


Article

AI and Text-Mining Applications for Analyzing Contractor's Risk in Invitation to Bid (ITB) and Contracts for Engineering Procurement and Construction (EPC) Projects

Su Jin Choi ¹, So Won Choi ¹, Jong Hyun Kim ² and Eul-Bum Lee ^{1,3,*} 

¹ Graduate Institute of Ferrous & Energy Materials Technology, Pohang University of Science and Technology (POSTECH), Pohang 37673, Korea; sujinchoil@postech.ac.kr (S.J.C.); smilesowon@postech.ac.kr (S.W.C.)

² WISEiTECH, Seoul 13486, Korea; jonghyun@wise.co.kr

³ Department of Industrial and Management Engineering, Pohang University of Science and Technology (POSTECH), Pohang 37673, Korea

* Correspondence: dreblee@postech.ac.kr; Tel.: +82-054-279-0136

Abstract: Contractors responsible for the whole execution of engineering, procurement, and construction (EPC) projects are exposed to multiple risks due to various unbalanced contracting methods such as lump-sum turn-key and low-bid selection. Although systematic risk management approaches are required to prevent unexpected damage to the EPC contractors in practice, there were no comprehensive digital toolboxes for identifying and managing risk provisions for ITB and contract documents. This study describes two core modules, Critical Risk Check (CRC) and Term Frequency Analysis (TFA), developed as a digital EPC contract risk analysis tool for contractors, using artificial intelligence and text-mining techniques. The CRC module automatically extracts risk-involved clauses in the EPC ITB and contracts by the phrase-matcher technique. A machine learning model was built in the TFA module for contractual risk extraction by using the named-entity recognition (NER) method. The risk-involved clauses collected for model development were converted into a database in JavaScript Object Notation (JSON) format, and the final results were saved in pickle format through the digital modules. In addition, optimization and reliability validation of these modules were performed through Proof of Concept (PoC) as a case study, and the modules were further developed to a cloud-service platform for application. The pilot test results showed that risk clause extraction accuracy rates with the CRC module and the TFA module were about 92% and 88%, respectively, whereas the risk clause extraction accuracy rates manually by the engineers were about 70% and 86%, respectively. The time required for ITB analysis was significantly shorter with the digital modules than by the engineers.

Keywords: artificial intelligence; invitation-to-bid (ITB) document; engineering-procurement-construction (EPC); information retrieval; machine learning; named-entity recognition (NER); phrase-matcher; natural language processing (NLP); Python; spaCy; text mining



Citation: Choi, S.J.; Choi, S.W.; Kim, J.H.; Lee, E.-B. 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* **2021**, *14*, 4632. <https://doi.org/10.3390/en14154632>

Academic Editor: Audrius Banaitis

Received: 21 June 2021

Accepted: 28 July 2021

Published: 30 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 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 (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Engineering, procurement and construction (EPC) projects are carried out for long periods by various participating entities such as project owners, vendors, contractors, or sub-contractors. They face many risks at each stage from bidding to maintenance [1–3]. Particularly, there are cases in which the uncertain project risks cannot be predicted in advance due to the contractor's lack of project experience and limited schedule in the project bidding and planning stages [4,5]. For example, in an offshore plant project based on a letter of intent to bid, significant losses occurred because the cost of equipment, the scale of workforce required, or the project risk was incorrectly predicted [6]. Insufficient risk prediction often leads to a significant loss to the project and the contractor's project

damage increases in proportion to the size of the project [7]. In particular, a project owner who orders a project sometimes tends to transfer project risk to the contractors by utilizing the characteristic of the unique EPC contract method (Lump-sum turn-key) in which the contractor is responsible for project execution [7,8]. In addition, for project risk analysis, feedback should be made within a limited time, such as during the bidding step. It is important how accurately the level of risk can be reviewed in EPC analysis, but it is difficult to solve in the field because the review period in the bidding process is set tight. These difficulties depend on the experience of experienced EPC experts. If the company's structural problems cause a shortage of skilled EPC experts, the accuracy of EPC analysis will be greatly affected. Accordingly, the need for a systematic project risk management system for the entire project cycle, including the bidding stage, was constantly emerging for EPC Contractors [9].

Therefore, in this study, the research team developed an algorithmic model that automatically extracts and analyzes contract risks by acquiring EPC contract Invitation to Bid (ITB) documents from contractors who have performed multiple EPC projects. At first, for algorithm development, the preceding research was divided into two steps. The research trend on risk analysis of the EPC project was examined, and the current development status of the natural language processing (NLP) technology for analyzing unstructured text data and areas of improvement were examined.

Based on previous research, two modules were developed for automatic contract risk analysis: The Critical Risk Check (CRC) module and the Terms Frequency Analysis (TFA) module. First, the CRC module supports user decision-making by automatically searching for project risk clauses based on machine learning (ML) algorithms and presenting the results. An algorithm that extracts the original text in the ITB documents and generates the result in the data frame format (comma-separated values: CSV) was applied. The TFA module learns EPC risk entities using named-entity recognition (NER) technology. Various similar phrases, which differ according to the project characteristics and regions, are extracted so that the risk can be evaluated according to the frequency of appearance. The NER model was customized according to the characteristics of the EPC project, and a system integrator (SI) was built so that the risk frequency results can be visualized and presented to users. In addition, a group of experts and engineers with experience in EPC projects participated from the beginning of this study to enhance the validation of development modules. In particular, the Subject-Matter Expert (SME) group built a framework for the development direction and performed validation work for the pilot test result for the modules.

Through the above algorithmic modules (CRC and TFA), the research team proposes a theoretical foundation and system supporting project risk analysis and decision-making by automatically extracting and evaluating risks for EPC projects.

Section 2 provides a literature review, and Section 3 gives an overview of the entire Engineering Machine learning Automation Platform (EMAP) system. Section 4 briefly describes each study step. In Sections 5 and 6, detailed research contents of CRC and TFA algorithm development are explained, respectively. The system integration for establishing a decision support system, which is the goal of the project receiving funding, is described in Section 7, and Section 8 shows the results of the verification of the system implemented as a platform. Sections 9 and 10 summarize the conclusions and future work, respectively.

2. Literature Review

In order to carry out this study, the following studies and precedent cases were investigated and reviewed. Similar research characteristics and trends, limitations of existing research projects, etc., were analyzed and bench-marked in this study through previous studies. Prior research was performed step by step by dividing into two categories as follows. First, the EPC project risk analysis cases are described, followed by document pre-processing for unstructured data and NLP technology application cases.

2.1. EPC Project Risk Studies

In this study, the researchers initially focused on EPC risk definitions and databases to build a decision-making system by using the rules made in text mining techniques. This section summarized the literature review focused on identifying project risk through text mining.

Yu and Wang [10] tried to systematize the quantitative risk analysis by evaluating the contractor risk (11 types) of the EPC project using the interpretive structural modeling (ISM) method. Shen et al. [1] analyzed the contractor's risk and the causes of the claims in the EPC projects through an EPC contractor case study in China and tried to verify that the EPC project risk is a direct influence factor of the claim by modeling structural equations. However, this study has a limitation of a small sample size for case studies. Kim et al. [11] developed a model (Detail Engineering Completion Rating Index System: DECRIS) to support the optimization of the construction schedule while minimizing the rework of EPC contractors during the offshore plant EPC project. They performed the project schedule, cost performance calculation, and validation for thirteen completed offshore projects in the DECRIS model. Mohebbi and Bislimi [8] presented a methodology for project risk management by studying parameters affecting EPC project risk and the general risk management model through an oil and gas plant project in Iran. Ullah et al. [12] established a theoretical framework by analyzing the causes of schedule delay and cost overrun of an EPC project. Gunduz and Maki [13] listed 39 attributes that negatively influenced the cost overrun of a construction project through a review of the project documentation. After that, these attributes were ranked through the Frequency-Cost Adjusted Importance Index (FCAII) technique.

As global demand for liquefied natural gas (LNG) rapidly increased as an energy source, Animah and Shafiee [14] reviewed risks of plant facilities such as floating production storage and offloading (FPSO) units, floating storage and regasification units (FSRU), LNG ships, and terminals. Son and Lee [15] applied text-mining technology to study the construction period prediction, key risk components of offshore EPC projects. Critical terms (CTs) were selected using the R program for the client's technical specification (scope of work) document, and the Schedule Delay Risk Index (SDRI) of the project was prepared. The schedule delay was predicted through regression analysis. Son and Lee's study can be used for the technical documents of the project bidding stage, but the conditions of the EPC main contract, where the project contract conditions are specified, were not included in the study subject. Chua et al. introduced a neural network technique to find key project management factors for successful budget performances of construction projects. Their study was an early case of applying a neural network model to a construction project and utilizing construction knowledge for solving schedule- and cost-related problems in construction sites [16]. Ho et al. proposed a Building Information Modeling-based Knowledge Sharing Management (BIMKSM) system that shares construction knowledge using Building Information Modeling (BIM) technology [17]. Sackey and Kim developed ESCONPROCS (expert system for construction procurement selection) to support decision-making on contract and procurement system selection in the construction industry based on extensive literature review [18]. Lee et al. developed an AI-based ITB risk management model to analyze ITB documents at the bidding stage of EPC projects [19]. Their model was implemented in the IBM's Watson Explorer AI architecture environment, and a testbed was also built for the pilot study [20]. It was different from the research of our research team in that commercialized Application Programming Interface (API) was purchased and used for research.

2.2. Pre-Processing for Unstructured Data and Application of NLP Technology

In EPC projects, most documents, such as ITB, contract documents, and technical specifications, are composed of unstructured documents (e.g., numbers, units, text formats). The current status and trend of natural language processing (NLP) technology were reviewed in the field of pre-processing and analysis of unstructured data. NLP is an artificial

intelligence (AI) technique that enables computers and humans to interact efficiently by providing a computer language and listener repertoire [21]. NLP is often used in the fields of information extraction (IE) and information retrieval (IR). IE refers to the process of extracting information that matches certain patterns by defining the type or pattern of information extracted in advance. IR is the task of finding the necessary information in a large repository of information, for example, a searching engine that searches for and ranks information according to user queries. NER for tagging people's names or geographical names is one type of IR [22].

In 2016, Tixier et al. [23] proposed a framework to extract meaningful empirical data from a vast database of digital injury reports on construction site accidents. A model was built to analyze the text of the construction injury report using NLP. They also introduced methods of overcoming the decoding of unstructured reports and building a system through the repetition of coding and testing. Tixier et al. [23] used structured data as a prerequisite when applying statistical modeling techniques. Lim and Kim [22] sorted documents according to the contents of documents and used a text mining technique, which is also referred to as text analysis or document mining, for extracting meaningful information from documents. They classified major research fields of construction automation by keywords and summarized the most frequently cited studies in construction automation. However, due to the collection of papers in the National Digital Science Library, there was a limitation that some papers might be omitted when analyzing the number of citations. In 2014 Williams and Gong [24] proposed a model that combines numerical and text data of construction projects using data mining and classification algorithms and predicted the possibility of cost overrun when bidding on a project. They used a stacking ensemble model in predicting construction cost overruns.

Mrzouk and Enaba [25] analyzed the text of a construction project contract using text mining. By integrating project correspondence within BIM and monitoring project performance, they developed the Dynamic Text Analytics for Contract and Correspondence model, a text analysis model based on project correspondence. Their study showed a methodology for extracting meaningful patterns from construction project documents. Zoua et al. [26] introduced a search methodology through two NLP techniques (vector space model and semantic query expansion) for the effective extraction of risk cases from the construction accident case database. Li et al. [27] proposed NER's unsupervised learning method using annotated texts in an encyclopedia. They stated that the approach to learning the NER model without manually labeled data shows better performance than the model fully trained with news source data.

After reviewing the cases and papers as described above, a lack of studies was revealed on systems that automatically extract and analyze the project risk from the letter of intent (LOI) and contract for the EPC project using ML. Therefore, this study focused on developing automatic risk extraction and analysis modules from the LOIs and contract documents provided by project owners for a machine learning-based decision-making system for EPC projects.

3. Overview of Engineering Decision-Support System

In this study, a machine learning-based integrated decision-support system was developed to provide EPC contractors the maximized efficiency on risk analysis of EPC contracts.

The engineering decision-support system, Engineering Machine learning Automation Platform (EMAP), consists of three modules: (1) ITB Analysis, (2) Engineering Design, and (3) Predictive Maintenance, as shown in Figure 1.

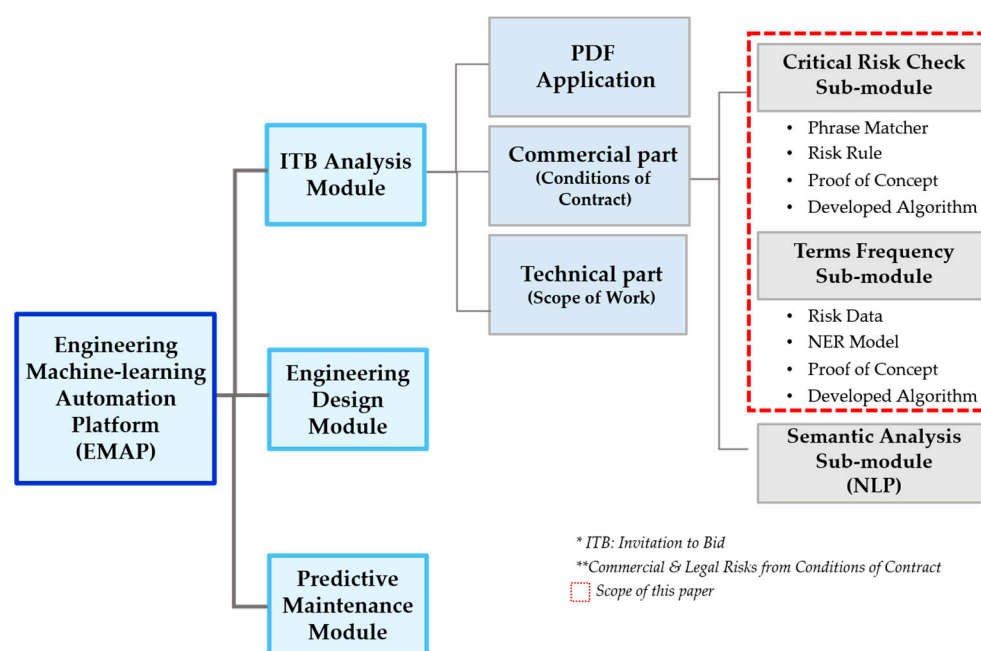


Figure 1. Overview of Engineering Machine learning Automation Platform (EMAP).

The Engineering Design sub-modules are Cost Estimate, Error Check, and Cost Change and Control. The cost estimate sub-module developed a design cost prediction model for estimating design time and cost by selecting variables and standardizing data through correlation analysis. The cost change and control check sub-module developed a prediction model that applied the scaling technique. The Cost Change and control check sub-module developed a prediction model that performed regression analysis by making the information of the existing project the variables.

Predictive maintenance is a module that develops a model for predicting maintenance and parts demand for equipment. Three types of analysis were performed: turbo fan engine, gearbox, and wastewater pump. The predictive maintenance module aims to provide a reference model after collecting data about the equipment. The prediction results for each reference model are visualized and implemented as a chart and presented in an integrated solution.

Considering the module characteristics and analysis types, various ML algorithm techniques, such as phrase/context-matching and random forest, were implemented in each analysis model. First, the ITB Analysis module analyzes ITB documents and identifies project risks in advance through machine learning-based automatic extraction of risk components and key parameters for ITB and contract documents. Second, the Engineering Design module selects variables through correlation analysis and standardizes information, so that it predicts project design time and cost estimates. Third, the Predictive Maintenance module predicts values of equipment such as turbofan engine, gearbox, and wastewater pump. This paper focused on the description of two sub-modules, Critical Risk Check (CRC) and Term Frequency Analysis (TFA) in the ITB Analysis module, and describes the detailed processes implementing NER and phrase matcher techniques in a syntactic analysis.

4. Analysis Framework for ITB Module

The framework of the ITB Analysis module consists of the following five steps as shown in Figure 2. The detailed description of each step is described in the following sections:

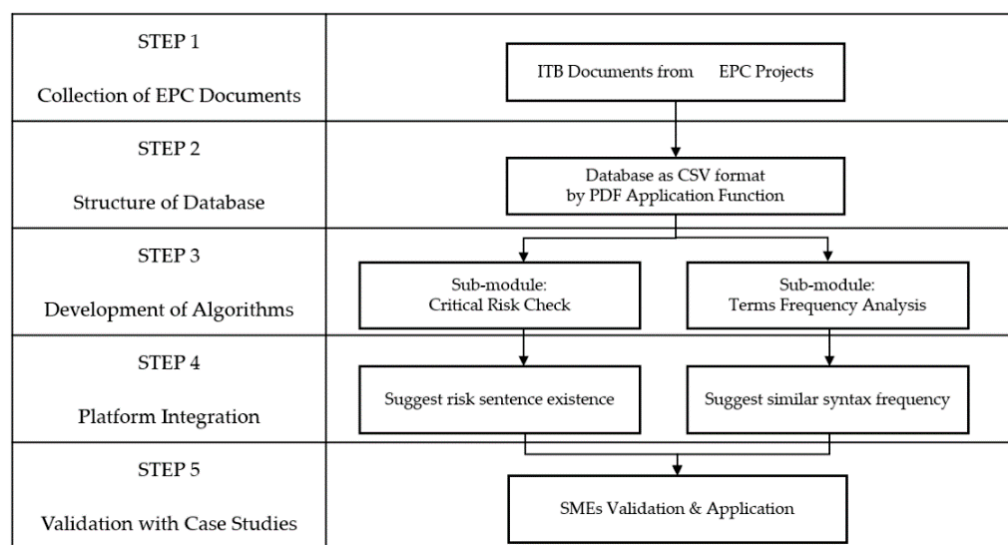


Figure 2. Analysis framework of the ITB Module (CRC and TFA sub-modules).

4.1. STEP 1: Collection of EPC Documents Subsection

The 25 EPC ITB documents were collected through the EPC contractors for various EPC fields such as offshore, onshore plants, and infrastructure projects for this study. In the 25 projects, the EPC ITB information was converted into 21,683 rows in a database (DB).

In the first step, a DB was created through a pre-processing step so that it could be used for risk analysis by converting complex sentences into short sentences. Table 1, below, shows the list of the collected EPC ITB documents. The project titles were anonymized in consideration of the relationship with the project owners.

4.2. STEP 2: Structure of Database

Pre-processing data is a major step to improve the accuracy of the selected data by applying data-mining techniques. The result of pre-processing has a substantial influence on the analysis quality of the ITB documents. A Portable Document Format (PDF)-importing function was built to collect ITB documents as a sub-module in the ITB Analysis module. This function recognizes the text on PDF documents, creates a CSV file containing the texts tagged with codes that indicate the text positions, and tracks the table of contents from the PDF documents.

Tagging the texts during the standardization stage enabled the module to find the sentences' affiliation in the contract during the analysis result. In this study, the ITB documents were converted into a DB and used to develop an automatic risk extraction algorithm. Table 2 shows an example of the exception of the EPC ITB document, and the following pre-processing steps were performed for accurate text recognition with Python programming [28]:

- First, recognize texts in page order on the ITB documents.
- Second, extract texts as sub-class each line and convert the text data into a CSV format.
- Third, add position codes to the line-based text contents in a CSV file.
- Fourth, convert capital letters to lower-case letters, tokenize words by classes, and eliminate stop words and white spaces.

Proper interpretation by experts was important because of the complex sentence structure (subject-verb-object: SVO), which is a common feature of project contract documents. Therefore, text-mining and NLP analysis techniques were applied for reducing errors and improving the machine's ability to understand the context of the sentence. In the text analysis of the EPC ITB documents, errors were reduced, and meaningful analysis results were achieved by collaborating with SMEs and reviewing the systematic processes.

Table 1. List of the collected EPC contracts in the database.

Project Category	No.	Project Name (Anonymized)	Project Type	Location
Offshore	1	P-1	FLNG ¹	Brazil
	2	C-1	Fixed Platform	Angola
	3	S-1	FPSO ²	Nigeria
	4	P-2	Drillship	for chartering
	5	E-1	Semi-submersible	Gulf of Mexico
	6	S-2	Fixed Platform	North Sea (Norway)
	7	I-1	FPSO	Australia
	8	C-2	Booster Compression Facilities	Gulf of Thailand
	9	T-1	TLP ³	Congo
	10	T-2	FPSO	Angola
	11	T-3	FPSO	Nigeria
	12	T-4	FPSO	Angola
Onshore	13	S-3	Petrochemical	Saudi Arabia
	14	A-1	Combined Cycle Power Plant	Kuwait
	15	A-2	Refinery	Kuwait
	16	C-3	LNG Terminal	USA
	17	P-3	Refinery	Peru
	18	O-1	Oil Collecting Station	India
	19	A-3	Coal-fired Power Plant	Chile
Infrastructure	20	W-1	Tunnel	Mexico
	21	P-4	Highway	Australia
	22	M-1	Tunnel	Australia
	23	R-1	Rail	Australia
	24	V-1	Tunnel	Australia
	25	N-1	Rail	Australia

¹ FLNG (floating liquified Natural gas): a floating production storage and offloading facility that conducts LNG operations from offshore gas wells. ² FPSO (floating production, storage and offloading): a ship-shaped floating production facility that process, store, and offload the hydrocarbon production from offshore oil and gas wells. ³ TLP (tension leg platforms): A floating production system typically that are buoyant production facilities vertically moored to the seafloor by tendons for offshore oil/gas wells.

As a general pre-processing technology for text recognition, various computing technologies such as stop words removal, stemming, and the term frequency/inverse document frequency (TF/IDF) algorithms are used [29]. In the text pre-processing step, unnecessary or obstructive texts, such as stop-word, white-space, and delimiters, were deleted or corrected, and the smallest unit phrases were identified. This sentence simplification processes were necessary due to the sentence complexity in the EPC ITB documents.

Collectively changing the capital letters on the ITB documents to the lowercase letters improved the accuracy rate in rule tagging.

As shown in Table 2, the sentences in the EPC ITB document were divided into titles, subheadings, and sentences. The title and subtitle were automatically recognized and processed as separate lines in Python programming. Each sentence was recognized and

extracted based on the delimiters, such as (a), (b) or (i), (ii). Table 2 shows the example of the position codes and the simple sentences after the pre-processing step.

Table 2. An example of the text contents before and after pre-processing (excerption).

The Original Text Contents before Pre-Processing		The Processed Text Contents after Pre-Processing		
No.	Content	No.	Position Code	Content
1	36 LIQUIDATED DAMAGES	1	36	LIQUIDATED DAMAGES
	36.1 Liquidated Damages for Late Completion (a) If contractor fails to complete the relevant part of the work by the relevant Completion Date then contractor will pay liquidated damages to Company in accordance with Exhibit B.	2	36.1	Liquidated Damages for Late Completion
	(b) Subject to Company's rights and remedies provided for under Article XX, sub- Articles YY to YY and Article ZZ, the payment of liquidated damages under sub-Article YY are the sole and exclusive financial remedy of Company in respect of contractor's failure to complete the relevant part of the work by the Completion Date.	3	(a)	If contractor fails to complete the relevant part of the work by the relevant Completion Date then contractor will pay liquidated damages to Company in accordance with Exhibit B.
		4	(b)	Subject to Company's rights and remedies provided for under Article XX, sub- Articles YY to YY and Article ZZ, the payment of liquidated damages under sub-Article YY are the sole and exclusive financial remedy of Company in respect of contractor's failure to complete the relevant part of the work by the Completion Date.

Recognizing and classifying each sentence were intended to improve the recognition rate of risk keywords through the pre-processing step. Tagging risk keywords with complex sentences was not properly performed and showed low-performance results due to errors and omissions. It was important to define the breakpoints of sentences in the pre-processing step. Sentences are often separated by separators or periods, but special symbols such as (:) and (;) are also shown as a sentence separator. The various cases of sentence breakpoints were collected in the ITB documents collected for this study, and the case rule for sentence breakpoints was developed to resolve irregularities of separating sentences. The parser technique was applied in Python programming to analyze the sentence structures and define the relationship among phrases in complex sentences. The parser technique allows checking the grammar of a series of strings and building meaningful tokens with parse trees for further utilization [30].

Through the pre-processing step described above, 21,683 lines of text contents were generated in the DB from the ITB documents. The PDF application function analyzes each page of the PDF file imported and stores information in a metadata format. Once a PDF file is uploaded, a header section, a footer section, and watermarks are eliminated as shown in Figure 3, and the text contents are stored in a CSV file.

Table 3 shows an example of the table generated in the DB from the PDF application function in the pre-processing step. The table consists of four columns: project title, class, sub-class, and content. Class and sub-class are the title and the sub-title in the contents, respectively. Each sub-class is associated with one sentence, which is the smallest unit in the document structure.

In this study, automatic extraction and CSV (comma-separated values) file format were achieved so that the ITB documents can be broken up sentence by sentence by using the parser technique. A CSV file is an application document that uses plain text and is a file that is widely used. Rather than using the existing commercialized technology, the sub-module was developed with specialized customizations in EPC analysis. In addition, unstructured data, such as numbers, units, and text formats of scans and hard copies, were digitized by recognizing texts using the optical character recognition (OCR) technique. OCR is a technique that converts characters into simple text that can be scanned or analyzed by a machine [31]. It is a highly usable technology field that can generate information read

by a machine as an output file. The digitized files are stored in the cloud using the marina DB service [32], and the cloud service was built from the viewpoint of user convenience in the field.

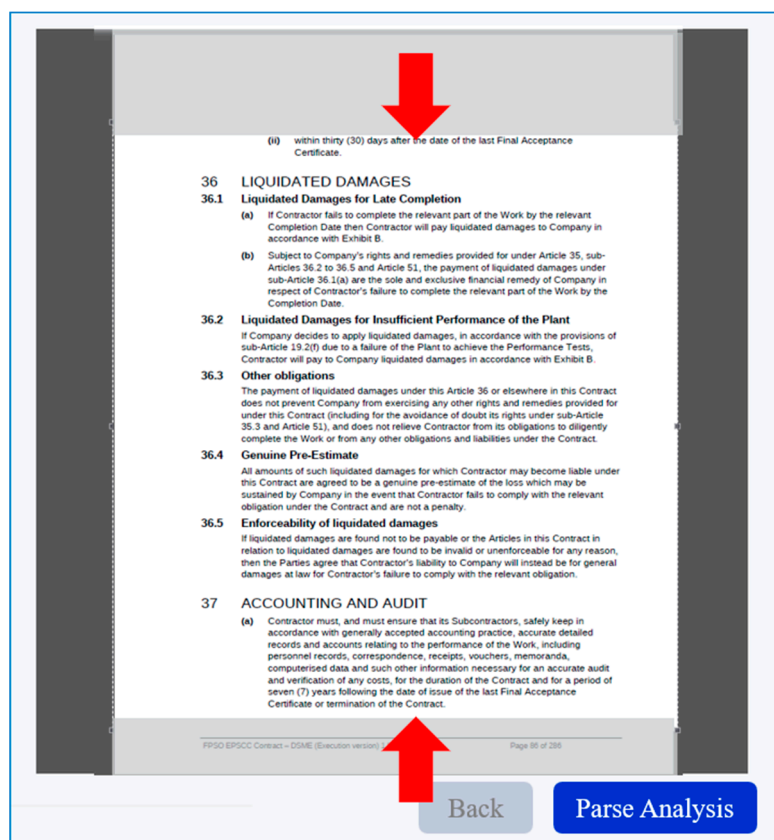


Figure 3. Screenshot of PDF application function (Example of header/footer elimination).

Table 3. An example of the table generated in the database.

Project	Class	Sub-Class	Content
Cheniere_ Sabine_ Pass_ LNG	2. RELATIONSHIP OF OWNER, CONTRACTOR AND SUBCONTRACTORS	2.1 Status of Contractor.	The relationship of Contractor to Owner shall be that of an independent contractor.

4.3. STEP 3: Development of Algorithm

This section briefly describes the CRC module and the TFA module. The CRC module uses a phrase-matcher, which is a rule-based technique for identifying risk clauses. The TFA module includes a machine learning algorithm for automatically extracting frequency in large amounts of data (so called 'big data'). The detailed development processes for the CRC module and the TFA module are described in Sections 5 and 6, respectively. Several advanced techniques, including text-mining, AI, machine learning, and information retrieval (IR), are applied in a complex interconnection to achieve the best performance of the application.

Table 4 summarizes the technique, input data types, program language, and target result for both the CRC and the TFA modules. The AI technique was applied to the CRC module and the IR module was applied to the TFA module. AI is a system and infrastructure technology in which machines implement human capabilities (learning, reasoning, perception) as computer programs [19]. IR is a technology for searching or tagging information, and there is NER as a sub-technology [33].

Table 4. Summary of algorithm.

	Critical Risk Check	Terms Frequency Analysis
Key Library	spaCy's PhraseMatcher [34]	spaCy's en_core_web_sm [35]
Technology	Artificial Intelligence (AI)	Information Retrieval (IR)
Input Data	EPC contract	EPC contract
Language Program	Python 3.7	Python 3.7
Target Result	This module aims to analyze the existence of EPC project risk using rule-based technique.	This module aims to tag similar syntax through a training model suitable for the EPC project.

The CRC module recognizes and standardizes the rules and automatically detects the events matching with the rules. The CRC module is useful to determine if a risk clause exists in the entire ITB document. The TFA module extracts the frequency of the risk clauses and similar syntaxes and identifies the risk that could not be found in the rule-based CRC modules. The TFA module presents the frequency of risk clauses and similar syntaxes to the user with visual outputs (charts and images). Both modules use the same file, but the output is a standardized CSV file for the CRC module. The TFA module provides a variety of image files, web format HTML, and table format CSV files. According to the purpose of the algorithm and technology, ITB Analysis was composed of three sub-analysis modules with several ML technologies for different purposes. It was divided into CRC, TFA, and NLP sub-modules with a syntactic approach, and each submodule provides algorithm development and solutions suitable for the analysis purpose.

4.3.1. Critical Risk Check (CRC) Module

The Critical Risk Check (CRC) module detects risks according to specific rules using PhraseMatcher [34]. PhraseMatcher is a function of spaCy [36], a Python open-source library that extracts the terms related to user-specified rules. First, pre-processing was performed to separate the sentences of ITB using spaCy's part-of-speech (POS) tagging and dependency parsing techniques prior to the main analysis. After pre-processing, ITB extracts risk clauses through CRC rules. For example, the rules were developed in this study to extract liquidated damages (LD), a key contract clause in ITBs. These rules were developed by analyzing the ITBs previously collected. The 35 CRC rules were configured from those EPC contracts. Among the list for CRC, Level 1 indicates major risk items in the contract, and Level 2 indicates the keywords' associated risks, as Table 6 in Section 5.1 below shows 'Liquidated Damages (LD)' and the related keywords in the contract risk list for CRC. If a user wants to find a Level 1 risk clause called 'Liquidated Damages', the LD-related sentences are extracted when a sentence containing one of the keywords listed in Level 2 appears. When a user inputs the ITB to be analyzed into the CRC module, the sentence containing the keyword for the LD is extracted and presented through the CRC rule. If the relevant risk clause does not exist, 'No detected message' is printed and displayed on the module interface. The detailed development process is described in Section 5.

4.3.2. Terms Frequency (TFA) Module (M2)

The Terms Frequency Analysis (TFA) module, the second sub-module in the ITB analysis module, consists of an analysis module using spaCy's NER model. The NER model learns the entity label from the collected sentence data and analyzes whether there is a keyword belonging to the entity label. It utilizes a statistical prediction technique pre-learned from the new data using the learned model [37]. The research team developed an algorithm that tags the position using the NER model, calculates the frequency, and generates a graph. For the TF analysis, the critical risk words in the EPC contract were first designated as NER level. After collecting the sentences with the corresponding level from the contract, it was written as a JavaScript Object Notation (JSON) file, which was used as training data for NER for learning entities. When a user enters the contract to be

analyzed, the location of the learned entity and its frequency are analyzed, and the result is displayed as an HyperText Markup Language (HTML) file. A total of 21,683 sentences were collected from the contracts for the TF module in this study. This study implemented the TF algorithm using NER using Python. The detailed explanation is described in Section 6.

4.4. STEP 4: Integrated Platform

A decision support system for EPC contract risk analysis was developed in this study. The purpose of the integrated platform was to enable users to analyze EPC contract risk on the web-based tool effectively. The cloud service of the integrated platform stores the data, analyzes it, and visualizes the results on the user's computer screen.

After the risk clauses detected from the 25 ITB documents were identified and stored in the marina Database Management System (DBMS) [32], they are linked with the modules during the analysis. The EPC contract analysis decision support system is described in detail in Section 7.

4.5. STEP 5: Case Analysis of the EPC Contracts

In STEP 5, the concept of risk in the EPC industry and the built lexicon were compared with the EPC risk database. The completed risk database and lexicon were used as a comparison database (DB) in which the algorithm model was built. A team of experts, with experience in bidding, contracting and performing multiple EPC projects, participated from the beginning of the development to plan and review the algorithm models during the framework stage. A case analysis of contracts obtained from the EPC contractor was conducted, and the rules were established through the concept of EPC risk and lexicon construction. The risk keywords commonly used in the EPC industry were grouped and established as a lexicon. A lexicon is also called a wordbook in linguistics, and a computing lexicon is a group of words created in a programming language (Python). Machine learning-based lexicons were developed and utilized in previous studies [5,35]. Lee, et al. analyzed FIDIC (International Federation of Consulting Engineers), an international standard contract, and defined problematic clauses that should not be missed and essential clauses that should not be omitted in the review process for overseas construction projects. They were named PCI (problems caused by included information) and PCNI (problems caused by not included information) to reflect the characteristics of risk determination [5]. Kang, et al. extracted design information from the piping and instrumentation diagram document of the EPC industry and converted it into a database [38].

In the Risk Keyword Lexicons, built through collaboration with the EPC experts, the following are several examples:

- Liquidated damages (LD) are considered the most serious risk in EPC projects. LD can be classified into delay liquidated damages (DLD) or performance liquidated damages (PLD). In both cases, if the EPC contractor fails to meet the contract delivery date or does not meet the performance required by the contract, it is a contract clause that compensates the project owner for losses.
- PLD also has the risk of incurring losses of up to five percent of the total contract amount. If the contract delivery date is delayed due to reasons beyond the responsibility of the EPC contractor or the contract performance has not been met, the contract is deferred from the client, and the DLD or PLD will be judged based on the extended contract delivery date or revised contract performance. Therefore, from the standpoint of the contractor, the biggest risk of DLD and PLD is not only whether the requirements included in the ITB document for DLD and PLD are appropriate, but also how reasonable the client is to delay delivery or change performance beyond the responsibility of the EPC contractor.
- There are cases in which delays due to a project owner's faults (design changes, various inspections/approvals, delays in decision-making, etc.) during the construction execution of the EPC contractor, or other reasons for various delivery extensions during construction, are not sufficiently recognized.

The research team conducted a study on the intersection rule through a comprehensive rule and a case study to selectively extract specific risks. In addition, in order to provide the application system integrator (SI) service to EPC contractors, the result of this research carried out, as a part of the project, also performs SI validation in addition to the validation stage. At the time of PoC validation of the ITB analysis module, it was conducted for DLD and PLD-related risks. The validation result is described in detail in Section 8.

5. ITB Analysis Module: Critical Risk Check (CRC) Sub-Module

The CRC module was developed to automatically detect the risk clauses, applying a machine learning algorithm. The process flow of the CRC sub-module is described in Figure 4. First, an ITB document to be analyzed is uploaded to the Risk Extraction Algorithm module (A1). Second, the algorithm checks whether a clause with DB risk exists through the pre-processing process (A2). Third, the algorithm investigates the existence of risk and, if present, extracts the sentence or generates a notification that there is no result value if no risk rule in the DB is tagged (A3).

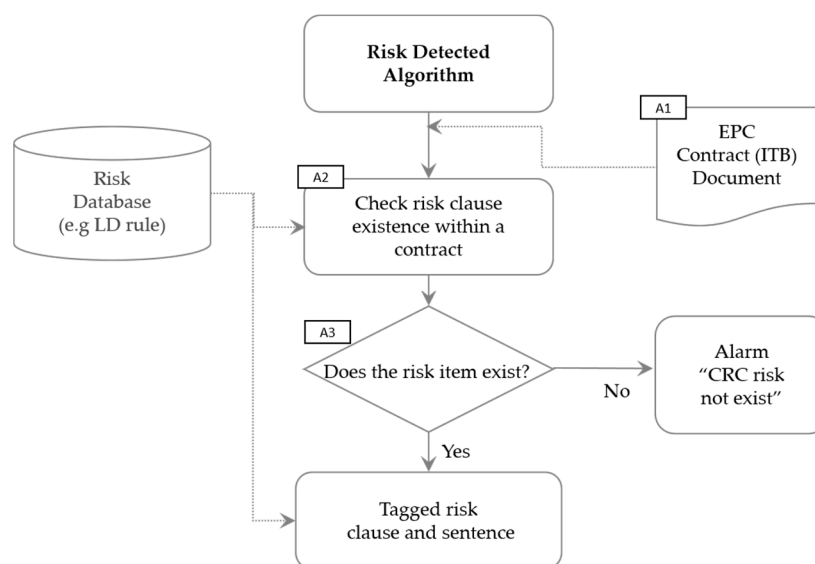


Figure 4. Algorithm flow of the Critical Risk Check module.

5.1. Risk Database of CRC Module

The CRC module enables the review of EPC ITB documents through the risk database established by the research team. By comparing the text of the target contract document with the risk database, the existence of EPC risk sentences was identified and extracted. The existence of risk can be identified when comparing the texts of the target contract with the Risk DB in the CRC module. The Risk DB is divided into the single rule (lower group) and the multi-rule (higher group). The single rule is 1:1 matching of rule and keyword. The intersection of the single rule was defined as the multi-rule.

In the CRC module, the risk rule is input in the form of a token in Python. For example, to extract word combinations in the form of “Token-1, Token-2, and Token-3”, [Attribute-1: Token-1], [Attribute-2: Token-2], [Attribute-3: Token-3] are entered into Python. The details can be specified by entering the appropriate attribute type in parentheses and set multiple attributes as needed. For example, a token called such liquidated damages is entered as [Liquidated Damages: such liquidated damages] in the single rule attribute called Liquidated Damages. In the EPC document, if there is a sentence with LD risk, as shown in Table 5, it is possible to check the existence of a sentence through keyword tagging that is underlined instead of the entire sentence. According to the risk DB, when all keywords belonging to “Liquidated Damage”, “Date”, and “Fail” exist as the intersection in the contract text, it is defined as a Delay Liquidated Damage (DLD) clause label. The

collected ITB documents were analyzed by the expert group, and based on the result of the judgment, the combination of attribute multi-rules was confirmed and converted into a DB. Detailed examples of the single rule and the multi-rule are presented in Tables 6 and 7.

Table 5. An example of liquidated damages sentences from ‘I-1’EPC Project Contract.

LD Sentence
All amounts of such liquidated damages for which contractor may become liable under this contract are agreed to be a genuine pre-estimate of the loss which may be sustained by Company in the event that contractor fails to comply with the relevant obligation under the contract and are not a penalty.

Table 6. An example of the single rules in the Critical Risk Check.

Single Label	Keyword
Liquidated Damages	such liquidated damages Liquidated Damages Damages
Liability	genuine pre-estimate liable
Fail	fails to comply with Fails
Penalty	not a penalty penalty

Table 7. An example of Critical Risk Check multi-rule.

Multi Label	Keyword (Single Label)
Delay Liquidated Damages Clause	Liquidated_Damages, Liability, Fail, Penalty Liquidated_Damages, Time_Barring, Payable

To extract the combination of LD in the ITB documents, five types of multi-labels were selected through collaboration with SMEs, as presented in Table 8. There were ten or more keyword intersection combinations under each label.

Table 8. An example of multi-rule label in the Critical Risk Check.

Multi Label
Liquidated Damages Clause
Delay Liquidated Damages Clause
Performance Liquidated Damages Clause
Exclusive Remedy 1 Clause
Exclusive Remedy 2 Clause

Risk DB was constructed by borrowing the keyword established as a lexicon. The rules applied to the algorithm were constructed in the JavaScript Object Notation (JSON) format and used for analysis. The JSON file format is an open standard format commonly used when creating DB in computing technique research [39]. Therefore, the JSON format was appropriately used for the lexicon, having various text structures such as type, name, and terms. An example of an application is shown in Figure 5.


```

{
  "type": "PHRASH",
  "name": "General_Damages",
  "terms": [
    "General Damage",
    "General Damages",
    'GD',
    "actual loss or damage",
    "Actual damages actual loss or damage caused by such delay",
    "actual loss or damage causedby such breach",
    "Direct Damage",
    "Direct Damages",
    "direct phisical loss or damage",
  ]
}

```

Figure 5. An example of using the JSON format.

5.2. Risk Detection with Phrase-Matcher

The CRC module analyzes the ITB document, automatically extracts the label and similar syntax, and presents it in a data frame format. An algorithm was developed using phrase-matcher technology to automatically analyze and extract risk rules in Python by embedding them. Phrase-matcher is one of the packages provided by spaCy and determines whether to match the token sequence based on the document. By using the Critical Risk Check module using the phrase-matcher technique, a large data list can be efficiently matched, and the matched patterns can be represented in the form of document objects [34].

The CRC module recognizes texts in the ITB document in a PDF format and detects the presence or absence of a keyword defined as a risk clause within the texts. The detected risk clause displays the original text, including the keyword, and is presented with the tagged keyword. The risk clause was converted into a DB and can be extended to the lexicon. In addition, if similar rules of various cases through rule-based analysis are embedded and expanded into a DB, a rule-based technology can relatively secure risk detection reliability.

5.3. Analysis Results of CRC Module

The result values were automatically classified and extracted in a CSV format according to the risk DB. The order of multi-label, single label, risk group, and sentence is automatically generated.

The content and the purpose of extraction of the header in Table 9 are described here:

- Multi-label is a group of single label combinations and is extracted when all combinations are included.
- In a single label, the keyword is the risk that exists in the tagged document. For example, the risk keyword 'failure' was extracted, and 'failure' belongs to the single label of fail. It facilitates risk analysis by organizing the extracted keywords and the labels that belong to them and presenting them to users.
- A risk group can be defined as a clause with a specific rule when a specific combination is tagged among multi-label combinations. In addition to a straightforward keyword tagging for users, the research team defined criteria that can easily judge risks.
- Sentence extracts the risk statement existing in the ITB document as it is. The reason is to support the user's intuitive analysis result analysis.

Table 9. An example of the Critical Risk Check sub-module result.

Multi-Label	Single Label (Label: Keyword)	Risk Group	Sentence
['Liquidated Damages', 'Liability', 'Fail', 'Penalty']	['Liquidated Damages: such liquidated damages', 'Liquidated Damages: a genuine pre-estimate of the loss', 'Liability: liable', 'Fail: fails', 'Penalty: not a penalty.']	['Delay Liquidated Damages Clause']	All amounts of such liquidated damages for which contractor may become liable under this contract are agreed to be a genuine pre-estimate of the loss which may be sustained by Company in the event that contractor fails to comply with the relevant obligation under the contract and are not a penalty.
['Fail', 'Remedy', 'Waiver', 'Contractor']	['Fail: failure', 'Remedy: remedy', 'Waiver: waiver', 'Contractor: PARTY']	['Exclusive Remedy 1 Clause']	The failure to exercise or delay in exercising a right or remedy under the CONTRACT shall not constitute a waiver of the right or remedy, or a waiver of any other rights or remedies, unless such waiver is set out in writing and executed by such PARTY's authorized representative and duly notified to the other PARTY.

In Table 9, Exclusive Remedy 1 (ER1) clause is a risk, which is defined as a clause on workers' compensation. The DLD clause states that if damage (loss) caused by the EPC contractor's schedule delay is compensated for by only DLD, further liability is exempted. If this DLD clause does not exist in the ITB document, it is exclusive remedy (ER) that there is a risk for the project owner to claim actual loss over the DLD due to actual (general) damage. ER risk can be a risk that cannot be easily detected during EPC analysis by a junior engineer with little project experience.

A file in the CSV format, as shown in Table 9, based on the final result standard of the CRC module's risk DB and algorithm, is created. Through the risk group presented in the results, even users with little experience in performing EPC contracts can check whether the risk exists or not. SI can be used as a report for decision-making by checking and downloading the standardized analysis results on the web page, as shown in Figure 6.

Conditions of contract: Critical risk Check

ITB analysis Selected Rule

☒ ☒

Critical Checking : RULE 1

If selected wanted rows, selected downloaded is available

<input type="checkbox"/>	Risk Label	Risk Keyword	Risk Group
<input type="checkbox"/>	Contract_Price Verb Payable Liquidated_Damages	Contract_Price : Contract Price Verb : aggregate Payable : payable Contract_Price : Initial Contract Price Liquidated_Damages : Initial Contract Price Contract_Price : Contract Price Contract_Price : final Contract Price Liquidated_Damages : final Contract Price Contract_Price : Contract Price	Performance Liquidated Damages Clause
<input type="checkbox"/>	Liquidated_Damages Liability	Liquidated_Damages : damages Liability : liability	Performance Liquidated Damages Clause
<input type="checkbox"/>	Liquidated_Damages Remedy Object	Liquidated_Damages : damages Remedy : remedy Object : obligations Object : obligations	Exclusive Remedy 2 Clause
<input type="checkbox"/>	Liquidated_Damages Remedy Object	Liquidated_Damages : damages Remedy : remedy Object : obligations Object : obligations	Exclusive Remedy 2 Clause
<input type="checkbox"/>	Remedy Liquidated_Damages	Remedy : remedies Liquidated_Damages : liquidated damages Liquidated_Damages : liquidated damages for late completion Liquidated_Damages : damages	Exclusive Remedy 2 Clause
<input type="checkbox"/>	Amount	Amount : amount	

Figure 6. An example of CRC results in the cloud service.

6. ITB Analysis Module: Terms Frequency Analysis (TFA)

NER is a technique that recognizes, extracts, and classifies an entity corresponding to a project owner, place, or time from a document through a dictionary definition [37]. NER is actively used in NLP technology and IR. The EPC risk labels used in the TFA were selected through collaboration with the SMEs, and the labels are composed of similar keywords as shown in Table 10. The TFA module collects risk data and creates and utilizes a training model to perform analysis. It is the task of the TFA module to build a training model specialized for EPC risk and implement the tagged frequency value in the algorithm model. The TFA process is shown in Figure 7, and the process is as follows:

- B1: Upload the ITB document to be analyzed.
- B2: Based on the risk DB, the document is analyzed in the NER model to which the NER package is applied.
- B3: Through model analysis, risk entities are tagged, and frequency is extracted.
- B4: The extracted results are visualized and presented as a frequency chart, a HTML rendering file, and a word cloud.

Table 10. Keyword list of fit for purpose label.

Label	NO	Family Keyword
Fit for Purpose	1	fit for purpose
	2	fit for the use
	3	fit for their intended purpose
	4	fit for the purposes
	5	fit for its purpose
	6	fit the purpose
	7	fit for its intended purpose
	8	fit for that purpose

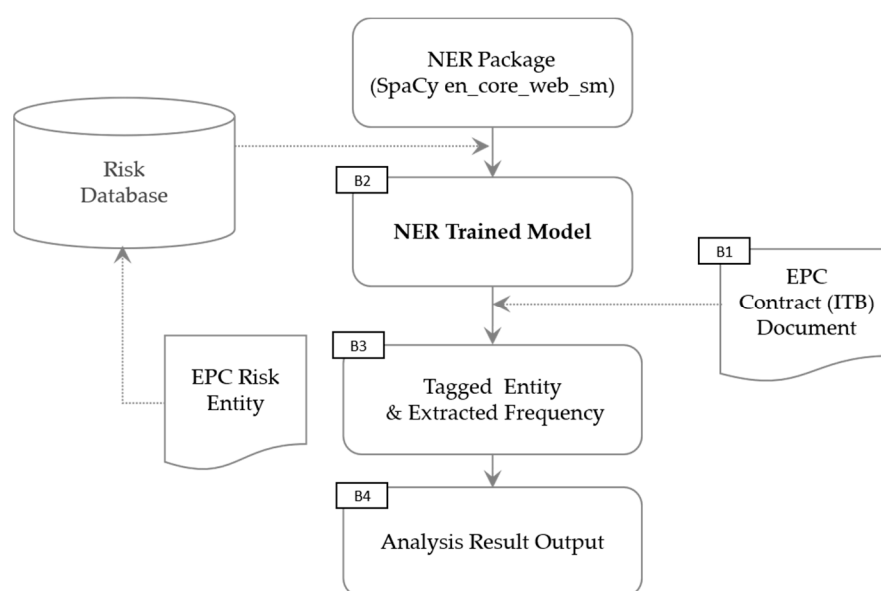


Figure 7. Algorithm flow of Terms Frequency Analysis module.

The collected EPC risk data are generated as training data in the JSON format. The module was completed by learning JSON training data in the NER package. The prototype of the NER model used spaCy's 'en_core_web_sm' model and was completed by customizing it to fit the EPC risk through validation. The 'en_core_web_sm' model is one of the libraries provided by spaCy and provides a pipeline optimized for English sentences [35]. It is possible to recognize basic written text such as blogs, news, comments, vocabulary, and entities type by using technologies such as tok2vec, tagger, lemmatizer, and parser.

Various modules exist and serve a similar purpose ('en_core_web_md', 'en_core_web_lg', 'en_core_web_trf'). When the EPC ITB document to be analyzed is uploaded to the completed model, the analysis result is extracted according to the model learning result, and a graph is generated by calculating the frequency of the entity. In addition, when interworking with the service platform, the model was saved in pickle format to provide accurate analysis values. Python's pickle implements a binary protocol of object structure to store and retrieve data with a small capacity [40].

6.1. Risk Database of TF Module

NER learning data were constructed by selecting sentences for risk clauses from 21,683 lines of sentences collected in the ITB documents. The NER learning data set was created as a JSON file, and 14 kinds of NER learning data JSON files were embedded in the integrated analysis system.

As shown in Table 10, EPC risk data start with training data consisting of eight similar phrases, such as 'fit for purpose.' Similar phrases include plural and singular, and as a result of analyzing the collected EPC ITB documents, they were grouped in the form of family keyword by applying the sentence format commonly used in contracts.

The process of generating NER learning data goes through the following three steps:

- Risk DB classifies only fit for purpose-related sentences from the ITB documents collected from the EPC contractor. In the sentence, the expert filters whether the EPC contractor's liability risk exists or not, and if there is no risk, the plaintext case is excluded.
- Assign a risk entity to each sentence. An example would be the sentence "Contractor warrants that the Plant will be fit for the purpose". In this example sentence, *fit for the purpose* belongs to the risk label. Risk entities start from the 43rd and belong to the 62nd character, so the string numbers are assigned as "43" and "62". The computing character counting method starts with 0 from the first digit.
- Each sentence's given entities are structured as a JSON file. As shown in Table 11, sentence: contract sentence, entities: risk word, and the label are listed in order. By being given as the standard of EPC risk data, it is possible to tag a keyword that has an entity similar to the training data (not learned, but similar to the training data) according to the learning result.

Table 11. An example of risk data JSON structure.

["sentence": "Contractor warrants that the Plant will be fit for the purpose", "entities": (43,62), "FFP"]
--

The NER learning data, created through the process described in Figure 7, are grouped under a group name called Label. In this study, a total of 14 labels applied to the TFA module are shown in Table 12 below. For the NER label, a keyword was selected in consideration of discrimination according to the application of the NER model through collaboration with SMEs. Each label consists of at least 50 to 200 sentences of collected data. When analyzing the TFA module, the model was saved in the pickle format to exclude the phenomenon of biased learning effect according to the data rate, and the saved model was linked to the cloud.

Table 12. An Example of NER label.

NER Label (14)
Damages (Liquidated Damages, Delay Liquidated Damages), dispute, change order (variation), fit for purpose, shall not be liable, limitation of liability, indemnify (indemnification), governing law (applicable law), deem, bank guarantee, Cost, Schedule, Safety, Quality

6.2. Calibration of the NER Model

Since the spaCy library provides only the model learned about the general object name, it is necessary to define the document's contractor risk and apply it to the model [35]. SpaCy's 'en_core_web' model shows the performance of Precision: 0.86, Recall: 0.85, and F-score: 0.86 in NER task [37]. The research team obtained, calibrated, and applied the model's parameter values, epoch and batch size, which are described in detail below.

In ML, hyperparameters are used to control the learning process [41]. Examples of hyperparameters are learning rate and mini-batch size. The time required to train and test the model may vary depending on the hyperparameters. Random numbers generated in a computer program generate a sequence that appears to be random by means of an algorithm determined by the computer. The starting number of a random sequence is called a seed. Once generated, the random seed becomes the seed value for the next generation. In this study, the following values were determined by fixing the random seed (see Table 13). The setting was fixed to the test value that gave the optimal result suitable for EPC risk:

- Epoch: In ML processes, the algorithm goes through each process from input to output using parameters, and epoch 1 can be seen as it completing the entire dataset once. Epoch = 1 means that the whole data set was used once and learned.
- Batch size: The sample size of data given per mini-batch is called batch size. In most cases, due to memory limitations and slowdown in speed, the data are divided into one epoch. The loss derived by inputting each datum of the mini-batch into the model is calculated, and the gradient value of the average of this loss is calculated and updated to the weight. The batch size gradually increases with epoch K, up to a maximum of 256.
- When epoch is K, Batch size, B(K), is expressed as Equation (1):

$$B(K) = \begin{cases} 256 & (K \geq \log_{1.01} 4 + 1) \\ 32 \times (1.01)^{K-1} & (K < \log_{1.01} 4 + 1) \end{cases} \quad (1)$$

- Optimizer: As the stochastic gradient descent (SGD) algorithm, it finds the optimal weight while adjusting the weight in the opposite direction to the gradient direction of the current weight.
- Learning rate: A coefficient that multiplies the gradient value when adding the gradient value of each layer calculated through the loss function and back propagation to the existing layer.
- Dropout: Dropout refers to a technique to reduce the amount of computation and overfitting by inactivating some random nodes among all nodes of the neural network. The dropout rate value means the ratio of the nodes to be deactivated among the nodes of the entire neural network.

Table 13. Calibration values of the NER sub-model.

Parameter	Value
Batch size	128
Dropout rate	0.75
Epoch	2000
Learning rate	1.001
Optimizer	SGD ¹

¹ Stochastic gradient descent: A method of using slope to minimize the value of the loss function, which defines the difference between the result value and the actual value from the network.

6.3. Analysis Results of TFA Module

By using the visualized model analysis result, the risk severity on the EPC project can be estimated to have an impact. The label and keyword frequency were automatically extracted and imaged, and the output types were expressed in three ways as follows:

- Frequency (bar chart): Automatically generated in graph format by calculating the existence and frequency of labels (Figure 8a).
- Word Cloud Image: Automatic extraction of frequency and importance through co-word analysis (Figure 8b).
- HTML file with rendering technique: Highlight and display on the web page (Figure 8c).



Figure 8. An example of TFA module results' visualization.

Figure 8 is the result of 'I-1 project' applied to the pilot project among the ITB documents, and the pre-learned entity was statistically converted through a complete model analysis and was automatically extracted as an image file. If we look at the bar chart in Figure 8a, we can see that the frequency of the Law label is tagged 62 times. Similar phrases learned in the Law label include governing law and applicable law, and local law that was not learned in the I-1 project result was additionally tagged. As such, it was verified that the appropriate EPC risk was extracted by analyzing the articles tagged in the validation step. In Figure 8b, word cloud result is also called tag cloud, and it means a structure that is visually arranged in consideration of importance, popularity, alphabetical order, etc., by analyzing the result values obtained from metadata. Each label, resulting from the term's frequency result and the frequency of similar phrases, was converted into a word cloud and used as visual result data. Figure 8 shows an example of (c) an HTML file with rendering technology application. It is possible to highlight using the rendering function by entity tagging through the learning model. By converting the ITB document into a text file, a user can highlight words tagged as EPC risk and specify the risk group together to generate a resulting value.

7. System Integration and Application

In this study, a machine learning-based integrated decision-making support system as a cloud-service platform, EMAP, was developed to maximize the convenience of EPC contractors based on the current research results on EPC risk analysis. The integrated system is composed mainly of three modules (ITB Analysis, Engineering Design, Mainte-

nance Analysis). Rather than simple integration, it was designed to build and operate an information system that meets the purpose. This research team referred to GE Predix and DNV's Veracity platform examples for system configuration [42,43].

7.1. Features of IT

NLP is a field of artificial intelligence where machines understand and respond to human language systems to help human activities [44]. In order to effectively utilize this key technology, the selection and use of a digital assistant are important. Python has the advantage of being concise and highly productive among programming languages. As it is the most used in the world, it has many useful libraries. Due to these characteristics of Python, Python was mainly used for algorithm development. SpaCy is an open-source software for advanced natural language processing written in a programming language (Python or Cython) [28,45]. Additionally, spaCy provides libraries for numeric data and text processing [44]. In this study, PoC was implemented as an algorithm module by borrowing spaCy libraries of Python.

This research team developed an ITB analysis module for EPC risk analysis by building an EPC risk extraction algorithm based on Python packages. In addition, it was embedded in a cloud-based integrated decision support platform through collaboration with a professional solution company with an independently developed machine learning engine [46]. As shown in Figure 9, the integrated platform is divided into three modules (ITB Analysis, Engineering Design Analysis, Predictive Maintenance), and each module has an algorithm function suitable for the purpose of analysis. HTML, Cascading Style Sheet and JavaScript were used to develop the integrated platform.

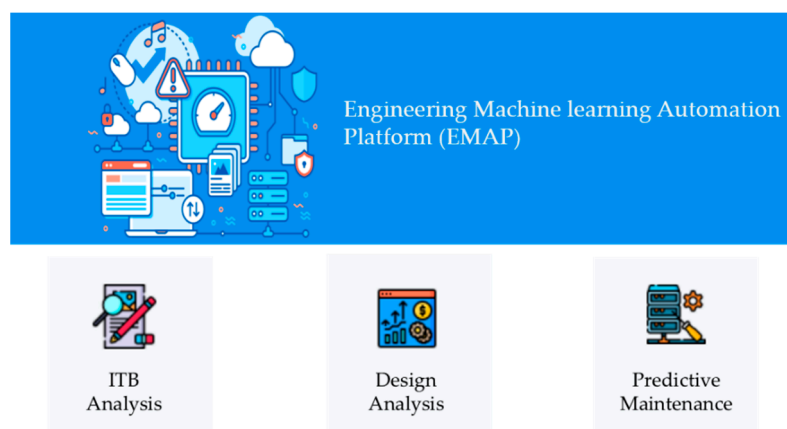


Figure 9. Engineering Machine learning Automation Platform (EMAP).

The integrated platform of this study was named Engineering Machine learning Automation Platform (EMAP), and it was based on a software as a service (SaaS) service that can be immediately used by providing an application solution. Oracle, one of the world's largest software companies, provides integrated services by dividing cloud services into SaaS, platform as a service (PaaS), and infrastructure as a service (IaaS) [47]. The database management system (DBMS) constructed in this study allows authorized users to access the DB through the server. By constructing an accessible DBMS, extensibility is given so that clients can use 'user functions.' The implementation of the 'user management function' made it possible to use the previously embedded DB structure as a new data set by assigning each user-id as a default and copying the modified default value. Through open-source MySQL DBMS, a local server was created, connected, and built to connect to the developed Python package DB. Oracle's MySQL is the world's most widely used relational DB management system [48].

The purpose of the integrated platform is to maximize user convenience on the web. The cloud service, applied in the integrated platform, stores the data, analyzes it, and visualizes the results within the platform so that the user can easily check it on the screen.

7.2. SI of ITB Module

SI can be defined as the overall process of developing a platform. Based on the algorithm described in Sections 5 and 6, a system for each module was constructed as shown in Figure 9 above. If a user clicks ITB analysis on the integrated platform, a user can see five sub-module menus, as illustrated in Figure 10. The CRC and TFA modules are the parts covered in this study. The package of each sub-module is linked to the cloud-integrated solution, and it can be extended to an administrator account.

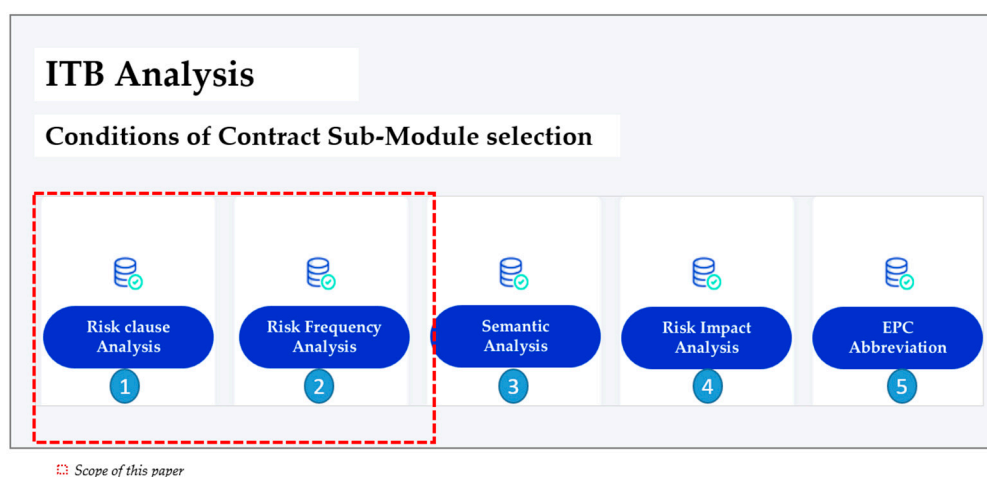


Figure 10. Screenshot of the Contract sub-module selection function from the ITB Analysis module.

In addition, to increase user utilization, the system is configured so that the user can add the desired rule on the final screen. Along with the existing risk DB accumulated through collaboration with an expert team, it provides a function that can be applied to risk analysis by adding arbitrary keywords and labels that the user needs.

Extensibility was achieved with a DB construction and the addition of user management functions. A user can upload and read the Excel format file on the User Interface (UI), apply the rules, and view the results in the existing analysis module. In addition, by conducting a software validation test in the process of implementing the system, errors or bugs that occurred during execution by developers and prospective users were reflected in advance.

8. ITB Analysis Module Validation through PoC

This section describes the validation process and results of the developed modules (CRC and TFA). Module validation was conducted to confirm the applicability of the modules for risk analysis in an actual EPC project. In order to evaluate the reliability and accuracy of the module, the expert group evaluated the ITB documents used in the module development. By engaging experts with specialized knowledge and execution experience from the beginning of development to discuss development ideas for modules and evaluate the analysis results, it was made to be a practical and unbiased validation.

8.1. Definition of Validation Method

The purpose of the module validation was to check whether the modules extract accurate information by converting the unstructured information of the ITB documents into the filtered data format in the DB. The PoC method approach was used to evaluate the quantitative level of the review results. In other words, the performance results (risk detection accuracy and time efficiency) executed by the machine learning-based automatic

risk extraction modules (CRC and TFA) were compared with the evaluation results from the EPC project engineers. The verification methodology followed the processes introduced by Lee et al. [5] and Kang et al. [38].

First, two engineers with experience in performing at least one EPC project participated as the main body of PoC implementation. As the analysis target project, ‘I-1 project (FPSO_Australia)’ was applied equally to the CRC module and the TFA module. SMEs with more than 15 years of experience participated in the verification of PoC performance results (TPV_Third Party Verification). As shown in Table 14, the SMEs participating in the validation process have experience in performing a number of EPC projects in bidding and contracting. PoC proceeded sequentially according to the following four steps:

- Distributed the ‘I-1project’ document (PDF) to two engineers and received replies within a limited period (3 days). In order to derive mutually independent results without mutual interest, data sharing and result replies were conducted individually.
- Conducted cloud-based module analysis using development modules (CRC and TFA) with the same materials distributed to the engineers (‘I-1project’ contract and selection rules).
- The analysis results and system module analysis results performed by the engineer were delivered to SMEs to evaluate and verify the analysis process and contents. The review process for the development module was conducted in a non-face-to-face manner, and the review process was conducted individually by each expert so that independent verification could be made.
- The research team summarized the quantitative evaluation results by converting the SMEs analysis results based on their own knowledge and project execution experience into numerical values.

Table 14. Information of SMEs participating.

Expert	Project Experiences	Specialty
SME 1	20 years	ITB analysis/Contract Management
SME 2	15 years	ITB analysis/Project Management
SME 3	17 years	EPC Project Business Management
Engineer 1	2 years	EPC Project
Engineer 2	5 years	EPC Project

A validation evaluation index was used to quantify the results. The evaluation index was divided into variables called True Positive (TP), False Positive (FP), and Total Risk, and the risk extraction accuracy of each module was quantitatively evaluated by calculating the numerical value using the above three variables. The validation accuracy calculation formula is as shown in Equation (2).

- TP: the risk sentence is extracted as risk.
- FP: the risk sentence was not extracted.
- Total Risk: total risk (TP + FP) for an ITB or contract document of a project.

$$\text{Accuracy} = \text{TP} / (\text{TP} + \text{FP}) \quad (2)$$

8.2. Model Validation Results

8.2.1. CRC Module Validation

To verify the CRC module, PoC was conducted by selecting “LD”, one of the risk keywords that can have a huge impact on the project if it occurs during bidding or execution of the EPC project. LD is a clause that the EPC contractor must pay to the ordering party when the contract delivery date is not met, or the performance required by the contract is not met. It can be divided into DLD or performance liquidated damages (PLD). In addition, EPC risk was divided into single rule and multi-rule. The single rule in the CRC module means risk keyword, and multi-rule is the intersection of every single rule. Therefore,

20 keywords of single rule and LD, which is a representative multi-rule, were selected as analysis targets. Twenty single keyword types can be seen in Table 15.

Table 15. Single keyword list.

Single Keyword (20)
Certificate, Guarantee, Dispute, Contractor, Completion, Period, Payable, Deem, Agreement, Waiver, Law, Liability, Novation, Remedy, Discretion, Taxes, Refund, Amount, Indemnify, Default

The CRC module analysis process is as follows. First, a bidding specification (I-1project) is selected. Then, the prepared bid specification is uploaded to the CRC module. After performing the analysis, the number of risk detections is quantified numerically. The CRC accuracy evaluation results are shown in Table 16.

Table 16. CRC sub-module results.

Risk Category	Subject	Total Sentence	Time Span (Hours)	Extracted Sentence	SMEs Validation		Accuracy (%)
					TP	FP	
Single Rule	Module	1122	0.6	230	205	25	89
	Engineer	1122	30	120	120	0	100
Multi-Rule (LD)	Module	1122	0.5	36	34	2	94
	Engineer	1222	24	28	28	0	100

- **Single Rule:** When comparing the CRC module analysis results as shown in Table 16, the engineer risk extraction accuracy (100%) is relatively higher than the machine learning-based CRC automatic extraction module accuracy (89%). However, the engineer extracts 120 sentences with EPC risk, whereas the CRC module detects 230 sentences, so the module has a relatively high risk detection capacity. In addition, 205 sentences out of the 230 sentence extraction results of the SMEs verification result module were verified as TP. An example sentence in Table 17 below is judged by SMEs as FP. SMEs suggested that the sentence belongs to the Definition chapter of the EPC ITB document and is simply a sentence that defines the term Contract Price. The difference between the risk sentence extraction results of the Engineer and the CRC module is the time limit and the large number of documents. The developed algorithm module is capable of extracting risk sentences by screening a large amount of text in a short time, but an omission occurred because the engineer reviewed the entire document within a limited time. In addition, it was found that some differences occurred in the result values of risk sentence judgment due to personal experience and subjective judgment among POC performing engineers.
- **Multi-rule:** As shown in Table 16, the engineer extracted 28 sentences with EPC risk, whereas the CRC module detected 36 sentences. As a result of SMEs' verification, 34 sentences were verified as TP. An engineer can see that the risk keyword detection accuracy (82%) drops by detecting only 28 out of 34 LD risk sentences verified by SME. Additionally, the module accuracy of the multi-rule (94%) is higher than the module accuracy of the single rule (89%). The multi-rule is an intersection of risk keywords, and although the number that exists in one project is small, it is selective and has higher accuracy.

Table 17. An example of contract sentence.

Contract Price means the aggregate of all sums payable under the contract calculated in accordance with Exhibit B as may be modified by Change Orders; it being understood that the initial Contract Price is that known at the Contract Date and the final Contract Price is that known after final assessment under the contract as described in sub-Article 34.6.
--

8.2.2. TFA Module Validation

First, for the verification of the TFA module, the target label was selected as damage (including LD and DLD), and the phrase label, FFP (Fit for Purpose), was selected through expert collaboration. Based on the target label, an accuracy evaluation of risk detection was performed for the module. Since the NER module is a technique for extracting keywords defined as entities such as people, regions, and organizations, the extracted target is defined as Keywords. The analysis process is similar to the verification process of the CRC module, and the accuracy results are shown in Table 18 below.

Table 18. Accuracy of Terms Frequency Analysis sub-module result.

Target Label	Subject	Contract (Project)	Time Span (Hours)	Extracted Keyword	SMEs Validation		Accuracy (%)
					TP	FP	
Damages	Module	I-1	1	98	90	8	92
	Engineer	I-1	24	82	82	0	100
Fit for Purpose (FFP)	Module	I-1	0.9	6	5	1	83
	Engineer	I-1	22	4	4	0	100

When comparing the analysis results of the TFA risk module, the machine learning-based automatic TFA risk extraction model can significantly reduce the risk extraction rate and the time required for EPC risk analysis (see Table 18). The engineer analysis result was 82 for the Damages label and 4 for the FFP label. In the TFA Module, 98 and 6 were tagged, respectively, and 8 and 1 keywords were determined as FP, respectively. The module validation results of each target label are 92% and 83%, respectively. Although 82 and 4 were extracted from the engineer validation, respectively, the keywords determined as FP as a result of the module validation were 90 and 5, and the engineers' analysis result accuracy was 91% and 80%, respectively.

As a result of examining the keyword determined by FP by an expert, it can be determined that it is not related to the target label. The main cause is the incompleteness of the knowledge-based learning model, which requires machines to understand human language. In order to minimize FP tagging errors, continuous expansion of learning data and advancement of machine learning functions through repetitive learning are required.

9. Conclusions

This paper proposes the technology to check the existence of risk clauses for the ITB documents that require prior analysis when bidding or executing EPC projects and to detect and manage project risk sentences. Accordingly, two algorithm modules were developed for automatic risk extraction and analysis in the EPC technical specification based on ML techniques. In the CRC module, EPC risk-related clauses were defined, and the rules were selected by structuring it in a lexicon and embedded in the CRC module. An automatic risk detection function was developed to find the risk specified in the lexicon from the collected DB. The TFA module collects and rearranges EPC risk sentences to build a learning data set and constructs the NER model through individual learning of EPC risk sentences. The ITB document was analyzed using a pre-learned model, and the frequency of EPC risk was calculated, so that it could be used as a decision support tool. Therefore, this study performed the model development work according to the following steps. First, the model collected ITB documents and classified them by risk types. 25 EPC projects, previously conducted from EPC companies, were collected, and risk sentences (21,683 lines) were classified by risk type. These sentences were used as basic data for model development.

Second, the research team developed an algorithm and model that can detect major risks of contracts through natural language processing of sentences included in the ITB documents. The data preprocessing and algorithm step removes unnecessary data, and its logic was configured with ML, according to the characteristics of the two modules. Risk detection technology was developed through keyword grouping by applying NLP's phrase-matching technology to the CRC algorithm. In the case of the TFA algorithm, it can

be used as an analysis tool by analyzing the ITB document using a pre-learned model and calculating the frequency of the EPC risk. In the system application stage, the algorithm was dash-boarded on the ML platform to visualize the analysis results. When the pattern of sentences in the ITB document matches the developed rules, the mechanism for extracting information is activated. The model is implemented using Python, and automatically extracts and reviews risk sentences when the user enters the ITB document. This supports the users' decision-making by viewing the model results. In addition, a user function has been added so that the model reflecting the rules desired by the user can be analyzed. The developed model is presented as a cloud-based integrated decision support platform, EMAP, considering user convenience.

Third, to enhance and verify the reliability of the developed model performance, collaboration with EPC project experts was conducted from the beginning of development, and the model performance and usability were reviewed by the experts. The experts who participated in the validation have experience in performing a number of EPC projects and allow independent validation to be performed. As a result of the pilot test, the CRC module-based risk sentence extraction accuracy result was 92%, and the TFA module-based risk sentence extraction accuracy result was 88%, whereas the engineer's risk sentence extraction accuracy result was 70% and 86%, respectively. This is the result of risk extraction of the model reflecting some types of errors found in the validation process. The model extraction results show higher performance than the results detected by PoC subjects (engineers), and the time required for analysis is also significantly different. According to the validation conducted in this study, the ITB document analysis is prone to deviations depending on the capabilities of each manager or engineer. The significance of this study can be summarized in two ways. The first is that a technical system has been established to support the ITB document risk sentence detection task more effectively. Second, a model capable of preemptive risk management based on ML was developed. It automatically extracts major risks by incorporating machine learning technology into an analysis algorithm and presents them in an evaluation index.

By applying various AI technologies in the EPC field, the research team developed a scientific tool to automatically identify risk clauses that engineers are likely to miss, reducing potential risks to contractors and significantly shortening contract review time. In this process, it was confirmed that AI and ML technology were successfully applied to the EPC field.

10. Limitation and Future Work

Despite the above study results, there are still some areas that need improvement when conducting related studies in the future. First, a rule that defines and embeds keyword risk needs to be improved at the level of a syntactic structure by targeting the conditions of a contract in the ITB document analysis procedure. Although the level of risk detection accuracy of the CRC module is stable, there are some areas in which unexpected errors or typos are not completely screened, other than the standard that is practically universal. For example, in the case of LD risk sentence validation, the degree of the ratio increased sharply in the single keyword sentence analysis, as was judged as an FP sentence among the module results. By analyzing multiple rules at once, they are extracted with a higher percentage of plain text or meaningless sentences. Continuous and elaborate knowledge-based rule updates are required to extract sentences that contain EPC Risk Keyword in the sentence. Accordingly, the authors will continue to conduct research to improve the reliability of analysis through linking NLP technology and technology updates that automatically pre-process documents in an atypical state.

Second, evaluating the recognition rate for automatic extraction of the TFA module, there are cases in which similar phrases are incorrectly tagged in the case of a label with insufficient learning. To reduce these errors, learning and training through big data are necessary, which is a common problem in the rule-based information extraction model. To overcome this, there is a need for continuous updates. The lexicon is continuously

expanding, data on various types of ITB documents are constructed, and the rules are implemented through ML. In future research, it is expected that the universality of the integrated platform can be secured through the introduction of advanced technologies and the expansion of training data through the collection of extensive ITB documents not only in the plant field but also in other industries.

Third, in this study, it was implemented through the construction of a lexicon including the contracts and the keywords in English only. In order to apply to another language in addition to English, further processes are necessary to collect contracts in the corresponding language, establishing the keywords and the rules in the database in the corresponding language, although the AI and ML technique would still work. The expected difficulty is that in addition to the existing English contract, it is necessary to collect data for the corresponding language, and when recognizing the texts from the contract through the program, the development of a recognition algorithm for each language should proceed.

Fourth, apart from this study, this research team is conducting research on the automatic risk extraction model through semantic analysis and the knowledge-based automatic risk extraction model for technical specifications. When construction of the individual modules is combined into an integrated ML decision support platform (Engineering Machine learning Automation Platform: EMAP) (to be completed in 2021), it will be able to provide practical direct and indirect benefits to EPC contractors.

Author Contributions: The contribution of the authors for this publication article are follows: conceptualization, S.J.C. and E.-B.L.; methodology, S.J.C., S.W.C. and E.-B.L.; software, S.J.C. and S.W.C.; validation, E.-B.L. and J.H.K.; formal analysis, S.J.C. and S.W.C.; writing—original draft preparation, S.J.C.; writing—review and editing, J.H.K. and E.-B.L.; visualization, S.J.C. and S.W.C.; supervision, E.-B.L.; project administration, E.-B.L.; and funding acquisition, E.-B.L. and J.H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Korea Ministry of Trade Industry and Energy (MOTIE) and the Korea Evaluation Institute of Industrial Technology (KEIT) through the Technology Innovation Program funding for “Artificial Intelligence and Big-data (AI-BD) Platform Development for Engineering Decision-support Systems” project (grant number = 20002806).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Kim, C.M. (a senior researcher at University of California—Davis) for his academic feedback on this paper, Lee, S.Y. (a senior researcher in POSTECH University) for his support on the manuscript editing work, Baek, S.B. (a graduate student in POSTECH University) for his assistance to Python coding, and Kim, C.Y. (a graduate student in POSTECH University) for her help on the manuscript revision. The views expressed in this paper are solely those of the authors and do not represent those of any official organization.

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

Abbreviations

The following abbreviations and parameters are used in this paper:

AI	Artificial Intelligence
API	Application Programming Interface
CRC	Critical risk check
CSV	Comma-separated values
CT	Critical Term
DBMS	Database Management System
DLD	Delay Liquidated Damage
EMAP	Engineering Machine learning Automation Platform
EPC	Engineering, Construction, and Construction
ER1	Exclusive Remedy 1
FP	False Positive

HTML	Hyper-Text Markup Language
IaaS	Infrastructure as a Service
IE	Information Extraction
IR	Information Retrieval
DBMS	Database Management System
ITB	Invitation to bid, invitation for bid or sealed bid
JSON	JavaScript Object Notation
LDs	Liquidated Damages
NER	Named Entity Recognition
NLP	Natural language processing
OCR	Optical Character Recognition
PaaS	Platform as a Service
PLD	Performance Liquidated Damages
PoC	Proof-of-concept
SaaS	Software as a Service
SI	System Integrator
SMEs	Subject-Matter Experts
TF	Terms Frequency

References

- Shen, W.; Tang, W.; YU, W.; Duffield, C.F.; Hui, F.K.P.; Wei, Y.; Fang, J. Causes of contractors' claims in international engineering-procurement-construction projects. *J. Civ. Eng. Manag.* **2017**, *23*, 727–739. [\[CrossRef\]](#)
- Du, L.; Tang, W.; Liu, C.; Wang, S.; Wang, T.; Shen, W.; Huang, M.; Zhou, T. Enhancing engineer–procure–construct project performance by partnering in international markets: Perspective from Chinese construction companies. *Int. J. Proj. Manag.* **2016**, *34*, 30–43. [\[CrossRef\]](#)
- Nurdiana, A.; Susanti, R. Assessing risk on the engineering procurement construction (EPC) project from the perspective of the owner: A case study. In Proceedings of the IOP Conference Series: Earth and Environmental Science, Surabaya, Indonesia, 1–2 October 2020; Volume 506, p. 012040. [\[CrossRef\]](#)
- Doloi, H.; Sawhney, A.; Iyer, K.C.; Rentala, S. Analysing factors affecting delays in indian construction projects. *Int. J. Proj. Manag.* **2012**, *30*, 479–489. [\[CrossRef\]](#)
- Lee, J.H.; Yi, J.S.; Son, J.W. Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based NLP. *J. Comput. Civ. Eng.* **2019**, *33*. [\[CrossRef\]](#)
- Korea Agency for Infrastructure Technology Advancement. *2013 General Report on the Analysis of Technology Level of Land Transportation (KAIA)*; Korea Agency for Infrastructure Technology Advancement: Anyang, Korea, 2013.
- Gunduz, M.; Yahya, A.H.A. Analysis of project success factors in construction industry. *Technol. Econ. Dev. Econ.* **2015**, *24*, 67–80. [\[CrossRef\]](#)
- Mohebbi, A.H.; Bislimi, N. Project Risk Management: Methodology Development for Engineering, Procurement and Construction Projects a Case Study in the Oil and Gas Industry. Master's Thesis, Karlstad University, Faculty of Economic Sciences, Communication and IT, Karlstad, Sweden, February 2012.
- Rijtema, S.; Haas, R.D. Creating sustainable offshore developments in the ultra-deep water. In Proceedings of the Offshore Technology Conference, Houston, TX, USA, 4–7 May 2015. [\[CrossRef\]](#)
- Yu, N.; Wang, Y. Risk analysis of EPC project based on ISM. In Proceedings of the 2nd IEEE International Conference on Emergency Management and Management Sciences, Beijing, China, 8–10 August 2011. [\[CrossRef\]](#)
- Kim, M.H.; Lee, E.B.; Choi, H.S. Detail engineering completion rating index system (DECRIIS) for optimal initiation of construction works to improve contractors' schedule-cost performance for offshore oil and gas EPC projects. *Sustainability* **2018**, *10*, 2469. [\[CrossRef\]](#)
- Ullah, K.; Abdullah, A.H.; Nagapan, S.; Suhoo, S.; Khan, M.S. Theoretical framework of the causes of construction time and cost overruns. *IOP Conf. Ser. Mater. Sci. Eng.* **2017**, *271*, 012032. [\[CrossRef\]](#)
- Gunduz, M.; Maki, O.L. Assessing the risk perception of cost overview through importance rating. *Technol. Econ. Dev. Econ.* **2018**, *24*, 1829–1844. [\[CrossRef\]](#)
- Animah, I.; Shafiee, M. Application of risk analysis in the liquefied natural gas (LNG) sector: An overview. *J. Loss Prev. Process. Ind.* **2020**, *63*. [\[CrossRef\]](#)
- Son, B.Y.; Lee, E.B. Using text mining to estimate schedule delay risk of shore oil and gas EPC case studies during the bidding process. *Energies* **2019**, *12*, 1956. [\[CrossRef\]](#)
- Chua, D.K.H.; Loh, P.K.; Kong, Y.C.; Jaselskism, E.J. Neural networks for construction project success. *Expert Syst. Appl.* **1997**, *13*, 317–328. [\[CrossRef\]](#)
- Ho, S.P.; Tserng, H.P.; Jan, S.H. Enhancing knowledge sharing management using BIM technology in construction. *Sci. World J.* **2013**, 1–10. [\[CrossRef\]](#) [\[PubMed\]](#)

18. Sackey, S.; Kim, B.S. Development of an expert system tool for the selection of procurement system in large-scale construction projects. (ESCONPROCS). *KSCE J. Civ. Eng.* **2018**, *22*, 4205–4214. [CrossRef]
19. Lee, D.H.; Yoon, G.H.; Kim, J.J. Development of ITB risk Mgt. model based on AI in bidding phase for oversea EPC projects. *J. Inst. Internet Broadcast. Commun. (JIIBC)* **2019**, *19*, 151–160. [CrossRef]
20. Watson Explorer; IBM: Armonk, NY, USA. Available online: <https://www.ibm.com/docs/en/watson-explorer/11.0.1?topic=components-product-overview> (accessed on 30 June 2021).
21. Cherpas, C. Natural language processing, pragmatics, and verbal behavior. *Anal. Verbal Behav.* **1992**, *10*, 135–147. [CrossRef]
22. Lim, S.Y.; Kim, S.O. A text mining analysis for research trend about information and communication technology in construction automation. *Korean J. Constr. Eng. Manag.* **2016**, *17*, 13–23. [CrossRef]
23. Tixier, A.J.-P.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Autom. Constr.* **2016**, *62*, 45–56. [CrossRef]
24. Williams, T.P.; Gong, J. Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Autom. Constr.* **2014**, *43*, 23–29. [CrossRef]
25. Marzouk, M.; Enaba, M. Text analytics to analyze and monitor construction project contract and correspondence. *Autom. Constr.* **2019**, *98*, 265–274. [CrossRef]
26. Zoua, Y.; Kiviniemib, A.; Jonesa, S.W. Retrieving similar cases for construction project risk management using natural language processing techniques. *Autom. Constr.* **2017**, *80*, 66–76. [CrossRef]
27. Li, M.; Yang, Q.; He, F.; Li, Z.; Zhao, P.; Zhao, L.; Chen, Z. An unsupervised learning approach for NER based on online encyclopedia. *Web Big Data* **2019**, 329–344. [CrossRef]
28. Python; Python Software Foundation: Wilmington, DE, USA. Available online: <https://www.python.org/> (accessed on 21 June 2021).
29. Vijayarani, S. Preprocessing techniques for text mining—An overview. *Int. J. Comput. Sci. Commun. Netw.* **2015**, *5*, 7–16.
30. Manning, C.; Schütze, H. *Foundations of Statistical Natural Language Processing*; Cambridge University Press: Cambridge, UK, 1999.
31. Shinde, A.A.; Chougule, S.R. Text pre-processing and text segmentation for OCR. *Int. J. Comput. Sci. Eng. Technol.* **2012**, *2*, 810–812.
32. Marina, D.B. The Open Source Relational Database. Available online: <https://mariadb.org/> (accessed on 21 June 2021).
33. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008. [CrossRef]
34. PhraseMatcher, spaCy API Documentation. Explosion, Berlin, Germany. Available online: <https://spacy.io/api/phrasematcher> (accessed on 21 June 2021).
35. enModels Model Documentation, Explosion, Berlin, Germany. Available online: <https://spacy.io/models/en> (accessed on 21 June 2021).
36. spaCy, Explosion, Berlin, Germany. Available online: <https://spacy.io/> (accessed on 21 June 2021).
37. NamedEntityRecognition, Training Pipelines and Models, Explosion, Berlin, Germany. Available online: <https://spacy.io/usage/training#quickstart> (accessed on 21 June 2021).
38. Kang, S.O.; Lee, E.B.; Baek, H.K. A digitization and conversion tool for imaged drawings to intelligent piping and instrumentation diagrams (P&ID). *Energies* **2019**, *12*, 2593. [CrossRef]
39. JavaScript Object Notation, Introducing JSON. Available online: <https://www.json.org/json-en.html> (accessed on 21 June 2021).
40. Pickle, Python Object Serialization. Available online: <https://docs.python.org/3/library/pickle.html> (accessed on 21 June 2021).
41. Amir-Ahmadi, P.; Matthes, C.; Wang, M.C. Choosing prior hyperparameters: With applications to time-varying parameter models. *J. Bus. Econ. Stat.* **2018**, *38*, 124–136. [CrossRef]
42. Predix Platform, General Electronic, Boston, MA, USA The Industrial IoT Platform. Available online: <https://www.predix.io/> (accessed on 21 June 2021).
43. Veracity, DNV, Sandvika, Norway. Available online: <https://www.veracity.com/> (accessed on 21 June 2021).
44. Vasiliev, Y. *Natural Language Processing with Python and SpaCy*; No Starch Press: San Francisco, CA, USA, 2020.
45. Behnel, S.; Bradshaw, R.; Dalcín, L.; Florisson, M.; Makarov, V.; Seljebotn, D.S. *Cython C-Extensions for Python*; Redmond: Washington, DC, USA; Available online: <https://cython.org/> (accessed on 21 June 2021).
46. Wiseprophet, *Machine Learning Automated Platform*; WISEiTECH: Seoul, Korea. Available online: <http://prophet.wise.co.kr/#/intro> (accessed on 21 June 2021).
47. Oracle; Integrated Cloud Application: Austin, TX, USA. Available online: <https://www.oracle.com/index.html> (accessed on 21 June 2021).
48. MySQL; Oracle Corporation: Austin, TX, USA. Available online: <https://www.mysql.com/> (accessed on 21 June 2021).