were proposed to improve the accuracy and interpretability of automated construction contract review.
Zhang et al. [18]	FIDIC-based QA dataset	NLP/LLM-based QA model	Contract question answering	Proposed a QA system that enables automatic understanding and answering of contract-related queries, extending beyond traditional classification tasks
Xiao et al. [19]	Construction disclosure documents	Generative knowledge-guided system	Automated review	Showed generative systems' potential in contract/document review.
Tang et al. [20]	Construction contract management literature	NLP + LLM review	Contract management support	Identified research–practice gap in LLM-enabled contract management.
Nasser et al. [21]	Contract texts/literature	NLP + text mining + deep learning	Contract taxonomy/selection factors	Demonstrated semantic taxonomy development for contract selection.

2.4. Research Gap and Motivation

Despite the growing body of research on NLP-based analysis of construction documents, several limitations remain in the existing literature. First, many earlier studies rely on traditional text representation methods such as TF–IDF, Bag-of-Words, and Word2Vec, which often fail to capture contextual relationships within complex legal texts such as construction contracts. Although these approaches have demonstrated the feasibility of automated contract analysis, they may not adequately represent semantic dependencies embedded in contractual language.

Second, while recent studies have increasingly adopted transformer-based language models such as BERT and RoBERTa, systematic comparisons of multiple contextual embeddings within a unified analytical framework remain relatively limited. As a result, the comparative effectiveness of different representation models for construction contract clause classification is not yet fully understood.

Third, much of the existing literature focuses on narrowly defined tasks, including risk detection, vague clause identification, requirement extraction, or dispute classification. However, construction contracts inherently contain multidimensional information related to obligations, operational actions, and key project management dimensions, such as cost, time, and quality, which are not adequately captured by task-specific approaches.

These limitations highlight the need for more comprehensive analytical frameworks that can capture contextual semantics and support multi-dimensional contract analysis. In this context, transformer-based models offer significant advantages due to their ability to model contextual dependencies through attention mechanisms, enabling more accurate representation of complex contractual language. Furthermore, integrating these representations with both machine learning and deep learning models allows for systematic evaluation of different modeling strategies within a unified framework.

While LLM-based approaches offer strong capabilities in reasoning and semantic understanding, their high computational cost, limited interpretability, and dependency on large-scale training data may restrict their direct applicability in domain-specific contract classification tasks. Therefore, this study focuses on transformer-based embeddings combined with efficient classification models.

Accordingly, this study proposes a multi-class, multi-dimensional classification framework that combines multiple transformer-based embeddings with machine learning and deep learning models for automated analysis of construction contract clauses. By systematically comparing different embeddings and classification architectures, the study aims to provide deeper insights into the effectiveness of advanced NLP techniques in construction contract analysis while contributing to improved transparency, reduced disputes, and more sustainable decision-making in construction project management.

2.5. Sustainability and Contract Governance

Sustainability in construction is increasingly discussed beyond environmental performance and also includes lifecycle value, resource efficiency, and governance-related considerations [22,23]. Effective contract management plays a central role in achieving these objectives by defining responsibilities, risk allocation, and decision-making structures. Previous studies have highlighted that poor contract clarity and inefficient contract administration contribute to project delays, disputes, and resource inefficiencies, which negatively affect sustainable project outcomes.

Digital technologies, including artificial intelligence and natural language processing, have the potential to improve contract governance by enabling automated analysis of contractual documents. Such technologies support faster identification of contractual risks and obligations, thereby enhancing decision-making processes and reducing inefficiencies. However, the integration of advanced NLP techniques into sustainable construction management remains limited.

3. Research Objectives and Methodological Framework

The primary objective of this study is to investigate the automated classification of construction contract clauses by systematically comparing multiple text representation techniques and classification algorithms. To achieve this goal, contextual embeddings generated by transformer-based language models—including BERT and its variants (ALBERT, DistilBERT, and RoBERTa)—are applied to construction contract texts derived from widely used standard form contracts.

These embeddings are evaluated in combination with various machine learning algorithms (Logistic Regression, Support Vector Machines, Random Forest, and Naïve Bayes) and deep learning architectures (RNN, GRU, and LSTM). The performance of the

proposed models is assessed using widely used classification metrics including accuracy, precision, recall, and F1-score in order to identify the most effective representation–model combinations for contract clause classification.

Beyond its methodological framework, this study provides several contributions to the construction management literature. First, the study introduces a structured dataset derived from widely used standard form construction contracts such as FIDIC and JCT. Unlike many previous studies that focus primarily on project-specific documents such as reports, claims, or correspondence, this research directly analyzes the core contractual documents that define responsibilities, obligations, and project management procedures.

Second, the study presents a systematic comparison of multiple text representation techniques and classification models for construction contract analysis. By integrating transformer-based contextual embeddings with both traditional machine learning and deep learning models, the study provides a comprehensive evaluation framework for analyzing complex contractual texts.

Third, the research introduces a multi-class classification framework that categorizes contractual clauses according to key construction project management dimensions, including obligations and time–cost–quality-related provisions. This approach extends previous research that typically focuses on a single task such as risk detection or contractor responsibility identification.

Finally, the proposed framework contributes to improving automated contract interpretation processes, which can support decision-making and enhance transparency in construction contract management. By facilitating more efficient analysis of contractual clauses, the study also contributes to more sustainable governance and management of construction projects.

4. Materials and Methods

This section describes the datasets used in the study, the text preprocessing procedures, the text representation techniques, the machine learning and deep learning models employed for classification, and the evaluation metrics used to assess model performance.

Generative AI tools were used as supportive tools for specific figure-preparation and manuscript-writing tasks. For schematic/illustrative figures, Google NotebookLM (Gemini 1.5 Pro) was used in the preparation of Figures 1–4. The authors first prepared the textual content and conceptual descriptions underlying these figures and uploaded these materials to Google NotebookLM (Gemini 1.5 Pro). Draft schematic/illustrative figure outputs were then generated based on author-prepared textual contents. Subsequently, ChatGPT (GPT-5.5) was used to prepare additional author-directed prompts for revising the generated draft visuals, improving their visual structure, and editing their textual elements. The alternative outputs were critically evaluated by the authors, the final figures to be included in the manuscript were selected by the authors, and all final versions were reviewed and approved by the authors.

For manuscript text, ChatGPT (GPT-5.5) was used in relation to the Experiments and Results section (Section 5) and the Discussion section (Section 6). These sections were first prepared by the authors in Turkish based on the authors' own analyses, experimental results, and interpretations. ChatGPT (GPT-5.5) was then used to support translation into academic English. Subsequently, AI-assisted suggestions were used to improve the wording, organization, clarity, and flow of these sections based on the authors' own analyses, results, and interpretations. In some cases, the AI tool provided wording alternatives and draft text suggestions to support the refinement and structuring of these sections. These suggestions were critically reviewed, revised, verified, and approved by the authors before inclusion in the manuscript.

AI tools were not used for data generation, data collection, annotation or labeling decisions, model training, statistical computation, generation of experimental results, independent data analysis, independent scientific interpretation, generation of data-driven result figures, or independent determination of the scientific content of any figure. All AI-assisted outputs were reviewed, revised, verified, and approved by the authors, who take full responsibility for the accuracy, integrity, and final content of the manuscript.

4.1. Research Workflow

The overall workflow of the proposed framework is illustrated in Figure 1. The process consists of several stages including contract document collection, sentence extraction, manual annotation, text preprocessing, representation using both traditional and transformer-based embeddings, classification using machine learning and deep learning models, and performance evaluation.

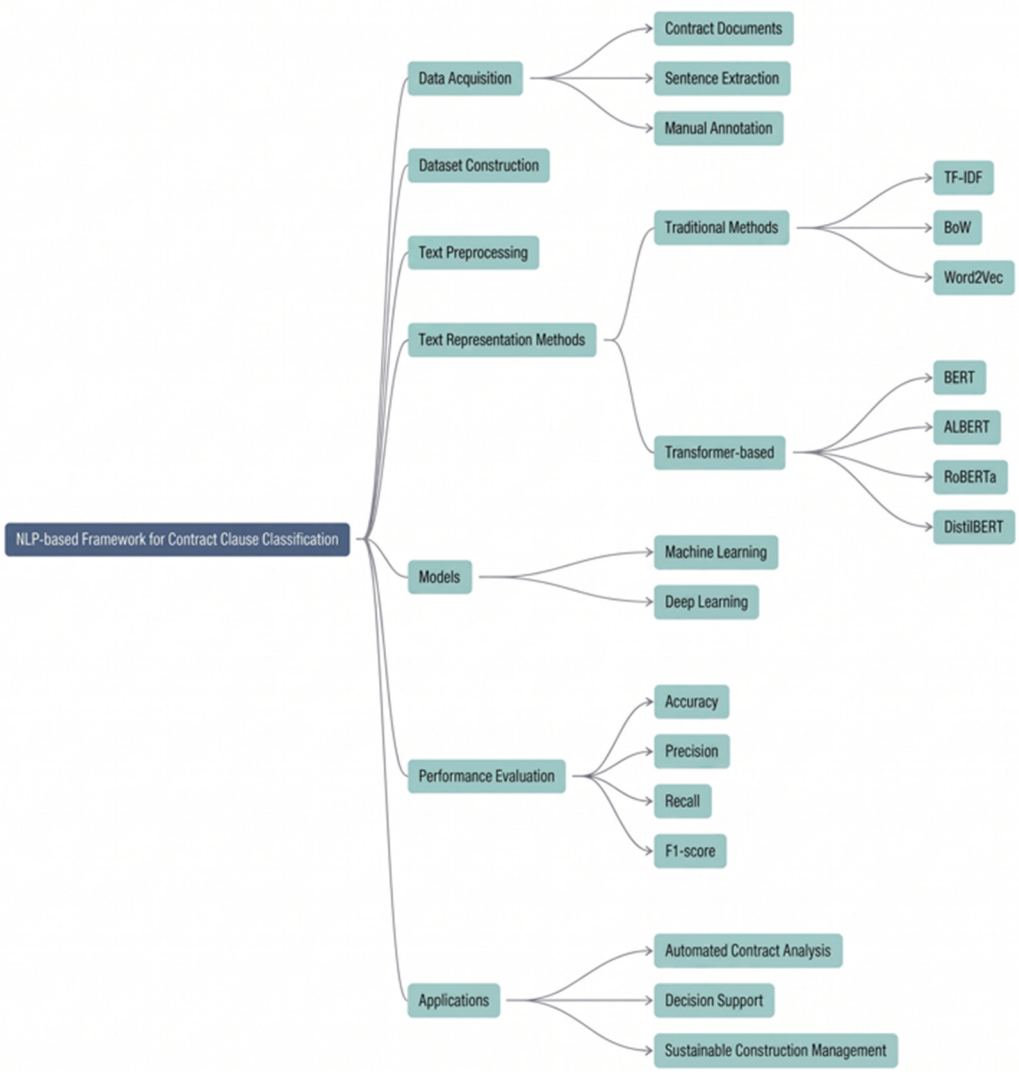


Figure 1. Overall workflow of the proposed NLP-based framework for automated construction contract clause classification and decision support in sustainable construction management.

4.2. Dataset Construction

The datasets used in this study were derived from two widely used standard form construction contracts:

- FIDIC Conditions of Contract for Design, Build and Operate.
- JCT Design and Build Contract.

These documents represent internationally recognized contractual frameworks used in large-scale construction projects. The use of standard form contracts allows the analysis to focus on widely applicable contractual structures rather than project-specific documents.

Datasets used in this study were constructed by manually extracting sentences from the contract texts and annotating them by the authors. The annotation process involved assigning each sentence to predefined thematic categories related to construction project management. All annotation decisions were made solely by the authors.

4.3. Dataset Structure

Two different datasets were constructed to investigate different dimensions of construction contract analysis.

4.3.1. Dataset 1: Contractual Obligation Categories

The first dataset focuses on identifying different types of contractual statements related to responsibilities and actions in construction contracts. Sentences were classified into four categories: *Obligations, Operational Actions, Optional Provisions,* and *Other General Statements*. The classification criteria were defined based on commonly observed linguistic patterns and functional roles within construction contract texts.

- Obligations

The "Obligations" category includes statements that define binding responsibilities and mandatory duties of the contractor, employer, or other parties under the contract. These provisions typically express non-negotiable requirements through modal verbs such as *"shall"* and *"must,"* indicating enforceable commitments. This category encompasses contractual responsibilities to be fulfilled during project execution, as well as legal obligations arising from applicable laws, regulations, and standards. It also includes requirements related to insurance, occupational health and safety, and quality compliance, all of which are essential for ensuring legal conformity and effective project governance.

Clauses were assigned to the Obligations category when they imposed a binding contractual duty, responsibility, prohibition, or compliance requirement on one of the parties. Typical linguistic indicators included modal and mandatory expressions such as "shall", "must", "is required to", "is responsible for", "is obliged to", and "shall not". However, the final label was determined based on the primary contractual function of the clause rather than on keywords alone.

- Operational Actions

The "Operational Actions" category refers to operational actions and procedural tasks that must be performed during project execution. Unlike the "Obligations" category, which emphasizes abstract responsibilities, this category focuses on concrete and executable activities, typically expressed through action-oriented phrases such as "to carry out," "to complete," and "to provide." These statements include tasks related to design, construction, and project delivery, as well as documentation, reporting, and communication processes required for the effective coordination and implementation of construction activities.

The distinction between the "Obligations" and "Operational Actions" categories is based on the contextual meaning of the sentences, the verbs they employ, and their functional role within the contract. While both categories indicate required actions, "Obligations" primarily reflect legal or contractual responsibilities, whereas "Operational Actions" statements represent specific operational steps. This distinction enables clearer

semantic separation between abstract duties and executable actions, thereby supporting more accurate classification in NLP-based analysis.

Clauses were assigned to the Operational Actions category when their primary function was to describe concrete and executable project activities, procedural steps, or implementation-related tasks. These included actions such as submitting documents, providing equipment or assistance, giving notice before inspection, arranging tests, examining affected works, and coordinating construction activities. Although many of these clauses contain mandatory modal verbs such as "shall" or "must", they were distinguished from the Obligations category based on their operational function rather than the presence of mandatory language alone.

- Optional Provisions

The "Optional Provisions" category includes clauses that describe discretionary actions or provisions that may be undertaken depending on specific project conditions. These statements do not impose mandatory requirements but instead introduce flexibility by allowing alternative approaches or conditional actions. Such clauses are typically identified through linguistic indicators including conditional expressions (e.g., *"if," "if any"*), advisory language, and modal verbs such as *"may"* and *"can,"* which indicate permission rather than obligation. In many cases, these provisions offer guidance on how project processes may be optimized without enforcing strict requirements.

Clauses were assigned to the Optional Provisions category when their primary function was to provide a discretionary right, conditional permission, optional arrangement, or alternative course of action rather than to impose a binding duty. Typical indicators included expressions such as "may", "at its option", "at its discretion", "subject to approval", "with consent", "if applicable", and "if any". However, the classification was not based solely on the presence of these terms. The final label was determined according to the primary contractual function of the clause, particularly whether the provision created flexibility or discretion for one of the parties.

- Other General Statements

This category includes general contractual statements that do not directly correspond to obligations, operational actions, or optional provisions. These clauses typically provide contextual, descriptive, or background information within the contract.

Clauses were assigned to the Other General Statements category when their primary function was to provide contextual, definitional, interpretive, or framework-related information rather than to impose a binding duty, describe an executable action, or provide an optional right. This category included clauses defining contractual terms, explaining the hierarchy of contract documents, identifying governing law or communication language, describing general risk categories, and clarifying how contract documents should be interpreted. Some clauses in this category may contain mandatory modal expressions such as "shall"; however, the classification was based on the primary contractual function of the clause rather than on modal verbs alone.

Due to copyright restrictions associated with the standard-form contract documents from which the clauses were derived, the full clause-level dataset cannot be publicly released. However, representative anonymized labeled examples are provided in Appendix A to improve transparency. The data supporting the findings of this study may be made available by the corresponding author upon reasonable request, subject to copyright and research-use restrictions.

The clauses were manually labeled by the authors according to predefined category definitions. To reduce annotation subjectivity, the labeling process was conducted through iterative discussions and consensus among the authors based on predefined

classification criteria. Each clause was examined in terms of its primary contractual function and assigned to the most appropriate category. Since the final labels were determined through consensus rather than independent annotation rounds, a formal inter-annotator agreement score, such as Cohen's kappa, was not calculated. This represents a limitation of the present study. Future work will employ independent multi-annotator labeling and report inter-annotator agreement to further strengthen annotation reliability.

The resulting dataset contains 450 sentences, distributed as shown in Figure 2.

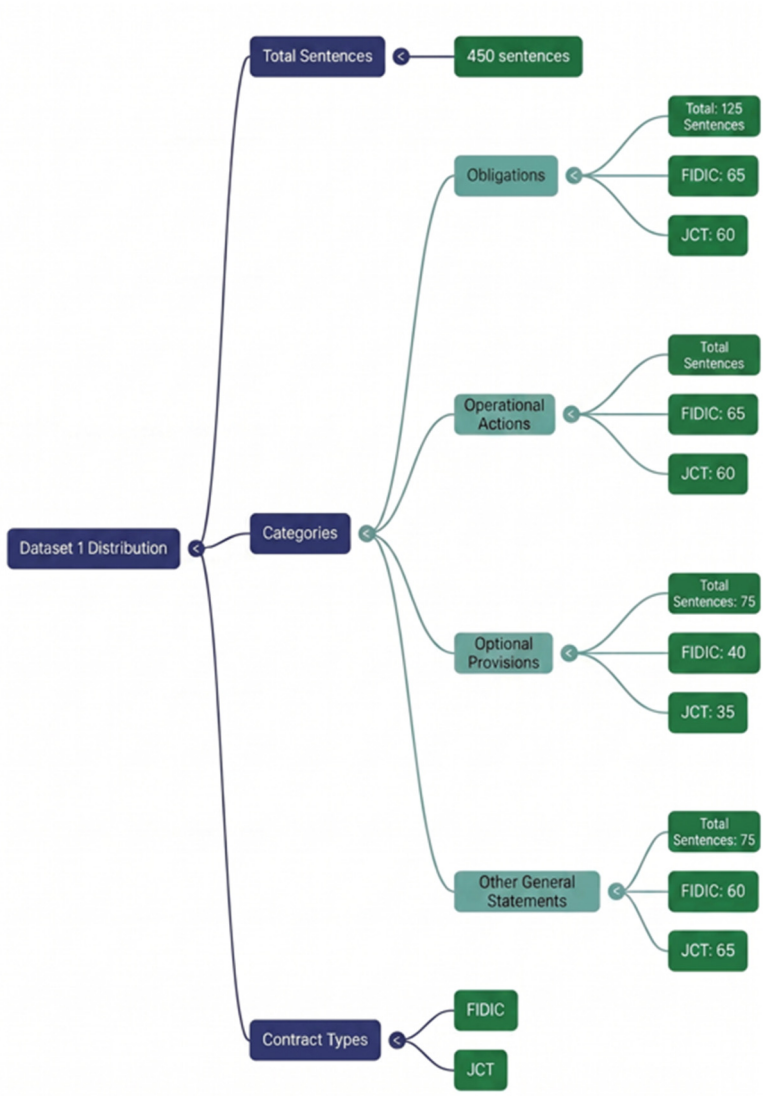


Figure 2. Dataset 1 distribution.

4.3.2. Dataset 2: Project Management Dimensions

The second dataset focuses on the time–cost–quality (TCQ) dimensions commonly used in construction project management. Sentences were classified into three categories: *Cost*, *Time*, and *Quality*, based on their relevance to key project performance factors.

- Cost

The "Cost" category includes clauses that describe financial aspects influencing the overall project cost. These statements cover cost variations arising from changes in materials, labor, or project scope, as well as financial implications of delays and schedule adjustments. The category also includes provisions related to procurement and subcontracting, insurance and risk management costs, and legal or administrative

requirements such as permits and licenses. In addition, clauses defining payment mechanisms—such as advance payments, interim payments, and final settlements—are included due to their central role in managing the project's financial lifecycle.

Clauses were assigned to the Cost category when their primary contractual function was related to financial consequences, payment administration, price adjustment, reimbursement, deductions, retention, insurance-related monetary exposure, or allocation of additional costs. Typical indicators included expressions such as "contract price", "contract sum", "cost", "payment", "interim payment", "final payment", "retention", "deduction", "reimbursement", "loss and expense", "damages", and "additional cost". However, the label was not assigned based solely on keywords. For clauses involving multiple project dimensions, such as delay-related compensation or defect-related remedial costs, the dominant contractual function was used to determine the label.

- Time

The "Time" category includes clauses related to scheduling and temporal aspects of the project. These statements define project timelines, including commencement, progress, and completion periods, as well as deadlines for fulfilling contractual obligations. The category also encompasses payment schedules, delays, and associated penalties, along with provisions governing suspension, continuation, or termination of work. Furthermore, it includes clauses specifying timeframes for legal and dispute resolution processes, such as mediation or arbitration, where procedural deadlines are explicitly defined.

Clauses were assigned to the Time category when their primary contractual function concerned scheduling, commencement, completion, progress monitoring, delay, notice periods, suspension periods, or extension of time. Typical indicators included expressions such as "commencement date", "time for completion", "completion date", "contract period", "programme", "progress", "delay", "extension of time", "within X days", and "prior to commencement". However, the classification was not based solely on temporal keywords. For clauses involving multiple project management dimensions, the dominant contractual function was used to determine the label. For example, a delay-related clause was labeled as Time when it primarily regulated the extension of the completion period, whereas it was labeled as Cost when it primarily concerned compensation or financial recovery.

- Quality

The "Quality" category includes clauses that define standards and requirements affecting construction quality. These statements address quality assurance and quality control processes, material specifications, workmanship standards, and design requirements. The category also includes provisions related to inspections, testing procedures, certifications, and compliance with technical specifications. Additionally, it covers mechanisms for identifying and correcting defects or non-conforming work, as well as requirements related to health and safety, environmental considerations, and sustainability. Clauses linking quality with project performance—such as timely delivery and adherence to specifications—are also included, as they directly influence overall project outcomes.

Clauses were assigned to the Quality category when their primary contractual function concerned technical compliance, design adequacy, material and workmanship standards, inspection, sampling, testing, commissioning, defect correction, remedial works, health and safety compliance, environmental protection, or personnel competence. Typical indicators included expressions such as "technical standards", "Employer's Requirements", "design", "Contractor's Documents", "materials", "workmanship", "inspection", "testing", "defects", "remedy", "as-built records", "qualified personnel",

and "compliance". However, the classification was not based solely on keywords. For clauses involving multiple project management dimensions, the dominant contractual function was used to determine the label. For example, a defect-related clause was labeled as Quality when it primarily concerned correction and conformity of the works, whereas it was labeled as Cost when it primarily allocated the financial burden of remedial action.

The resulting dataset contains 827 sentences, distributed as shown in Figure 3.

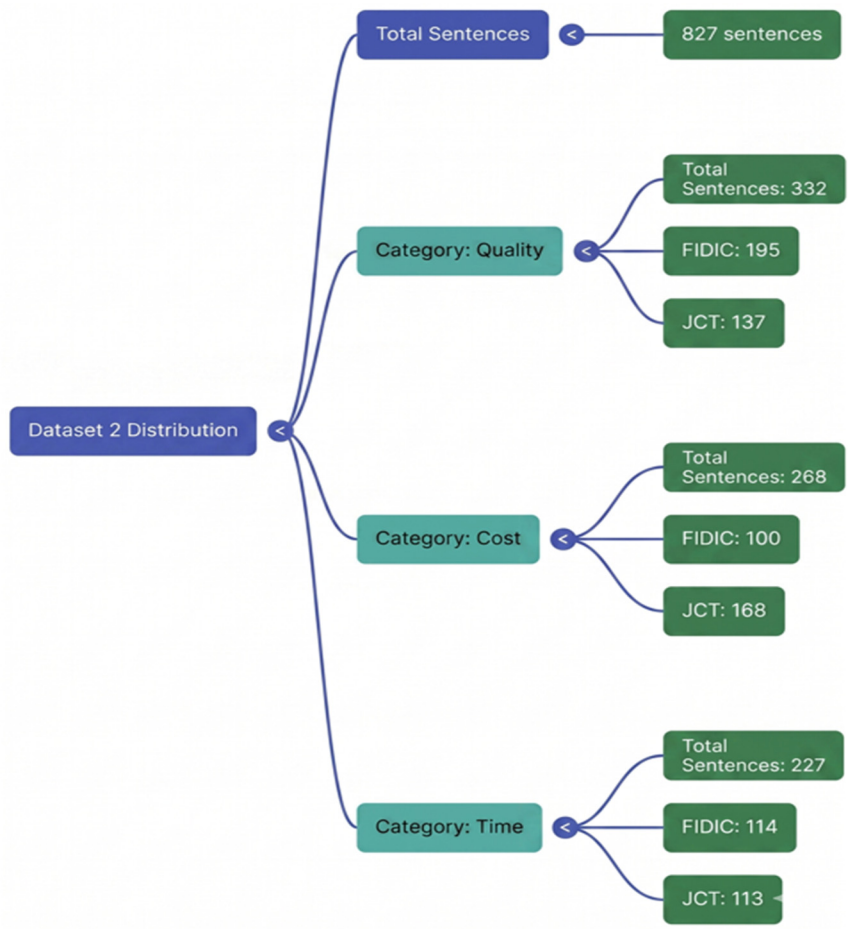


Figure 3. Dataset 2 distribution.

4.4. Text Preprocessing

Before classification, the textual data were preprocessed using the Python Natural Language Toolkit (NLTK) library.

The preprocessing steps included:

- Removal of punctuation marks and special characters;
- Conversion of all text to lowercase;
- Tokenization;
- Normalization of words to their root forms.

These steps ensure that textual data are converted into a consistent format suitable for machine learning algorithms.

4.5. Text Representation Techniques

Machine learning algorithms cannot directly process textual data. Therefore, textual information must be converted into numerical feature vectors through text representation techniques.

In this study, three main categories of text representation techniques were used:

- Frequency-based representations;
- Prediction-based embeddings;
- Transformer-based contextual embeddings.

Traditional representations such as TF-IDF and Word2Vec capture statistical relationships between words in documents. However, these approaches often fail to capture contextual dependencies in complex legal texts.

To address this limitation, the study also employs transformer-based pre-trained language models.

The transformer architecture, originally introduced by Vaswani et al. [24], enables contextual language modeling using attention mechanisms.

Transformer architectures are based on encoder–decoder neural network structures that process textual sequences through self-attention mechanisms rather than recurrent operations. In transformer-based language models, each token in a sentence is represented not only by its lexical meaning but also by its contextual relationships with other tokens in the sequence. This contextual representation is achieved through the self-attention mechanism, which enables the model to dynamically assign importance weights to different words within a sentence.

A key component of the transformer architecture is multi-head attention, where multiple attention operations are performed in parallel to capture different semantic and syntactic relationships between words. In addition, positional encoding is incorporated into the input embeddings to preserve sequential information, since transformers do not inherently process text in sequential order like recurrent neural networks.

Among transformer architectures, encoder-based models such as BERT and its variants are particularly effective for text classification tasks because they generate bidirectional contextual embeddings by simultaneously considering both left and right contextual information within a sentence. In this study, transformer-based embeddings extracted from BERT, RoBERTa, DistilBERT, and ALBERT were used as contextual feature representations for downstream machine learning and deep learning classifiers.

Although traditional representations such as TF-IDF and Word2Vec are discussed to provide methodological background and contextual comparison with transformer-based approaches, the experimental framework of this study focused primarily on contextual embeddings generated by transformer-based language models.

4.6. Transformer-Based Pre-Trained Models

Pre-trained language models are neural network models trained on large corpora prior to being adapted for specific tasks.

In this study, the following transformer-based models were used:

- BERT;
- RoBERTa;
- DistilBERT;
- ALBERT.

These models generate contextual embeddings that capture semantic relationships between words within sentences.

4.6.1. RoBERTa

RoBERTa is an optimized version of the BERT architecture that improves training procedures and model performance through larger datasets and longer training times [25].

4.6.2. DistilBERT

DistilBERT is a compressed version of BERT developed using knowledge distillation techniques, enabling faster training while maintaining competitive performance [26].

4.6.3. ALBERT

ALBERT reduces the number of parameters in the BERT architecture through parameter sharing and embedding factorization strategies [27].

4.6.4. Transformer Embedding and Classification Procedure

Algorithm 1 presents the overall transformer-based embedding extraction and classification workflow used in this study.

Algorithm 1. Transformer-Based Contract Clause Classification Framework.

```
Input:
Contract clause dataset D
Pre-trained transformer model T
Classifier model C
Output:
Predicted clause labels Y
1: Load construction contract datasets
2: Apply text preprocessing
- lowercase conversion
- punctuation removal
- tokenization
3: FOR each sentence s in dataset D DO
4: Tokenize sentence using transformer tokenizer
5: Add special tokens [CLS] and [SEP]
6: Convert tokens into input IDs and attention masks
7: Feed inputs into transformer model T
8: Extract contextual embedding corresponding to [CLS] token
9: Store embedding vector e
10: END FOR
11: Construct embedding matrix E from all vectors e
12: Split dataset into training and testing subsets
13: Train classifier model C using E_train
14: Generate predictions on E_test
15: Compute evaluation metrics:
Accuracy, Precision, Recall, F1-score, Specificity
16: Return predicted labels and performance results
```

In the proposed framework, the [CLS] token representation extracted from the final hidden layer of each transformer model was used as a fixed-length contextual embedding for downstream classification tasks. These embeddings were subsequently provided as inputs to both traditional machine learning classifiers and sequential deep learning architectures such as RNN, GRU, and LSTM.

4.7. Machine Learning Models

Several supervised machine learning algorithms were used for classification. These include:

- Logistic Regression (LR);
- Support Vector Machines (SVM);
- Random Forest (RF);
- k-Nearest Neighbors (KNN);
- Naïve Bayes (NB).

These algorithms were selected because they represent widely used baseline classifiers for text classification tasks [28,29].

4.8. Deep Learning Models

To capture sequential dependencies within contractual sentences, deep learning models based on recurrent neural networks were also employed.

The models used include:

- Recurrent Neural Networks (RNN);
- Gated Recurrent Units (GRU);
- Long Short-Term Memory networks (LSTM).

These models are particularly suitable for sequence-based textual data because they can capture long-range dependencies between words within sentences [30–32].

4.9. Model Evaluation Metrics

The performance of the classification models was evaluated using commonly used metrics in text classification:

- Accuracy;
- Precision;
- Recall;
- F1-score;
- Specificity.

These metrics provide complementary insights into classification performance by evaluating both overall prediction accuracy and class-specific predictive capabilities [33].

Accuracy measures the overall proportion of correctly classified instances among all predictions generated by the model and is defined as [33]:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

where (TP) represents true positives, (TN) true negatives, (FP) false positives, and (FN) false negatives. Although accuracy is one of the most widely used evaluation metrics, it may not fully reflect classification quality in cases involving imbalanced datasets or uneven class distributions [33].

Precision evaluates the proportion of correctly predicted positive observations among all observations predicted as positive by the model and is calculated as:

In the context of contract clause classification, precision reflects the reliability of predicted clause categories and indicates the extent to which false positive classifications are minimized [34].

Recall, also referred to as sensitivity, measures the proportion of actual positive observations correctly identified by the model:

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

High recall values indicate that the classification framework can successfully detect relevant contractual clauses without missing important category instances [33,34].

Specificity measures the proportion of correctly identified negative observations and was calculated as:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

For the multi-class classification tasks, specificity values were calculated using a one-vs-rest approach for each class and then averaged across all classes.

The F1-score represents the harmonic mean of precision and recall and provides a balanced evaluation metric when both false positives and false negatives are important:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Since construction contract clauses may contain semantically overlapping expressions, the F1-score offers a more comprehensive assessment of classification effectiveness compared to using accuracy alone [33,34].

4.10. Implications for Construction Contract Management

Beyond methodological evaluation, the proposed framework aims to support automated interpretation of contractual clauses in construction projects.

Improved classification of contractual provisions may assist practitioners in identifying key obligations, project management dimensions, and operational actions within complex contract documents. This capability may assist decision-making processes and support transparency in construction contract governance, thereby providing potential process-level contributions to sustainable project management practices.

From a sustainability perspective, the ability to automatically classify contractual clauses can support more efficient resource allocation, reduce administrative workload, and improve the consistency of contract interpretation. These improvements may contribute to economic sustainability and institutional efficiency in construction project management, although their practical effects should be validated through future implementation studies.

5. Experiments and Results

In this study, multi-class classification experiments were conducted on two separate datasets constructed from construction contract texts using transformer-based natural language processing techniques. After extracting textual representations through pre-trained approaches, these representations were subsequently used as input features for conventional machine learning algorithms, including Logistic Regression (LR), K-Nearest Neighbors (KNN), Random Forest (RF), Naïve Bayes (NB), and Support Vector Machines (SVM), as well as deep learning algorithms such as Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), and Gated Recurrent Units (GRU).

In addition to traditional approaches, contextual embeddings obtained from pre-trained transformer models were also utilized. Specifically, DistilBERT, RoBERTa, and ALBERT models were employed to extract deep semantic representations from the text. Using the Hugging Face Transformers library, each sentence in the dataset was tokenized and processed through the selected transformer model. The embedding corresponding to the [CLS] token from the final hidden state was extracted and used as a fixed-length feature vector for classification tasks. This embedding extraction process was implemented using PyTorch (version 2.11) in a Google Colab Pro+ environment with GPU acceleration to improve computational efficiency. The resulting embedding vectors were subsequently provided as inputs to various classifiers for training and evaluation.

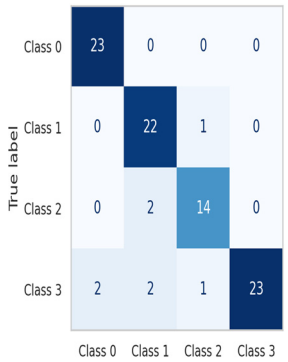
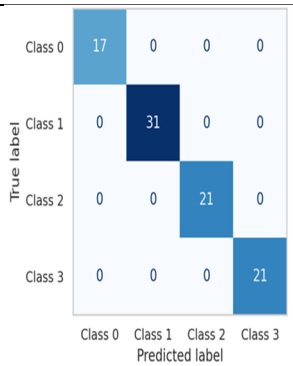
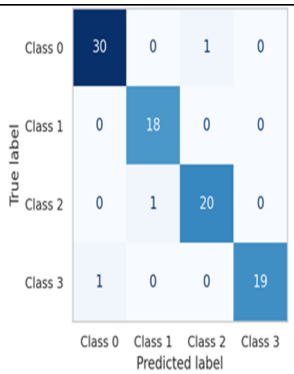
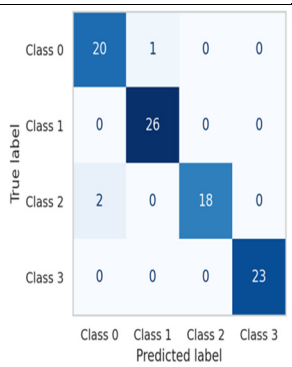
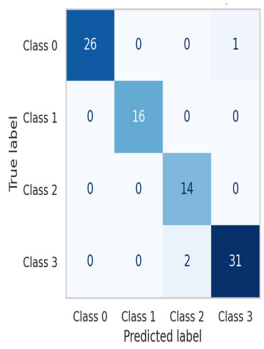
All experiments were conducted using two separate datasets generated from construction contract texts. The datasets were divided into training and testing subsets with an 80–20% split. The train–test split was performed at the sentence level after the annotation process using a random 80–20 partition. Since the sentences originated from the same standard-form contract documents, some contextual similarity may exist between training and testing samples. Therefore, the reported results should primarily be interpreted as comparative performance evaluations among different NLP architectures rather than fully generalized cross-contract prediction performance. Future studies will investigate stricter validation strategies, including contract-level and cross-domain

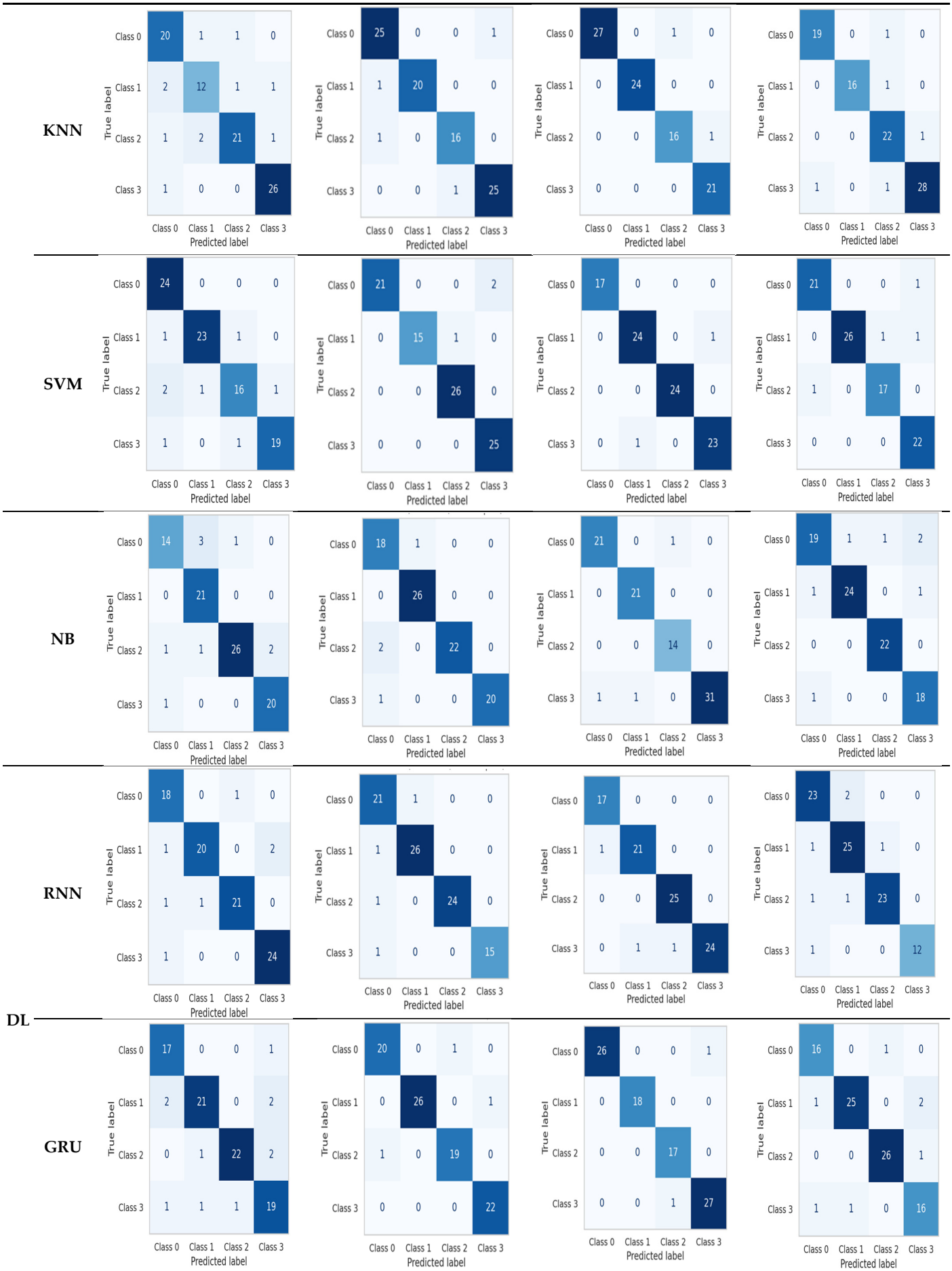
splitting protocols, to further reduce potential information leakage. Standard preprocessing techniques, including the removal of special characters and irrelevant tokens, were applied prior to model input. Model performance was evaluated using several metrics, including Accuracy, Precision, Recall, F1-score, and Specificity. The confusion matrices, from which these performance metrics were derived, are presented in Tables 2 and 3 for DS1 and DS2, respectively.

Although several models achieved accuracy values exceeding 98%, these results should be interpreted carefully, considering the relatively limited dataset size, the standardized structure of construction contract language, and the sentence-level train–test splitting strategy adopted in this study. To improve evaluation robustness, cross-validation procedures were employed during the experimental process. However, broader statistical validation and comparative significance testing across all model architectures were beyond the scope of the present study.

Therefore, the reported results should primarily be interpreted as comparative benchmarking outcomes within the constructed datasets rather than definitive generalized performance indicators. Future studies should incorporate larger and more heterogeneous datasets, repeated cross-validation experiments, statistical significance analysis, and external validation protocols to more rigorously assess model stability and generalization across different contract domains.

Table 2. Confusion Matrix Results of ML and DL Models with Transformer-Based Embeddings on DS1. Darker colors represents higher results, lighter colors represents lower results.

Model	BERT	RoBERTa	DistilBERT	alBERT
ML – LR				
ML – RF				


KNN
SVM
NB
RNN
DL
GRU
True label
Predicted label
Class 0
Class 1
Class 2
Class 3

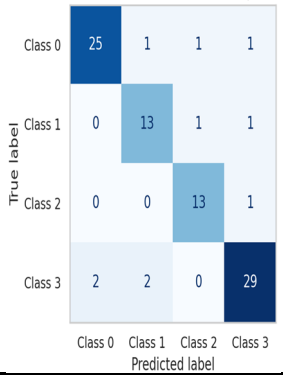
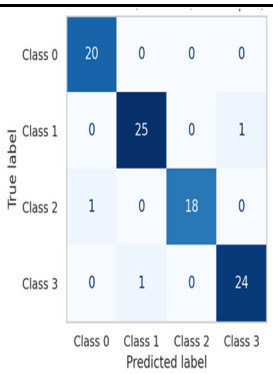
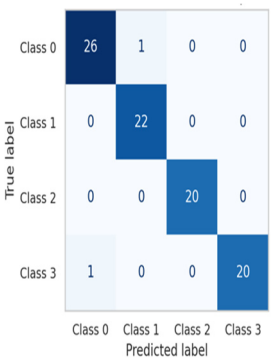
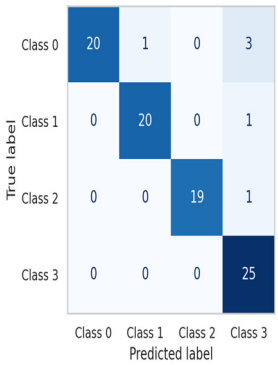
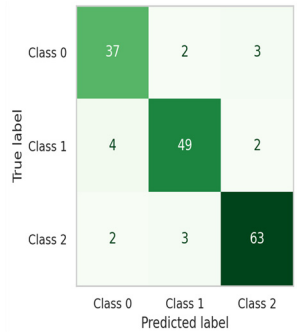
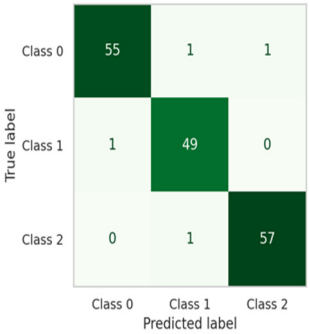
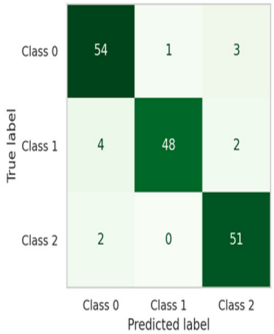
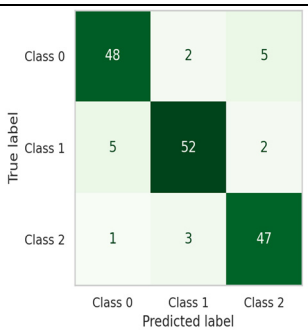
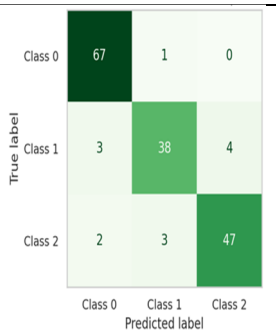
LSTM				

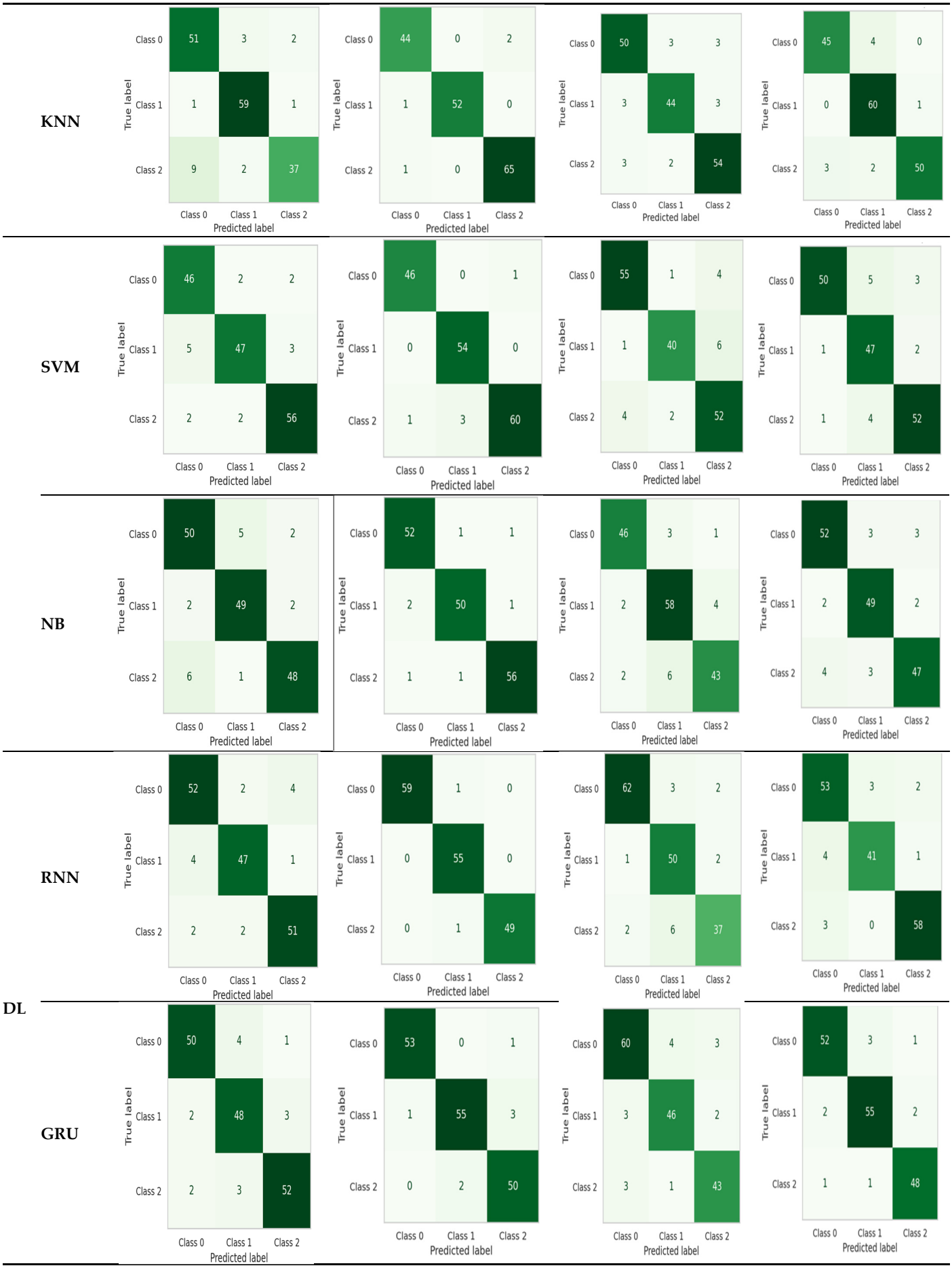
Table 2 compares the classification performances of machine learning (ML) and deep learning (DL) models using transformer-based embedding methods on the DS1 dataset based on confusion matrix results. Overall, DistilBERT- and RoBERTa-based representations appear to yield higher correct classification rates across several models. In particular, the high diagonal values observed in the LR, RF, and SVM models indicate that the classes were largely predicted correctly. By contrast, greater inter-class confusion was observed in some models using BERT- and ALBERT-based representations.

Among the ML models, SVM and RF produced comparatively more balanced results, whereas KNN and NB exhibited increased misclassification rates for certain classes. Regarding the DL models, RNN, GRU, and LSTM architectures generally demonstrated strong performance; however, their effectiveness varied depending on the type of transformer embedding employed. In particular, the pronounced increase in diagonal density in the DistilBERT-based GRU and LSTM models suggests that these combinations performed class discrimination more effectively.

Overall, the findings indicate that transformer-based embedding methods have a substantial impact on model performance, and that DistilBERT- and RoBERTa-based representations achieve superior results in both ML and DL models.

Table 3. Confusion Matrix Results of ML and DL Models with Transformer-Based Embeddings on DS2 (80–20 Train–Test Split). Darker colors represents higher results, lighter colors represents lower results.

	Model	**BERT**	**RoBERTa**	**DistilBERT**	**alBERT**
ML	**LR**				
	RF				


KNN
SVM
NB
RNN
GRU
DL
True label
Predicted label
Class 0
Class 1
Class 2

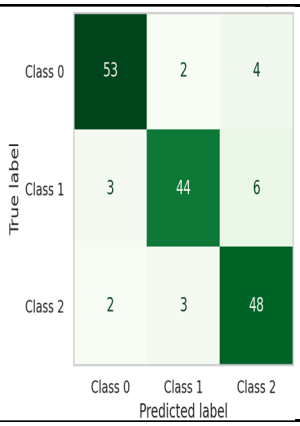
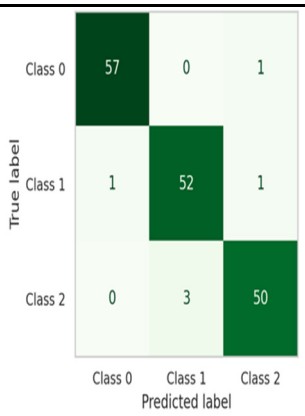
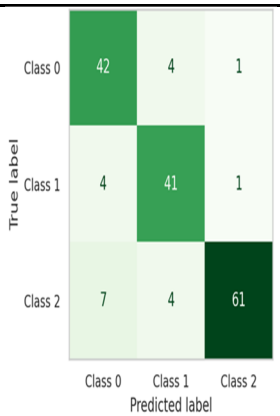
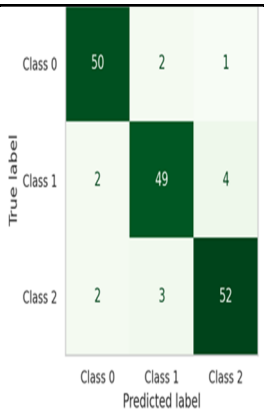
LSTM				

Table 3 presents the confusion matrix results obtained on the DS2 dataset using an 80% training and 20% testing split. Overall, the results indicate that transformer-based embedding methods have a substantial impact on classification performance. In particular, the considerably high diagonal values observed in most RoBERTa-based models demonstrate that the classes were accurately predicted.

Among the ML models, the LR, RF, and KNN algorithms achieved superior results when combined with RoBERTa embeddings, with notably low misclassification rates. Although DistilBERT-based models also exhibited high overall performance, a limited degree of confusion was observed among certain classes. BERT and ALBERT embedding methods yielded successful results in some models; however, increased error rates were observed for specific classes.

On the deep learning side, the LSTM model, particularly when used with RoBERTa and ALBERT embeddings, produced highly balanced results with high classification accuracy. In addition, although the DistilBERT-based LSTM model also achieved strong correct classification performance, some inter-class confusion persisted.

Overall, the findings presented in the table reveal that ML and DL models combined with transformer-based embedding methods achieved high classification performance on the DS2 dataset, with RoBERTa-based representations providing a more stable and robust performance in particular.

When Tables 2 and 3 are evaluated together, it is evident that transformer-based embedding methods substantially affect the performance of both machine learning and deep learning models on the DS1 and DS2 datasets. While greater inter-class confusion was observed among the models on the DS1 dataset, the markedly increased diagonal values in the DS2 dataset indicate that the models produced more accurate and balanced predictions.

In particular, RoBERTa-based embeddings achieved high performance on both datasets, although they yielded more stable results on DS2. DistilBERT embeddings also demonstrated strong performance, especially when combined with DL models. By contrast, greater inter-class confusion was observed in some models employing BERT- and ALBERT-based representations.

Overall, the results obtained on the DS2 dataset indicate higher accuracy and lower error rates compared with DS1, suggesting that the data distribution and the training–testing split contributed positively to model performance.

The performance metrics derived from these confusion matrices are presented in Tables 4–6.

Table 4. Performance Results of ML and DL Models with BERT (80–20 Train–Test Split).

		Model	Accuracy	F1	Precision	Recall	Specificity
DS1	ML	LR	0.9556	0.954	0.955	0.953	0.982
		RF	0.9639	0.962	0.964	0.961	0.986
		KNN	0.9500	0.948	0.949	0.947	0.980
		SVM	0.9611	0.960	0.962	0.958	0.985
		NB	0.9444	0.943	0.945	0.941	0.978
	DL	RNN	0.9583	0.956	0.957	0.955	0.984
		GRU	0.9667	0.965	0.966	0.964	0.987
		LSTM	**0.9722**	**0.970**	**0.972**	**0.969**	**0.989**
DS2	ML	LR	0.8833	0.881	0.885	0.880	0.940
		RF	0.9056	0.903	0.906	0.901	0.950
		KNN	0.8722	0.870	0.873	0.868	0.935
		SVM	0.8972	0.894	0.896	0.892	0.945
		NB	0.8500	0.848	0.851	0.846	0.925
	DL	RNN	0.9139	0.911	0.914	0.910	0.952
		GRU	0.9250	0.922	0.924	0.920	0.958
		LSTM	0.9361	0.934	0.935	0.933	0.962

Table 4 presents the performance results of the ML and DL models obtained after integrating the base BERT representations on the DS1 and DS2 datasets. For the DS1, the highest accuracy was achieved by the LSTM model with 97.22%, followed by GRU (96.67%) and RF (96.39%). These results indicate that DL architectures exhibit strong learning capabilities after BERT-based feature extraction and particularly excel in handling sequential data structures. Among the ML models, RF and SVM also demonstrated notable performance, achieving accuracy rates above 96%. In contrast, the NB model showed the lowest performance with an accuracy of 94.44%.

When moving to DS2, a noticeable decrease in performance was observed across all models. This decline can be attributed to the more complex and imbalanced structure of the DS2. On the DL side, the LSTM model again achieved the best performance with an accuracy of 93.61%, followed by GRU (92.50%) and RNN (91.39%). Among the ML models, RF (90.56%) and SVM (89.72%) produced the highest accuracy values. The NB model again demonstrated the lowest performance on this dataset, with an accuracy of 85.00%.

Overall, the results indicate that DL-based models outperform ML models on both DS1 and DS2. This finding suggests that deep learning architectures are capable of effectively learning from language model-based embeddings such as BERT and provide more stable performance in classification tasks. Furthermore, it is evident that the quality and structural characteristics of the dataset have a direct impact on model performance.

Table 5. Performance Results of ML and DL Models with ALBERT (80–20 Train–Test Split)

		Model	Accuracy	F1	Precision	Recall	Specificity
DS1	ML	LR	0.9639	0.962	0.964	0.961	0.987
		RF	0.9694	0.968	0.969	0.967	0.989
		KNN	0.9556	0.949	0.955	0.952	0.984
		SVM	0.9667	0.965	0.966	0.964	0.988
		NB	0.9500	0.948	0.949	0.947	0.981
	DL	RNN	0.9667	0.965	0.966	0.964	0.987
		GRU	0.9732	0.970	0.971	0.969	0.990
		LSTM	**0.9778**	**0.976**	**0.977**	**0.975**	**0.992**

DS2	ML	LR	0.9361	0.934	0.935	0.933	0.970
		RF	0.9444	0.942	0.944	0.941	0.973
		KNN	0.9278	0.925	0.926	0.924	0.968
		SVM	0.9389	0.937	0.938	0.936	0.971
		NB	0.9111	0.908	0.910	0.907	0.960
	DL	RNN	0.9472	0.945	0.946	0.944	0.975
		GRU	0.9583	0.956	0.957	0.955	0.980
		LSTM	0.9639	0.962	0.963	0.961	0.984

Table 5 presents a comparative evaluation of the ML and DL model performances obtained after integrating ALBERT on the DS1 and DS2 datasets. On DS1, the highest accuracy was achieved by the LSTM model with 97.78%. This was followed by the GRU model with an accuracy of 97.32% and the RF model with 96.94%. Overall, after the integration of ALBERT, a noticeable improvement was observed in the performance of both ML and DL models compared to the baseline BERT representations. In particular, DL architectures that are sensitive to sequential data structures, such as LSTM and GRU, demonstrated high learning performance, benefiting from ALBERT's compact and lightweight architecture.

When examining the results for DS2, a slight decrease in performance was observed across all models compared to DS1. This reduction may be attributed to the more challenging structure of the dataset or the increased complexity in class separation. Nevertheless, DL models continued to achieve high accuracy on DS2. The LSTM model produced the best result with an accuracy of 96.39%, followed by GRU with 95.83% and RNN with 94.72%. Among the ML models, RF (94.44%) and SVM (93.89%) achieved the highest performance. The NB model exhibited comparatively lower performance on both datasets, remaining behind the DL models and the other ML algorithms.

The results indicate that despite its lower parameterization, ALBERT effectively provides contextual representations and significantly improves classification performance, particularly when combined with DL architectures. Considering both accuracy and other evaluation metrics, integrating ALBERT with ML and DL models enables the generation of more stable and robust classification outcomes.

Table 6. Performance Results of ML and DL Models with RoBERTa (80–20 Train–Test Split).

		Model	Accuracy	F1	Precision	Recall	Specificity
DS1	ML	LR	0.9694	0.968	0.969	0.967	0.989
		RF	0.9750	0.973	0.974	0.972	0.991
		KNN	0.9639	0.962	0.963	0.961	0.987
		SVM	0.9732	0.970	0.971	0.969	0.990
		NB	0.9583	0.956	0.952	0.955	0.984
	DL	RNN	0.9742	0.970	0.971	0.969	0.990
		GRU	0.9778	0.976	0.977	0.975	0.992
		LSTM	**0.9806**	**0.979**	**0.980**	**0.978**	**0.993**
DS2	ML	LR	0.9744	0.972	0.973	0.971	0.990
		RF	0.9806	0.979	0.980	0.978	0.993
		KNN	0.9694	0.967	0.968	0.966	0.988
		SVM	0.9778	0.976	0.977	0.975	0.992
		NB	0.9639	0.961	0.962	0.960	0.986
	DL	RNN	0.9750	0.973	0.974	0.972	0.991
		GRU	0.9833	0.982	0.983	0.981	0.994
		LSTM	**0.9861**	**0.985**	**0.986**	**0.984**	**0.995**

Table 6 presents the performance comparisons of ML and DL models obtained using the RoBERTa-based representation method on the DS1 and DS2 datasets. On DS1, the LSTM model achieved the highest performance with an accuracy of 98.06%. This was followed by GRU with 97.78%, RF with 97.50%, and SVM with 97.32%. The superior performance can be attributed to RoBERTa being pre-trained on a larger corpus and its ability to provide richer contextual representations. When combined with DL models, these deep contextual embeddings significantly improved classification performance. ML models also benefited from RoBERTa-based representations; methods such as Logistic Regression, KNN, and SVM achieved accuracy values above 96%. The NB model, however, showed relatively lower performance with an accuracy of 95.83%.

For the DS2, the impact of RoBERTa became even more pronounced. On the DL side, the LSTM model achieved an accuracy of 98.61%, followed by GRU with 98.33% and RNN with 97.50%, producing the best results compared with all previously used embedding methods. Among the ML models, RF and SVM also reached notably high accuracy levels. These results indicate that RoBERTa's strong contextual representation capability can be effectively integrated not only with DL architectures but also with traditional ML algorithms.

Overall, RoBERTa improved the performance of all models and achieved the highest accuracy levels compared with other BERT-based variants. Particularly when used with sequential DL architectures such as LSTM and GRU, RoBERTa-based classification frameworks demonstrated strong generalization capability and successfully captured inter-class distinctions.

The superior performance of RoBERTa- and DistilBERT-based representations may be explained by their ability to capture contextual semantic relationships embedded in construction contract language. Construction contracts frequently contain long and syntactically complex sentences involving conditional expressions, modal verbs, legal obligations, procedural actions, and domain-specific terminology. In such contexts, contextual embedding models capable of representing semantic dependencies across entire sentences become particularly advantageous.

RoBERTa benefits from optimized transformer training strategies, including dynamic masking, larger training corpora, and longer training duration, which improve its ability to learn contextual semantic patterns compared to the original BERT architecture. This capability is especially important for construction contract analysis, where subtle linguistic variations may alter contractual meaning and responsibility allocation.

DistilBERT, although computationally lighter than BERT, preserves much of the semantic representation capability of larger transformer models through knowledge distillation. Its comparatively strong performance suggests that compressed transformer architectures can still effectively capture legal–technical language structures while providing computational efficiency advantages.

The findings further indicate that transformer-based contextual embeddings are more effective than traditional lexical representations in distinguishing semantically overlapping contractual categories. In particular, contextual attention mechanisms appear to improve the identification of nuanced differences between obligation-oriented, procedural, and project-management-related clauses commonly encountered in construction contracts.

Table 7 presents a comparative evaluation of the ML and DL model performances obtained after integrating DistilBERT into the DS1 and DS2 datasets. On DS1, the highest accuracy was observed for the LSTM model with 98.33%, followed by GRU with 98.06% and RF with 97.78%. These results demonstrate that DistilBERT, despite being a smaller and faster transformer model, is capable of providing strong contextual representations

even with a reduced number of parameters and can be effectively integrated with both ML and DL models. Among the ML algorithms, RF, SVM, and LR achieved accuracy rates above 97%, indicating strong performance. The NB model, however, produced comparatively lower results with an accuracy of 96.39%.

Table 7. Performance Results of ML and DL Models with DistilBERT (80–20 Train–Test Split).

		Model	Accuracy	F1	Precision	Recall	Specificity
DS1	ML	LR	0.9712	0.971	0.972	0.970	0.990
		RF	0.9778	0.976	0.977	0.975	0.992
		KNN	0.9694	0.967	0.968	0.966	0.988
		SVM	0.9750	0.973	0.974	0.972	0.991
		NB	0.9639	0.961	0.962	0.960	0.986
	DL	RNN	0.9750	0.973	0.974	0.972	0.991
		GRU	0.9806	0.979	0.980	0.978	0.993
		LSTM	**0.9833**	**0.982**	**0.983**	**0.981**	**0.994**
DS2	ML	LR	0.8833	0.881	0.885	0.880	0.940
		RF	0.9056	0.903	0.906	0.901	0.950
		KNN	0.8722	0.870	0.873	0.868	0.935
		SVM	0.8972	0.894	0.896	0.892	0.945
		NB	0.8500	0.848	0.851	0.846	0.925
	DL	RNN	0.9139	0.911	0.914	0.910	0.953
		GRU	0.9250	0.922	0.924	0.920	0.958
		LSTM	0.9361	0.934	0.935	0.933	0.962

On the DS2, the overall performance decreased noticeably across all models. Nevertheless, DL models continued to outperform ML approaches. The LSTM model achieved an accuracy of 93.61%, followed by GRU with 92.50% and RNN with 91.39%. Among the ML models, RF (90.56%) and SVM (89.72%) showed the best performance, whereas NB again demonstrated the lowest accuracy (85.00%). These findings suggest that despite the more complex or imbalanced structure of DS2, DL architectures combined with DistilBERT embeddings provide more stable classification results.

Overall, although DistilBERT models are designed to be lightweight and computationally efficient, they achieved high accuracy and classification performance on both DS1 and DS2 datasets. This makes DistilBERT a highly suitable and efficient option for applications with limited computational resources. When used with DL architectures, DistilBERT produced particularly strong results with models such as LSTM and GRU.

In this study, four transformer-based language models—BERT, ALBERT, RoBERTa, and DistilBERT—were integrated with both traditional ML and DL models and evaluated on two different datasets (DS1 and DS2). The results demonstrate that all four language models significantly improved classification performance. In particular, RoBERTa and DistilBERT achieved the highest accuracy and evaluation metrics when combined with both ML and DL models. On the DS1, the LSTM model achieved the best overall performance among all models, reaching an accuracy of 98.06% with RoBERTa and 98.33% with DistilBERT. GRU and RNN models also produced results that were very close to these top performances.

From the ML perspective, Random Forest, Support Vector Machine, and Logistic Regression models achieved accuracy rates above 95% in most cases, whereas the Naïve Bayes model consistently demonstrated relatively lower performance across both datasets. Despite its compact architecture, ALBERT produced stable and competitive results, particularly when combined with DL models capable of processing sequential

data, such as LSTM and GRU. RoBERTa, benefiting from its richer pre-training strategy, consistently achieved the highest performance across both DS1 and DS2 datasets.

DistilBERT also produced highly competitive results despite its reduced parameter size and achieved the highest accuracy when combined with LSTM on the DS1. Although DS2 presented a more challenging structure for all models, DL architectures maintained consistent and strong performance on this dataset as well. Overall, when embedding methods were combined with DL models, higher classification performance was achieved compared to ML models. This finding suggests that contextual representations can be learned more effectively when integrated with sequential DL architectures.

Overall, the results indicate that RoBERTa- and DistilBERT-based representations provide the strongest classification performance when combined with LSTM and GRU classifiers. While RoBERTa achieved the highest performance on DS2, DistilBERT provided a competitive and computationally efficient alternative, particularly on DS1. These findings suggest that model selection should consider both classification performance and computational efficiency, especially for practical contract management applications with limited computational resources.

The superior performance of deep learning models, particularly LSTM and GRU, can be attributed to their ability to capture sequential dependencies and contextual relationships within high-dimensional transformer embeddings. Unlike traditional machine learning algorithms, which treat features as independent inputs, sequential architectures are capable of modeling complex semantic interactions between tokens.

Furthermore, the relatively lower performance of Naïve Bayes can be explained by its strong independence assumptions, which are incompatible with dense contextual embeddings where features are highly interdependent. This highlights the importance of aligning classifier architecture with the nature of the embedding space.

The observed performance differences between DS1 and DS2 can be interpreted in terms of dataset complexity and class imbalance. DS2 presents a more challenging classification environment, where overlapping class boundaries and uneven data distribution reduce model performance. However, transformer-based embeddings combined with deep learning architectures demonstrate strong generalization capability, particularly in the case of RoBERTa, which achieved improved performance even under complex conditions.

From a practical perspective, these findings suggest that model selection should be aligned with dataset characteristics and computational constraints. While RoBERTa-based models provide the highest accuracy for complex datasets, lightweight models such as DistilBERT and ALBERT offer competitive performance with reduced computational costs, making them suitable for real-time or resource-constrained applications.

Cross-Validation Results

To further evaluate the robustness and generalization capability of the proposed model, 5-fold cross-validation was performed on the best-performing configurations for both datasets. The results are presented in Tables 8 and 9 for DS1 and DS2, respectively.

Table 8. 5-Fold Cross Validation Results for DS1 (RoBERTa + LSTM)

Fold	Accuracy	F1 Score	Precision	Recall	Specificity
1	0.978	0.976	0.977	0.975	0.992
2	0.979	0.977	0.978	0.976	0.993
3	0.977	0.975	0.976	0.974	0.991
4	0.980	0.978	0.979	0.977	0.994
5	0.978	0.976	0.977	0.975	0.992

Average	0.9784	0.9764	0.9774	0.9754	0.9924

As shown in Table 8, the RoBERTa + LSTM model demonstrates consistently high performance across all folds for the DS1. The accuracy values range between 0.977 and 0.980, with an average accuracy of 0.9784. Similarly, F1-score, precision, and recall values remain stable across folds, indicating balanced classification performance. The low variation among fold results suggests stable performance across the cross-validation folds within the constructed dataset. However, this should not be interpreted as evidence of generalization to unseen contract types or external contract corpora.

Table 9. 5-Fold Cross Validation Results for DS2 (RoBERTa + LSTM).

Fold	Accuracy	F1 Score	Precision	Recall	Specificity
1	0.980	0.978	0.979	0.977	0.993
2	0.979	0.977	0.978	0.976	0.992
3	0.981	0.979	0.980	0.978	0.994
4	0.977	0.975	0.976	0.974	0.991
5	0.978	0.976	0.977	0.975	0.992
Average	0.9790	0.9770	0.9780	0.9760	0.9924

Table 9 presents the cross-validation results for the DS2. Similar to DS1, the RoBERTa + LSTM model achieves highly consistent performance across all folds. The accuracy values vary between 0.977 and 0.981, with an average accuracy of 0.9790. The consistency of F1-score, precision, and recall values indicates stable performance across the evaluated folds. These findings indicate that the proposed approach maintains stable performance across different data partitions of the constructed datasets.

Overall, the cross-validation results indicate that the proposed model achieved stable performance across the evaluated folds. However, because the experiments were conducted on datasets derived from a limited number of standard-form contracts, these results should be interpreted as internal validation findings rather than evidence of broad generalizability. The minimal variation across folds indicates stable learning behavior, while the consistently high-performance metrics demonstrate the model's ability to effectively capture discriminative patterns in construction contract texts. Moreover, the slightly higher performance observed in DS2 may be attributed to the presence of more distinctive linguistic features, which facilitate classification.

6. Discussion

6.1. Model Performance

To synthesize the extensive experimental results and highlight the most effective classification strategies, the top-performing 'champion' model combinations for both DS1 and DS2 are summarized in the Table 10 below. This consolidated overview serves as a critical reference point for comparing the effectiveness of different contextual embedding methods and classification architectures when applied to complex, multi-class contractual texts.

Table 10. Top 5 Model Combinations by Datasets.

Rank	Dataset	Model Combination (Embedding + Classifier)	Accuracy	F-1 Score	Precision
1	DS2	RoBERTa + LSTM	98.61%	0.985	0.986
2	DS1	DistilBERT + LSTM	98.33%	0.982	0.983

3	DS2	RoBERTa + GRU	98.33%	0.982	0.983
4	DS1	RoBERTa + LSTM	98.06%	0.979	0.980
5	DS1	DistilBERT + GRU	98.06%	0.979	0.980

The summary data clearly indicates that configurations coupling RoBERTa and DistilBERT with sequential deep learning architectures—specifically LSTM—consistently achieve the highest performance levels, with accuracy and F1-scores exceeding 98% in several experimental settings. This superior performance confirms that while traditional machine learning models remain competitive when paired with high-quality embeddings, sequential neural architectures are significantly more adept at capturing the intricate linguistic structures and long-range semantic dependencies inherent in construction contract clauses. Furthermore, the fact that the lightweight DistilBERT-LSTM model maintains parity with the more parameter-heavy RoBERTa model suggests a highly efficient pathway for deploying these tools in resource-constrained project governance environments.

The experimental results demonstrate that transformer-based contextual embeddings significantly improve the performance of both machine learning and deep learning classifiers in construction contract clause classification tasks. In this study, four transformer-based language models—BERT, ALBERT, RoBERTa, and DistilBERT—were integrated with multiple classification algorithms and evaluated on two datasets representing different contractual analysis tasks. The results show that contextual embeddings play a critical role in capturing semantic relationships within contractual sentences and improving classification accuracy.

Across both datasets, deep learning models consistently outperformed traditional machine learning classifiers. In particular, sequential architectures such as LSTM and GRU achieved the highest classification performance when combined with transformer-based embeddings. For example, in the DS1, which focuses on contractual obligation-related categories, the RoBERTa–LSTM combination achieved an accuracy of 98.06%, while the DistilBERT–LSTM configuration achieved the highest accuracy of 98.33%. These results indicate that deep learning models are particularly effective in learning sequential linguistic patterns present in contractual texts.

The performance gap between deep learning and traditional machine learning models can be explained by the nature of construction contract language. Contractual statements often contain subtle contextual differences, particularly between categories such as obligations, to-do actions, and optional clauses. Deep learning architectures are better suited to capturing these contextual dependencies because they can model sequential relationships between tokens within a sentence. Consequently, when contextual embeddings generated by transformer models are combined with sequential neural architectures, the resulting classification framework becomes highly effective for complex legal and technical text analysis.

In contrast, classical machine learning models such as Logistic Regression, Random Forest, and Support Vector Machines produced slightly lower performance scores, although they still achieved strong results overall. In several configurations, Random Forest and SVM models achieved accuracy levels above 95%, demonstrating that traditional classifiers remain competitive when paired with high-quality contextual embeddings. However, Naïve Bayes consistently showed the lowest performance among the tested algorithms, suggesting that probabilistic approaches relying on strong independence assumptions may be less suitable for complex contractual language.

The second dataset (DS2), which includes clauses related to time, cost, and quality dimensions of construction project management, generally produced slightly lower classification performance compared with DS1. This performance difference likely results

from the more complex semantic boundaries and dense semantic overlap between DS2 categories. Unlike DS1, where categories like 'Obligations' and 'Optional' are often distinguished by distinct modal verbs (e.g., 'shall' vs. 'may'), the categories in DS2—cost, time, and quality—frequently share a common technical vocabulary and interdependent procedural descriptions. For instance, a clause regarding 'delay liquidated damages' involves both time-related milestones and cost-related financial penalties, creating a higher level of semantic ambiguity that traditional models struggle to resolve. For example, contractual clauses referring to time and quality may share similar linguistic structures, making classification more challenging. This increased complexity is reflected in the 4–5% performance drop observed in most ML classifiers; however, the ability of transformer-based embeddings to capture high-dimensional contextual nuances allowed the DL models to maintain superior class separation despite these overlapping boundaries. Despite this increased complexity, deep learning models again demonstrated strong performance, particularly when combined with RoBERTa and DistilBERT embeddings.

Among the tested transformer models, RoBERTa and DistilBERT consistently produced the strongest results across both datasets. RoBERTa benefits from a more extensive pretraining process and improved training strategies compared with the original BERT architecture, enabling it to capture richer contextual representations. DistilBERT, although designed as a lighter and more efficient model, achieved performance levels comparable to RoBERTa while requiring fewer parameters and lower computational resources. These results suggest that both models are highly suitable for automated analysis of construction contract texts.

Overall, the experimental findings highlight the importance of combining contextual language representations with sequential neural architectures for construction contract analysis. The integration of transformer embeddings with deep learning classifiers provides a powerful framework for identifying contractual obligations, operational clauses, and project management provisions within complex construction contracts. These findings also confirm that advances in NLP and transformer-based language models can significantly enhance automated document analysis in construction management applications.

Another important factor contributing to the high performance of the proposed framework is the alignment between the characteristics of transformer-based embeddings and the linguistic structure of construction contract texts. Transformer models are particularly effective in capturing contextual dependencies and subtle semantic variations, which are essential for distinguishing between closely related contractual categories. When these embeddings are combined with sequential deep learning architectures such as LSTM and GRU, the models are able to further exploit token-level dependencies and sentence-level semantic structures. This synergy between contextual representation and sequential modeling plays a critical role in achieving robust and high-accuracy classification performance.

Although the datasets used in this study are relatively limited in size, this characteristic is consistent with the nature of construction contract analysis, where high-quality labeled data are difficult and time-consuming to obtain due to the need for domain expertise and careful annotation. Contract clauses often require expert interpretation, which restricts the availability of large-scale annotated datasets in this domain.

In addition, the relatively high classification performance observed in this study can be attributed to the structured and domain-specific nature of construction contract language. Many clauses contain distinctive linguistic patterns and terminology that can be effectively captured by advanced text representation methods, particularly

transformer-based embeddings. This allows models to learn discriminative features even from moderately sized datasets.

Furthermore, the use of contextual embeddings combined with both machine learning and deep learning models enhances the model's ability to capture semantic nuances and improves classification performance. However, it should be noted that the results may be influenced by dataset size and domain specificity, and future studies should validate the proposed framework using larger and more diverse contract datasets to further assess its generalizability.

6.2. Comparison with Previous Studies

The findings of this study align with and extend the growing body of research on natural language processing (NLP) applications in construction contract analysis. Previous studies have demonstrated the feasibility of applying machine learning and deep learning techniques to extract contractual requirements, identify risk clauses, and analyze construction-related documents. However, many of these studies have focused on relatively narrow analytical tasks or have relied on traditional text representation methods.

Early research in construction contract and bidding document analysis relied on a combination of rule-based NLP and conventional machine learning techniques. For instance, Lee et al. [5] developed a rule-based text mining framework to automatically identify potentially problematic clauses in construction contracts, demonstrating the feasibility of automated clause extraction based on predefined linguistic patterns. In contrast, Lee and Yi [4] employed text mining and machine learning methods to analyze unstructured bidding documents and predict project risk, highlighting the value of textual information relying on feature engineering and statistical representations. Despite their effectiveness, these early approaches were largely dependent on syntactic patterns, keyword-based features, or manually designed rules, limiting their flexibility and generalizability across diverse contract structures. Similarly, Hassan and Le [1] and Hassan et al. [6] employed natural language processing and supervised machine learning techniques to analyze construction contract documents, demonstrating the effectiveness of data-driven approaches for both requirement identification and classification tasks, with different algorithms achieving optimal performance under varying conditions. While these studies confirmed the practical value of NLP in construction management, they primarily relied on feature-based representations such as TF-IDF, n-grams, and Word2Vec, which provide limited capability in capturing deeper contextual dependencies and complex semantic relationships within contractual text.

More recent research has increasingly adopted transformer-based language models for construction contract analysis. For instance, Moon et al. [8] used a BERT-based model to classify contractual risk clauses and reported strong classification performance across multiple risk categories. Likewise, Zhou et al. [2] applied BERT together with deep learning architectures to detect missing clauses in construction contracts, demonstrating the effectiveness of contextual language representations in contract analysis tasks. These studies confirm that transformer-based embeddings significantly outperform traditional feature-based approaches when analyzing complex legal and technical texts.

The results of the present study are consistent with these findings but extend them in several important ways. First, unlike many prior studies that focus on a single classification task—such as risk detection, dispute identification, or vague clause detection—this study evaluates multi-class classification tasks across two different analytical dimensions. DS1 addresses contractual obligations and operational clauses, while DS2 focuses on key project management dimensions, including cost, time, and quality. This dual analytical framework provides a broader perspective on automated

construction contract analysis compared with earlier studies that typically address only one category of contractual information.

Second, this research systematically compares multiple transformer-based embeddings within a unified experimental framework. While previous studies often rely on a single contextual language model such as BERT, the present study evaluates four transformer models—BERT, ALBERT, RoBERTa, and DistilBERT—together with both traditional machine learning classifiers and sequential deep learning architectures. The results demonstrate that RoBERTa and DistilBERT consistently provide the highest classification accuracy, particularly when integrated with LSTM and GRU networks. This comparative evaluation provides deeper insights into the relative effectiveness of different contextual representation models for construction contract analysis.

Third, the classification performance obtained in this study exceeds the accuracy levels reported in several previous studies in the literature. For example, Gao et al. [14] introduced a context-aware deep learning framework that integrates RoBERTa and BiGRU to capture both sentence-level and multi-sentence contextual information for construction risk identification. Their model achieved an F1-score of 87.1% and an accuracy of 86.7%, highlighting the importance of contextual modeling in improving classification performance. In contrast, the transformer–deep learning combinations evaluated in this study achieved classification accuracies exceeding 98% in several configurations, particularly when RoBERTa or DistilBERT embeddings were combined with LSTM networks. While differences in datasets and classification tasks make direct numerical comparisons challenging, these results nonetheless highlight the effectiveness of combining contextual embeddings with sequential deep learning models for complex contractual text classification.

Another important distinction of this study lies in the use of real standard-form construction contracts, specifically FIDIC and JCT contract documents. Many previous studies rely on synthetic datasets, simplified contractual examples, or project-specific documents. By contrast, the dataset used in this research is derived from widely used international contract standards, which enhances the practical relevance and potential applicability of the proposed framework in real construction project environments.

Finally, recent literature has begun exploring the use of large language models (LLMs) for construction contract management and automated contract review. Studies such as Wong et al. [16], Zheng et al. [17], Tang et al. [20], and Zhang et al. [18] highlight the growing potential of large language model (LLM)-based systems in construction contract analysis, while addressing different task-specific challenges. Wong et al. [16] focus on contract risk identification by integrating LLMs with domain knowledge to mimic expert-level reasoning in detecting risky clauses. Zheng et al. [17] propose a knowledge graph-enhanced LLM framework that leverages retrieval-augmented reasoning to improve contract review accuracy and interpretability. Tang et al. [20] provide a comprehensive overview of LLM capabilities in the construction domain, emphasizing their potential for semantic understanding and complex knowledge processing. In contrast, Zhang et al. [18] develop a domain-specific question answering (QA) system based on supervised fine-tuning and reinforcement learning, enabling users to query contract knowledge in natural language. Despite these advancements, existing studies primarily focus on individual task-oriented applications rather than offering systematic comparisons across different modeling approaches or evaluation settings. The present research complements these emerging directions by providing empirical evidence on how contextual embeddings interact with both machine learning and deep learning classifiers in multi-class construction contract classification tasks.

Overall, the comparative analysis indicates that integrating transformer-based contextual embeddings with deep learning classifiers provides a powerful and flexible

approach for automated construction contract analysis. The proposed framework not only achieves high classification performance but also offers a comprehensive analytical structure, thereby contributing to the growing literature on AI-supported construction contract management.

6.3. Sustainability Implications

The sustainability implications of the proposed framework should be interpreted at the process level rather than as direct evidence of project-level reductions in disputes, delays, or cost overruns. This study evaluates classification performance and does not empirically measure downstream project outcomes. Therefore, the results do not prove that the proposed framework directly reduces project delays, contractual disputes, or administrative costs. Instead, they indicate that transformer-based clause classification may support sustainability-oriented contract governance by improving the speed, consistency, and structure of contract information processing.

The proposed NLP-based framework contributes to this objective by enabling automated classification and analysis of construction contract clauses. By integrating transformer-based contextual language models with machine learning and deep learning classifiers, the framework can identify contractual obligations, operational actions, and cost-, time-, and quality-related provisions embedded within contract documents. This capability may reduce manual screening workload and improve the consistency of clause identification; however, these practical effects should be empirically validated through future time-saving experiments, expert-based evaluations, and real project case studies.

From a sustainability perspective, the high classification performance of the best-performing models can be linked to measurable intermediate indicators rather than direct project-level outcomes. These indicators may include the proportion of clauses automatically classified with high confidence, the reduction in manual clause screening workload, faster retrieval of cost-, time-, and quality-related provisions, improved consistency of clause categorization across reviewers, and earlier identification of contract provisions relevant to risk allocation and decision-making. These indicators provide a more measurable bridge between NLP model performance and sustainability-oriented contract governance.

In addition, the proposed framework may support institutional transparency by providing a structured and repeatable mechanism for categorizing contractual clauses. Automated classification can help organize contractual responsibilities and project management provisions in a more consistent manner, which may assist stakeholders in reviewing obligations, payment-related clauses, time-related provisions, and quality requirements. However, the framework should be regarded as a decision-support tool rather than a substitute for expert legal or contractual judgment. Its contribution to governance lies in supporting more consistent information retrieval and clause organization, rather than independently ensuring transparency or accountability.

Another sustainability-related implication concerns the role of digital technologies in construction sector transformation. As the construction industry increasingly adopts digital project management systems, NLP-based contract analysis tools could be integrated into document management workflows to support faster clause retrieval and more structured contract review. Such systems may assist project managers in locating cost-related clauses, schedule constraints, and quality requirements more efficiently, although the actual time savings and operational benefits should be quantified in future implementation studies.

Finally, the integration of advanced NLP techniques into construction contract management contributes to the broader goal of sustainable infrastructure delivery. By improving the clarity, accessibility, and interpretability of contractual information,

automated contract analysis tools can support more informed decision-making throughout the project lifecycle. In this sense, the proposed framework not only advances research on NLP-based construction document analysis but also contributes to the development of more transparent, efficient, and sustainable construction project management practices.

In addition to improving operational efficiency, the proposed framework contributes to the broader sustainability goals of the construction industry by supporting data-driven governance. By enabling automated and consistent interpretation of contractual clauses, the framework reduces uncertainty in decision-making processes and enhances transparency among project stakeholders. This contributes to more resilient and sustainable project delivery systems.

Accordingly, the sustainability-related contribution of this study lies not in demonstrating direct reductions in delays, disputes, or administrative costs, but in providing a technically evaluated classification framework that can serve as a basis for future empirical assessment of contract governance efficiency.

As illustrated in Figure 4, the proposed framework may contribute to sustainable construction contract governance through process-level improvements rather than directly measured project-level outcomes.

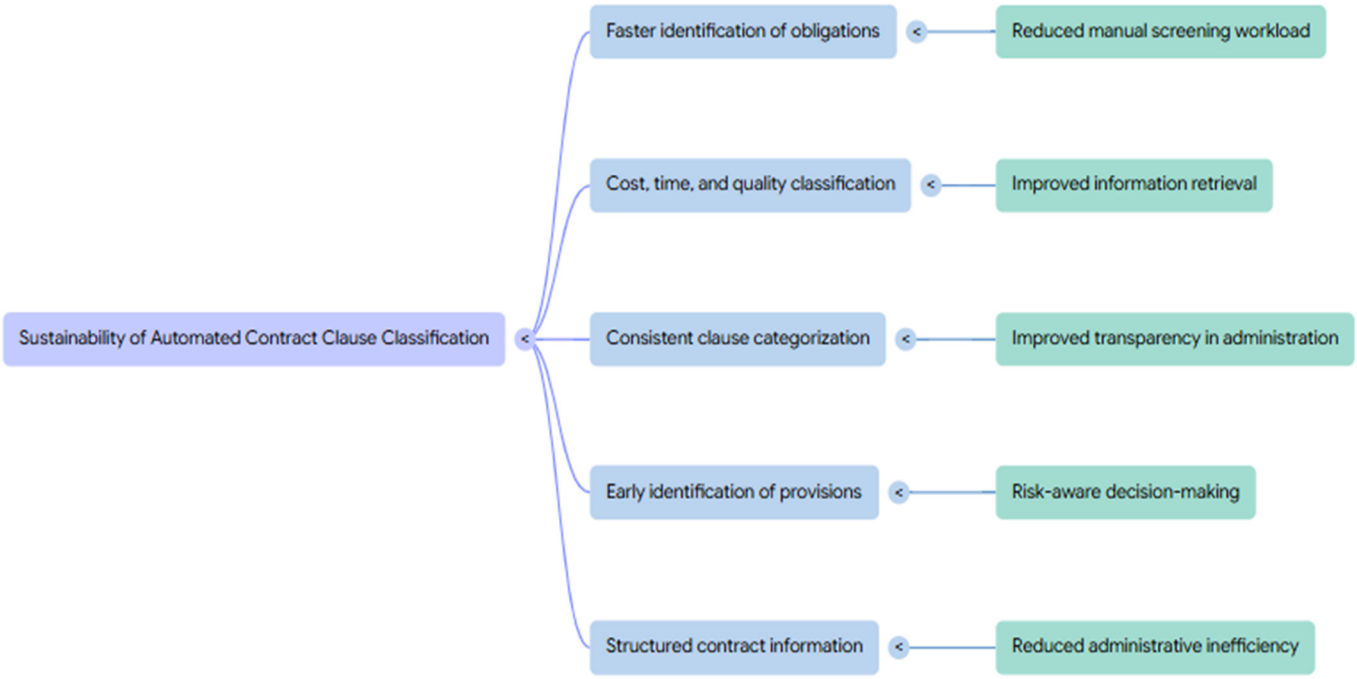


Figure 4. Conceptual Link Between Automated Contract Clause Classification and Potential Process-Level Contributions to Sustainable Construction Contract Governance.

Figure 4 summarizes the potential process-level sustainability implications of the proposed framework. It should be noted that this study does not directly measure project-level reductions in disputes, delays, or cost overruns. Instead, the framework provides a classification mechanism that may support sustainability-oriented contract governance by improving the speed and consistency of clause identification. In practical terms, automated classification of obligations and cost-, time-, and quality-related clauses may reduce manual screening workload, improve contract information retrieval, and support more transparent contract administration. These process-level outputs may serve as intermediate indicators for future empirical studies examining the relationship between NLP-based contract analysis and sustainable construction project governance.

Despite the strong classification performance achieved in this study, several limitations should be acknowledged. First, the datasets are relatively small and derived from standard-form contract documents, which may affect the generalizability of the results. Additionally, although high accuracy values were achieved, further validation using larger and more diverse contract datasets is required. Second, this study evaluates classification performance rather than actual reductions in review time, disputes, delays, or administrative costs. Therefore, future studies should validate the proposed framework using larger and more diverse contract datasets, expert-based evaluations, time-saving experiments, and real project case studies. Future studies may also explore different contract types and multilingual datasets to improve robustness.

From a practical sustainability perspective, the proposed framework may contribute to measurable improvements in construction project governance and administrative efficiency. Automated identification and classification of contractual clauses can reduce the time required for manual contract review, accelerate the detection of critical obligations and risk-related provisions, and improve consistency in contractual interpretation across project stakeholders. These improvements may help reduce administrative workload, minimize contract-review-related delays, and support faster decision-making processes during project execution.

In large-scale infrastructure projects, delays in identifying contractual obligations or inconsistencies in contract interpretation frequently contribute to claims, disputes, schedule overruns, and additional administrative costs. Therefore, high-accuracy transformer-based clause classification systems may indirectly support sustainability objectives by reducing procedural inefficiencies, improving resource utilization, and enhancing institutional transparency within construction project governance. Although the present study does not directly quantify these operational impacts, the findings provide a foundation for future research investigating measurable sustainability indicators such as reduction in contract review time, dispute frequency, administrative processing effort, and project delay risks.

6.4. Limitations and Future Research

Although the proposed framework achieved promising classification performance, several limitations should be acknowledged. First, the datasets were derived from a limited number of standard-form construction contracts, which may restrict the generalizability of the findings across different contractual contexts and jurisdictions. Second, the annotation process was conducted manually by the authors, and formal inter-annotator agreement metrics were not calculated. Third, the train–test split was performed at the sentence level, which may introduce contextual similarity between training and testing samples, potentially leading to optimistic performance estimates.

In addition, although cross-validation was employed to improve the internal stability assessment of the experimental evaluation, the study did not include an external held-out validation dataset or detailed error analysis. Therefore, while the proposed models demonstrated strong comparative performance within the constructed datasets, the generalizability and failure behavior of the models across unseen contract domains remain uncertain. Future studies should incorporate external validation datasets, confusion-pattern analysis, and explainable AI approaches to better evaluate model stability, interpretability, and misclassification behavior.

Another important direction for future research involves benchmarking the proposed framework against general-purpose large language models (LLMs) such as ChatGPT and GPT-4. Although transformer-based embeddings were evaluated extensively in this study, direct comparison with instruction-tuned LLMs may provide additional insight into semantic reasoning capabilities in construction contract analysis.

Furthermore, investigating cases where LLMs fail to correctly classify contractual clauses could help identify common sources of ambiguity, contextual misunderstanding, and domain-specific limitations in legal–technical language interpretation.

Another limitation of the study relates to the use of a multi-class single-label classification framework for Dataset 2 (DS2). In practice, construction contract clauses may simultaneously involve multiple project management dimensions such as cost, time, and quality. For example, clauses concerning delays, penalties, or corrective actions may contain overlapping semantic characteristics across different categories. However, to maintain a consistent experimental framework and enable systematic comparison among different transformer-based embeddings and classification architectures, each sentence in DS2 was assigned a single dominant label during the annotation process.

Another important limitation concerns the relatively small size of the datasets used in this study. DS1 and DS2 contain 450 and 827 annotated sentences, respectively, which may increase the risk of overfitting, particularly when using high-capacity transformer-based models such as RoBERTa. Although pre-trained transformer embeddings can mitigate data scarcity issues through transfer learning and contextual representation learning, limited dataset size may still affect model generalizability and robustness.

To reduce overfitting risk, cross-validation and comparative evaluation across multiple machine learning and deep learning architectures were employed during the experimental process. Nevertheless, the reported results should be interpreted primarily as comparative benchmarking results within the scope of the constructed datasets. Future studies should validate the proposed framework using larger, more diverse, and cross-domain construction contract corpora.

Although this simplification improved methodological consistency, it may not fully capture the multidimensional semantic structure of complex contractual provisions. Therefore, future studies should investigate multi-label classification approaches capable of representing overlapping contractual concepts and more complex semantic relationships within construction contracts.

Future studies should focus on larger and more diverse datasets, cross-contract validation strategies, and the integration of large language models (LLMs) for more advanced semantic reasoning and automated contract interpretation.

7. Conclusions

This study presented a comprehensive framework for the automated classification of construction contract clauses by integrating transformer-based contextual embeddings with both machine learning and deep learning models. Construction contracts are inherently complex documents characterized by dense legal and technical language, making manual interpretation time-consuming and prone to inconsistency. By leveraging recent advances in natural language processing, the proposed approach addresses these challenges through a systematic and scalable analytical pipeline.

The experimental results demonstrate that transformer-based embeddings significantly enhance classification performance across different contract analysis tasks. In particular, the combination of contextual representations with sequential deep learning architectures, such as LSTM and GRU, consistently achieved the highest performance levels. Among the evaluated models, RoBERTa–LSTM and DistilBERT–LSTM configurations delivered the highest classification performance within the constructed datasets, achieving accuracies exceeding 98% in multiple experimental settings. These findings confirm that the synergy between contextual language modeling and sequential neural architectures is highly effective for capturing the semantic complexity of contractual texts.

Beyond performance improvements, this study contributes to the literature by proposing a multi-class, multi-dimensional classification framework that extends beyond the task-specific focus of many previous studies. By simultaneously analyzing contractual obligations and key project management dimensions such as cost, time, and quality, the proposed framework provides a more comprehensive understanding of construction contract structures. Furthermore, the systematic comparison of multiple transformer embeddings and classification models offers valuable insights into the relative effectiveness of different NLP approaches for contract analysis.

From a practical perspective, the proposed framework has the potential to support automated contract review in construction project environments. The ability to classify contractual clauses rapidly and consistently may assist experts in reviewing obligations, operational actions, and cost-, time-, and quality-related provisions. Rather than replacing expert judgment, the framework should be viewed as a decision-support mechanism that may improve the organization and retrieval of contractual information.

From a sustainability standpoint, the findings suggest that automated clause classification may provide process-level support for economic efficiency, institutional transparency, and contract governance. By enabling more structured and consistent identification of contractual information, the proposed approach may help reduce manual review workload and support decision-making processes. However, direct impacts on resource efficiency, contractual conflicts, project delays, or administrative costs should be empirically measured in future studies.

Despite the promising results, several limitations should be acknowledged. The relatively limited dataset size may affect the generalizability of the findings, and further validation using larger and more diverse contract datasets is necessary. Future research may also explore more advanced transformer architectures and multilingual or cross-lingual contract datasets to improve model robustness across different legal and linguistic contexts.

Another promising direction involves improving the interpretability of NLP-based contract analysis systems. Integrating explainable artificial intelligence (XAI) techniques, such as SHAP or LIME, could enhance transparency in model decision-making processes, which is particularly critical in legal and contractual domains where interpretability and trust are essential. In addition, future studies may investigate the practical deployment of the proposed framework through API-based services and real-time applications, enabling integration with digital project management platforms. Exploring hybrid architectures that combine transformer-based embeddings with alternative neural structures may also further extend the applicability and performance of automated construction contract analysis systems.

Overall, this study demonstrates that combining transformer-based NLP techniques with deep learning models provides a promising approach for automated construction contract clause classification. The proposed framework advances research in construction text mining and offers a basis for future empirical studies on how automated clause classification may support contract governance and sustainable construction project management.

Author Contributions: Conceptualization, A.D. and L.O.U.; methodology, A.D. and L.O.U.; software, A.D.; validation, A.D. and L.O.U.; formal analysis, A.D.; investigation, A.D. and L.O.U.; resources, A.D.; data curation, A.D. and L.O.U.; writing—original draft preparation, A.D.; writing—review and editing, A.D. and L.O.U.; visualization, A.D.; supervision, L.O.U.; project administration, L.O.U.; funding acquisition, A.D. and L.O.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The full clause-level dataset cannot be made publicly available due to copyright restrictions associated with the standard-form contract documents from which the clauses were derived. Representative anonymized labeled examples are provided in Appendix A. The dataset supporting the findings of this study may be made available by the corresponding author upon reasonable request, subject to copyright and research-use restrictions.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-5.5) for language editing and polishing, translation support, summarization, paraphrasing, citation support, accessing and organizing scientific literature, and limited assistance in drafting, refining, and structuring the presentation of the results and discussion. Google NotebookLM (Gemini 1.5 Pro) was used to generate draft schematic/illustrative figures based on textual content prepared by the authors, and ChatGPT (GPT-5.5) was used to support the subsequent editing and refinement of these figures according to the authors' own instructions and prompts. All AI-assisted outputs were reviewed, revised, verified, and approved by the authors, who take full responsibility for the accuracy, integrity, and final content of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

This appendix presents representative anonymized examples of contract clauses together with their manually assigned labels used during the annotation process.

Table A1. Representative Anonymized Labeled Examples.

No.	Representative Anonymized Clause Example	Label	Rationale for Labeling
1	The contractor shall maintain the required insurance policies throughout the contract period and ensure that the works, personnel, and site materials are adequately covered.	Obligations	This clause imposes a binding contractual duty on the contractor. The use of "shall maintain" indicates a mandatory requirement rather than an optional or descriptive statement.
2	The contractor must notify the employer's representative immediately after identifying any error, discrepancy, or event that may affect the proper execution of the works.	Obligations	The clause creates a compulsory notification duty. The modal verb "must" and the time-sensitive requirement "immediately" show that the contractor is legally required to act.
3	The contractor shall submit a detailed work programme to the employer's representative within the period specified in the contract documents.	Operational Actions	This clause describes a concrete procedural action: preparing and submitting a work programme. Although it uses "shall", the main function is not to define a broad legal responsibility but to specify an executable project management task.
4	The contractor shall provide the necessary equipment, documents, personnel, and assistance required for carrying out the specified tests.	Operational Actions	The clause identifies practical resources and support that must be provided for a testing activity. It is classified as Operational Actions because it refers to the implementation of a specific site or quality-control procedure.
5	The employer may postpone the handover of the site or any relevant part of it for a limited period if such postponement is permitted under the contract particulars.	Optional Provisions	This clause gives the employer a discretionary right to postpone site possession under specified conditions. The use of "may" and the conditional reference to the contract particulars indicate that the

			action is permitted but not automatically mandatory.
6	The employer may, with the contractor's consent, use or occupy part of the works before formal completion has been certified.	Optional Provisions	The clause describes a conditional permission rather than a compulsory requirement. It depends on the contractor's consent and therefore represents an optional arrangement available under certain circumstances.
7	The documents forming the contract shall be read together and interpreted as mutually explanatory parts of the same agreement.	Other General Statements	This clause explains how contract documents should be interpreted. It does not impose a specific operational task or discretionary option; instead, it provides a general rule for contractual interpretation.
8	The order of priority among the contract documents shall be determined according to the sequence stated in the contract agreement.	Other General Statements	The clause defines the hierarchy of contract documents. It is classified as a general statement because its main function is to clarify document precedence rather than assign a specific duty to a party.
9	The contract price shall include the amounts payable for the design, execution, completion, and operation-related services, together with any adjustments permitted under the contract.	Cost	This clause is classified as Cost because its primary function is to define the financial scope of the contract price and the possibility of price adjustments.
10	If the contractor incurs additional cost as a result of an employer-related delay or instruction, the contractor may be entitled to reimbursement or an adjustment to the contract sum.	Cost	The clause focuses on financial compensation arising from delay or employer instruction. Although delay is mentioned, the dominant function is the recovery of additional costs.
11	The commencement date shall be the date notified by the employer's representative as the formal starting point for the execution of the works.	Time	This clause is classified as Time because its main function is to define the formal start date of the project. It establishes a temporal reference point for subsequent contractual periods and deadlines.
12	The contractor shall complete the works, or any specified section of the works, within the time for completion stated in the contract data, including any approved extension of time.	Time	The clause directly regulates the completion period and refers to possible extensions of time. Its dominant function is scheduling rather than cost allocation or quality control.
13	The contractor shall ensure that the design, technical documents, materials, and workmanship comply with the employer's requirements and the applicable technical standards.	Quality	This clause is classified as Quality because its primary function is to regulate compliance with design, material, workmanship, and technical standards.
14	The contractor shall prepare and keep updated a complete set of as-built records showing the works as actually executed on site.	Quality	The clause concerns documentation accuracy and technical record-keeping. It is labeled as Quality because as-built records support verification of whether the completed works conform to contractual and technical requirements.

References

1. Hassan, F.U.; Le, T. Automated requirements identification from construction contract documents using natural language processing. *J. Leg. Aff. Disput. Resolut. Eng. Constr.* **2020**, *12*, 04520009. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000379.
2. Zhou, H.; Gao, B.; Tang, S.; Li, B.; Wang, S. Intelligent detection on construction project contract missing clauses based on deep learning and NLP. *Eng. Constr. Archit. Manag.* **2025**, *32*, 1546–1580. https://doi.org/10.1108/ECAM-02-2023-0172.

3. Deza, J.I.; Ihshaish, H.; Mahdjoubi, L. A machine learning approach to classifying construction cost documents into the international construction measurement standard. *arXiv* **2022**, arXiv:2211.07705. https://doi.org/10.48550/arXiv.2211.07705.
4. Lee, J.; Yi, J.-S. Predicting project's uncertainty risk in the bidding process by integrating unstructured text data and structured numerical data using text mining. *Appl. Sci.* **2017**, *7*, 1141. https://doi.org/10.3390/app7111141.
5. Lee, J.; Yi, J.-S.; Son, J. Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based NLP. *J. Comput. Civ. Eng.* **2019**, *33*, 04019003. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000807.
6. Hassan, F.U.; Le, T.; Tran, D.-H. Multi-class categorization of design–build contract requirements using text mining and natural language processing techniques. In *Construction Research Congress 2020: Computer Applications, Proceedings of the Construction Research Congress 2020, Tempe, AZ, USA, 8–10 March 2020*; American Society of Civil Engineers: Reston, VA, USA, 2020. https://doi.org/10.1061/9780784482889.135.
7. Jafari, P.; Al Hattab, M.; Mohamed, E.; Abourizk, S. Automated extraction and time-cost prediction of contractual reporting requirements in construction using natural language processing and simulation. *Appl. Sci.* **2021**, *11*, 6188. https://doi.org/10.3390/app11136188.
8. Moon, S.; Chi, S.; Im, S.-B. Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (BERT). *Autom. Constr.* **2022**, *142*, 104465. https://doi.org/10.1016/j.autcon.2022.104465.
9. Candaş, A.B.; Tokdemir, O.B. Automated identification of vagueness in the FIDIC Silver Book Conditions of Contract. *J. Constr. Eng. Manag.* **2022**, *148*, 04022007. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002254.
10. Candaş, A.B. Multipurpose Semantic Analysis of Construction Text Documentation. Ph.D. Thesis, Middle East Technical University, Ankara, Türkiye, 2022.
11. Zhong, B.; Shen, L.; Pan, X.; Zhong, X.; He, W. Dispute classification and analysis: Deep learning-based text mining for construction contract management. *J. Constr. Eng. Manag.* **2024**, *150*, 04023151. https://doi.org/10.1061/JCEMD4.COENG-14080.
12. Kazemi, M.H.; Alvanchi, A. Application of NLP-based models in automated detection of risky contract statements written in complex script system. *Expert Syst. Appl.* **2025**, *259*, 125296. https://doi.org/10.1016/j.eswa.2024.125296.
13. Dikmen, I.; Eken, G.; Erol, H.; Birgonul, M.T. Automated construction contract analysis for risk and responsibility assessment using natural language processing and machine learning. *Comput. Ind.* **2025**, *166*, 104251. https://doi.org/10.1016/j.compind.2025.104251.
14. Gao, N.; Touran, A.; Wang, Q.; Beauchamp, N. Construction risk identification using a multi-sentence context-aware method. *Autom. Constr.* **2024**, *164*, 105466. https://doi.org/10.1016/j.autcon.2024.105466.
15. Lee, W.; Lee, S. Development of a knowledge base for construction risk assessments using BERT and graph models. *Buildings* **2024**, *14*, 3359. https://doi.org/10.3390/buildings14113359.
16. Wong, S.; Zheng, C.; Su, X.; Tang, Y. Construction contract risk identification based on knowledge-augmented language models. *Comput. Ind.* **2024**, *157–158*, 104082. https://doi.org/10.1016/j.compind.2024.104082.
17. Zheng, C.; Wong, S.; Su, X.; Tang, Y.; Nawaz, A.; Kassem, M. Automating construction contract review using knowledge graph-enhanced large language models. *Autom. Constr.* **2025**, *175*, 106179. https://doi.org/10.1016/j.autcon.2025.106179.
18. Zhang, M.; Xu, C.; Gan, Y.; Wang, Y.; Fu, Y.; Chen, Y. Automating construction contract question answering using large language model and fine-tuning. *Expert Syst. Appl.* **2026**, *297*, 129493. https://doi.org/10.1016/j.eswa.2025.129493.
19. Xiao, H.; Zhuang, J.; Yang, B.; Han, J.; Yu, Y.; Lai, S. Generative knowledge-guided review system for construction disclosure documents. *Adv. Eng. Inform.* **2025**, *68*, 103618. https://doi.org/10.1016/j.aei.2025.103618.
20. Tang, Y.; Zheng, C.; Kassem, M.; Xie, X.; Su, X. Bridging the research–practice gap in construction contract management with NLP and LLMs. *Autom. Constr.* **2026**, *181*, 106614. https://doi.org/10.1016/j.autcon.2025.106614.
21. Nasser, M.; Hamdy, A.; Marzouk, M. Analyzing construction contract selection factors: A semantic analysis using NLP. *Adv. Eng. Inform.* **2026**, *69*, 103810. https://doi.org/10.1016/j.aei.2025.103810.
22. Kibert, C.J. *Sustainable Construction: Green Building Design and Delivery*, 4th ed.; Wiley: Hoboken, NJ, USA, 2016.
23. Love, P.E.D.; Matthews, J.; Simpson, I.; Hill, A.; Olatunji, O. A benefits realization management building information modeling framework for asset owners. *Autom. Constr.* **2014**, *37*, 1–10. https://doi.org/10.1016/j.autcon.2013.09.007.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30, Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
25. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692.

26. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108.
27. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942. https://doi.org/10.48550/arXiv.1909.11942.
28. Aggarwal, C.C.; Zhai, C. A survey of text classification algorithms. In *Mining Text Data*; Aggarwal, C.C., Zhai, C., Eds.; Springer: Boston, MA, USA, 2012; pp. 163–222. https://doi.org/10.1007/978-1-4614-3223-4_6.
29. Jurafsky, D.; Martin, J.H. Speech and Language Processing, 3rd ed.; online manuscript. 2023. Available online: https://web.stanford.edu/~jurafsky/slp3/ (accessed on 6 June 2026).
30. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
31. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1724–1734. https://doi.org/10.3115/v1/D14-1179.
32. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
33. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.
34. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Int. J. Mach. Learn. Technol.* **2011**, *2*, 37–63.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.