Lee J.H., Yi J.-S., Son J.W., and Jang, Y.-E. (2017). “Pre-bid Clarification for Construction Project Risk Identification using Unstructured Text Data Analysis”. In: *LC3 2017: Volume I – Proceedings of the Joint Conference on Computing in Construction (JC3)*, July 4-7, 2017, Heraklion, Greece, pp. 219-227. DOI: https://doi.org/10.24928/JC3-2017/0028.

PRE-BID CLARIFICATION FOR CONSTRUCTION PROJECT RISK IDENTIFICATION USING UNSTRUCTURED TEXT DATA ANALYSIS

JeeHee Lee[1], June-Seong Yi [2], and JeongWook Son[3], and Ye-Eun Jang[4]

Abstract: This paper analysed construction bidding information in order to define risk factors that can be appeared in the bid documents of construction projects. For this purpose, text analysis was conducted on bidders’ inquiry information (Pre-bid RFI), which inquires uncertain information and omissions in the bid documents in order for pre-bid clarification. From the results of the analysis, what types of risk factors exist in the bid documents and what parts of the bid documents can be pre-reviewed to proactively respond to the uncertain owner’s requirements. The results are expected to be used as important information for pre-bid clarification of bid documents. Moreover, this study can be meaningful in that it provides a comprehensive way to grasp a large amount of 1,054 documents without analysing the contents of individual documents directly through analysis of bidding information of construction projects using text mining.

Keywords: Bid documents, pre-bid clarification, pre-bid RFI, text mining

1 Introduction

There are many conflicts and uncertainties in construction bid documents (contracts, drawings, specifications, etc.), such as wrong information, omitted information, and information discrepancies, which cause project risks and cost overrun. Since contractors can understand project objectives and owner’s design intent from the bid documents, pre-bid clarification is a primary work in the project initial phase (Lee et al. 2015).

The construction bidding process contains a large amount of document-based text information written in text (Mohemad et al. 2011). Moreover, since construction bid period is usually not enough to thoroughly review complicated documents, bidders are easy to fail to resolve many of the uncertainties and risks associated with the projects hidden in the bid documents within a limited time. If the quality of bid documents is poor, the uncertainties and risks will increase. In order to clarify the uncertainties in the project, which is written in bid documents, before the project is carried out, pre-bid clarification process is very crucial.

In general, bidders send pre-bid inquiries to project owners before bid opening in order for clarifying owner’s requirements. When owners receives the bidder’s inquiries, the owners should reply to the questions in time and open the results of the inquiries to every

[1] Ph.D Candidate, Department of Architectural and Urban Systems Engineering, Ewha Womans University, Seoul, South Korea, jeehee@ewhain.net

[2] Professor, Department of Architectural and Urban Systems Engineering, Ewha Womans University, Seoul, South Korea, jsyi@ehwa.ac.kr

[3] Assistant Professor, Department of Architectural and Urban Systems Engineering, Ewha Womans University, Seoul, South Korea, jwson@ewha.ac.kr

[4] Graduate Student, Department of Architectural and Urban Systems Engineering, Ewha Womans University, Seoul, South Korea, kdkmn@naver.com

bidder so as bidders to figure out the clarification. Pre-bid clarification process is not simply a formal process for bidding, instead it is also an institutional device that improves the accuracy of bidders' estimates by clarifying uncertain and ambiguous information in the bid documents and prevents future design changes and claims. In addition, bidders' inquiries data can be considered as one of the important information to be generated at the bidding stage because they will be attached to the bid documents later which is contractual information. Daoud and Allouche (2003) argued that the more amendments introduced to the documents as a result of the pre-bid inquiries, the lower was the quality of the bid documents. The quality of bid documents determines project's future performance and cost overrun during project life cycle. In addition, the bidders' inquiries are likely to include various types of potential risk factors that may influence project's performance. Therefore, it will be possible to extract key risk factors commonly pointed out in the bid documents by analyzing the bidders' inquiries.

Especially, in recent international construction projects, it is emphasized that errors related to bid documents not previously reviewed by bidders at the bidding stage cannot be recognized as change order. However, it is not easy to review vast amounts of bid documents during short bidding period. Current approach for reviewing bid documents is impractical and time consuming when the evaluators need to identify, aggregate and synthesize salient information of these criteria manually (Mohemad et al. 2011).

Besides, previous risk management researches have mainly concentrated on high level risk factors, such as country risks, economic risks, and owner's risks, etc., bid documents related risk factors were not clearly examined in the risk management area. In other words, in the previous risk management research, only the level of '*ambiguity of bid documents*' was discussed about the risk related to the bid documents, but the research on the specific risk factors causing the '*ambiguity of bid document*' was lacking.

In this study, to define risk factors that can be appeared in the bid documents of international construction projects, we investigated the bidders' inquiry documents prepared by bidders to find out what factors should be examined for pre-bid clarification. In addition, we tried to effectively analyse a large number of data by using an unstructured text data analysis technologies, text mining, which can innovatively analyse a vast amount of text documents in a short time. The results of this study can provide a good reference for future bidders to review the bid documents and provide knowledge on the information to look for in pre-bid clarification.

2 RELATED WORKS

Several researches have been conducted on the quality of construction bid documents. Daoud and Allouche (2003) examined the information of bid inquiries (requests for information; RFI) created by bidders during the bidding process and presented what type of error occurred in the bid documents of the construction project through a case study. From the case study, queries for design drawings were found to be the most among the bid documents. Lu et al. (2016) analysed the completeness of contract documents which is very important information of construction bid documents. They explored the effect of contract completeness on contractors' opportunistic behavior and presented four dimensions indicating contract completeness: issue inclusiveness, term specificity, contingency adaptability, and contractual obligatoriness. Erdis and Ozdemir (2013) studied technical specification related disputes between the employer and contractor in construction industry. They examined large amounts of specification which is part of construction documents, and court records and found out equivocal expressions written in specification could cause construction disputes. In other words, this shows that since the information written in construction bid documents could cause projects risks and further disputes, pre-bid clarification is very crucial for project risk management aspects.

However, studies on construction risk management so far have not dealt with the risk factors related to bid documents in depth. Mustafa and Al-Bahar (1991) proposed 32 risk factors including 6 different risk classification. They classify *design risks* as one of the risk category in a construction project and the category has 5 different risk factors: incomplete design scope, defective design, errors and omissions, inadequate specifications, and design changes. However, there is a lack of analysis on how each risk factor is expressed in the bid documents and the specific type of risk factors.

Thus, we investigates real world project's bid information using pre-bid inquiries and try to show what type of risk factors exist in the bid document and which information in the bid document should be reviewed in advance.

3 Text Mining

Recently, *Big Data* technology has been rapidly emerging, and technologies for dealing with unstructured data such as text, image, and voice data, as well as technologies for handling vast amount of data, are rapidly developing. The vast majority of information in the construction industry is also in the form of unstructured data. In general, text mining is widely used as a method of analysing text data. Text mining is a technique based on NLP that allows computers to understand the natural language, by extracting patterns and relationships from unstructured data in natural language to find meaningful information. In the field of construction area, some studies using text mining technology for automatic classification of construction documents have been conducted. Caldas et al. (2002) developed a prototype of an automatic classification system for construction documents through text mining based machine learning to automatically classify a large number of construction documents stored in PMIS (project management information system). In addition, studies using text mining to interconnect information related to content in the document and document categories. Mao et al. (2007) used a metadata model of unstructured data to link individual content and related information in a document. Thus, research for analysing unstructured text data in the construction field is being carried out in various aspects, and there is a movement to integrate the unstructured data and the structured data to improve the efficiency of the information management.

4 Research Methodology

4.1 Research Method and Process

Unstructured text data analysis methods such as text mining, information retrieval (IR), and natural language processing (NLP) were used in this study to efficiently analyse the information of text-based bid documents. Through R programming, a free software environment for statistical computing and text data analysis, unstructured text documents were structured, analysed and visualized.

The research process is shown in Figure 1. Before performing the text mining process bidders' inquiry information was extracted from the bid documents of the previously performed projects. As mentioned above the bidders' inquiries could contain questions about various types of risks hidden in the bid documents. Moreover, since the bidders' inquiries comprise risk factors related to the different kinds of bid documents, such as contract documents, notice to bidder, drawings, and specifications, it is very useful information to comprehensively understand the types of risks that can occur in the bid documents. After obtaining the bidders' query information, the text data pre-processing which is essential for text mining should be performed. The pre-processing process prepares the computer to automatically recognize the documents written in text form. The pre-processing method is variable depending on the situation of the text data. The analysis of the text data is carried out with the pre-processed data. Since the purpose of the study is to identify risk factors associated with the bid documents, we focused on three analysis methods: word frequency, term association rules, and text topic modelling. A detailed explanation of each analysis method will be given in the next section.

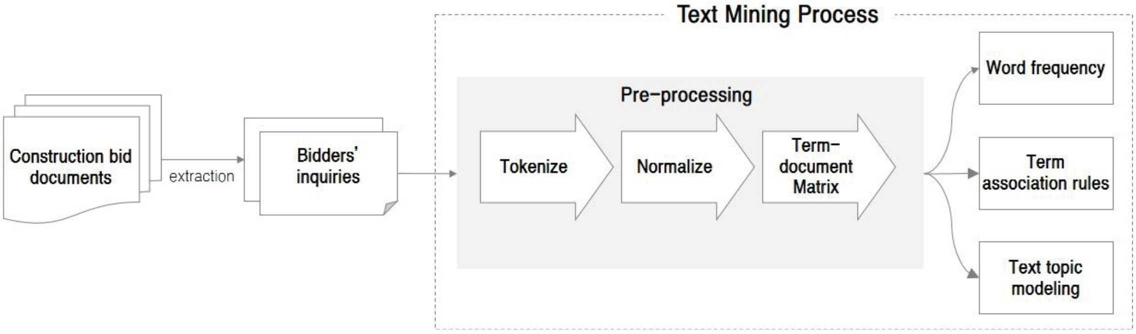

Figure 1 Research process

4.2 Data Collection

In order to analyse the bid query information of construction projects, this study collected bidders' inquiries data on public construction projects that were ordered by the California Department of Transportation (Caltrans) in the US for the past three years. Caltrans construction bid documents include special provisions, notice to bidders, specifications, drawings, and bid items. Caltrans places hundreds or more projects each year and has the advantage of data accessibility because it is open to the public for all bidding information, including bidders' inquiries.

The analysis data is a total of 1,054 bid inquiry documents generated from 211 infrastructure public projects. All data were analysed through text data pre-processing. Table 1 is a part of the bid inquiry documents, which consists of the bidders' inquiry and the owner or engineer's answer.

5 Text Mining Analysis

5.1 Pre-processing of unstructured text data

In order to perform text mining analysis, it is necessary to remove unnecessary information and convert the unstructured data into the structured data. Thus, pre-processing of unstructured text data is very important for successful text mining.

There are various types of words in text documents such as nouns, adjectives, and verbs, and unnecessary words exist in analysis such as punctuation, symbols, and spaces. In order to eliminate the distinction between uppercase and lowercase letters as a part of eliminating unnecessary information on 1,054 bidders' inquiries, all the words were converted to lowercase. After that, all punctuation marks (punctuation, comma, semicolon, colon, etc.) are removed from the document and words that do not take up a great deal of weight in describing the contents such as article, preposition, conjunction, etc. In addition, words such as *contractor*, *inquiry, construction* etc., which appear repeatedly in the construction bid documents, though not abbreviated words, are removed together because they are not meaningful analysis objects in themselves. In order to remove this unnecessary information, the *tmMap()* function of the tm package in R was used to the elements of all documents in the Corpus. The process of eliminating unnecessary information through R programming is summarized as follows:

```
> Corpus <- Corpus(DirSource("C:/Users/TextFile"), readerControl=list(language="en"))
> Corpus<-tm_map(Corpus, removeNumbers)
> Corpus<-tm_map(Corpus, tolower)
> Corpus<-tm_map(Corpus, stemDocument)
> Corpus<-tm_map(Corpus, removePunctuation)
> Corpus<-tm_map(Corpus, stripWhitespace)
> Corpus<-tm_map(Corpus, removeWords,      stopwords("english"))
> myStopwords<-c(stopwords("english"), "contractor", "inquiry", "Caltrans", "construction")
```

Table 1: Part of bidders' inquiries

Project Code	Inquiry (Bidder)	Response (Owner or Engineer)
01-0B3004	Detail Pile No. 4 to 24 on sheet 44 shows removal of lean concrete and replacement with structural concrete. Does the reinforcing steel shown in this detail extend below the transition elevation?	On plan sheet 44 of 55, see section A-A. No bar reinforcement is shown in section A-A, so bar reinforcement is not required in that area of the soldier piling.
01-0B5004	Please clarify under what item is the water trucking for the plants will be paid. Should there be an item for developing water supply? How far is the closest water source from the jobsite? Is Caltrans paying for water?	The Department will not be paying for water as a separate bid item. The contractor is responsible for providing the offsite water source. Your attention is directed to section 20-1.02B, section 20-3, and section 20-4 of the Revised Standard Specifications. Full compensation for water is included in Bid Item No. 59, PLANT ESTABLISHMENT WORK.
01-0E5704	Based on the requirement for exclusionary devices for work between March 1 and August 31 (per Section 14-6.02C(5) and Section 14-6.10A(1)), are exclusionary devices required for work outside of the nesting season (September 1 to February 29), as there is no specification stipulation to install exclusionary devices during the September 1 to February 29 period. Please clarify.	An addendum has been issued to address this bidder inquiry. Please refer to Addendum No. 2, issued on Thursday, May 5, 2016. Please bid per the current contract documents.

As a result of pre-processing, it was found that documents consisting of 11,874 words were reduced by 50% to 5,196 words after the removal of unnecessary information. In the text data pre-processing, the minimum length of the word is specified as 3 characters.

In order to make text data in a form that can be automatically recognized by a computer, the text data must be converted into structured data. For this purpose, a text-based document is represented by a vector space model. To represent the text data as a vector space model, a term-document matrix should be created for the documents from which unnecessary information has been removed. The term-document matrix is a mathematical matrix structure that describes the frequency of words occurring in a set of documents. In the term-document matrix, rows represent documents, columns represent words, and the elements of each matrix represent the frequency at which characteristic words occur in a particular document. However, when creating a term-document matrix for 1,054 documents and 5,196 words, a matrix of considerable size is created. Therefore, to increase the efficiency of the analysis, it is necessary to reduce the size of the matrix. In other words, it is necessary to assign a weight to relatively important words to select words with high priority. There are many ways to calculate this weighting value, the most well-known of which is the *Tf-Idf* (term frequency-inverse document frequency) weighting method. *Tf-Idf* is a concept that considers not only the frequency of words in each document but also the frequency with which words occur in various documents. The value of *Tf-Idf* is based on the concept that the importance of a word is lowered as the word appears frequently in all documents. In this study, the weights of existing words in the document are applied through the *Tf-Idf* weighting method, and the formula is as follows:

$$w_{t,d} = \log(1 + tf) \times \log(\frac{\mathrm{N}}{df})$$

$w_{t,d}$: *tf-idf* weight; *tf:* term frequency;
N: total number of documents; *df*: document frequency

5.2 Word frequency

The text mining results of construction bid questionnaires can be used to identify which words and terms are used frequently and understand what type of risk is mentioned based on this. As a result of word frequency, we could figure out that the words related to the submission of bidding documents such as *bid* (1,693 times) and *submit* (1,198 times) are most frequently found, and many terms such as *contract* (1,014 times) and *section* (904 times) which related contract documents are found. In addition, words such as *special* (374 times) and *provision* (318 times) indicate that the query for special contract conditions was made in a number of bidding inquiries. Furthermore, there are many words that can be guessed that a problem occurred in design documents such as *specification* (283 times), *plan* (272 times), and *sheet* (318 times). Figure 2 shows the percentage of top 15 words in the total documents.

5.3 Term association rules

Although it is possible to grasp the words repeatedly used in the bid questionnaires through the analysis of frequent words, it is difficult to understand the meaning of each word in the sentences by frequency analysis alone. Therefore, association rule analysis was performed to extract patterns for words relationship. Association rule analysis is a data mining technique, also known as market basket analysis, which means a rule or condition expressing how often an event occurs at the same time. The analysis of the association rules for the bid inquiries was conducted for the purpose of identifying what words are commonly found when a specific word appears in the same bid questionnaire. For example, to understand the meaning of the word '*specification*' in the document, we tried to grasp the meaning of the sentence more precisely by analyzing the association rules. A total of 5,167 association rules were generated by analyzing association rules based on analysis data.

Fig. 3 shows the distribution of support, reliability, and degree of lift of a total of 5,167 association rules derived. As a result, the confidence of the rule was higher than 0.1 in most cases, resulting in a relatively high confidence level.

However, all the association rules are not meaningful and reliable. Therefore, only a few rules with high explanatory power were extracted in this study, and some of them were summarized in Table 2. From the rules presented in the analysis, how the words are used together with the words in the document can be found.

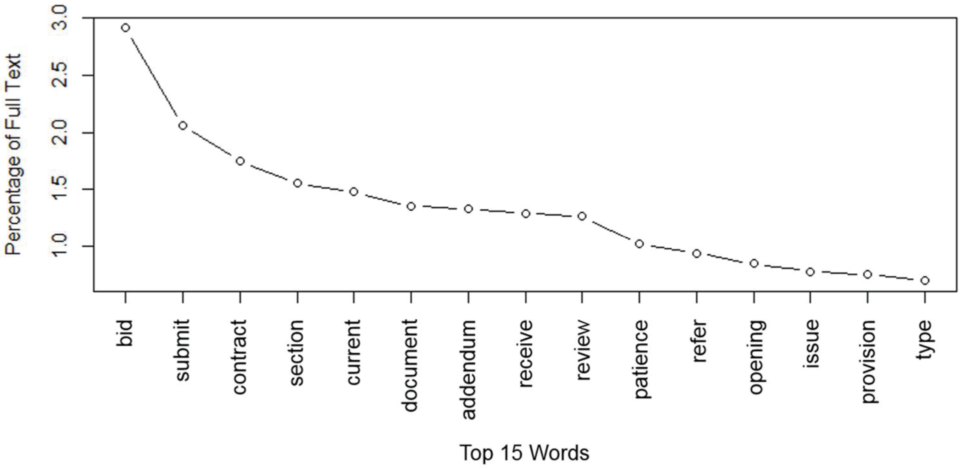

Figure 2 Top 15 words of bidders' inquiries

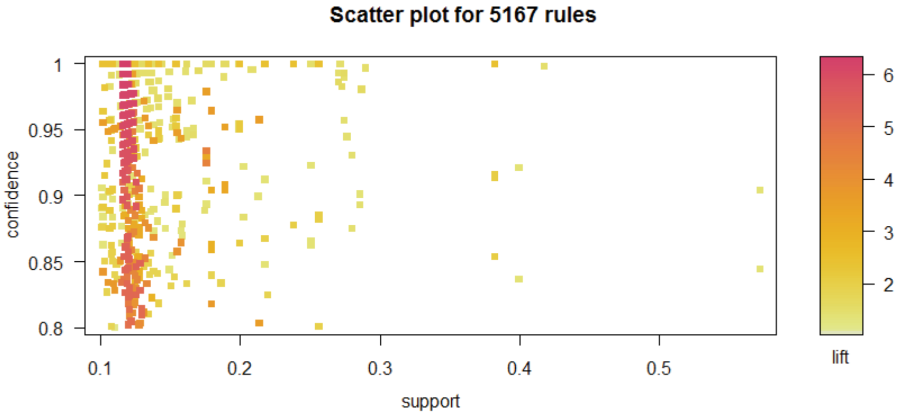

Figure 3 Scatter plot of text association rules

5.4 Text Topic Modelling

Based on the understanding of the relationship between words, this section conducted topic modeling of the documents in order to understand what subjects each topic could be tied to by using text mining on the bid inquiries. In this study, we use the LDA (Latent Dirichlet Allocation) method, one of the topic modeling methods, to identify the topics that are commonly discussed in documents. Unlike the existing text classification methods, the LDA method is a method of distinguishing types through Unsupervised Learning. In other words, since the data is analyzed without knowing how many types of analysis data can be classified, it can be useful when there is no prior information on how the analysis data can be classified as in the present study. In LDA, it is assumed that not all documents belong to one topic, but two or three different topics are mixed, and the words in the same topic group are related to each other.

Table 2: Result of text association rules

Rules		Support	Confidence	lift
{Specification}	{Standard}	0.1204	0.6939	4.0412
{Standard}	{Section}	0.1546	0.9005	1.9733
{Item, section}	{Bid}	0.1138	0.8888	1.3158
{Date, note, provide}	{Due}	0.1176	1	6.3113
{Contract, date, document, role}	{Due}	0.1157	1	6.3113

As a result of topic modeling through R programming, we found 5 topics. In order to select 5 topics for 1,054 bid documents, several trials and errors were made and the case where the most similar words were selected as one topic was finally selected. The words listed in each topic are words that have high explanatory power on the topic.

First, in the case of Topic 1, it is seen that words related to specifications are selected as words with high explanatory power among bid documents. Therefore, problems related to specifications can be viewed as a question group. In addition, it can be understood through the words such as '*material*', '*type*' and '*requirement*' that problems (errors, conflicts, omissions, etc.) regarding the material type and requirements described in the specification are written in the bid inquiries. Topic 2 is about construction site conditions and operation information. The words such as '*water*', '*site*', and '*traffic*' shows that there were many inquiries about the matters to be secured in advance in order to carry out construction work. Topic 3 is the topic of bid item included in the bid documents. It could be guessed that here is a problem with the quantity of bid items when the word '*quantity*' is included. Topic 4 is a group of topics related to the information of design drawings. It is understood that the majority of the questionnaires created to confirm the document discrepancies among design documents. Topic 5 is a topic group related to the contract clauses, and it is judged that many inquiries were made to confirm the toxicological provisions and omissions in the contract.

The types of contents of the questionnaire for each topic group are summarized as follows:

- Topic 1: Requirements of clarification for construction materials and types described in specification
- Topic 2: Questions on construction general information and site conditions
- Topic 3: Requirements to clarify the quantities of bidding items
- Topic 4: Questions about design drawing information
- Topic 5: Request for coordination of special terms and conditions of the contract clauses

6 CONCLUSIONS AND LIMITATIONS

In this study, text mining was performed on the information of the bidders' question in construction projects and it was possible to understand the types of words, the relation between words, and the subject type of documents as a results of text mining analysis. In other words, in the past construction work, what part of the bidders reviewed and inquired when reviewing the bidding documents could be identified. Besides, this study can be meaningful in that it provides a comprehensive way to grasp a large amount of 1,054 documents without analysing the contents of individual documents directly through analysis of bidding information of construction projects using text mining.

However, the results of the text mining conducted in this study are limited by the fact that the subjective judgment of the researcher can be intervened because it is greatly influenced by the data pre-processing and refining process. In addition, it is analyzed only the bid information of the construction occurred in the US construction market due to the difficulty of collecting information. Therefore, if the case data of various regions are acquired in the future, it will be possible to present various comprehensive results.

7 ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1D1A1A02061864)

8 REFERENCES

Caldas, C., Soibelman, L., and Han, J. (2002). Automated Classification of Construction Project Documents. *Journal of Computing in Civil Engineering*, 16(4), 234-243.

Daoud, O. E. and E. N. Allouche. (2003). “Bid queries as a gauge for quality control of documents." Proc. of the 5th Construction Specialty *Conference of the Canadian Society for Civil Engineering*, Monction, Nouveau-Brunswick, Canada.

Erdis, E., and Ozdemir, S. A. (2013). Analysis of technical specification-based disputes in construction industry. *KSCE Journal of Civil Engineering*, 17(7), 1541-1550.

Lee, J. H., Son, J. W. and Yi, J. S. (2015). “Multilevel project-oriented risk-mining approach for overseas construction project’s preemptive action.” *Proceedings of the 32nd CIB W78 Conference*, Eindhoven, The Netherlands. 336-343.

Lu, W., Zhang, L., & Zhang, L. (2016). Effect of Contract Completeness on Contractors’ Opportunistic Behavior and the Moderating Role of Interdependence. *Journal of Construction Engineering and Management*, 142(6), 04016004.

Mohemad, R., Hamdan, A. R., Othman, Z. A., & Mohamad Noor, N. M. (2011). Ontological-Based Information Extraction of Construction Tender Documents. In *E. Mugellini, P. S. Szczepaniak, M. C. Pettenati, & M. Sokhn (Eds.), Advances in Intelligent Web Mastering – 3: Proceedings of the 7th Atlantic Web Intelligence Conference, AWIC 2011*, Fribourg, Switzerland, January, 2011 (pp. 153-162). Berlin, Heidelberg: Springer Berlin Heidelberg.

Mustafa, M. A., & Al-Bahar, J. F. (1991). Project risk assessment using the analytic hierarchy process. *IEEE transactions on engineering management*, 38(1), 46-52.

Soibelman, L., Wu, J., Caldas, C., Brilakis, I., & Lin, K.-Y. (2008). Management and analysis of unstructured construction data types. *Advanced Engineering Informatics*, 22(1), 15-27.

Tanaka, T. (1988). *Analysis of claims in U.S. construction projects*. Master thesis, Massachusetts Institute of Technology, Boston.