A Digitalized Design Risk Analysis Tool with Machine-Learning Algorithm for EPC Contractor’s Technical Specifications Assessment on Bidding

Min-Ji Park 1, Eul-Bum Lee 1,2,* , Seung-Yeab Lee 1 and Jong-Hyun Kim 3

Abstract: Engineering, Procurement, and Construction (EPC) projects span the entire cycle of industrial plants, from bidding to engineering, construction, and start-up operation and maintenance. Most EPC contractors do not have systematic decision-making tools when bidding for the project; therefore, they rely on manual analysis and experience in evaluating the bidding contract documents, including technical specifications. Oftentimes, they miss or underestimate the presence of technical risk clauses or risk severity, potentially create with a low bid price and tight construction schedule, and eventually experience severe cost overrun or/and completion delays. Through this study, two digital modules, Technical Risk Extraction and Design Parameter Extraction, were developed to extract and analyze risks in the project’s technical specifications based on machine learning and AI algorithms. In the Technical Risk Extraction module, technical risk keywords in the bidding technical specifications are collected, lexiconized, and then extracted through phrase matcher technology, a machine learning natural language processing technique. The Design Parameter Extraction module compares the collected engineering standards’ so-called standard design parameters and the plant owner’s technical requirements on the bid so that a contractor’s engineers can detect the difference between them and negotiate them. As described above, through the two modules, the risk clauses of the technical specifications of the project are extracted, and the risks are detected and reconsidered in the bidding or execution of the project, thereby minimizing project risk and providing a theoretical foundation and system for contractors. As a result of the pilot test performed to verify the performance and validity of the two modules, the design risk extraction accuracy of the system module has a relative advantage of 50 percent or more, compared to the risk extraction accuracy of manual evaluation by engineers. In addition, the speed of the automatic extraction and analysis of the system modules are 80 times faster than the engineer’s manual analysis time, thereby minimizing project loss due to errors or omissions due to design risk analysis during the project bidding period with a set deadline.

Keywords: decision support; engineering procurement and construction (EPC); technical specifications; technical risk extraction; risk phrase extraction; phrase matcher; machine learning algorithm; text and data mining; terms frequency; artificial intelligence

1. Introduction

The Engineering, Procurement, and Construction (EPC) project is a technology-intensive industry that requires advanced manufacturing technology and knowledge services in design, manufacturing, and installation fields. It is one of the large complex
industries, with various stages comprising maintenance to repair [7]. However, the EPC-type project contract can be a one-sidedly advantageous contract form for project owners [7]. For example, when an EPC project is brought to competitive bidding, project owners may compete in bidding at a lower price. In other words, there are cases in which large-scale losses occur in EPC projects because contractors sometimes receive projects at a lower price in competitive bidding because project risks are overlooked. One of the critical steps required to prevent loss and ensure profit in an EPC project is the engineering phase. In particular, the ability to perform engineering is a significant factor for project success. Although the design cost accounts for 7 to 10 percent of the total construction cost [7] since the design is a factor that directly affects the entire project process, including construction, maintenance, repair, systematic technical specification analysis and evaluation, it has a direct impact on project bidding and success [7]. In other words, to gain an edge on the project bidding competition, it is necessary to minimize contract risk by bidding at a suitable price rather than aiming to lower the bid price.

The study started with the goal, "what if engineers can perform risk analysis at an equivalent level of decades of experience in the EPC field while they review bidding documents within a limited review period?" In order to achieve the study goal, a comprehensive risk analysis tool that could identify multiple types of risk factors at a glance in the EPC field was necessary. The risk factors that needed to be analyzed in EPC are the associated time and cost. This study developed an integrated system tool to identify and eliminate risk factors in the bidding documents before a contract is executed. When technical risks are identified and eliminated in technical specifications through an automated and integrated tool before a contract is executed, contractors can reduce the extra cost and avoid the delinquency of construction completion.

In this study, EPC contract analysis technology was developed based on machine learning (ML) algorithms and standardized analysis data by pre-processing data using text data loading and sentence classification technology for 25 technical specifications obtained from project owners or contractors. Two modules, Technical Risk Extraction (TRE) and Design Parameter Extraction (DPE), were developed for design risk extraction and the analysis of technical specifications. ML is a technique that learns based on empirical data, makes predictions, and improves its performance. It is a suitable technology to research and build algorithms for this purpose. It takes the approach of building specific models to make predictions or decisions based on input data. An overview of the two modules (TRE and DPE) developed as follows is described.

First, the TRE module was developed to automatically detect and evaluate design risk clauses that are easy to miss during bidding or project execution due to time constraints or a lack of personal competence. Based on the design risk keywords selected by the expert group with experience over ten EPC projects, a lexicon was built and grouped by severity. When inputting engineering documents such as technical specifications that require analysis, the evaluation factor (risk score) for the new project severity was suggested through the phrase matcher technique between each clause and risk keywords in the lexicon [7]. In other words, users could detect and prevent risks in advance by automatically detecting risk items. The international codes, such as the American Society of Mechanical Engineers (ASME) Code, were used as the engineering standards [7]. Standard design parameters (SDP) for each analysis target were established, and similar expressions of the relevant requirements were
detected. A comparative analysis was then performed with the design standard using the context manager technique of analyzing technical specifications that a user needs to compare. Through the context manager technique developed in this study, only the items to be compared among various requirements were extracted and analyzed. In other words, when there is a corresponding numerical requirement or similar expression, it can analyze the context to find a design standard, compare it with the embedded engineering standard, and present it to support a user’s decision making.

A pilot test was performed to verify the effectiveness and suitability of the developed modules (TRE and DPE). In the TRE module, the design risk extraction accuracy of the system module was more than 50 percent higher than the risk extraction accuracy of the subject to be verified by engineers. In the DPE module, the design risk extraction accuracy of the system module had a relative advantage of 20 percent or more compared to the risk extraction accuracy of the subject to be verified by engineers. The most important part of verification is the analysis time. It took 42.5 h for the subject to be verified by engineers, but it took only less than 0.5 h for the system module developed in this study. In other words, it was confirmed that the module’s risk analysis time was reduced by 98.8 percent, compared to the engineer’s risk analysis time, and at the same time, it showed a performance advantage for analysis detection.

As mentioned above, through the two system modules developed through this study, it is possible to detect risk clauses and requirements in the technical specifications automatically and to extract and evaluate risks, thereby preventing design risks from project bidding and execution. This is a significant contribution that lays the foundation for minimizing contractors’ losses in EPC projects. An application platform that customizes modules according to the characteristics of the EPC contractors through the editing function and project-based data accumulation was developed for practical implementation in the EPC field. If the data accumulation for the characteristics of other industrial fields beyond the EPC plant field is carried out in the future, it will be possible to develop more versatile technology for practical implementation in various engineering fields.

Section ?? summarizes the literature review and Section ?? describes the methodology and model development processes. Section ?? discusses the validation methods and results and Section ?? discuss the conclusion and the future research direction.

2. Literature Review

This study aimed to identify technical risks that can be reviewed in the bidding stage of the EPC project. For this purpose, previous studies in related fields are thoroughly reviewed and summarized in this section. The summary of the previous research consists of two parts. First, studies on technical risk analysis and countermeasures of the EPC project were analyzed. Not only the EPC plant projects but also the construction projects were included in the literature review. Next, research that applied ML technologies to the EPC projects and developing systems using ML technologies was reviewed.

2.1. Technical Risk Analysis in Engineering Projects

Studies for risk reduction from the project owners’ perspective have been preceded for a long time in the EPC project. Micheli et al. [?] proposed an efficient cost-saving approach where owners select suppliers. Memon et al. [?] conducted a study on risk factors in developing countries, especially among construction projects worldwide. Doloi et al. [?] also analyzed risk factors in the construction sector of India. One case study of risk analysis from the contractors’ perspective was carried out by Kerur et al. [?], but studies from the project owners’ perspective have also been actively conducted. This study aimed to reduce design risks in EPC projects to balance perspectives between project owners and contractors (engineers and managers).

The risks of EPC projects are associated with project schedule delay and cost overrun, and the research that suggests models as solutions have been preceded [? ?]. However, it was difficult for users to utilize the proposed models practically. In other studies, Wang
et al. and Heravi et al. presented models to predict labor productivity for construction period reduction. Even more recently, studies to identify the cause of schedule delays in large construction projects have been preceded. However, there have been insufficient studies to suggest models to directly solve the causes of risks in EPC projects. Therefore, this study aimed to develop algorithm-based models to solve the causes of risks in EPC projects.

The previous studies found that project performance is directly related to risk through time and cost management. Risks arise due to time constraints and a lack of experience in the design phase. Yi et al. performed Monte Carlo simulations for all major critical activities to analyze the temporal impact of each phase of the onshore EPC project, and as a result, they found that the design phase affects the overall project schedule up to ten times more than the procurement and construction phases. Recent studies focused on the contract of EPC Invitation to Bid (ITB), so it was necessary to study technical specification analysis at the design stage. In this study, the technical specifications of ITB for the EPC project were targeted.

Risk studies in the bidding stage for successful project implementation were recently conducted. Zekri et al. conducted a study to improve the performance of the bidding process information system. Recently, Gunduz et al. found that decision making at the design and contract stage was considered the critical factor for successful project management, and systematic decision making has not yet been made. In other fields, models for risk-appropriate bidding were proposed. Although research on risk management was conducted in the EPC project field, the existing earned value management and earned schedule management methodologies have been expanded. Therefore, in this study, the aim was to develop a new algorithm tailored to the characteristics of the technical specifications in the EPC project field.

2.2. Application of Machine Learning Technology in an Engineering Sector

Studies on the use of ML technology in the engineering industry have been actively conducted recently, and various research results are being published. In this section, studies that applied ML technologies for the risk analysis of engineering projects are summarized. In addition to ML technology, the development of a decision support platform is also reviewed and summarized in this section. First, the case of Watson, a decision-making system that has received the most attention recently, is reviewed. The system has been applied and utilized in various fields, but it has not been verified whether it is suitable for EPC projects. Lin et al. proposed a passage partitioning approach based on domain ontology to improve the performance of information retrieval (IR) functions for technical documents in Architecture/Engineering/Construction (A/E/C) projects. Their technique was not suitable for technical specification documents, where multiple formats exist for different countries. Wang et al. proposed a decision-making model for owners’ capital management. Although entropy weight was used for bid evaluation, their study required a decision-making system in the design process, as a model was limited to the bidding process. Shin et al. analyzed data patterns using the lift sensor module and a storage device for tall buildings. Through their study, they developed a decision-making system that shortens the waiting time of users. In this study, rapid analysis time and accurate decision-making support were set as the top priorities to help solve the time constraint and lack of engineers’ experience.

Studies on extracting requirements using Natural Language Processing (NLP) have been conducted in several fields. In the DPE module developed in this study, the user’s requirements can be restored and modified. Podolski proposed a solution using artificial intelligence for project scheduling in construction projects. In this study, a way to shorten the schedule through project resource management was developed. Lee et al. developed a model to identify and extract risk clauses from contracts for onshore construction projects by applying NLP. Regarding Information Extraction (IE) technologies, the International Federation of Consulting Engineers (FIDIC) Rebook were
used as the analysis targets. As their study only analyzed the FIDIC Rebook, not the various contracts of the construction project, a limitation was that the accuracy of information extraction decreased when entering other types of contracts. The TRE module of this study was developed as a risk extraction system specialized for technical specification documents. In another study, Chozas et al. [??] improved the performance by using learning programming for the medical field and marketing field.

In this study, the research team reviewed studies on the current technical specifications. Saint-Dizier [??] presented a corpus of incoherence to solve overlap and incoherence in thousands of requirements within technical specifications. The classification was limited by pattern definition for each configuration method and requirement. Abualhaija et al. [??] proposed an ML approach to distinguish different requirements from free-form textual requirements specifications and conducted a study that trained and evaluated a dataset consisting of 30 industrial requirements specifications. However, risk analysis on the technical specifications or technical research requirements to extract the risk sentence in the technical specifications was insufficient. Sacky et al. [??] developed a procurement system for construction projects. It was developed as a decision-making system for procurement professionals. In this study, as a system for EPC engineers and managers, it was designed to be used by experts with a lot of experience and practitioners with insufficient experience. Marzouk and Enaba [??] developed a DTA-CC model that extracts important keywords by applying text mining techniques for the contract analysis of construction projects.

Following the fourth industrial revolution, the use of big data using information and communication technology (ICT) such as smart sensors increased [??], and studies on evolved ML using big data have been conducted [??]. Additionally, the point of this study was that the number of studies that grafted ML into contract documents, that is, legal documents, was increasing [??]. In addition, recently, Fantoni et al. [??] attempted to automatically detect, extract, split and assign textual information in documents for high-speed train research. The study, which applied knowledge-based state-of-the-art Computational Linguistic tools, was published with AnsaldoBreda S.p.A. The methodology was implemented for the project. Shah et al. [??] developed a technique for extracting Non-Functional Requirements through machine learning. However, most technologies using ML were only developed for technology and were not implemented on the platform. In other words, use of the commercialized decision-making program by actual users was limited. Zhuang [??] implemented a simulation to predict marine oil pollution on a platform. Based on these cases, this research team analyzed the technical specifications of the EPC project and developed a risk decision support system targeting engineers and managers of EPC contractors.

3. Methodology and Model Development

3.1. Research Methodology and Overall Process

The process of this study was carried out in four steps as follows: (1) database construction for algorithm analysis; (2) algorithm development; (3) system application; (4) the verification of analysis results and the analysis of results. The detailed explanation of each step is described in ???, and this section provides a brief introduction to understand the overall research flow.

The first step is the data collection and application stage, suitable for the two algorithm modules (TRE and DPE). The data lexicon was built by collecting design risk keywords from the TRE module, and it was collectively called the technical risk lexicon (TRL). When analyzing technical specifications, the database derives a score calculated from the frequency of risk keywords and presents the results in order of risk severity. The process for building the TRL is introduced in detail in Section ???. Next, in the DPE module, two types of data were required. The first was basic data to establish the engineering standard for the equipment or structure, such as the vessel and instruments to be analyzed, that is, the SDP. The process for building the SDP is introduced in detail in Section ???. It is displayed in a table format and includes parameters and information about the
corresponding equipment. The second required data were collectively called the synonym dictionary of SDP, synonyms for the first basic data. It played a role in enabling the detection and extraction of synonymous words and similar words.

As the second algorithm development stage, the logic construction suitable for the two module characteristics and the ML technology were used accordingly. By applying the phrase matcher technique [? ] of NLP to the TRE algorithm, it was possible to analyze the results according to the severity of risk through the grouping of keywords and scoring based on frequency. The risk sentence was extracted by matching the risk keyword with the phrase in the technical specifications through the phrase matcher technology. In the DPE module, an algorithm that derives the result by comparing the standard with the design standard of the corresponding technical specifications was applied through the context manager technique developed in this research project. Context manager means the process of learning the parameters of SDP in the context and reducing the influence score as the relevance of SDP decreases, or learning new parameters of SDP. The SDP of the equipment selected by the user can be learned, which serves to provide the data of the desired information at high speed.

In the third system application stage, the analysis result was visualized by implementing the algorithm in a dashboard on the ML platform. In other words, the module user could view the two modules (TRE and DPE) after selecting the technical specification analysis module. After uploading the selected technical specification document files and performing the analysis, the results were visually confirmed, and quantitative analysis was performed on the screen. The detailed description of the system configuration of each module is discussed in ???, respectively.

In the last step, a pilot test was performed to verify the system performance of the implemented two modules. The performance results are presented in Section ???. Two EPC engineers (verification subjects) participated as a comparison test group with the module in the pilot test. The third-party verification method was adopted for the verification of the pilot test results. Two subject matter experts (SMEs) with over 15 years of experience participated as verification evaluators, and the SMEs who participated in the verification had experience in performing many projects. First, for the pilot test, one EPC project technical specification was selected separately from 25 technical specifications. The review process was conducted individually by each expert so that independent verification was carried out. The risk analysis detection rate of the system module and EPC engineer was calculated, and the analysis time was measured. Based on the verification results, it was prepared in the form of a table of results for each module, and the final result was displayed as a Detection Performance Comparison Index (DCPI). Figure ?? shows the research steps and process methodology of this study.

**Figure 1.** Research steps and process methodology.
3.2. Overview of Automated Design Risk Analysis Module

3.2.1. Design Risk Analysis Parameters

This section introduces a brief overview of technical specification analysis modules for an EPC project. The analysis system consists of two modules (TRE and DPE), and the parameters related to the input and output of each module are described in Figure 2. Input data is a technical specification file for the EPC project. The TRE module extracts design risk clauses. Phrase matcher technology is used for the algorithm, and the data type is TRL. The DPE module is used to extract numerical requirements, and the algorithm uses context manager technology. The data type consists of two types of SDP’s synonym dictionary. Output is a severity evaluation table and graph in the TRE module. In the DPE module, it is a parameter comparison table for equipment. For the DPE module, a system platform targeting EPC contractors was built, and the TRE module aimed to shorten the review time for technical specification risk clauses and improve the detection capabilities of those in charge of actual EPC projects. In addition, the evaluation factor is presented for use in project evaluation. Engineers laid the foundation for reducing the analysis time for numerical requirements in technical specifications and minimizing engineering errors and omissions through the DPE module.

Figure 2. ML-based design risk automatic analysis module parameters.

3.2.2. Design Risk Analysis Process

The system development in this study consisted of three stages. In Stage 1, the PDF (Portable Document Format) to be analyzed was converted into the standardized data in the pre-processing module embedded on the platform. Next, the pre-processed data were uploaded by selecting the technical specification to be analyzed. Stage 2 is a full-fledged algorithm-based analysis stage. It applied a series of processes in which a user uploads data, and the actual analysis is performed through the analysis process. In Stage 3, the analysis result was presented visually on the platform, enabling the use of the output. Figure ?? shows each stage of the technical specification analysis system, and Figure ?? shows the stage screenshots of the system platform. The details of Figure ?? are described in ???. Once the module’s framework was built, the specific logic for the algorithm was developed. After linking the logic with the ML technique, the visualization interface of the analysis result was developed. Section ?? describes the detailed algorithm development process and analysis method for TRE and DPE in Section ??.
3.3. Technical Risk Extraction (TRE) Module

The TRE module presents an algorithm that automatically extracts design risk clauses by analyzing technical specifications. Using the phrase matcher technique in NLP, the risk keywords in the technical specifications were automatically extracted, the extracted
sentences were presented, and the evaluation factor and histogram were presented through the analysis results in a step-by-step manner.

The analysis process and algorithm of the TRE module were described in the activity sequence. As metadata consisting of tokens in sentence units was required to analyze PDF files, the definition of the analysis target in the technical specifications input file and the database embedded in the algorithm were compared in the TRE module.

3.3.1. PDF Data Pre-Processing—Sentence Tokenization

Data Pre-Processing refers to the process of converting unstructured data into structured data. This process must be preceded in order to perform the TRE and DPE modules because all documents must be classified in a language that the computer can parse. First, text data were loaded from a PDF using Tika Package. Tika is the library of the programming language (Python) and is used as a function to load (extract) the text of a PDF file. The loaded data were classified through the Sentence Tokenizer method. When the unit of a token was a sentence, this task classified sentences in a corpus, which is sometimes called sentence segmentation. A corpus is a collection of language samples used to study a natural language. A token is a grammatically indivisible language element. Text tokenization refers to the operation of separating tokens from the corpus.

When the analysis was performed by applying sentence tokenization by the period of the sentence as in the usual Sentence Tokenizer method, as shown in Figure 5, the period already appeared several times before the end of the sentence. It was not suitable for the classification of sentences in technical specifications that contained periods to express belonging to sentences, such as “5.1.5”.

![Figure 5. Sentence Tokenizer sentence result example for automatic analysis of technical specifications.](image)

The rules were directly defined according to the sentence format of the technical specifications among ITB documents or how special characters are used within the corpus. Through the ITB Sentence Tokenizer, sentences were divided into units that indicated the affiliation of the device. A binary classifier was used to handle exceptions in which a period appeared multiple times. When a period was a part of a word, as in “5.1.5”, and when a period played the role of dividing an actual sentence, it was divided into two classes and the ML algorithm was trained. The separated sentences were saved as standardized data in the comma-separated values (CSV) format. As shown in Figure 5, “5.2 Design Pressure”, it was confirmed that the affiliation (number) of the sentence was distinguished. Based on the following results, the items of technical specifications were classified and applied to both modules. Figure 5 shows the results of the PDF standardization module.
3.3.2. Definition of Input File

The document to be analyzed was the technical specification, one of the bid documents obtained from the project owners at the bidding stage. The technical specifications for most EPC projects describe the engineering requirements for major equipment and parts required for the project.

3.3.3. Definition of Database—TRL Construction

Based on the project’s technical specifications collected from EPC contractors, documents were analyzed, and necessary data were collected. For this study, 25 technical specifications for various EPC projects such as construction and plant fields were collected, with case-based design risk keyword extraction from the collected technical specification documents, the participation of experts (SMEs) with more than 15 years of experience in EPC project execution and a review to establish the TRL. TRL requires a continuous update in connection with the number of technical specifications for EPC projects continuously accumulating. Since documents for analysis mainly consist of unstructured data, converting them into structured data that a computer can recognize is required. Next, we look at the database formation process included in the algorithm based on the collected technical specification documents.

Lexicon terms means embedded data in the form of a database within the algorithm. This TRE module is configured in the form of a lexicon, that is, a lexicon for risk terms for design risk keyword extraction. TRL has the structure shown in Table ??.

In the 25 technical specifications, about 450 technical risk vocabularies were collected for each type of work, such as machinery, electricity, instrumentation, civil engineering, architecture, firefighting, HVAC and plumbing. As for the risk vocabulary, the advisory team was formed of three experts with over 15 years of experience in EPC projects, and they selected risk sentences that could affect the project by comprehensively judging the upstream and downstream engineering process. Additionally, the severity was classified into three groups based on the risk keyword. The classification method described in Dumbravă et al. [? ] was utilized as the evaluation method of the Impact Matrix.

Based on the review of five SMEs, they were classified into three severity groups and scored, and the classification criteria are as follows. The HH (High Impact/High Probability) group had the keywords that SMEs experts judged have a high probability of risk occurrence and significant impact. The HM (High Impact/Medium Probability) group was classified as the keywords that have a high risk or can have a significant impact but have a relatively lower risk occurrence than the HH group. Finally, the MM (Medium Impact/Medium Probability) group was classified into the factors with a degree of risk but relatively little impact. The H/H group had a risk score of 3 points, the H/M group had a risk score of 2 points, and the M/M group had a risk score of 1 point. As a risk classification criterion, the Impact Matrix was referred to, as shown in Figure ??.
in group A were the H/H group, the keywords included in group B were the H/M group, and the keywords included in group C were defined as the M/M group.

Table 1. Technical risk lexicon (example).

| Clause Type | High Impact/High Probability (3) | High (Medium) Impact/Medium (High) Probability (2) | Medium Impact/Medium Probability (1) |
|-------------|----------------------------------|-----------------------------------------------|-------------------------------------|
| all         | unless otherwise specified       | unless otherwise mentioned                    | in compliance with                   |
| throughout  | approved by                      | approved by                                   | in accordance with                   |
| owner       | not exceed                       | not exceed                                    | according to                         |
| Project Owners | not applicable                   | not applicable                                | shall comply with                   |
| no additional cost | not permitted                   | not permitted                                 | shall submit                        |
| by the bidder | not allowed                     | not allowed                                   | Discrepancy                         |
| by the Contractors | but not limited                 | shall not                                     | Still                                |
| Contractors shall include | permit                         | shall not                                     | Even                                 |
| Contractor’s responsibility | confirm to                     | shall not                                     | Except                               |
| Contractors shall provide | prohibited                     | permit                                        | Allowable                            |
| without any change order | shall be determined             | permit                                        | Adequate                             |
| un-priced | existing                         | permit                                        | even under                          |
| no-priced | modification                      | permit                                        | any other                           |
| contract price | shall be determined             | confirm to                                    | greater than                        |
| under Contractors scope | full                          | prohibited                                    | larger than                         |
| Project Owner’s approval | exceed                         | modifyed                                      | more than                           |
| without permission | exceed                         | modifyed                                      | less than                           |

1 See Impact Matrix for Technical Risk Keyword Classification.

Figure 7. Impact matrix for classifying technical risk keywords.

In the TRE module, the evaluation factor of the EPC project is presented in the summary results output as an evaluation index, and it is presented to users as an index to evaluate the project’s risk level. The evaluation factor is the value obtained by dividing the “total risk score” by the “total number of extracted clauses” and it is the core information of the summary results output.

The range designation of the evaluation factor was performed through the normalization process. As shown in Table ??, the range was specified through discussion by five SMEs based on the average values of the evaluation factors analyzed from 25 technical specifications. The average and standard deviation of the evaluation factors were 0.822 and 0.375 for the 23 technical specifications after excluding the minimum (0.26) and the maximum (2.71) values from the 25 technical specifications based on the trimmed average law. The five SMEs reviewed the 23 evaluation factors and determined the ranges. If the
evaluation factor exceeded 1.0, it was high risk, a value between 0.6 and 1.0 was defined as medium risk, and a value less than 0.6 was defined as low risk.

Table 2. Normalization result of evaluation scores for 25 projects.

| EPC Project ID | Sum of Risk Score | Number of Risk Sentence | Number of Sentence | Evaluation Factor |
|---------------|-------------------|-------------------------|--------------------|------------------|
| 1             | 3440              | 1245                    | 3183               | 1.08             |
| 2             | 1205              | 283                     | 444                | 2.71             |
| 3             | 1280              | 366                     | 895                | 1.43             |
| 4             | 881               | 394                     | 1537               | 0.57             |
| 5             | 1055              | 437                     | 2021               | 0.52             |
| 6             | 138               | 70                      | 536                | 0.26             |
| 7             | 1260              | 464                     | 1236               | 1.02             |
| 8             | 263               | 118                     | 592                | 0.44             |
| 9             | 237               | 92                      | 638                | 0.37             |
| 10            | 655               | 215                     | 637                | 1.03             |
| 11            | 972               | 408                     | 1596               | 0.61             |
| 12            | 2868              | 986                     | 2751               | 1.04             |
| 13            | 315               | 147                     | 712                | 0.44             |
| 14            | 1189              | 327                     | 764                | 1.56             |
| 15            | 3256              | 1098                    | 2878               | 1.13             |
| 16            | 259               | 94                      | 613                | 0.42             |
| 17            | 1276              | 522                     | 2391               | 0.53             |
| 18            | 2966              | 1083                    | 3004               | 0.99             |
| 19            | 486               | 137                     | 782                | 0.62             |
| 20            | 1237              | 592                     | 2177               | 0.57             |
| 21            | 267               | 95                      | 702                | 0.38             |
| 22            | 826               | 378                     | 728                | 1.13             |
| 23            | 1462              | 407                     | 974                | 1.50             |
| 24            | 1482              | 687                     | 1601               | 0.93             |
| 25            | 1017              | 521                     | 1753               | 0.58             |
| Total         | 30,292            | 11,166                  | 35,145             | 18.91            |

Mean of the Evaluation Factors 0.8220
Standard Deviation of the Evaluation Factors 0.375

1,2 The maximum and minimum values of were excluded from the calculation of mean and standard deviation.

3.3.4. TRL Terms

First, in the TRE module, the TRL was embedded in the system. It is composed of a list of keywords called terms in coding and of groups A, B, and C according to the degree of risk. The input file to be analyzed in correspondence with the corresponding terms was a PDF standardized technical specifications file. It is composed of a CSV file and is broken down into sentences. Breaking sentence by sentence is called sentence tokenization, and in the next step, keywords were matched through tokenization of words or phrases.

3.3.5. Word Tokenization—Count Vectorizer

In this case, the count vectorizer function of the feature-extraction sub-package of Scikit-learn was used to tokenize the sentence and divide it into words or phrases. This process is called word tokenization. Since the concept of this module is to automatically extract the risk clause through the frequency and severity score of the technical risk keywords, expressing the frequency as a vector value is necessary. This process is called word counting. Count vectorizer is used to generate count vectors, a Python class that converts documents into the Token Count Matrix [7]. The concept diagram of the TRE module shows the process of the word counting by count vectorizer and group counting, scoring, and sorting by phrase matcher in Figure ??.
Figure 8. Flowchart of the algorithm concept diagram of the TRE module.

3.3.6. Word Dictionary—Count Vectorizer

Count vectorizer learns all the words in the corpus and creates a count vector by matching each word to the number of times it appears in each sentence. The generated vectors are derived as a dictionary result, as shown in Figure 9. The dictionary is a data structure that can store key-value type values. Through this process, the keywords of each sentence are converted into a dictionary and can be used for analysis.

Figure 9. Corpus count vectorizing.

3.3.7. Grouping—Phrase Matcher

Phrase matcher is a spaCy technique that learns matching patterns and creates rule-based matches. Using the phrase matcher technique efficiently matches many keywords. An input file, that is, the data list loaded by default from a code is transformed into tokenized texts. Some languages, including English selected in this study, consist of both lowercase and uppercase letters. By designating the lowercase property in this transforming process, the phrase matcher technique can match without distinction between lowercase and uppercase letters. The group counting step extracts word frequencies from count vectors and grouping them into groups. The number of keywords for each sentence is
calculated from the count-vectorizing result, and the keyword frequency is calculated by grouping the keywords into groups. Phrase matcher refers to counting the number of occurrences by matching the TRL Terms stored in the system to each sentence of the technical specification documents. Match results are derived, for example, A (1), B (3), C (2), depending on the sentence.

3.3.8. Score Calculation and Sorting—Phrase Matcher

In the scoring and sorting step, the total score of the sentence was determined by the score of each group. Then, the ranking was given and sorted according to the score. As for the score for each group, according to the TRL as mentioned above, group A (HH) was given 3 points, group B (HM) 2 points, and group C (MM) 1 point. For example, in the case of sentences resulting from a grouping of A(1), B(3), and C(2), the total score was A-related 3 points, B-related 6 points, and C-related 2 points, resulting in a total risk score of 11 points.

3.3.9. Analysis Result Output—Output

The sorted sentences were calculated as the following four data frames. The TRE analysis module provides a summary table for four analysis results (main result, summary result, histogram data, and histogram image).

As for the output of the TRE module, the analysis result was derived as an Excel file. Based on the following results, the severity of the risk sentence was scored and displayed, and the risk sentence according to the ranking, the affiliation of the sentence, and the frequency of keywords could be checked. As the next output, the summarized result was presented as an Excel file, and the total number of sentences, the number of extracted risk sentences, and the evaluation factor were calculated from the result. In addition, a histogram of the analysis result was created in the form of an image, and the information of the histogram was also output and could be checked. The analysis result table among the total four outputs was implemented as an analysis result screen, as shown in Figure ???. The remaining three outputs are displayed on one screen. The summary table, histogram, and histogram information are displayed on the result summary screen, as shown in Figure ???.

![Figure 10. TRE analysis result screen.](image-url)
3.4. Design Parameter Extraction (DPE) Module

The DPE module analyzes the requirements and ranges of each equipment parameter by comparing the SDP of the technical specifications embedded in the system with the design standards of the object to be compared and presents the comparison results. Through the function of detecting design errors for numerical requirements in the technical specifications, it was confirmed whether the selected equipment met the minimum requirements. The comparison results present TRUE if the analysis result falls within the standard specification and its category, and FALSE if the result of the specification does not apply or deviates from the standard specification. This not only shortens the comparison time but also allows a user to identify various engineering requirements that may be omitted. In addition, by providing the ability to edit SDP data according to the characteristics of a user’s demand, the purpose of this was to prevent the design risk of contractors in advance and to establish a project-specific DB.

The basic data of the DPE module analysis algorithm were constructed in two ways (SDP or synonym dictionary). The first data were the SDP for equipment used in the EPC project. SDP are the standard data for comparative analysis and are composed of a data table format, as shown in Table ???. Table ?? shows pressure vessel equipment among plant equipment, and Table ?? shows SDP for instrument equipment, respectively. Parameter (PRM) consists of three steps and is defined in order of importance. PRM1 stands for Definition. Basic input data required for analysis include “design temperature,” “minimum thickness,” and “design pressure.” PRM2 can be a component of equipment. Like “Shell” or “Head,” it means each part of the pressure vessel. PRM3 is a sub-element to additional input information such as environment and status. Next, the attribute to designate the range of values and the range of technical specifications are presented, and finally, the unit for the number is input.

![HISTOGRAM RESULT](image)

**SUMMARY RESULT**

| SUMMARY ITEM          | QUANTITY |
|-----------------------|----------|
| Total Risk Score      | 3424     |
| Number of Risk Sentences | 1258    |
| Number of Sentences   | 3285     |
| Evaluation Factor     | 1.04     |

**[NOTE]**

* Project Evaluation Score = Risk Score / Total Project Sentence
* Project Evaluation Score Range
  - High > 1.0
  - Medium = 0.6 ~ 1.0
  - Low < 0.6

**Figure 11.** TRE result summary screen.
Table 3. SDP of pressure vessels in mechanical process discipline (excerpt).

| PRM1 Definition | PRM2 Component | PRM3 Sub Element | Attribute | Range 1 | Range 2 | Unit |
|-----------------|----------------|------------------|-----------|---------|---------|------|
| Design Temperature | Service | Chemical | greater than | 120 | - | °C |
| | | Steam | greater than | 120 | - | °C |
| | | Wet sour | max | 200 | - | °C |
| | | Chemical | greater than | 1.7 | - | MPa |
| | Unfired Steam Boiler | | greater than | 0.34 | - | MPa |

Table 4. SDP of 4 pieces of equipment in the instrument process.

| Equipment | PRM1 Definition | PRM2 Component | PRM3 Sub Element | Attribute | Range 1 | Range 2 | Unit |
|-----------|-----------------|----------------|------------------|-----------|---------|---------|------|
| Control Valve | Noise Level | Control Valve | | max | 85 | | dB |
| | Pressure Drop | | Flow Condition | at least | 50 | | % |
| | | | Maximum Flow | max | 1.1 | | time |
| | Nominal Size | Globe And Ball Valve | | at least | 1 | | in |
| Actuator | Operating Range | Diaphragm Actuator | | max | 4 | | MPa |
| | | Piston Actuator | Air Supply | at least | 4 | | MPa |
| Transmitter | Accuracy | Pneumatic Transmitter | Full Scale | at least | 1 | | % |
| | | Electronic Transmitter | Full Scale | at least | 0.25 | | % |
| Gauge | | Local Gauge | Full Scale | max | 1 | | % |

After the process of analysis and review by experts for each type of work, a hierarchy for a total of 10 types of equipment was constructed, as shown in Figure 12. The prepared thesaurus is shown in Table ??. SDP parameters are sometimes expressed differently in technical specifications of projects. Although the meaning is the same, they are sometimes used in various expressions. A synonym dictionary was built to secure the accuracy and reliability of the analysis. The two embedded data (SDP and synonym dictionary) were used for context manager analysis.

Figure 12. Hierarchy of standard design parameters—10 pieces of equipment.
Table 5. Synonym dictionary of DPE.

| SYM Category | SDP Word               | SYM Word         |
|--------------|------------------------|------------------|
| P (Parameter)| Design Metal Temperature| MDMT             |
| P            | Stainless Steel        | STS              |
| P            | Minimum Thickness      | smallest thickness|
| P            | Design Temperature     | Design temp      |
| A (Attribute)| Max                    | maximum          |
| A            | not be greater than    |                  |
| A            | at least               | minimum          |
| A            | or more                |                  |
| U (Unit)     | %                      | percent          |
| U            | °C                     | degree C         |
| U            | Mpa                    | N/m²             |
| U            | dBA                    | dB(A)            |

In the DPE algorithm, the user can select the equipment so that SDP data can be managed through the Database (DB). Data management through the DB shortens the analysis time of the selected equipment and facilitates management when data are converted or added later by the user through the SDP editing function. The DB is provided through the Database Management System (DBMS) management system built through the open source of MySQL. MySQL is the world’s most widely used relational DB management system. The digitized data are stored in the cloud using the MySQL DB service. DB construction for the DPE module created one database and two lower tables (SDP and synonym dictionary). As shown in Figure ??, the conditions for each table column are specified differently depending on the property.

- PK: Primary Key. Is it a property that can distinguish each row?
- NN: Not Null. Should it not be blank?
- UN: Unsigned type. Is it unsigned data?
- AI: Auto Increment. When checked, a larger value of 1 is automatically set for each row.

The data of each table were applied by loading the CSV files of SDP and synonym dictionary. It was confirmed that the SDP data were activated as a DB table and applied, as shown in Figure ??.
In order to utilize the above SDP and synonym data for analysis, the analysis code must access the server and receive data. A database server and a client communicate in a language called SQL (Structured Query Language). When a client sends a query (request), the server can create, modify, delete, and print data accordingly.

Since data must be received from the server, the corresponding SELECT statement is sent as a query. For the user account function, only the row whose USER_ID column value is the same as the current user ID is selected and loaded. The user account function is a function that allows users to modify the DB as a function to improve the usability of future platform users. Through this, it became possible to build a customized DBMS according to user demand. Figure 14 shows the algorithm concept diagram of the DPE module. The SDP and synonym dictionary construction process was previously described, and the description of the context manager applied as the analysis method of this module is described below.

In order to select SDP as the internal data clarification process of the DPE module, it is necessary to only leave the data corresponding to the work type and equipment specified...
in the checklist. Next, SDP and part of synonym data are reprocessed using the internal dictionary among the data types of the programming language (Python). Through this, the execution time can be shortened when searching for keywords or synonyms in sentences. The plural form of parameter was added as a synonym in the dictionary. Furthermore, sentence pre-processing, the step of generating an N-gram that views a phrase in which \( n \) tokens are bundled as a unit through the tokenization process, was performed. Section ?? to Section ?? are the same pre-processing steps described in Section ??.

3.4.1. Keyword Extraction

In the keyword extraction stage, keywords include parameter, attribute, range, and unit. Range is extracted with Python’s regular expression, and the rest are extracted using N-gram and Dictionary. The extracted keywords were used for the following algorithm analysis.

3.4.2. Context Update

In the context update stage, analysis time could be reduced because only the SDP data related to the sentence were selectively analyzed. The relevant DPE data were selected using the parameter context data. The context analysis algorithm was improved so that sentence analysis and comparison was possible even if there were no parameters in one sentence. Sentences containing keywords of SDP were converted into context scores and expressed according to their influence on the relevant paragraph. In calculating the context score, parameters 1, 2, and 3 were learned as context keywords and applied when calculating the context score of a paragraph. Up to three parameters were trained for each parameter type. The context analysis algorithm was designed so that the influence of the parameter could be extended to several sentences beyond one sentence. Therefore, even if all parameters in the sentence are not included, it is possible to perform a comparative analysis with SDP by checking the context.

In the context updating process, if the SDP keyword of the context is duplicated or if the keyword of the SDP of the same device is detected, the SDP extraction score of the corresponding context is increased. When the SDP keyword is detected in the sentence, it is added to the context keyword data, and a score is given to the paragraph. This score decreases each time the sentence is changed, and when the score reaches 0, the corresponding SDP keyword is deleted from the context data. Once the SDP keyword appears, the influence of the SDP keyword is maintained for several sentences after that. This context data were used to compare and analyze the attribute, range, and unit (ARU) of the embedded SDP parameters in the subsequent process.

3.4.3. ARU Set Selection

Each of several attributes, ranges, and units included in the sentence are grouped in the ARU Set Selection stage. The range may be two. For example, if there was a sentence “The length is greater than 1 m, and the weight is 5 kg maximum.” they were grouped into “greater than, 1 m” and “maximum, 5 kg”, respectively.

3.4.4. Standard Selection Context

In the Standard Selection Context step, SDP criteria are selected. The row containing the parameter stored in the context data in the embedded SDP Table is found. As one of the sentence conditions, if even one definition of range or unit for numerical requirements is omitted in the sentence, the analysis is performed except for the sentence. It has the advantage of increasing the analysis accuracy and speeding up the analysis.

3.4.5. Context Manager

The context manager technology developed in this module is applied from keyword extraction to the Standard Selection Context process. Only relevant SDP data were selected and analyzed through the context manager, and analysis was not performed if certain
conditions were not met. Through this, it was possible to remove unnecessary extraction results during comparative analysis and shorten the extraction time. The context manager technique was developed based on the concept of supervised learning in this study. Context manager sets the standard design parameters with fixed numerical values or ranges and uses them as standards to compare the input variables extracted from the target technical specifications in the DPE module. For example, in the clause, “pressure vessel’s pressure is 150 MPa”, context manager sets “150 MPa” as the fixed value of the standard design parameter “pressure” in the model training process. When the DPE module extracts design parameters and their values from the target technical specifications, context manager will compare the values in the design parameters with the fixed values of the SDP in the model and generates a result report. If context manager finds a value higher than “150 MPa” as the SDP, “pressure” in a target technical specification, context manager considers the clause to contain the design risk and includes it in a result report.

3.4.6. ARU Comparison

In the ARU Comparison step, a criterion to be automatically extracted from the sentence and compared with the selected the ARU set was selected from the embedded SDP Table. It included the context data or the position of the parameter in the sentence. The result value was derived by comparing the criteria selected in the ARU Set Selection step and the Standard Selection Context step with each other. It shows TRUE if the extracted range is included in the standard range, otherwise FALSE. It repeats the process from Keyword Extraction to ARU Comparison until the analysis is complete for all sentences in the technical specification. After analysis, the data frame is formed as the output. The data frame is displayed as an SDP comparison table in an Excel format. A list of equipment for each type of work is constructed. It can be confirmed that the parameters of the corresponding equipment are shown, and the standard ARU and the ARU of the analysis target are shown in turn. The corresponding output is shown on the analysis result screen, as shown in Figure 16.

![Figure 16. Result screen of the DPE analysis.](image)

4. Model Validation through a Case-Study

In this section, model validation was performed to evaluate the extraction accuracy of the development model (TRE and DPE) related to the technical specifications and to review the applicability in the EPC project. In this study, in order to secure the reliability and accuracy of the two modules (TRE and DPE), engineers (verification subjects) and SMEs (verification evaluators) with knowledge and performance experience on technical
specification analysis were involved in developing the module. The paper referenced for establishing the verification methodology and the definition of the evaluation index is in Lee et al. [7, 8]. Kang et al. [9] compared the performance results of this verification module compared to humans.

4.1. Validation of Model Design

The research team conducted a pilot test on the development model to evaluate the extent to which the two developed modules (TRE and DPE) quantitatively reached the level of detection and analysis results by understanding unstructured documents. A method of comparing the efficiency of risk detection and analysis among engineers with experience in EPC project execution compared to the ML-based risk analysis module (TRE and DPE) was adopted. According to the following procedure, verification of the module’s performance compared to the analysis performance of engineers was carried out.

4.1.1. Validation Methods

The five SMEs participated in order to establish the TRL and SDP at the beginning of this study. In the verification process, the risk analysis capabilities of the engineers and the developed models were tested. Two EPC engineers (Engineers A and B) conducted the risk analysis in the pilot experiment method without prior information on this study. Two SMEs (SME A and B) reviewed the risk analysis results made by the two engineers and the by the models in the model validation process. The SMEs A and B are experts who participated from the beginning of the study and have sufficient system knowledge required for fair validation. SME A reviewed the risk analysis result made by Engineer A and SME B reviewed the report made by Engineer B independently.

First, two engineers who were currently carrying out the EPC project participated as verification subjects to compare the test group with the module. The third-party verification method was adopted for the verification of the pilot test results. To this end, two SMEs with more than 15 years of experience participated as verification evaluators, and as shown in Table 6 below, the SMEs who participated in verification have experience in performing many projects, bidding, and engineering work. The experts participated through collaboration in the construction of the SDP. They participated from the early stage of the development of this research control so that detailed technology, including severity analysis, could be reviewed and verified.

Table 6. Information of verification participants.

| Expert    | Project Experiences | Specialty                        |
|-----------|---------------------|----------------------------------|
| SME A     | 20 yrs.             | ITB analysis/Engineering Management |
| SME B     | 17 yrs.             | ITB analysis/Project Management  |
| Engineer A| 2 yrs.              | EPC Project Engineer             |
| Engineer B| 5 yrs.              | EPC Project Engineer             |

The review process was conducted individually by each expert so that independent verification could be made, and the pilot test was performed sequentially according to the steps shown in Table 7.

4.1.2. Validation Data

Representative data for case verification is found in the ‘Offshore Plant EPC Project’ (Fixed platform India), which is one technical specification sheet of an EPC project in the offshore plant field. This study was also collected in addition to the 25 technical specifications collected for this study. As data for the verification process, it was selected based on the opinions of SMEs who directly performed the project. The same was applied to the TRE module and the DPE module.
Table 7. Description of verification for technical specification analysis module.

| Step | Verification Sequence | Description |
|------|-----------------------|-------------|
| 1–1  | Engineer’s Analysis   | One project’s technical specifications (PDF) were distributed to two engineers and results were delivered within a limited period (three days) (data distribution and result reply are conducted independently). With the same data distributed to the engineer (one project’s technical specification), this research team conducted the analysis using the system module (TRE and DPE). |
| 1–2  | Analysis of System Module | The above two analysis results were delivered to two SMEs to evaluate and verify the analysis process and content (the review process was conducted independently in a non-face-to-face manner). |
| 2    | Verification of SME    | The research team quantified the analysis results of SMEs, organized them, and evaluated the results. |
| 3    | Result                | |

1 Technical Risk Extraction module, 2 Design Parameter Extraction module.

4.1.3. Evaluation Index

In order to evaluate the verification result, it is important to select an appropriate evaluation index. The evaluation of IE results through NLP can be evaluated according to whether relevant information is extracted or not (irrelevant information). IE refers to the process of extracting information matching the pattern by defining the type or pattern of information to be extracted in advance. In this study, the following four indicators were defined as evaluation indicators:

- Extraction Target: The TRE module is described as TRL, and the DPE module as SDP as an index that is the standard for risk judgment when extracting the risk of the relevant module.
- Extraction Validation: Among the total number of clauses (sentences) that were classified as risk clauses (sentences) by modules and engineers, the number of cases where non-dangerous sentences were incorrectly extracted as risks is also included. It consists of the number of risk extraction cases classified based on the Extraction Target as an actual dangerous sentence by verifying the SMEs of the risk extraction result. It was defined as extraction validation in the result table.
- Extraction Accuracy: As the accuracy ratio of risk sentences extracted by modules and engineers, it was defined as the value obtained by dividing the number of total risk extraction sentences from the number of extraction validation sentences of SMEs.
- Detection Performance Comparison Index: As an evaluation index of modules and engineers’ relative risk analysis performance, the DCPI was defined by the relative evaluation of the module’s extraction validation and engineers’ extraction validation values.

4.2. Validation Results

The definition of the verification result is as follows. Intersection Risk Extraction Quantity means the risk (for example, technical risk clause or numerical requirement) commonly extracted by the engineer and system module. Difference Risk Extraction Quantity means the difference set between the engineer and the system module. SME only performed verification for the analysis results of the engineer and the system module, respectively. SME A verified the analysis result of Engineer A, and SME B verified the analysis result of Engineer B. The detailed quantitative verification results of the modules are as follows.

4.2.1. Validation of the TRE Module

The verification process is as follows. First, the ‘Offshore Plant EPC Project’ (Fixed platform India) technical specification was selected to verify the TRE module. The prepared project technical specifications were uploaded to the TRE module. The number of design risk detections after analysis was defined as risk extraction. In the case of the engineer, the sentence determined as a risk clause by analyzing the risk of the sample technical
specifications was defined as risk extraction, and the time taken for analysis was measured. The results of risk extraction of the two groups were reviewed for consistency by SMEs, and extraction validation values were derived based on this. The extraction validation value for total risk extraction was presented as the extraction accuracy value. Additionally, for performance evaluation, the number of extraction validation sentences of the module and the engineer was compared and expressed as the DCPI. Table ??, below, and in the following Section ?? show the TRE detection accuracy validation results through the following process.

Table 8. The accuracy results of the TRE module’s risk sentence extraction.

| Test Target   | Risk Sentence Extraction Quantity (Per Test Project) | Extraction Validation | DPCI (Detection Performance Comparison Index) |
|---------------|-----------------------------------------------------|-----------------------|---------------------------------------------|
|               | Total | Intersection | Difference | SME A | SME B | SME (A + B)/2 | Extraction Accuracy |                        |
| TRE Module    | 334   | 243         | 91         | 301   | 303   | 302          | 90%                  | 1.2                       |
| Engineer A    | 256   | 243         | 13         | 243   |       | 250.5        | 95%                  | 0.8                       |
| Engineer B    | 262   | 243         | 19         |       | 258   |             | 98%                  |                           |

The test target was the extraction result value using the TRE module and the extraction result value by the two engineers. Risk sentence extraction quantity consists of the risk extraction results of one technical specification such as total, intersection, and difference, and the extraction validation was proceeded in the order of verification by the SME. The SME’s validation values were compared, analyzed, and expressed as the DCPI.

Looking at the total (one project) value, 334 risk sentences were detected by the TRE module, 256 by Engineer A, and 262 risk sentences by Engineer B. The TRE module and engineers commonly extracted 243 technical risk sentences. Subtracting the intersection sentences from the total number of risk extraction sentences yields 91 risk sentences for the TRE module, 13 for Engineer A, and 19 for Engineer B. For the verification of the module, the average of the verification values of two SMEs was used. The results of Engineer A were reviewed by SME A, and the results of Engineer B were reviewed by SME B. A total of 301 and 303 sentences were extracted as risk sentences by SME A and SME B, respectively. The average of the number of the extracted sentences by SME A and B was 302 sentences, which showed extraction accuracy of 90 percent compared to the total value. The verification result of Engineer A was 254 sentences, showing 95 percent extraction accuracy compared to the total value of 256 sentences. The verification result of Engineer B was 258 sentences, showing 98 percent extraction accuracy compared to the total value of 262 sentences. The engineer’s extraction accuracy was about 7 percent more accurate than the TRE module. However, looking at the SME validation result, the TRE module’s risk extraction result was 302 sentences, and the engineer’s risk extraction result averaged 250.5 sentences. When analyzing the TRE module, about 50 additional risk sentences were detected. In terms of the detection rate of the risk sentences, it was confirmed that the DPCI was as high as 1.2 for the TRE module and 0.8 for the engineers. Figure ?? shows three example sentences out of 50 missing sentences. The reason for the omission regarded documentary factors, such as when the sentence was included in a detailed item or when the sentence was a complex sentence. In addition, environmental factors due to limited time and lack of experience also played a role.
Based on the above results, the analysis time required for the TRE module and the analysis time of Engineers A and B according to the risk analysis of one representative technical specification were evaluated. In the risk analysis of two engineers, the risk keyword was first checked, the total risk score was calculated, and the number of corresponding risk sentences was extracted. The analysis results of the TRE module and engineers are shown in Table 9.

**Table 9. Analysis time results of the TRE module.**

| Division   | Total Project Sentence | Risk Sentence | Risk Score | Evaluation Score | Time Used 1 |
|------------|------------------------|---------------|------------|------------------|-------------|
| TRE        | 3285                   | 1258          | 4324       | 1.04             | 0.5 h       |
| Engineer A | 3285                   | 1250          | 4200       | 1.28             | 45 h        |
| Engineer B | 3285                   | 1270          | 4100       | 1.25             | 40 h        |

1 Time taken to extract risk sentence and risk score.

Engineer B, with five years of experience in the related field and two years of experience as an engineer, had different criteria for judging the risk according to the knowledge and experience of each engineer in the process of extracting technical risk clauses. It was found that the analysis result of Engineer A may be different. In addition, when comparing the analysis results of the TRE risk module, the ML-based automatic TRE risk extraction model not only shows a relatively high-risk extraction detection result but also the time required for the module is more than 80 times that of 0.5 h compared to the average 42.5 h of engineers.

4.2.2. Validation of the DPE Module

Validation was performed on the DPE module. The procedure is the same as the above TRE module, and the extraction accuracy of the module was checked according to whether the numerical parameter was detected. Table ?? and Figure ?? show the results of the risk extraction analysis of the numerical requirements of the DPE module.

**Table 10. Technical specification DPE module risk requirements extraction accuracy results.**

| Test Target | Risk Sentence Extraction Quantity (Per Test Project) | DPCI (Detection Performance Comparison Index) |
|-------------|-----------------------------------------------------|----------------------------------------------|
|             | Total | Intersection | Difference | Validation | Extraction Accuracy |                             |
| DPE Module  | 252   | 187          | 65         | SME A | 238 | 240 | 239.5 | 94% | 1.1 |
| Engineer A  | 223   | 187          | 36         | SME B | 221 | 225 | 225 | 99% | 0.9 |
| Engineer B  | 231   | 187          | 44         | SME (A + B)/2 | 229 | 99% | | | |
4.2.2. Validation of the DPE Module

Validation was performed on the DPE module. The procedure is the same as the above TRE module, and the extraction accuracy of the module was checked according to whether the numerical parameter was detected. Table 10 and Figure 18 show the results of the risk extraction analysis of the numerical requirements of the DPE module.

### Table 10. Technical Specification DPE Module Risk Requirements Extraction Accuracy Results.

| Test Target | Risk Sentence Extraction Quantity (Per Test Project) | DPCI (Detection Performance Comparison Index) | Validation Extraction Accuracy |
|-------------|-----------------------------------------------------|---------------------------------------------|--------------------------------|
| DPE Module  | 252 187 65 238 241 239 5 94% 1.1                   |                                             |                                |
| SME A      | 223 187 36 221  225 99% 0.9                        |                                             |                                |
| SME B      | 231 187 44  229 99%                                  |                                             |                                |

A total of 252 risk sentences were detected by the DPE module, whereas Engineer A and B detected 223 and 231 risk sentences, respectively. The number of common risk sentences extracted by the DPE module and two engineers was 187. There was a difference between the excluding intersection values and the total value. A total of 65 risk sentences were detected by the DPE module, 36 by Engineer A, and 44 by Engineer B. The module verification result was 238 for SME A and 241 for SME B, indicating that an average of 239.5 risk sentences was detected by SME A and B. The verification result of Engineer A was performed by SME A, and it turned out to be 221 sentences. The verification result of Engineer B was performed by SME B, and it turned out to be 229 sentences.

In the DPE module, as an extraction function of objective numerical requirements, the module’s extraction accuracy was 239.5 sentences compared to the total 252 sentences, which is 94 percent. The extraction accuracy of Engineer A was 99 percent, with 221 sentences compared to 223 sentences. The extraction accuracy of Engineer B was 99 percent, with 229 sentences compared to the total 231 sentences. It was found that the engineers showed an accuracy of 5 percent or more when analyzing the objective numerical requirements.

However, looking at the result of SME risk extraction, it was found that the DPE module extracted about 15 more sentences, with 239.5 sentences, than the average value achieved by engineers, 225 sentences. In the difference between these 15 sentences, there were sentences that the engineer missed due to time constraints and a lack of experience. Figure 19 shows three example sentences out of 15 missing sentences. There was a high possibility that the engineer omitted the numerical parameters in the sentences where the detailed items were listed rather than in a table. As a result, it was confirmed that the DCPI of the DPE module was 1.1, which is higher than that of the engineers’ DCPI of 0.9.

![Graph showing validation results](image)

**Figure 18.** The validation results of the TRE and DPE modules, (a) TRE Module Validation; (b) DPE Module Validation.

A total of 252 risk sentences were detected by the DPE module, whereas Engineer A and B detected 223 and 231 risk sentences, respectively. The number of common risk sentences extracted by the DPE module and two engineers was 187. There was a difference between the excluding intersection values and the total value. A total of 65 risk sentences were detected by the DPE module, 36 by Engineer A, and 44 by Engineer B. The module verification result was 238 for SME A and 241 for SME B, indicating that an average of 239.5 risk sentences was detected by SME A and B. The verification result of Engineer A was performed by SME A, and it turned out to be 221 sentences. The verification result of Engineer B was performed by SME B, and it turned out to be 229 sentences.

In the DPE module, as an extraction function of objective numerical requirements, the module’s extraction accuracy was 239.5 sentences compared to the total 252 sentences, which is 94 percent. The extraction accuracy of Engineer A was 99 percent, with 221 sentences compared to 223 sentences. The extraction accuracy of Engineer B was 99 percent, with 229 sentences compared to the total 231 sentences. It was found that the engineers showed an accuracy of 5 percent or more when analyzing the objective numerical requirements.

However, looking at the result of SME risk extraction, it was found that the DPE module extracted about 15 more sentences, with 239.5 sentences, than the average value achieved by engineers, 225 sentences. In the difference between these 15 sentences, there were sentences that the engineer missed due to time constraints and a lack of experience. Figure 19 shows three example sentences out of 15 missing sentences. There was a high possibility that the engineer omitted the numerical parameters in the sentences where the detailed items were listed rather than in a table. As a result, it was confirmed that the DCPI of the DPE module was 1.1, which is higher than that of the engineers’ DCPI of 0.9.

**Figure 19.** Example of design parameter sentences missing by the subject of verification.
5. Conclusions and Future Works

5.1. Conclusions and Summary

In this study, two algorithms for the automatic extraction and analysis of ML-based technical specifications were used in order to confirm the existence of design risk clauses for technical specifications that require prior analysis and reflection when bidding or carrying out EPC projects and to suggest project risk. A severity management technology module was developed. The first is an ML algorithm-based TRE module, which is technology used to automatically detect and analyze technical risk clauses that are prone to omissions or errors in bidding due to time constraints or a lack of personal competence. By comparing the numerical requirements of the technical specifications, the user can automatically detect the requirements of project owners that are different or over the range from the standard technical specifications.

Therefore, this study was carried out according to the following steps. First, the risk types and severity of the technical specifications were derived through the collection of previously conducted EPC project bidding or execution cases (Lessons and Learned). To this end, technical specifications of 25 EPC projects performed in the past were collected from project owners and classified by risk sentence type, and these sentences were used as basic data for model development.

Second, we developed an algorithm and model that can identify major risks of contracts through natural language processing of sentences included in technical specifications. For NLP of risk sentences, in the pre-processing data stage to remove unnecessary data, and in the algorithm stage, the logic was configured according to two module characteristics, and ML technology was applied. NLP’s phrase matcher technology was applied to the TRE algorithm. It is possible to extract results according to the grouping of keywords and risk severity by applying them. In the DPE comparison algorithm, a technique was developed to compare the standard criteria with the design criteria of the corresponding technical specifications through context manager technology and draw the results. The analysis result was visualized in the system application stage by making the algorithm a dashboard on the ML platform. If the pattern of the text in the technical specifications matches the developed rules, the mechanism for information extraction is activated, and the automatic risk extraction model of the technical specifications implemented through the Python programming language automatically extracts the risk sentence when the user enters the technical specifications and review so that users can make decisions based on the review results of the model.

Third, to enhance the reliability and verify the performance of the developed model, collaboration with EPC project experts was carried out from the beginning of development, and the technical specifications were reviewed by the subject of verification (engineer) and the technical specifications by the development model. The results were compared with each other. The experts who participated in the verification have about 15 to 20 years of experience performing EPC projects, and by excluding mutual discussion about the verification contents, independent verification was made.

Looking at the verification results of the module, the accuracy of the extracted risk sentences shows that the extraction accuracy of the participating verification subjects (engineers) is high, but the development module shows a higher detection rate for the total amount of risk extraction, that is, the number of risk extraction sentences. In other words, the engineer shows more than 95 percent accuracy in the sentences judged as being a risk, but because there are many missing risk sentences because they extract fewer sentences than the module; it is difficult to judge the risk of the entire document. However, in the development module, the accuracy of the risk extraction sentence exceeds 90 percent, but the total amount of risk sentences extracted is relatively higher than that of the engineer. In other words, it is more efficient to use a module with high-risk detection performance when judging the overall project risk. In the TRE module, the risk of 302 sentences was detected compared to the 250 sentences of the risk clause found by the engineer. As a result, the module’s DCPI was 1.2, and the engineer’s DCPI was 0.8, confirming that the
module’s performance was more than 50 percent superior. In the DPE module, the risk of 240 sentences was detected compared to the risk requirements of 225 sentences found by the engineer. As a result, the module’s DCPI was 1.1, and the engineer’s DCPI was 0.9, confirming that the module’s performance was more than 20 percent relatively superior. In addition, the total amount of extracted sentences and the time required for analysis could be significantly reduced. In other words, the risk clauses analyzed by engineers over an average of 42.5 h were analyzed in less than 0.5 h for modules.

As a result, the EMAP tool allows engineers with little experience in EPC projects to achieve the risk analysis result at the equivalent level of the experts with decades of experience in the EPC field. Some EPC contractors use their own Excel spreadsheets to calculate cost estimates or check design risks, but these applications are far from the comprehensive solution that could automatically identify multiple types of technical risks and summarize the risk levels in the reports. EPC engineers can view the comprehensive risk analysis results through the report generation function in the digitalized design risk analysis tool, EMAP, which integrates the multiple risk analysis functions and executes a series of automatic risk analyses seamlessly.

There were many cases of direct impact on project execution due to risks, including engineering errors and omissions due to limited time and a lack of experience. At the same time, engineers manually analyzed vast amounts of technical specifications in the preceding stage of the EPC project. Through the automatic risk analysis module for engineering documents, including technical specifications developed through this study, it is possible to automatically extract and analyze technical specification design risk factors based on consistency. The TRE module can support the management’s decision making by evaluating the project risk in the bidding stage, and the DPE module is expected to be usefully utilized by practitioners in analyzing design requirements in the field. It was developed to update specialized modules according to the characteristics of EPC contractors by field through the user editing function and the case-based data accumulation of modules.

5.2. Limitation and Future Research

Despite the above research achievements and academic and practical contributions, this project has several limitations in the following aspects. It is necessary to consider this when using this study’s results and conducting related research in the future.

First, when collecting data, most of the data were accumulated in this study based on technical specifications from EPC plant bids. Although the format and detailed requirements of the technical specifications were collected by country and project owners, there is a limitation in that the versatility is poor when analyzing other technical specifications. Such a problem is a common problem in rule-based information extraction models. In order to overcome this problem, the lexicon (TRL) vocabulary is continuously expanded, and through machine learning, rules are established by building data on technical specifications in various fields. Suppose the risks of the vast technical specifications in the plant field as well as in other industries can be efficiently extracted, and the data accumulated through ML can be applied to each other organically between EPC contractors. In that case, the loss can be reduced by referring to exemplary cases in various fields. We hope that this will serve as a basis for minimizing and sharing alternatives.

Second, the TRE module calculates the risk of the technical specifications based on the word and frequency determined to be risky. Therefore, it is not easy to analyze sentences and paragraphs because the risk judgment criteria are limited to keywords. For example, there is a possibility that an error of judgment may occur because only the presence or absence of the word itself is checked, not the meaning of the word used. The level of analysis can be improved in a variety of ways. For example, NLP’s Sentiment Analysis can be used to determine the degree of affirmation/negation of a sentence containing a risk word and reflect it in risk assessment.
Furthermore, it is also possible to consider developing a model that learns the technical specifications for each sentence and grasps the risk level of the sentence itself. In addition, the TRE module requires a continuous update of the TRL based on the data accumulated by experts and practitioners in the field. Through this, it seems that evaluation factors that can be used for each field will be established. In conclusion, it is necessary to share and collaborate on successful and unsuccessful EPC projects by constructing big data for TRL’s risk keywords and the development of algorithms.

Third, in the DPE module, if many parameters, attributes, ranges, and units are mixed in one sentence, the ARU selection process or the SDP selection process cannot be performed accurately, so the accuracy of the results may be lowered. Because it is a rule-based system, it can respond to the formats of some technical specifications used in the model development, but there is a limit in performance degradation when other types of technical specifications are input. Therefore, if the criteria suitable for each field are presented through the learned algorithm, it will be possible to develop a more versatile technology and overcome the limitations of the rule-based system.

Finally, as part of the technological innovation program supported by the Korean government for the development of artificial intelligence and big data, the development of the Engineering Machine learning Automation Platform (EMAP), an integrated decision-making support system based on ML, was completed in 2021. Cases such as GE’s Predix and DNV’s Veracity were referred to in order to establish the integrated system. The integrated system comprises three modules (ITB Analysis, Engineering Design, Maintenance Analysis), and this paper focused on the ITB analysis system. For EMAP, classification, regression, and deep learning algorithms can be selected for supervised learning, and clustering algorithms are supported for unsupervised learning. Users can prevent risks by analyzing bidding documents in advance through the ITB analysis module developed through the supervised learning algorithms and identify the severity of design costs, design errors, and design changes through the design analysis modules developed based on unsupervised learning algorithms in the EMAP system.

Author Contributions: Conceptualization, M.-J.P. and E.-B.L.; methodology, M.-J.P., S.-Y.L. and E.-B.L.; software, J.-H.K.; validation, E.-B.L., S.-Y.L. and J.-H.K.; data collection, M.-J.P. and S.-Y.L.; writing—original draft preparation, M.-J.P. and S.-Y.L.; writing—review and editing, E.-B.L.; visualization, J.-H.K.; supervision, E.-B.L.; project administration, J.-H.K. and E.-B.L.; 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 sponsored 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: Special thanks to Chang-Mo Kim and So-Won Choi for the academic feedback on this paper, and Sung Bin Baek for his support on the Python coding.

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

| Abbreviation | Description |
|--------------|-------------|
| ARU          | Attribute, range, unit |
| CSV          | Comma-separated values |
| DB           | Database |
| DBMS         | Database Management System |
| DCPI         | Detection Performance Comparison Index |
| DPE          | Design Parameter Extraction |
| EPC          | Engineering, Procurement and Construction |
| ICT          | Information and communication technology |
| IE           | Information Extraction |
| ITB          | Invitation to Bid |
| ML           | Machine learning |
| NLP          | Natural Language Processing |
| O&M          | Operation and Maintenance |
| PDF          | Portable Document Format |
| PRM          | Parameter |
| SDP          | Standard Design Parameter |
| SME          | Subject Matter Expert |
| SQL          | Structured Query Language |
| TRE          | Technical Risk Extraction |
| TRL          | Technical Risk Lexicon |

References

1. Kim, M.H.; Lee, E.B.; Choi, H.S. Detail Engineering Completion Rating Index System (DECRIS) for Optimal Initiation of Construction Works to Improve Contractors Schedule-Cost Performance for Offshore Oil and Gas EPC Projects. *Sustainability* 2018, 10, 2469. [CrossRef]

2. Shen, W.X.; Tang, W.Z.; Yu, W.Y. Causes of contractors’ claims in international engineering-procurement-construction projects. *J. Clin. Endocrinol. Metab.* 2017, 23, 727–739. Available online: https://journals.vgtu.lt/index.php/JCEM/article/view/1164 (accessed on 10 June 2021). [CrossRef]

3. Kim, H.J.; Choi, J.H. Development of a Conceptual Estimate Methodology for Plant Construction Projects. *Korean Construct. Eng. Manag.* 2019, 20, 141–150. [CrossRef]

4. Putra, G.A.S.; Triyono, R.A. Neural Network Method for Instrumentation and Control Cost Estimation of the EPC Companies Bidding Proposal. *Procedia Manuf.* 2015, 4, 98–106. [CrossRef]

5. Oevermann, J. Semantic PDF Segmentation for Legacy Documents in Technical Documentation. *Procedia Comput. Sci.* 2018, 137, 55–65. [CrossRef]

6. The American Society of Mechanical Engineers. Setting the Standard. Available online: https://www.asme.org/ (accessed on 15 June 2021).

7. Micheli, G.J.L.; Cagno, E.; Giulio, A.D. Reducing the total cost of supply through risk-efficiency-based supplier selection in the EPC industry. *J. Purch. Supply Manag.* 2009, 15, 166–177. [CrossRef]

8. Memon, A.H.; Rahman, I.A. Preliminary Study on Causative Factors Leading to Construction Cost Overrun. *Int. J. Sustain. Construct. Eng. Technol.* 2011, 2, 57–71. Available online: https://www.researchgate.net/publication/256662503_Preliminary_Study_on_Causative_Factors_Leading_to_Construction_Cost_Overrun (accessed on 17 May 2021).

9. Doloi, H.; Sawhney, A.; Iyer, K.C. Analysing factors affecting delays in Indian construction projects. *Int. J. Proj. Manag.* 2012, 30, 479–489. [CrossRef]

10. Kerur, S.; Marshall, W. Identifying and Managing Risk in International Construction Projects. *Bloomsbury Qatar Found. J.* 2012, 8, 1–14. [CrossRef]

11. Al Haj, R.A.; El-Sayegh, S.M. Time–Cost Optimization Model Considering Float-Consumption Impact. *J. Constr. Eng. Manag.* 2015, 141, 1–10. [CrossRef]

12. Kim, E.R.; Kim, S.R.; Kim, Y.G. A component-based construction process control system for increasing modifiability. *J. Inst. Internet Broadcast. Commum.* 2015, 15, 303–309. [CrossRef]

13. Wang, Y.; Le, Y.; Dai, J. Incorporation of Alternatives and Importance Levels in Scheduling Complex Construction Programs. *J. Manag. Eng.* 2015, 31, 3–11. [CrossRef]

14. Heravi, G.; Eslamdoost, E. Applying Artificial Neural Networks for Measuring and Predicting Construction-Labor Productivity. *J. Constr. Eng. Manag.* 2015, 141, 3–11. [CrossRef]

15. Alsharif, S.; Karatas, A. A Framework for Identifying Causal Factors of Delay in Nuclear Power Plant Projects. *Procedia Eng.* 2016, 145, 1486–1492. [CrossRef]
16. Johansen, A.; Landmark, A.D.; Olshausen, F. Time Elasticity—who and what Determines the Correct Project Duration? *Procedia Comput. Sci.* 2016, 100, 586–593. [CrossRef]
17. Gunduz, M.; Yahya, A.M.A. Analysis of project success factors in construction industry. *Technol. Econ. Dev. Econ.* 2018, 24, 67–80. [CrossRef]
18. Tripathi, K.K.; Jha, K.N. Application of fuzzy preference relation for evaluating success factors of construction organisations. *Eng. Constr. Archit. Manag.* 2018, 25, 758–779. [CrossRef]
19. Chang, C.J.; Yu, S.W. Three-Variance Approach for Updating Earned Value Management. *J. Constr. Eng. Manag.* 2018, 144, 2–14. [CrossRef]
20. Yi, D.K.; Lee, E.B.; Ahn, J.Y. Onshore Oil and Gas Design Schedule Management Process Through Time-Impact Simulations Analyses. *Sustainability* 2019, 11, 1613. [CrossRef]
21. Son, B.Y.; Lee, E.B. Using Text Mining to Estimate Schedule Delay Risk of 13 Offshore Oil and Gas EPC Case Studies During the Bidding Process. *Energies* 2019, 12, 1956. [CrossRef]
22. 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.* 2019, 31, 151–160. [CrossRef]
23. Zekri, M.; Zahaf, S.; Gargouri, F. Specification of the data warehouse for the decision-making dimension of the Bid Process Information System. *Procedia Comput. Sci.* 2019, 159, 1190–1197. [CrossRef]
24. Gunduz, M.; Almuajebh, M. Critical Success Factors for Sustainable Construction Project Management. *Sustainability* 2020, 12, 1990. [CrossRef]
25. Stetter, C.; Piel, J.H. Competitive and risk-adequate auction bids for onshore wind projects in Germany. *Energy Econ.* 2020, 90, 104849. [CrossRef]
26. Votto, R.; Ho, L.L.; Bersanetti, F. Applying and Assessing Performance of Earned Duration Management Control Charts for EPC Project Duration Monitoring. *J. Constr. Eng. Manag.* 2020, 146, 10–12. [CrossRef]
27. Ferrucci, D.; Brown, E. Building Watson: An Overview of the DeepQA Project. *AI Mag.* 2010, 31, 59–79. [CrossRef]
28. Lin, H.T.; Chi, N.W.; Hsieh, S.H. A concept-based information retrieval approach for engineering domain-specific technical documents. *Am. Enterp. Inst.* 2012, 26, 349–360. [CrossRef]
29. Wang, Z.H.; Zhan, W. Dynamic Engineering Multi-criteria Decision Making Model Optimized by Entropy Weight for Evaluating Bids. *Syst. Eng. Procedia* 2011, 5, 49–54. [CrossRef]
30. Shin, J.; Kwon, S.; Moon, D. A Study on Method of Vertical Zoning of Construction Lift for High-rise Building based on Lift Planning & Operation History Database. *KSCE J. Civ. Eng.* 2018, 22, 2664–2677. [CrossRef]
31. Li, Y.; Guzman, E.; Tsiamoura, K. Automated Requirements Extraction for Scientific Software. *Procedia Comput. Sci.* 2015, 51, 582–591. [CrossRef]
32. Winkler, J.; Vogelsang, A. Automatic Classification of Requirements Based on Convolutional Neural Networks. In Proceedings of the IEEE 24th International Requirements Engineering Conference (REW 2016), Beijing, China, 12–16 September 2016; pp. 39–45. [CrossRef]
33. Podolski, M. Management of resources in multiunit construction projects with the use of a tabu search algorithm. *J. Clin. Endocrinol. Metab.* 2017, 23, 263–272. Available online: https://journals.vgtu.lt/index.php/JCEM/article/view/966 (accessed on 21 June 2021). [CrossRef]
34. Lee, J.H.; Yi, J.S.; Son, J.W. Construction Bid Data Analysis for Overseas Projects Based on Text Mining -Focusing on Overseas Construction Project’s Bidder Inquiry. *Korean J. Constr. Eng. Manag.* 2016, 17, 89–96. [CrossRef]
35. Chozas, A.C.; Memeti, S.; Plana, S. Using Cognitive Computing for Learning Parallel Programming: An IBM Watson Solution. *Procedia Comput. Sci.* 2017, 108, 2121–2130. [CrossRef]
36. Saint-Dizier, P. Mining incoherent requirements in technical specifications: Analysis and implementation. *Data Knowl. Eng.* 2018, 117, 290–306. [CrossRef]
37. Abualhaija, S.; Arora, C.; Sabetzadeh, M. A Machine Learning-Based Approach for Demarcating Requirements in Textual Specifications. In Proceedings of the IEEE 27th International Requirements Engineering Conference (RE 2019), Jeju, Korea, 23–27 September 2019; pp. 51–62. [CrossRef]
38. 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]
39. Marzouk, M.; Enaba, M. Text analytics to analyze and monitor construction project contract and correspondence. *Autom. Constr.* 2019, 98, 265–274. [CrossRef]
40. Seong, Y.H.; Jung, K. A Study on the Applications of Information and Communication Technology for 4th Industrial Revolution in Safety and Health of Workers. *Korea Saf. Manag. Sci.* 2019, 21, 17–23. [CrossRef]
41. Al-Sahaf, H.; Bi, Y.; Chen, Q. A survey on evolutionary machine learning. *R. Soc. N. Z.* 2019, 49, 205–228. [CrossRef]
42. Sleimi, A.; Ceci, M.; Sannier, N. A Query System for Extracting Requirements-Related Information from Legal Texts. In Proceedings of the IEEE 27th International Requirements Engineering Conference (RE 2019), Jeju, Korea, 23–27 September 2019; pp. 319–329. [CrossRef]
43. Fantoni, G.; Coli, E.; Chiarello, F. Text mining tool for translating terms of contract into technical specifications: Development and application in the railway sector. *Comput. Ind.* 2020, 124, 103357. [CrossRef]
44. Shah, U.S.; Patel, S.J.; Jinwala, D.C. Constructing a Knowledge-Based Quality Attributes Relationship Matrix to Identify Conflicts in Non-Functional Requirements. *J. Comput. Tech. Nanosci.* 2020, 17, 122–129. [CrossRef]
45. Zhuang, Y. Simulation of economic cost prediction of offshore oil pollution based on cloud computing embedded platform. *Microprocess. Microsyst.* 2021, 83, 103993. [CrossRef]
46. PhraseMatcher, spaCy API Documentation; Explosion Software: Berlin, Germany; Available online: https://spacy.io/api/phonematcher (accessed on 20 July 2021).
47. Do, Q.A.; Bhowmik, T.; Bradshaw, G.L. Capturing creative requirements via requirements reuse: A machine learning-based approach. *J. Syst. Softw.* 2020, 170, 110730. [CrossRef]
48. Python; Python Software Foundation: Wilmington, DE, USA; Available online: https://www.python.org/ (accessed on 20 July 2021).
49. Garnier, M.; Saint-Dizier, P. Improving the Use of English in Requirements. *Int. Requir. Eng. Board* 2016, 6, 1–14. Available online: https://hal.archives-ouvertes.fr/hal-01446902 (accessed on 3 June 2021).
50. Dumbrava, V.; Severian Iacob, V. Using Probability—Impact Matrix in Analysis and Risk Assessment Projects. *J. Knowl. Manag. Econ. Inf. Technol.* 2013, 3, 1–7. Available online: https://ideas.repec.org/a/spp/jkmeit/spi13-07.html (accessed on 22 May 2021).
51. MySQL; Oracle Corporation: Austin, TX, USA; Available online: https://www.mysql.com/ (accessed on 21 June 2021).
52. 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. Civil Eng.* 2019, 33, 04019003. [CrossRef]
53. Lee, J.H. Development of construction risk extraction model for overseas construction projects based on natural language processing (NLP). Ph.D. Thesis, Ewha Womans University, Seoul, Korea, 2018.
54. 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]
55. Choi, S.W.; Lee, E.-B.; Kim, J.H. The Engineering Machine-Learning Automation Platform (EMAP): A Big Data-Driven AI Tool for Contractors’ Sustainable Management Solutions for Plant Projects. *Sustainability* 2021, 13, 384. [CrossRef]
56. 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. [CrossRef]
57. Predix: The Industrial IoT Platform; General Electronic: Boston, MA, USA; Available online: https://www.ge.com/digital/iiot-platform (accessed on 16 June 2021).
58. Veracity Platform; DNV: Sandvika, Norway; Available online: https://www.veracity.com/ (accessed on 16 June 2021).
59. Bezdan, T.; Stoean, C.; Namany, A.A.; Bacanin, N.; Rashid, A.T.; Zivkovic, M.; Venkatachalam, K. Hybrid Fruit-Fly Optimization Algorithm with K-Means for Text Document Clustering. *Mathematics* 2021, 9, 1929. [CrossRef]