Canadian Journal of Marketing Research
ISSN: 0829-4836
https://canadian-jmr.com/index.php/cjmr

Research Article

AI/NLP-Based Predictive Project Risk Analytics: The Contract Risk Analyzer for Energy EPC Contracts

Harish Satyala[1*] and Vijaya Dixit[2]
[1]Research Scholar, Operations Management, Indian Institute of Management Ranchi (IIM-Ranchi), Jharkhand 835303, India
[2]Associate Professor, Operations Management, Indian Institute of Management Ranchi (IIM-Ranchi), Jharkhand 835303, India

***Corresponding Author**
Dr. Harish Satyala
(harish.s19eph@iimranchi.ac.in)

Article History
Received: 21.09.2026
Accepted: 18.10.2026
Published: 09.11.2026

DOI: 10.61336/cjmr.1604.36

Abstract: In business-to-business energy markets, much of the value of a project is won or lost at the tender stage, where long engineering, procurement and construction (EPC) contracts allocate cost, schedule and liability risk through dense legal text. This article develops and evaluates the Contract Risk Analyzer, a design-science artefact that converts contract sentences into a prioritized and explainable risk register. A corpus of 1,304 cleaned sentences drawn from four FIDIC contract books was labelled by sentence type (five classes) and responsible party (three classes). Twenty individual classifiers built on bag-of-words, TF-IDF, character n-gram and latent semantic representations were benchmarked against eight hybrid ensembles using an 80/20 hold-out design and five-fold cross-validation. The deployed stacking meta-learner achieved 89.7% accuracy, 88.7% macro-F1 and 98.9% ROC-AUC for sentence type, and an optimized soft-voting ensemble achieved 97.4% accuracy and 98.1% macro-F1 for responsible party. The best hybrid exceeded the best individual model by 4.7 macro-F1 points. Classifier outputs were combined with onerous-term rules and a 3,000-record quantified risk library and applied to a 176-page EPC works contract containing 2,240 sentences. The system registered 371 risk-bearing findings across 24 categories, including 102 Critical and 165 High findings, produced a Contract Risk Index of 79.2% and allocated 71.4% of findings to the contractor. A composite predictability index routed 30.5% of findings to expert review. The results show how text analytics turns contract documents into decision-ready intelligence for bid/no-bid, pricing and negotiation decisions, and they offer a reproducible template for hybrid text classification in marketing research.

Keywords: Contract Analytics, Text Mining, Natural Language Processing, Hybrid Ensemble Learning, B2B Tendering, Risk Intelligence, Energy EPC Projects

Copyright @ 2026: This is an open-access article distributed under the terms of the Creative Commons Attribution license which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use (NonCommercial, or CC-BY-NC) provided the original author and source are credited.

INTRODUCTION

Words are part of almost every marketplace exchange, and marketing scholars increasingly treat them as data. Automated text analysis has been used to explain consumer attitudes, forecast product success and anticipate credit risk from the language that market actors produce [1-3]. In business-to-business (B2B) markets the most consequential text is often the contract itself. For firms that sell engineering, procurement and construction (EPC) services to energy-sector clients, the contract is the instrument that fixes price, scope, schedule, liability and the allocation of uncertainty between buyer and seller. Decisions about whether to bid, how much risk premium to price, which clauses to negotiate and how to staff a project are therefore decisions about contract text.

Energy EPC contracts are long, heterogeneous and dense. A single set of general conditions can run to hundreds of pages, and the practical meaning of a clause on time bars, delay damages, payment, variations, site conditions or set-off is often dispersed across definitions, cross-references and special conditions. Bid teams typically have days rather than weeks to review them. Manual review is slow, varies by reviewer and tends to concentrate attention on familiar clauses,

so onerous provisions can be priced too low or not at all. Studies of EPC bidding show that information hidden in unstructured documents carries predictive signal: text extracted from invitations to bid, scopes of work and clarification records improves the prediction of schedule delay and bidding risk relative to structured data alone [4-6].

Recent work in construction informatics has shown that natural language processing (NLP) and machine learning (ML) can classify contract sentences, detect risk clauses and identify the party responsible for an obligation [7-9]. Three limitations remain. First, most systems stop at classification and do not translate predictions into a priced, owned and monitored risk register that a commercial team can act on. Second, model portfolios are usually small, so the incremental value of hybrid ensembles over strong individual classifiers is rarely quantified across many alternatives. Third, few systems make the uncertainty of each automated judgement visible, even though trust and calibrated confidence determine whether managers adopt analytic tools [10-11].

This article addresses these limitations by designing, implementing and evaluating the Contract Risk Analyzer, a hybrid NLP, ML and rule-based decision-support system for energy EPC contracts. Following a design-science approach [12], the system was trained on sentences from four FIDIC contract books, benchmarked across 20 individual classifiers and eight hybrid ensembles, and then applied to a real 176-page EPC works contract. The research is guided by four questions:

- **RQ1:** How accurately can contract sentences be classified by functional type and responsible party using individual and hybrid NLP models
- **RQ2:** Do hybrid ensembles improve hold-out classification performance relative to the best individual models
- **RQ3:** How can sentence classifications, onerous-language rules and a quantified risk library be combined into transparent, prioritized risk intelligence for a full contract
- **RQ4:** How can the uncertainty of automated outputs be measured and used to direct expert review

The article makes three contributions. Methodologically, it provides one of the broadest published comparisons of contract-sentence classifiers, covering 28 configurations across five text representations, and it reports the full metric profile, calibration and class-level behaviour of the selected models. Substantively, it shows that a hybrid pipeline can convert a 176-page contract into a ranked register of 371 risk-bearing findings with owners, mitigation actions and early-warning indicators. Managerially, it shows how text analytics supports the commercial decisions at the heart of B2B selling in project markets: bid/no-bid screening, risk pricing, negotiation agenda setting and handover from sales to delivery. For marketing researchers, the design offers a reusable template for hybrid text classification in which machine predictions, domain rules and human review are combined and their reliability is reported explicitly.

Literature Review

Text as Data for Business Decisions: Marketing research has moved from counting words to modelling language as evidence about the people and organizations that produce it. Berger *et al.* [1] distinguish text that reflects its producer from text that affects its receiver, and argue that automated analysis can unite consumer behaviour, strategy and quantitative modelling around a common data source. Humphreys and Wang [2] set out a staged workflow from construct definition to validation and warn that automated measures must be validated against human judgement. Netzer *et al.* [3] show that the words borrowers write in loan applications predict default beyond financial variables, demonstrating that text can carry risk signals that structured data miss. Contracts are a natural extension of this agenda: they are written by the buyer, read by the seller and fix the economic terms of the exchange. Yet contract text has received little attention in marketing research compared with reviews, social media and service conversations.

Classical information retrieval provides the representational foundation for most applied text classifiers. Term frequency–inverse document frequency (TF-IDF) weights terms by their within-document frequency and their rarity across a corpus [13], and n-gram features capture short phrases such as liquidated damages, extension of time or at his own cost. Latent semantic analysis reduces sparse term matrices to dense components that capture co-occurrence structure. These representations remain competitive for short, formulaic legal sentences, where vocabulary is constrained and phrase-level cues are strong.

Automated Contract Analysis in EPC and Construction

A growing stream of research applies NLP to construction and EPC contracts. Choi *et al.* [4] built text-mining and AI modules that extract contractor risks from invitations to bid and EPC contracts using phrase matching and named-entity recognition, and Choi and Lee [14] extended this with an ontological semantic model and bidirectional LSTM networks. Park *et al.* [15] developed a design-risk tool for EPC technical specifications, and Jafari *et al.* [16] combined NLP

Table 1: Representative Studies on Predictive Project Risk and Contract Analytics

Study	Context and data	Method	Main contribution	Remaining gap
Lee and Yi [5]	Construction bidding and clarification text	Text mining with structured data	Text improved bidding-risk prediction	No contract-level risk register
Son and Lee [6]	13 offshore oil and gas EPC projects	Text mining and regression	Schedule-delay risk from bid documents	Small case base, narrow model set
Owolabi *et al.* [22]	4,294 PPP projects	Regression, trees, RF, SVM, DNN	Benchmarked completion-risk models	Structured data only
Choi *et al.* [4]	EPC invitations to bid and contracts	Phrase matching, NER, ML	Automated contractor-risk extraction	Limited model comparison
Park *et al.* [15]	EPC technical specifications	ML extraction modules	Design-risk and parameter extraction	Technical, not commercial, risk
Jafari *et al.* [16]	Construction contracts	NLP and simulation	Reporting requirements and time–cost	Single risk type
Moon *et al.* [8]	Construction specifications	BERT	Risk-clause detection	No allocation or mitigation
Wong *et al.* [9]	Construction contracts	Knowledge-augmented language models	Contract-risk identification	Limited decision support
Dikmen *et al.* [7]	FIDIC and project contracts	12 NLP/ML models and ensemble	Sentence type and responsible party	Transfer to energy EPC, no scoring
This article	Four FIDIC books, 176-page energy EPC contract	20 individual and 8 hybrid models, rules, risk library	Scored, owned and monitored risk register with uncertainty flags	–

PPP = public–private partnership, NER = named-entity recognition, RF = random forest, SVM = support vector machine, DNN = deep neural network, BERT = bidirectional encoder representations from transformers

extraction of reporting requirements with time–cost simulation. Moon *et al.* [8] detected risk clauses in construction specifications with a BERT model, Wong *et al.* [9] proposed knowledge-augmented language models for contract-risk identification, Pham and Han [17] predicted risk-handling actions with multitask classification, and Kim *et al.* [18] generated detection rules automatically from contract documents. Fu *et al.* [19] machine-coded the control, coordination and adaptation functions of contract clauses. Closest to the present work, Dikmen *et al.* [7] benchmarked twelve NLP and ML models on FIDIC-based sentences and classified both sentence type and responsible party. Outside construction, the CUAD benchmark established expert-annotated clause review as an NLP task [20], and Graham *et al.* [21] classified deontic modalities such as obligations and permissions in contracts.

Parallel work on predictive project risk shows that risk can be learned from both structured and textual evidence. Owolabi *et al.* [22] benchmarked regression, tree, support-vector and neural models to predict completion risk across 4,294 publics–private partnership projects, and Son and Lee [6] used text mining on 13 offshore oil and gas EPC projects to estimate schedule-delay risk during bidding. Table 1 summarizes representative studies and the gap that each leaves for an integrated, energy-oriented decision-support system.

Ensemble Learning, Calibration and Explainability

Ensembles combine classifiers whose errors are imperfectly correlated. Stacked generalization trains a meta-learner on the out-of-fold predictions of base models [23], bagging-based methods such as random forests average decorrelated trees [24], and gradient-boosting systems add learners sequentially to correct residual errors [25-27]. Voting schemes aggregate labels or probabilities and can be weighted by validated performance. Ensembles typically gain most when classes are imbalanced and individual models fail on different minority classes, which is the case in contract text, where obligations dominate and explicit risk or right sentences are comparatively rare.

For decision support, accuracy alone is insufficient. Macro-averaged metrics give each class equal weight and are preferred for imbalanced problems [28]. Calibration determines whether a predicted probability of 0.8 corresponds to 80% observed accuracy [10], and local explanations increase user trust by showing why a case was classified as it was [11]. Work on legal prediction cautions that models trained on documents should be framed as screening aids rather than as predictions of adjudicated outcomes [29]. These insights motivate a design that reports class-level performance and calibration, attaches rule-based reasons to each flagged clause and routes uncertain cases to human experts.

Research Gap

Prior work establishes that contract sentences can be classified automatically, but it leaves four gaps that this article addresses. Model comparisons are narrow and rarely quantify hybrid gains across many alternatives. Classification outputs are seldom converted into scored, owned and mitigated risks at the level of a full contract. The reliability of individual automated judgements is rarely made visible. And applications to energy-sector EPC contracts, where commercial risk allocation drives bidding and pricing decisions, remain scarce. The system described next responds to each gap.

MATERIALS AND METHODS

Research Design

The research follows a pragmatic, design-science orientation in which an artefact is built to solve a practical problem and is then evaluated rigorously [12]. The artefact is the Contract Risk Analyzer, and the design follows an input–process–output logic (Figure 1). Evaluation evidence is generated at three levels: supervised evaluation of classifiers on labelled sentences, algorithmic integration of classifier probabilities, rule triggers and risk-library parameters into priority scores, and a field demonstration on a real EPC contract. The unit of analysis for classification is the contract sentence, the reporting unit for decision support is the risk-bearing clause matched to a risk-library record. Because the system screens contract language before execution, its outputs are interpreted as prioritization aids for professional review rather than as predictions of litigated outcomes.

Data

Labelled Contract Corpus: Training data were assembled from four FIDIC contract books: The Red Book (construction), Yellow Book (plant and design–build), Orange Book (design–build and turnkey) and Silver Book (EPC/turnkey). Each sentence record carries an identifier, book, language, page, clause, sentence text, sentence-type label, related-party label, label confidences, a review flag and the extraction method. The combined workbook contained 3,345 raw rows (Table 2). Optical character recognition artefacts, non-English fragments, duplicates and unreadable strings were removed, leaving 1,304 clean sentences. Sentence-type labels cover five classes: Obligation (810), Heading (257), Risk (96), Definition (72) and Right (69) (Figure 2). Related-party labels (Contractor, Employer, Shared) apply to the 975 sentences for which a party is meaningful.

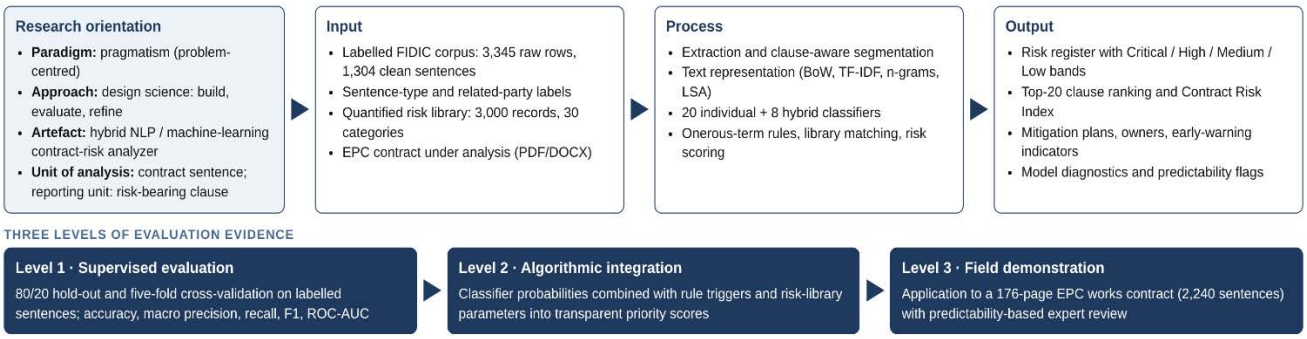


Figure 1: Research Design and Evaluation Levels of the Contract Risk Analyzer

The design combines a pragmatic research orientation with design-science build–evaluate cycles. Evaluation evidence is generated at three complementary levels

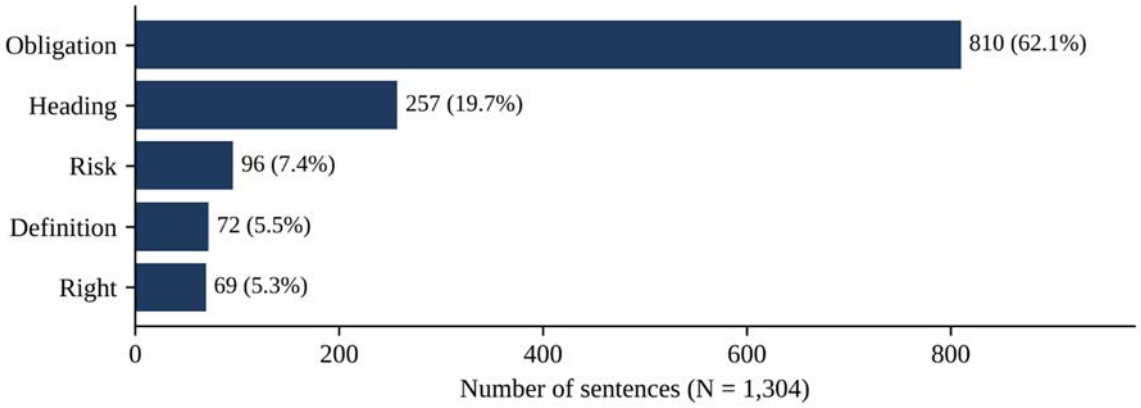


Figure 2: Sentence-Type Distribution in the Clean Modelling Corpus

Obligations account for 62.1% of sentences, so macro-averaged metrics are used to give minority classes such as Risk and Right equal weight

Table 2: Composition of the Labelled FIDIC Workbook

Sheet	Raw rows	Largest raw label classes
Combined (unique)	3,345	Obligation 2,926, Risk 158, Right 131
Red Book	1,392	Obligation 1,147, Risk 101, Right 72
Yellow Book	1,312	Obligation 1,074, Risk 103, Right 72
Orange Book	620	Obligation 575, Right 26, Heading 19
Silver Book	741	Obligation 720, Heading 21
Clean modelling corpus	1,304	Obligation 810, Heading 257, Risk 96, Definition 72, Right 69

After cleaning, 2,041 rows were excluded. Related-party labels apply to 975 of the 1,304 clean sentences

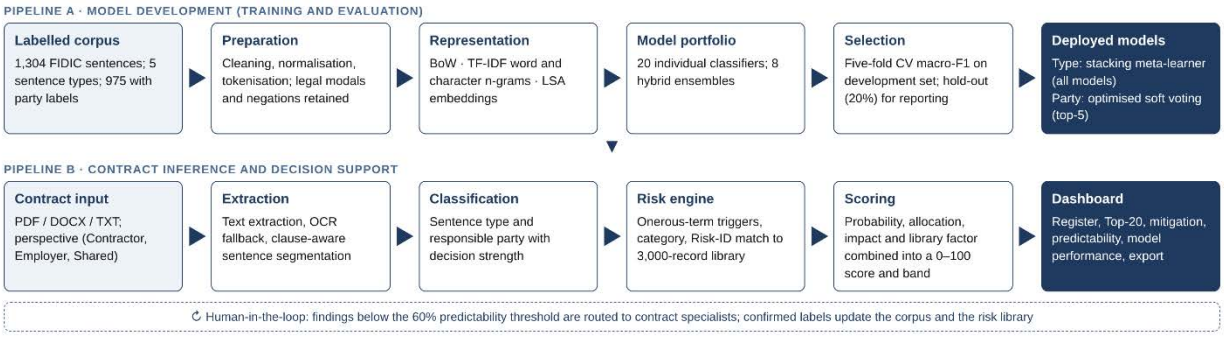


Figure 3: System Architecture and End-to-End Analytical Flow
Pipeline A trains and selects classifiers, Pipeline B applies them to a new contract. Findings below the predictability threshold are routed to specialists, whose confirmed labels update the corpus and risk library

Quantified Risk Library

A structured library of 3,000 energy-project risk records organized into 30 categories enriches each classified clause. Records carry a risk identifier, category, project type, energy segment, region, life-cycle phase, likelihood factor, exposure percentage, risk owner, impact areas, early-warning indicators and mitigation actions. The library turns a classified sentence into a managed risk with an owner and a monitoring plan.

Application Contract

The field demonstration used the general conditions of contract for EPC works issued by a public-sector construction agency in India. The document runs to 176 pages and yields 2,240 sentences after extraction and segmentation. It was analysed from the contractor's perspective, which corresponds to the position of a bidder deciding whether and at what price to tender.

System Architecture and Tools

The analyzer links two pipelines (Figure 2). Pipeline A develops and evaluates classifiers on the labelled corpus. Pipeline B applies the selected classifiers to a new contract, detects onerous wording, matches clauses to the risk library, scores and ranks findings and presents them in an interactive dashboard. The system was implemented in Python as a Streamlit web application, with scikit-learn estimators for text representation and classification and interactive Plotly charts for the dashboard. Six dashboard views structure the output: Risk Dashboard, Model Execution, Predictability, Model Performance, Mitigation and Method. All results can be exported as HTML and CSV reports for audit.

Text Extraction, Preprocessing and Representation

Documents in PDF, DOCX or TXT format are parsed to text, with optical character recognition for image-based pages, and segmented into sentences while page, heading and clause identifiers are retained for traceability (Figure 3). Preprocessing removes headers, footers and numbering artefacts, normalizes case and Unicode, and tokenizes text. Because legal meaning depends on modal and conditional words, terms such as shall, may, not, unless and subject to, as well as numerical references, are preserved rather than treated as stop words.

Five representation families were used: bag-of-words counts, TF-IDF word n-grams [13], character n-gram TF-IDF, which is robust to spelling variants and OCR noise, combined word and character TF-IDF, and latent semantic analysis (LSA) embeddings obtained by truncated singular value decomposition of the TF-IDF matrix. Each representation produces a feature matrix $X \in \mathbb{R}n\times p$ for n sentences and p features, paired with label vectors y for each task. To prevent information leakage, vectorizers and feature transformations were fitted only on training folds.

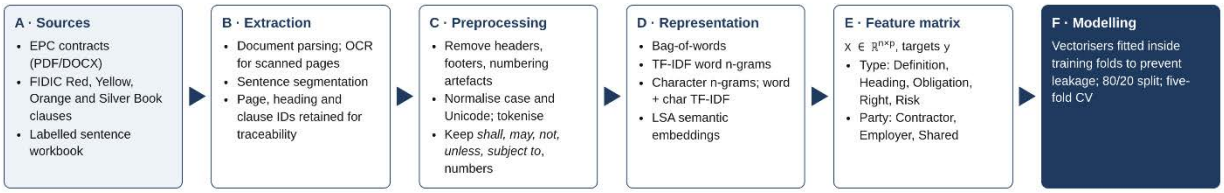


Figure 4: Text Extraction, Preprocessing and Feature-Engineering Pipeline

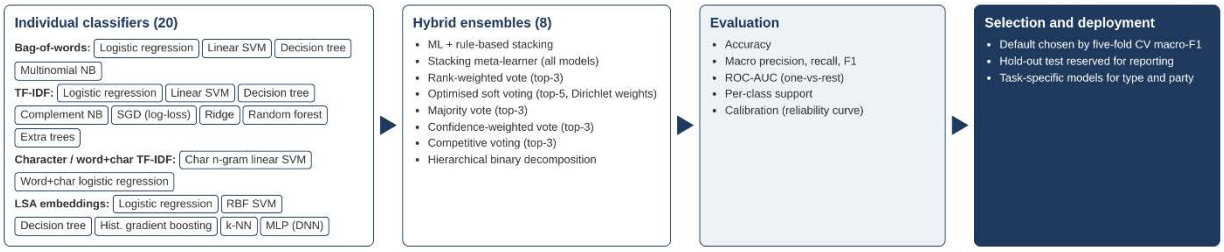


Figure 5: Machine-Learning Framework: Individual Classifiers, Hybrid Ensembles, Evaluation and Deployment

Table 3: Hybrid Ensemble Configurations

Hybrid configuration	Combination mechanism
ML + rule-based stacking	Meta-learner trained on base-model probabilities combined with rule-trigger features from the onerous-language rules
Stacking meta-learner (all models)	Meta-learner trained on out-of-fold class probabilities of all 20 base models (Wolpert, 1992)
Rank-weighted vote (top-3)	The three best models by cross-validated macro-F1 receive weights 3, 2 and 1 in a weighted soft vote
Optimized soft voting (top-5)	Non-negative weights for the five best models chosen by Dirichlet search to maximize out-of-fold macro-F1
Majority vote (top-3)	Each of the three best models casts one vote, ties are broken by summed probability
Confidence-weighted vote (top-3)	Each model adds its confidence to the class it predicts, the class with the largest confidence mass wins
Competitive voting (top-3)	Majority of the top three models if at least two agree, otherwise the prediction of the most accurate model
Hierarchical binary decomposition	The multiclass problem is decomposed into a sequence of binary decisions

Top-k membership is determined by five-fold cross-validated macro-F1 on the development set. Eight configurations were evaluated for sentence type and seven for related party

Classification Models

The individual model portfolio comprises 20 classifiers that combine the five representation families with linear, probabilistic, margin-based, tree-based, instance-based and neural learners (Figure 4): logistic regression, linear and radial-basis support vector machines, ridge and stochastic-gradient classifiers, multinomial and complement naive Bayes, decision trees, random forests, extra trees, histogram gradient boosting, k-nearest neighbours and a multilayer perceptron. Eight hybrid ensembles were then built from these base learners (Table 3). Two classification tasks were modelled separately because the linguistic cues for function and for responsibility differ: sentence type (Definition, Heading, Obligation, Right, Risk) and related party (Contractor, Employer, Shared).

Evaluation Protocol and Metrics

Data were split into an 80% development set and a 20% hold-out test set: 261 sentences for sentence type and 195 for related party (20% of the 975 party-labelled sentences). Five-fold cross-validation on the development set was used for tuning and for selecting the default deployed model, and the hold-out set was reserved for final reporting. For each class k, precision, recall and F1 were computed from true positives (TP), false positives (FP) and false negatives (FN):

$$\text{Precision}_k = TP_k / (TP_k + FP_k), \text{Recall}k = TP_k / (TP_k + FN_k) \quad (1)$$

$$F1_k = 2 \times \text{Precision}_k \times \text{Recall}_k / (\text{Precision}_k + \text{Recall}_k), \text{Macro-F1} = (1/K) \Sigma_k F1_k \quad (2)$$

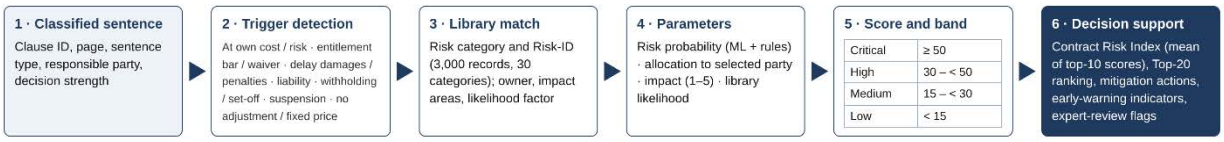


Figure 6: Risk Identification, Scoring and Prioritization Logic

Table 4: Risk-Score Bands and Priority Classes

Risk score (0–100)	Priority class	Recommended action
≥ 50	Critical	Negotiate, price explicitly or qualify before bid submission
30 to < 50	High	Price contingency and assign owner with early-warning indicators
15 to < 30	Medium	Monitor through the project risk register
< 15	Low	Accept, review at stage gates

Thresholds are configurable so that organizations can align them with their own risk appetite

Accuracy is the overall share of correct predictions, and support is the number of true test cases in a class. ROC-AUC was computed one-versus-rest and macro-averaged. Calibration was assessed with a reliability curve comparing binned predicted confidence with observed accuracy (Guo *et al.*, 2017). On the application contract, the agreement of each model m with the selected model was computed as the share of the N sentences on which their predicted labels coincide:

$$\text{Agreement}_m = (1/N)\ \Sigma_i\ 1[\hat{y}_{m,i} = \hat{y}_{sel,i}] \tag{3}$$

Risk Identification, Scoring and Mitigation

Risk identification combines classifier outputs with explicit rules (Figure 5). Rule triggers detect onerous wording such as at own cost or risk, entitlement bars and waivers, delay damages and penalties, liability, withholding and set-off, suspension and fixed-price or no-adjustment terms. Candidate clauses are matched to a category and a Risk-ID in the risk library, filtered by project context. A screening risk probability combines the ML probability with rule evidence. The priority score then reflects this probability together with the share of the risk allocated to the selected party, the impact level (1–5) and the library likelihood factor, and is expressed on a 0–100 scale. Scores are mapped to four configurable priority bands (Table 4). At contract level, the Contract Risk Index (CRI) summarizes the concentration of severe provisions:

$$\text{CRI} = (1/10)\ \Sigma_{j=1}^{10}\ \text{RS}_{(j)} \tag{4}$$

where RS(j) is the j-th highest clause-level risk score. Mitigation guidance is assembled from clause-specific trigger playbooks and the library's controls, owners and early-warning indicators. Where a total contract value is entered, the system also estimates expected monetary value and residual exposure using a configurable mitigation-effectiveness parameter (35% by default).

Predictability Index and Expert Review

To make the reliability of each automated judgement visible, a composite predictability index was computed for every classified sentence:

$$\text{Predictability} = 0.35\ A + 0.25\ C + 0.15\ (1 - H_{norm}) + 0.25\ P_{hist} \tag{5}$$

where A is the agreement of the constituent models with the predicted class, C is the confidence of the selected model, Hnorm = −Σk pk log pk / log K is the normalized entropy of the predicted class distribution, and Phist is the hold-out precision of the selected model for the predicted class. Sentences scoring below a configurable threshold of 60% are routed to contract specialists, who check the full clause, related provisions and the proposed library match. The threshold is an operational decision rule that balances review workload against coverage. Two further indicators are reported per clause: decision strength, the predicted-class probability of the sentence-type model, and party confidence, the corresponding probability of the party model.

RESULTS

Classifier Performance

Table 5 reports hold-out performance for the deployed models and for the best individual and hybrid sentence-type configurations. The deployed stacking meta-learner achieved 89.7% accuracy, 85.1% macro precision, 93.4% macro recall, 88.7% macro-F1 and 98.9% ROC-AUC on 261 test sentences. The deployed related-party model, optimized soft voting over the five best base models, achieved 97.4% accuracy and 98.1% macro precision, recall and F1, with a ROC-AUC of 99.8% on 195 test sentences. The highest hold-out sentence-type score was obtained by ML + rule-based stacking (92.3% accuracy, 89.5% macro-F1), 4.7 macro-F1 points above the best individual model, bag-of-words logistic regression (84.8%). Answering RQ1 and RQ2, contract sentences can be classified with high accuracy, and hybridization adds a material gain for the harder five-class sentence-type task.

Table 6 and Figure 7 extend the comparison to all 28 sentence-type configurations. Seven of the eight hybrids outperformed every individual classifier on macro-F1, only hierarchical binary decomposition (82.6%) fell inside the individual range. Among individual models, linear learners on sparse lexical features dominated: bag-of-words and TF-IDF logistic regression, the ridge classifier and combined word and character TF-IDF all reached 84.4% to 84.8% macro-F1. Models on dense LSA embeddings were more variable, ranging from 84.0% for the RBF support vector machine to 52.8% for the LSA decision tree, which indicates that dimensionality reduction discards phrase-level cues that matter in legal text. Naive Bayes variants and k-nearest neighbours performed worst, consistent with their sensitivity to class imbalance.

Table 5: Hold-Out Classification Performance of Deployed and Best-Performing Models

Task and model	n	Accuracy	Macro p	Macro R	Macro-F1	ROC-AUC
Sentence type: stacking meta-learner (deployed)	261	89.7	85.1	93.4	88.7	98.9
Sentence type: ML + rule-based stacking (best hybrid)	261	92.3	85.9	94.1	89.5	99.3
Sentence type: BoW + logistic regression (best individual)	261	85.4	84.7	86.3	84.8	97.4
Related party: optimized soft voting, top-5 (deployed)	195	97.4	98.1	98.1	98.1	99.8

Values are percentages on the 20% hold-out set. Macro P = macro precision, Macro R = macro recall. The deployed default was selected by five-fold cross-validated macro-F1 on the development set, so it need not have the highest hold-out score

Table 6: Hold-Out Performance of All 28 Sentence-Type Configurations

Rank	Model	Type	Accuracy	Macro P	Macro R	Macro-F1	ROC-AUC
1	ML + rule-based stacking	Hybrid	92.3	85.9	94.1	89.5	99.3
2	Stacking meta-learner (all models)	Hybrid	89.7	85.1	93.4	88.7	98.9
3	Rank-weighted vote (top-3)	Hybrid	87.4	87.8	89.2	87.8	97.8
4	Optimised soft voting (top-5)	Hybrid	87.0	84.6	87.8	85.7	97.6
5	Majority vote (top-3)	Hybrid	85.4	86.4	85.4	85.1	97.9
6	Confidence-weighted vote (top-3)	Hybrid	85.4	86.4	85.4	85.1	97.2
7	Competitive voting (top-3)	Hybrid	85.4	86.4	85.4	85.1	97.8
8	BoW + logistic regression	Individual	85.4	84.7	86.3	84.8	97.4
9	Word+char TF-IDF + logistic regression	Individual	87.7	84.8	85.0	84.7	97.8
10	TF-IDF + ridge classifier	Individual	86.2	85.3	85.0	84.7	96.8
11	TF-IDF + decision tree	Individual	86.2	82.6	86.7	84.4	91.2
12	TF-IDF + logistic regression	Individual	85.1	82.9	86.5	84.4	96.6
13	LSA + RBF SVM	Individual	86.6	86.4	82.0	84.0	94.1
14	Char n-gram + linear SVM	Individual	88.1	88.8	80.6	83.4	98.0
15	BoW + linear SVM	Individual	84.7	88.0	82.1	83.3	97.2
16	TF-IDF + SGD (log-loss)	Individual	85.1	85.2	81.9	83.2	96.0
17	Hierarchical binary decomposition	Hybrid	86.2	86.8	80.5	82.6	98.0
18	BoW + decision tree	Individual	83.1	78.6	85.4	81.5	91.0
19	LSA + logistic regression	Individual	83.9	78.0	86.4	81.2	96.8
20	LSA + MLP (DNN)	Individual	83.5	82.8	78.4	80.3	95.7
21	TF-IDF + linear SVM	Individual	83.9	85.4	78.0	80.3	96.5
22	TF-IDF + random forest	Individual	84.3	88.1	76.1	79.6	97.8
23	TF-IDF + extra trees	Individual	84.3	85.3	75.3	78.4	97.0
24	LSA + hist. gradient boosting	Individual	80.8	80.7	65.7	71.6	95.0
25	BoW + multinomial naive Bayes	Individual	76.6	73.7	70.0	68.7	91.4
26	TF-IDF + complement naive Bayes	Individual	74.3	64.8	66.6	64.8	89.7
27	LSA + k-nearest neighbours	Individual	69.4	69.9	51.6	56.0	83.4
28	LSA + decision tree	Individual	69.7	52.0	54.0	52.8	74.8

n = 261 hold-out sentences, values in percent. BoW = bag-of-words, TF-IDF = term frequency–inverse document frequency, LSA = latent semantic analysis, SVM = support vector machine, S

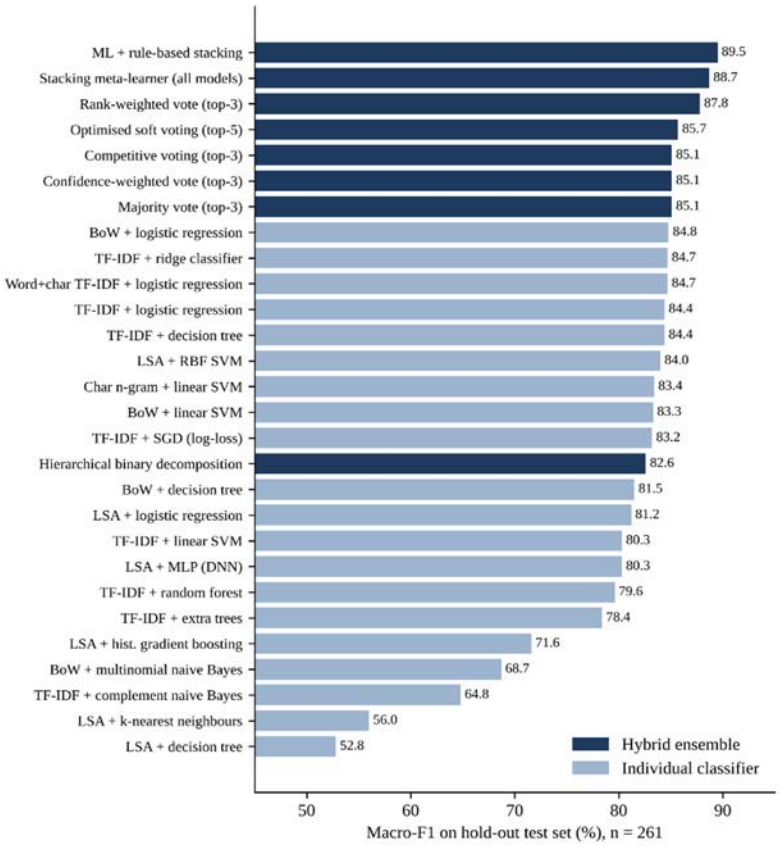


Figure 7: Sentence-Type Models Ranked by Hold-Out Macro-F1
Dark bars denote hybrid ensembles and light bars individual classifiers (n = 261)

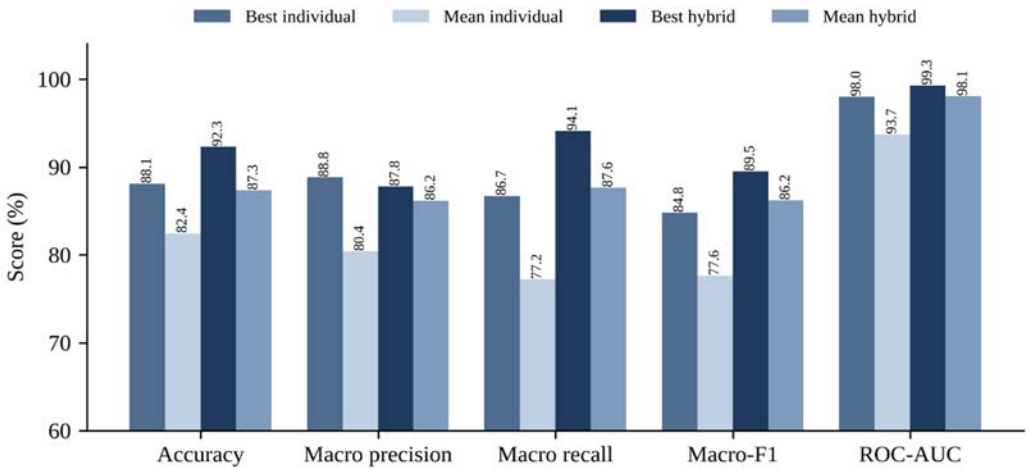


Figure 8: Best and Mean Performance of Individual and Hybrid Models by Metric
Computed from the 20 individual and 8 hybrid sentence-type configurations on the hold-out set (n = 261)

Figure 8 summarizes the gain profile. Relative to the best individual model on each metric, the best hybrid improved accuracy from 88.1% to 92.3%, macro recall from 86.7% to 94.1%, macro-F1 from 84.8% to 89.5% and ROC-AUC from 98.0% to 99.3%, while macro precision was comparable (88.8% vs. 87.8%). The mean hybrid outperformed the mean individual model by 8.6 macro-F1 points (86.2% vs. 77.6%), showing that ensembling also reduces exposure to a poor

Table 7: Class-Level Hold-Out Performance of the Deployed Sentence-Type Model

Class	Precision	Recall	F1	Support
Definition	1.000	1.000	1.000	14
Heading	0.800	0.923	0.857	52
Obligation	0.972	0.870	0.919	162
Right	0.765	0.929	0.839	14
Risk	0.720	0.947	0.818	19
Macro average	0.851	0.934	0.887	261

Model: stacking meta-learner (all models)

Table 8: Comparison of Individual and Hybrid Models by Classification Task

Indicator	Sentence type	Related party
Hold-out observations	261	195
Individual models evaluated	20	20
Hybrid models evaluated	8	7
Best individual model	BoW + logistic regression	BoW + decision tree
Best individual macro-F1	84.8%	98.1%
Best hybrid model	ML + rule-based stacking	ML + rule-based stacking
Best hybrid macro-F1	89.5%	98.5%
Hybrid gain (macro-F1 points)	+4.7	+0.4
Deployed model	Stacking meta-learner (all models)	Optimized soft voting (top-5)

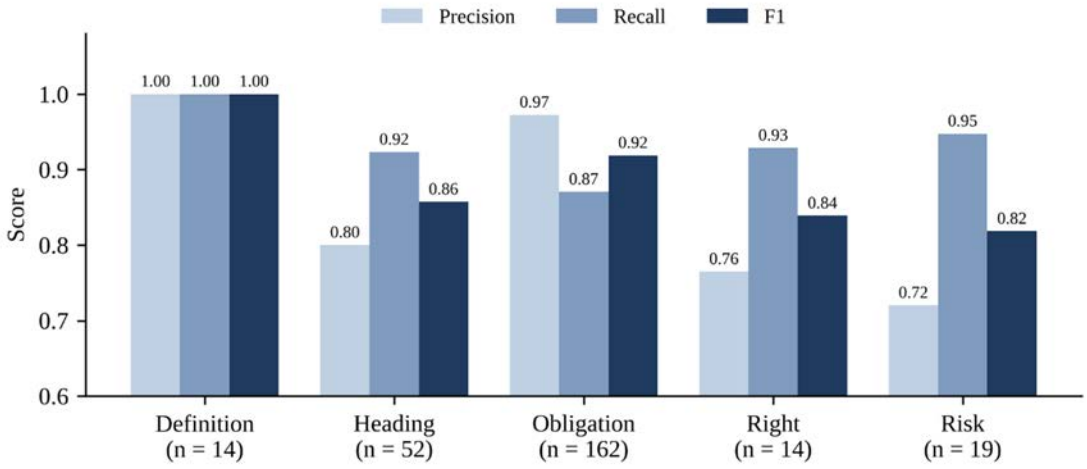


Figure 9: Precision, Recall and F1 by Sentence Type for the Deployed Model

choice of base model. The gain is concentrated in recall, which matters for screening because missing a risk clause is usually costlier than reviewing a false alarm.

Class-level results for the deployed sentence-type model are reported in Table 7 and Figure 9. Every class reached an F1 of at least 0.82. Definitions were classified perfectly, obligations reached an F1 of 0.92, and the Risk class combined very high recall (0.95) with lower precision (0.72). This profile suits a screening tool: nearly all labelled risk sentences are detected, and the cost is a manageable number of false positives that the review layer can filter.

For related-party classification the task was easier and individual models were already strong (Table 8). The best individual model reached 98.1% macro-F1, and the best hybrid improved this by only 0.4 points, to 98.5%. Hybridization therefore adds most value where the label space is larger and classes are more imbalanced.

Model Execution on the Application Contract

All 20 individual and 8 hybrid models classified every one of the 2,240 sentences of the application contract. Table 9 reports, for each model, the number of sentences labelled Risk, Obligation and Right, the mean decision strength, agreement with the selected stacking model and hold-out accuracy. The comparison reveals large differences in how models behave on unseen text. Decision-tree models flagged many more risk sentences (119–137) than most linear models (16–38) and did so with very high decision strength (96.8%–98.5%), whereas random forest and extra trees flagged almost none. Agreement with the selected model was highest for the strongest lexical models, combined word and character logistic regression (88.7%) and optimized soft voting (87.8%), and lowest for naïve Bayes, k-nearest

Table 9: Model Execution Results on the 2,240-Sentence Application Contract

Model	Risk	Oblig.	Right	Decision Strength	Agreement	Hold-Out Acc.
BoW + logistic regression	38	1211	58	89.3	81.2	85.4
BoW + linear SVM	22	1271	56	79.0	82.2	84.7
BoW + decision tree	127	1144	71	96.8	79.1	83.1
TF-IDF + logistic regression	35	1437	58	69.5	81.9	85.1
TF-IDF + linear SVM	26	1522	45	77.0	82.8	83.9
TF-IDF + decision tree	119	1176	72	98.5	81.3	86.2
LSA + logistic regression	195	1429	118	81.4	78.2	83.9
LSA + RBF SVM	82	1612	67	81.1	81.4	86.6
LSA + decision tree	137	1376	152	97.1	66.3	69.7
BoW + multinomial naive Bayes	69	1944	64	87.5	63.2	76.6
TF-IDF + complement naive Bayes	105	1636	62	38.1	61.1	74.3
TF-IDF + SGD (log-loss)	36	1505	50	76.5	81.1	85.1
TF-IDF + ridge classifier	33	1559	51	75.5	82.6	86.2
Char n-gram + linear SVM	16	1600	27	83.0	86.3	88.1
Word+char TF-IDF + logistic regression	52	1514	65	77.7	88.7	87.7
TF-IDF + random forest	0	1319	37	78.8	80.2	84.3
TF-IDF + extra trees	2	1404	28	78.6	82.9	84.3
LSA + hist. gradient boosting	32	1689	55	96.2	80.1	80.8
LSA + k-nearest neighbours	39	1375	18	73.3	61.3	69.3
LSA + MLP (DNN)	65	1657	71	95.4	77.0	83.5
Majority vote (top-3)	28	1256	63	93.4	82.6	85.4
Competitive voting (top-3)	28	1256	63	92.3	82.6	85.4
Confidence-weighted vote (top-3)	28	1257	63	94.0	82.7	85.4
Rank-weighted vote (top-3)	31	1273	66	81.5	83.8	87.4
Optimised soft voting (top-5)	33	1398	60	75.7	87.8	87.0

Decision strength is the mean probability of each model's predicted class (%), agreement is the share of sentences on which the model agrees with the selected stacking meta-learner (%), hold-out accuracy is from the labelled test set (%). Counts for Heading and Definition are omitted. The three stacking and decomposition hybrids are not listed separately; the selected stacking model is the reference for agreement

Table 10: Contract-Level Risk Indicators for the Application Contract

Indicator	Value
Pages / sentences analysed	176 / 2,240
Perspective	Contractor
Risk-bearing findings registered	371
Risk categories represented	24 of 30
Critical / High / Medium / Low	102 / 165 / 104 / 0
Share in Critical or High bands	72.0%
Contract Risk Index (mean of top-10 scores)	79.2%
Mean screening risk probability	64.8%
Allocated to Contractor / Shared / Employer	265 / 101 / 5
Mean predictability of registered findings	≈ 71%
Findings below 60% predictability (expert review)	113 (30.5%)

neighbours and the LSA decision tree (61.1%–66.3%). High decision strength therefore does not imply agreement or accuracy, which motivates the composite predictability index used for review routing.

Contract-Level Risk Profile

Applied to the 176-page contract from the contractor's perspective, the analyzer registered 371 risk-bearing findings across 24 of the 30 library categories (Table 10, Figures 10 and 11). Of these, 102 were rated Critical and 165 High, so 72.0% of findings fell in the two highest bands, and 104 were Medium. No finding was rated Low, which reflects the screening logic: clauses are registered only when classifier evidence and rule triggers jointly indicate risk. The Contract Risk Index was 79.2%, well above the Critical threshold of 50, and the mean screening risk probability was 64.8%. Combined time and cost effects were the dominant impact areas.

Responsibility allocation is strongly one-sided (Figure 12). The party model assigned 265 findings (71.4%) to the contractor, 101 (27.2%) to shared responsibility and only 5 (1.3%) to the employer. For a bidder, this concentration is commercially decisive: it indicates that most risk-bearing provisions would have to be priced, qualified or negotiated rather than passed back to the client. The probability–impact cross-tabulation (Figure 13) shows that 284 findings (76.5%) carry a Major impact level of 4, and that 160 findings (43.1%) combine a probability above 60% with Major or Severe impact. The modal cell, 104 findings, combines Major impact with a probability of 40%–60%.

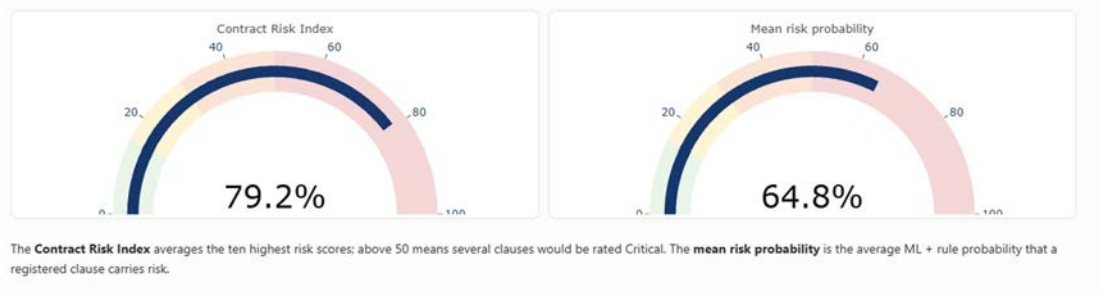


Figure 10: Risk Dashboard: Contract Risk Index and Mean Risk Probability
Screenshot from the Contract Risk Analyzer dashboard for the application contract

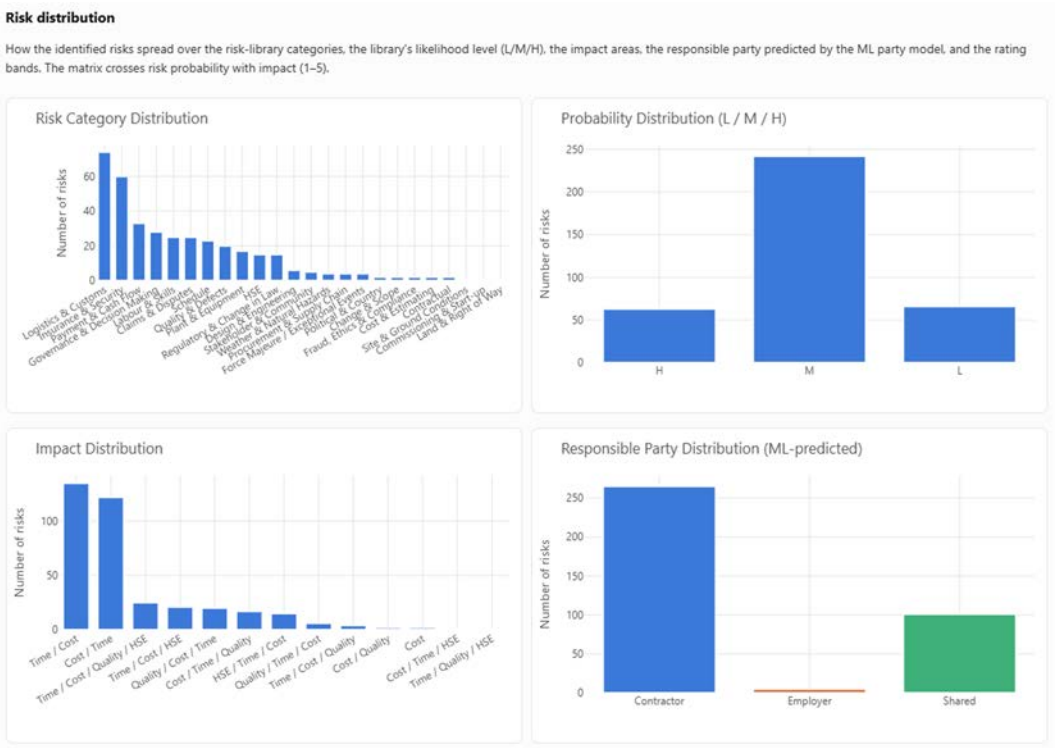


Figure 11: Risk Dashboard: Distribution by Category, Library Likelihood, Impact Area and Predicted Responsible Party
Screenshot from the Contract Risk Analyzer. Responsible party is predicted by the deployed related-party ensemble

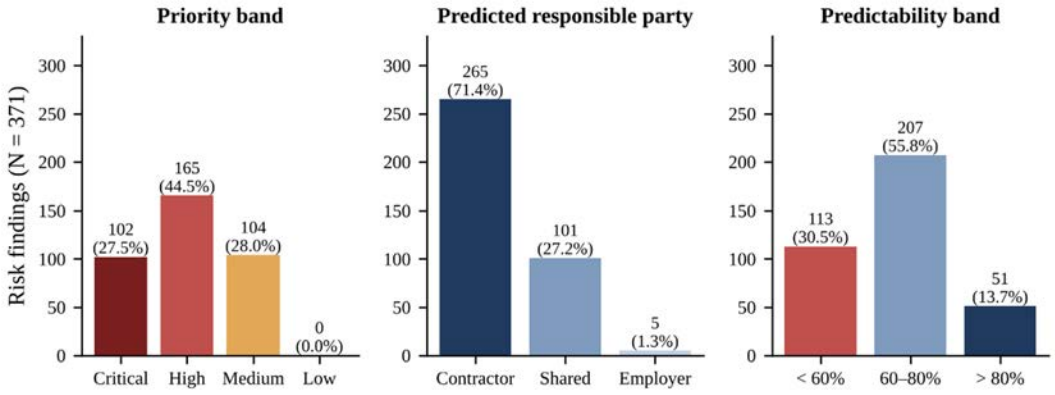


Figure 12: Composition of the Risk Register by Priority Band, Responsible Party and Predictability
N = 371 registered findings

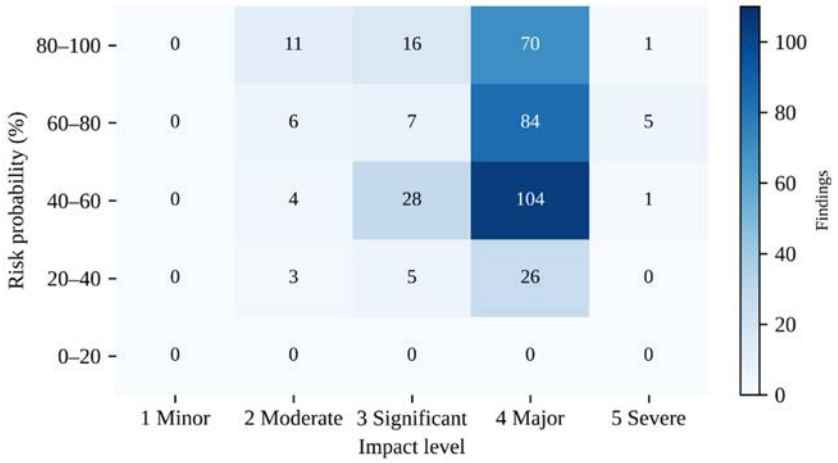


Figure 13: Probability × Impact Matrix of Registered Findings
Cell values are counts of findings (N = 371). Row totals from 20–40% to 80–100%: 34, 137, 102 and 98. Column totals for impact levels 2 to 5: 24, 56, 284 and 7

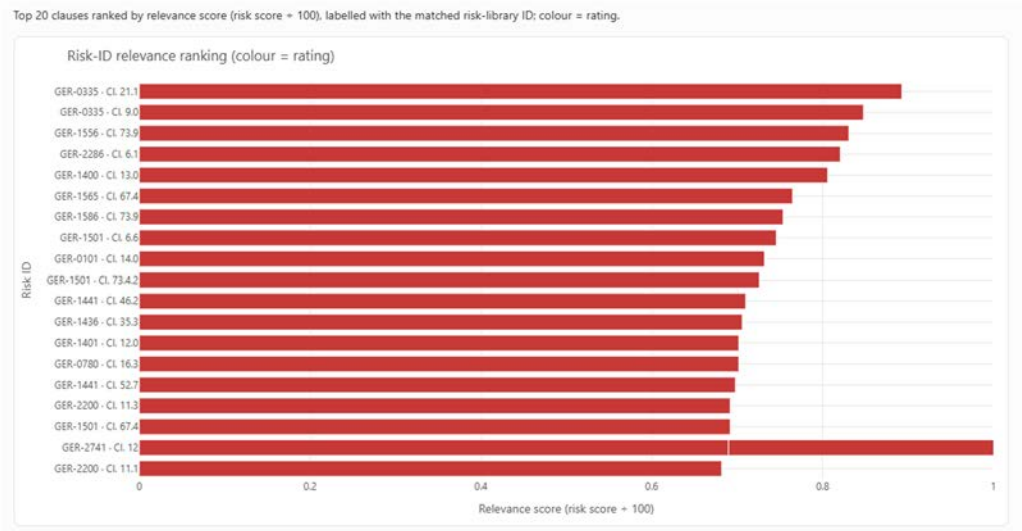


Figure 14: Risk-ID Relevance Ranking of the Top-20 Clauses
Screenshot from the analyzer. Relevance = risk score ÷ 100, bar colour denotes the priority band

Table 11: Ten Highest-Ranked Risk Findings in the Application Contract

Clause	Category	Risk-ID	Score	Prob.	Impact	Alloc.	Pred.	Triggers
21.1	Design and engineering	GER-0335	89.3	92	5	96	48	At own cost, entitlement bar / waiver
9.0	Design and engineering	GER-0335	84.8	99	4	95	75	Liability
73.9	Labour and skills	GER-1556	83.1	99	4	93	68	Delay damages / penalties, liability
6.1	Change and scope	GER-2286	82.1	97	4	93	75	Classifier and library match
13.0	–	GER-1400	80.6	–	–	–	–	–
67.4	Labour and skills	GER-1565	76	95	4	86	63	Withholding / set-off, liability
73.9	Labour and skills	GER-1586	75	87	4	97	61	Classifier and library match
6.6	Labour and skills	GER-1501	75	87	4	96	56	Withholding / set-off
14.0	Schedule	GER-0101	73	89	4	90	70	Entitlement bar, suspension, no adjustment
73.4.2	Labour and skills	GER-1501	73	84	4	97	77	Delay damages, withholding / set-off

Score = risk score (0–100), Prob. = screening risk probability (%), Impact = impact level (1–5), Alloc. = share allocated to the contractor (%), Pred. = predictability (%). Detailed review-card values for rank 5 were not exported. Findings 1 and 8 fall below the 60% predictability threshold and were routed to expert review

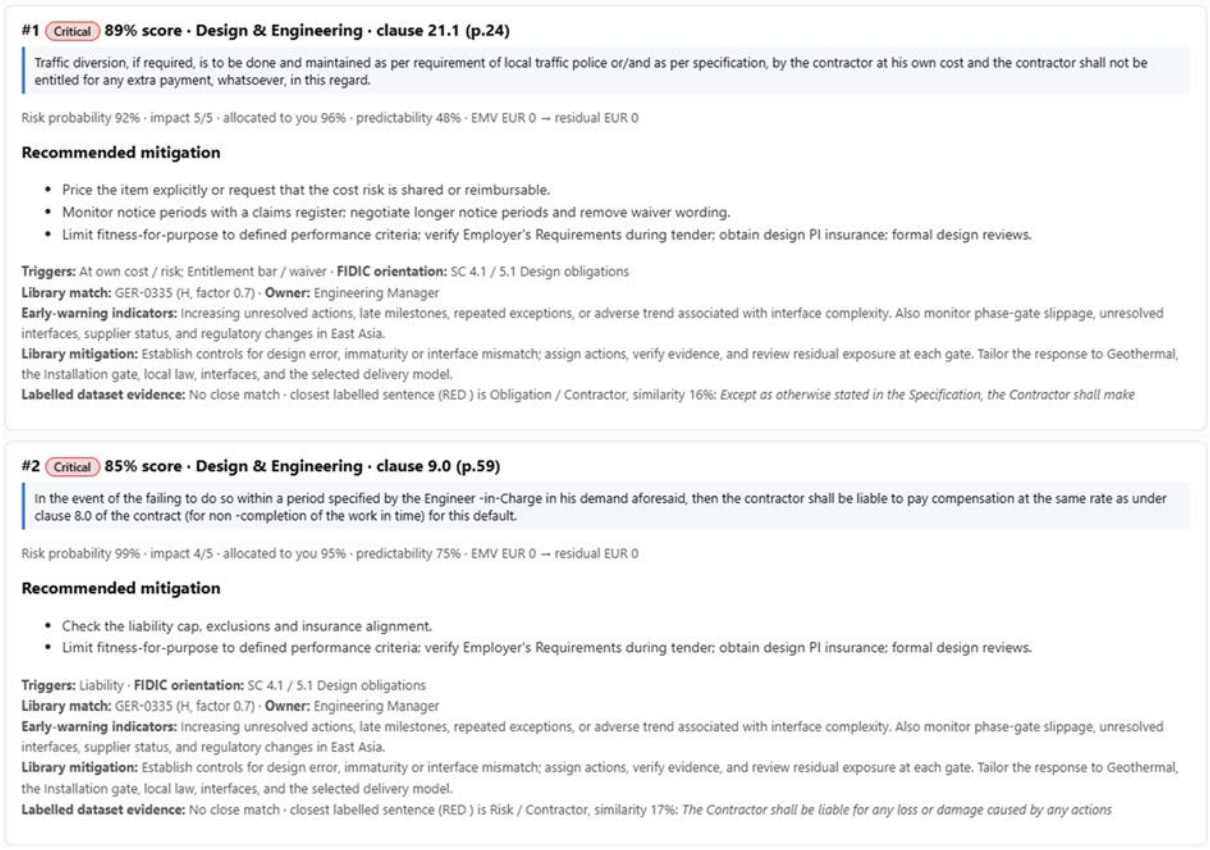


Figure 15: Clause-Level Review Cards with Recommended Mitigation (Ranks 1 and 2)
Each card shows the source sentence, page, category, risk probability, impact, allocation, predictability, triggers, FIDIC orientation, risk-library match and owner, early-warning indicators and the closest labelled training sentence. Monetary values are populated once a total contract value is entered. Further cards are reproduced in Appendix A

Table 12: Predictability Indicators

Indicator	All analysed sentence records	Registered risk findings
Records	1,593	371
Mean predictability	82%	≈ 71%
Mean model agreement	80%	–
Below 60% threshold (expert review)	165 (10.4%)	113 (30.5%)
Selected-model hold-out accuracy	89.7%	89.7%

Ranked Clauses and Mitigation

The Top-20 view converts each score to a 0–1 relevance scale and labels it with the matched Risk-ID and clause (Figure 14). Table 11 details the ten highest-ranked findings. The top finding, clause 21.1, obliges the contractor to carry out and maintain traffic diversion at its own cost and excludes any entitlement to extra payment. It combines a 92% risk probability, the maximum impact level and 96% allocation to the contractor, yielding a score of 89.3. The analyzer recommends pricing the item explicitly or requesting cost sharing or reimbursement, maintaining a claims register for notice periods and removing waiver wording, and it assigns the risk to the engineering manager with early-warning indicators such as unresolved interface actions and milestone slippage (Figure 15). Other high-ranked provisions concern compensation for delay at the delay-damages rate (clause 9.0), statutory penalties of up to 5% of the estimated cost (clause 73.9), time extensions limited to deviations (clause 6.1), final and binding deductions decided by the engineer (clause 67.4) and the exclusion of compensation for suspension (clause 14.0).

Predictability, Calibration and Expert Review

The predictability module evaluated 1,593 sentence records (Table 12, Figure 16). Mean predictability was 82% and mean model agreement 80%, and 165 records (10.4%) fell below the 60% threshold. Among the 371 registered risk

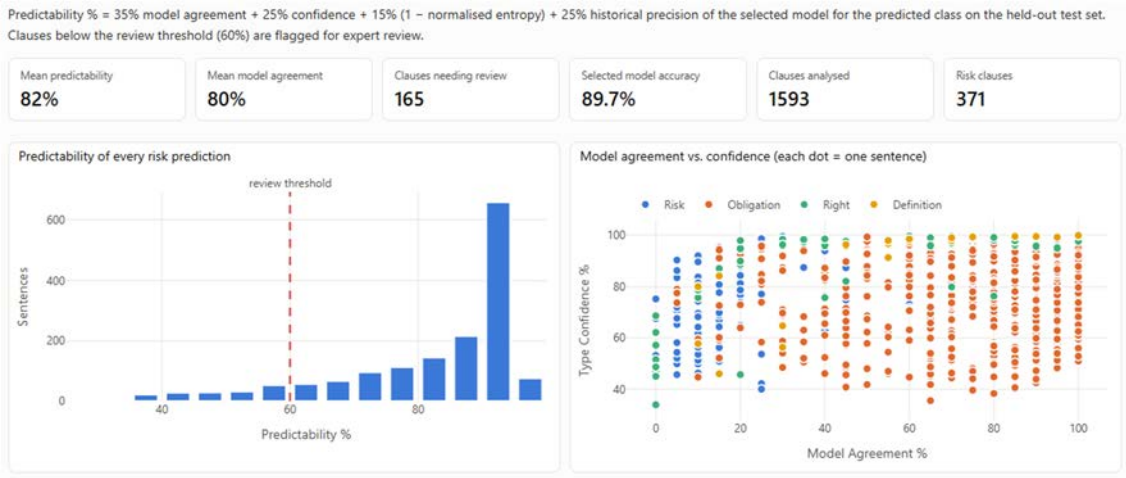


Figure 16: Predictability Diagnostics: Distribution with Review Threshold and Model Agreement Versus Confidence
Screenshot from the analyzer's Predictability view. The dashed line marks the 60% review threshold, each dot in the scatter plot is one sentence, coloured by predicted type

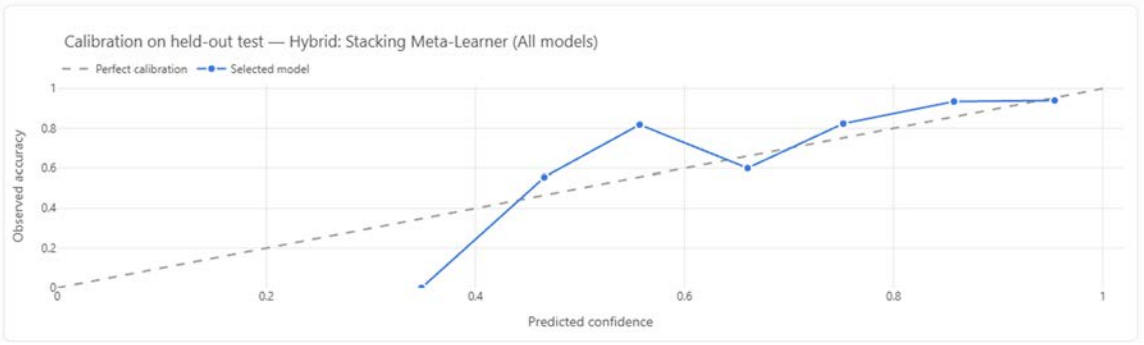


Figure 17: Calibration of the Deployed Stacking Meta-Learner on the Hold-Out Test Set
Screenshot from the analyzer. The dashed line represents perfect calibration

findings, mean predictability was lower, at approximately 71%: 113 findings (30.5%) fell below the threshold, 207 (55.8%) scored 60%–80% and 51 (13.7%) scored above 80%. Risk-bearing clauses are thus classified with less certainty than ordinary sentences, and the review layer concentrates specialist attention exactly where automated evidence is weakest. The top-ranked finding illustrates the point: clause 21.1 has the highest risk score but a predictability of only 48%, so it is flagged for confirmation before action. The scatter of model agreement against classification confidence (Figure 16) shows that the two indicators carry distinct information, which justifies combining them in a composite index.

The reliability curve of the deployed stacking model on the hold-out set (Figure 17) tracks the diagonal closely for predicted confidence above 0.7, where most predictions lie. Below 0.6 the curve is irregular because few test sentences fall in these bins. This pattern supports using model confidence as one input to review routing, but not as the sole criterion, which is consistent with the composite design of the predictability index.

DISCUSSION

This article set out to show how contract text can be converted into decision-ready risk intelligence for energy EPC tendering. The results answer the four research questions. Contract sentences can be classified by type and responsible party with high accuracy (RQ1). Hybrid ensembles yield a material gain on the harder sentence-type task, adding 4.7 macro-F1 points at the top and 8.6 points on average, and a smaller gain where individual models are already near ceiling (RQ2). Combining classifier probabilities with onerous-term rules and a quantified risk library produces a transparent, ranked register in which every finding carries a reason, an owner, mitigation actions and early-warning indicators (RQ3). A composite predictability index makes uncertainty visible and directs expert attention to the 30.5% of findings that most need it (RQ4).

Table 13: Validity Considerations and Planned Extensions

Validity Dimension	Current Support	Planned Extension
Construct	FIDIC-derived sentence labels, explicit rule categories	Expert-adjudicated labels for 300–500 clauses
Internal	Five-fold cross-validation, locked hold-out set, leakage control in folds	Evaluation held out by book and by contract
External	Application to a real 176-page EPC contract	Additional FIDIC books, contract forms and energy segments
Outcome	Risk prediction as a screening measure of project success	Linkage to later claims, delays and disputes
Reliability	Predictability index, calibration and agreement diagnostics	Inter-rater agreement between experts and system
Reproducibility	Versioned code, library, metrics, split indices and exported reports	Public release of an anonymized benchmark

Theoretical and Methodological Implications

The findings extend the text-as-data agenda in marketing [1-2] from consumer-generated text to the governance documents of B2B exchange. Just as the language of loan applications reveals default risk [3], the language of a contract reveals how risk will be distributed between buyer and seller before a project starts. Methodologically, the broad model comparison shows that, for short and formulaic legal sentences, sparse lexical representations with linear learners form a strong baseline and that dense LSA embeddings lose discriminative phrase information. The largest gains come from combining heterogeneous learners with domain rules. The finding that decision strength, agreement and accuracy diverge on unseen text cautions against equating a model's confidence with its reliability, and it supports reporting class-level, calibration and agreement diagnostics alongside headline accuracy. More broadly, the design shows how marketing researchers can build text classifiers that pair machine predictions with explicit rules and human validation and that report the reliability of each judgement.

Managerial Implications

For energy-sector EPC contractors, the analyzer supports the commercial decisions that determine value capture in project markets. In bid/no-bid screening, the Contract Risk Index and the share of Critical findings provide a fast, comparable signal of how onerous a tender is. In pricing, the register identifies which provisions require explicit contingency or qualification, for example uncompensated traffic diversion or uncapped statutory penalties. In negotiation, the ranked list and the allocation profile, here 71.4% of findings allocated to the contractor and only 1.3% to the employer, form an evidence-based agenda for clarification requests and proposed amendments. At contract award, the register with owners and early-warning indicators supports a structured handover from the sales and bid team to project delivery and forms an audit-ready record. For clients and project sponsors, the same view reveals where a tender transfers risk in ways that may raise bid prices or reduce competition, which informs procurement strategy. The predictability flags allow legal and commercial specialists to focus their limited review time where automated evidence is weakest.

Limitations and Future Research

Several boundaries qualify the findings and point to future research (Table 13). The labels were derived from FIDIC reference books, and evaluation used a sentence-level split, document-disjoint testing on additional contracts with expert-annotated clauses would strengthen generalization claims. The field demonstration covers one contract, so comparative analysis across contract forms, regions and energy segments is a natural next step. Risk scores are screening priorities rather than predictions of realized claims, delays or losses, linking them longitudinally to claims and dispute outcomes would establish predictive validity at the project level. Finally, transformer-based language models could be added to the portfolio and benchmarked against the hybrid design, with attention to cost, transparency and calibration.

CONCLUSION

This article designed, implemented and evaluated Contract Risk Analyzer, a hybrid NLP, machine-learning and rule-based system that turns energy EPC contract text into a prioritized, explainable and reviewable risk register. Across 28 model configurations, hybrid ensembles delivered the strongest classification performance, with 88.7% macro-F1 for the deployed sentence-type model and 98.1% for the related-party model. Applied to a 176-page contract, the system registered 371 risk-bearing findings, identified a Contract Risk Index of 79.2%, revealed a strongly contractor-weighted risk allocation and routed uncertain findings to expert review. By connecting text analytics to the commercial decisions of bidding, pricing and negotiation, the research shows how contract documents can become a source of market intelligence and offers marketing researchers a transparent template for hybrid text classification.

REFERENCES

1. Berger, J. *et al.* "Uniting the Tribes: Using Text for Marketing Insight." Journal of Marketing, vol. 84, no. 1, 2020, pp. 1–25. https://doi.org/10.1177/0022242919873106.
2. Humphreys, A. and R. J.-H. Wang. "Automated Text Analysis for Consumer Research." Journal of Consumer Research, vol. 44, no. 6, 2018, pp. 1274–1306. https://doi.org/10.1093/jcr/ucx104.
3. Netzer, O. *et al.* "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications." Journal of Marketing Research, vol. 56, no. 6, 2019, pp. 960–980. https://doi.org/10.1177/0022243719852959.
4. Choi, S. J. *et al.* "AI and Text-Mining Applications for Analyzing Contractor's Risk in Invitation to Bid (ITB) and Contracts for Engineering Procurement and Construction (EPC) Projects." Energies, vol. 14, no. 15, 2021, article 4632. https://doi.org/10.3390/en14154632.
5. Lee, J. and J.-S. Yi. "Predicting Project's Uncertainty Risk in the Bidding Process by Integrating Unstructured Text Data and Structured Numerical Data Using Text Mining." Applied Sciences, vol. 7, no. 11, 2017, article 1141. https://doi.org/10.3390/app7111141.
6. Son, B.-Y. and E.-B. Lee. "Using Text Mining to Estimate Schedule Delay Risk of 13 Offshore Oil and Gas EPC Case Studies during the Bidding Process." Energies, vol. 12, no. 10, 2019, article 1956. https://doi.org/10.3390/en12101956.
7. Dikmen, I. *et al.* "Automated Construction Contract Analysis for Risk and Responsibility Assessment Using Natural Language Processing and Machine Learning." Computers in Industry, vol. 166, 2025, article 104251. https://doi.org/10.1016/j.compind.2025.104251.
8. Moon, S. *et al.* "Automated Detection of Contractual Risk Clauses from Construction Specifications Using Bidirectional Encoder Representations from Transformers (BERT)." Automation in Construction, vol. 142, 2022, article 104465. https://doi.org/10.1016/j.autcon.2022.104465.
9. Wong, S. *et al.* "Construction Contract Risk Identification Based on Knowledge-Augmented Language Models." Computers in Industry, vols. 157–158, 2024, article 104082. https://doi.org/10.1016/j.compind.2024.104082.
10. Guo, C. *et al.* "On Calibration of Modern Neural Networks." Proceedings of the 34th International Conference on Machine Learning, vol. 70, 2017, pp. 1321–1330. https://proceedings.mlr.press/v70/guo17a.html.
11. Ribeiro, M. T. *et al.* "'Why Should I Trust You?': Explaining the Predictions of Any Classifier." Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, 2016, pp. 97–101. https://doi.org/10.18653/v1/N16-3020.
12. Hevner, A. R. *et al.* "Design Science in Information Systems Research." MIS Quarterly, vol. 28, no. 1, 2004, pp. 75–105. https://doi.org/10.2307/25148625.
13. Salton, G. and C. Buckley. "Term-Weighting Approaches in Automatic Text Retrieval." Information Processing & Management, vol. 24, no. 5, 1988, pp. 513–523. https://doi.org/10.1016/0306-4573(88)90021-0.
14. Choi, S.-W. and E.-B. Lee. "Contractor's Risk Analysis of Engineering Procurement and Construction (EPC) Contracts Using Ontological Semantic Model and Bi-Long Short-Term Memory (LSTM) Technology." Sustainability, vol. 14, no. 11, 2022, article 6938. https://doi.org/10.3390/su14116938.
15. Park, M.-J. *et al.* "A Digitalized Design Risk Analysis Tool with Machine-Learning Algorithm for EPC Contractor's Technical Specifications Assessment on Bidding." Energies, vol. 14, no. 18, 2021, article 5901. https://doi.org/10.3390/en14185901.
16. Jafari, P. *et al.* "Automated Extraction and Time-Cost Prediction of Contractual Reporting Requirements in Construction Using Natural Language Processing and Simulation." Applied Sciences, vol. 11, no. 13, 2021, article 6188. https://doi.org/10.3390/app11136188.
17. Pham, H. T. T. L. and S. Han. "Natural Language Processing with Multitask Classification for Semantic Prediction of Risk-Handling Actions in Construction Contracts." Journal of Computing in Civil Engineering, vol. 37, no. 6, 2023. https://doi.org/10.1061/JCCEE5.CPENG-5218.
18. Kim, J. *et al.* "Inherent Risks Identification in a Contract Document through Automated Rule Generation." Automation in Construction, vol. 172, 2025, article 106044. https://doi.org/10.1016/j.autcon.2025.106044.
19. Fu, Y. *et al.* "Control, Coordination and Adaptation Functions in Construction Contracts: A Machine-Coding Model." Automation in Construction, vol. 152, 2023, article 104890. https://doi.org/10.1016/j.autcon.2023.104890.
20. Hendrycks, D. *et al.* "CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review." Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, vol. 1, 2021.
21. Graham, S. G. *et al.* "Natural Language Processing for Legal Document Review: Categorising Deontic Modalities in Contracts." Artificial Intelligence and Law, vol. 33, no. 1, 2025, pp. 79–. https://doi.org/10.1007/s10506-023-09379-2.
22. Owolabi, H. A. *et al.* "Predicting Completion Risk in PPP Projects Using Big Data Analytics." IEEE Transactions on Engineering Management, vol. 67, no. 2, 2020, pp. 430–453. https://doi.org/10.1109/TEM.2018.2876321.
23. Wolpert, D. H. "Stacked Generalization." Neural Networks, vol. 5, no. 2, 1992, pp. 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1.
24. Breiman, L. "Random Forests." Machine Learning, vol. 45, 2001, pp. 5–32. https://doi.org/10.1023/A:1010933404324.
25. Chen, T. and C. Guestrin. "XGBoost: A Scalable Tree Boosting System." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785.
26. Ke, G. *et al.* "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." Advances in Neural Information Processing Systems, vol. 30, 2017.

27. Prokhorenkova, L. *et al.* "CatBoost: Unbiased Boosting with Categorical Features." Advances in Neural Information Processing Systems, vol. 31, 2018.
28. Saito, T. and M. Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets." PLOS ONE, vol. 10, no. 3, 2015, article e0118432. https://doi.org/10.1371/journal.pone.0118432.
29. Medvedeva, M. *et al.* "Rethinking the Field of Automatic Prediction of Court Decisions." Artificial Intelligence and Law, vol. 31, 2023, pp. 195–212. https://doi.org/10.1007/s10506-021-09306-3.