Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P030-cd

Using NLP for Automated Contract Review and Risk Assessment

Gorkem Eken

Civil Engineering, Middle East Technical University, Ankara Türkiye, gorkemeken@gmail.com

Irem Dikmen

Construction Management and Eng., University of Reading, Reading, United Kingdom, i.dikmen@reading.ac.uk

M. Talat Birgonul

Civil Engineering, Middle East Technical University, Ankara Türkiye, birgonul@metu.edu.tr

Introduction:
Construction projects are notoriously risky due to the involvement of multiple parties having different objectives, limited project time and budget, high organizational and technological complexity, and vulnerability due to dynamic macroenvironmental conditions. Contracts are legal documents that define the responsibilities of the parties and allocate risks. To create an adequate risk management plan, contractors must conduct tedious contract review processes to identify the risks retained by them.
Problem Statement:
Although legal professionals try to assess risks in documents in detail, the possibility of errors due to unrecognized or misinterpreted risk elements remains as in-depth review of contracts is usually not possible during the short bidding period. Therefore, there is a growing need for intelligent systems that automatically analyze contracts to ensure that clauses in contracts are accurately defined and categorized with minimal human intervention. Automated analysis of contracts can be a solution for early detection of contract risks.

Methodology:
This research project involves the development of an automated text analysis model based on natural language processing (NLP) and supervised machine learning (ML) to improve the contract review process in the bidding stage. To demonstrate the applicability of the model, the FIDIC standard form of contract was selected, and all sentences were labeled with the sentence type and risk ownership in order to create a training dataset. Sentence type consists of Risk, Right, Obligation, Heading and Definition labels. The risk ownership consists of Contractor, Employer and Shared labels. In addition, the test dataset was created using a real contract of a construction project. The selected real contract has been prepared based on FIDIC Silver Book for an airport project. Preprocessing methods such as lemmatization and stop word removal were employed. After the preprocessing steps, the number of sentences in the training dataset created from the FIDIC Red, Silver, and Yellow Book decreased from 5346 sentences to 2268 sentences when repeated sentences were removed. On the other hand, the number of sentences in the test data set created from the real contract decreased from 1305 sentences to 1217 sentences when only unique sentences were kept. The labels in the training and test datasets were validated with the help of expert meetings with six participants who were working in departments of contract. One of them has a Ph.D. degree, and three of them have an M.Sc. degree. Half of them have more than 10 years of work experience. Randomly selected 10% of sentences in each dataset were relabeled by experts for both sentence type and risk ownership. Expert labels were compared with the labels given by researchers and the deviation between the two sets was calculated as 3%. Datasets used to train and evaluate 12 ML models.
12 machine learning models were built based on Bag of Words, Term Frequency-Inverse Document Frequency, pre-trained Spacy and Glove word embeddings and Bidirectional Encoder Representations from Transformers (BERT) word embedding techniques and logistic regression, support vector machine, decision tree, recurrent neural network and BERT algorithms. The classification models were evaluated based on four parameters: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The confusion matrix for binary classification and multi-class classification were created, respectively. The performance of a machine learning model is usually measured by six metrics: Accuracy, Precision, Recall, F1 score, Specificity, and Area Under the Curve, which are calculated using TP, FN, FP, and TN values. The most appropriate performance metric for the study is accuracy. However, the accuracy measures may not perform as expected if the dataset has a non-uniform distribution across different classes. Therefore, accuracy and f1 score, which is the harmonic mean of precision and recall and calculated as 2xTP/(2xTP+FP+FN), were used as evaluation metrics in this study.

Results:
The BERT model achieved 87% accuracy for sentence type classification, which has 5 labels as heading, definition, risk, right and obligation, and 80% accuracy for risk ownership classification, which has 3 labels as contractor, employer and shared. The best tree models were ensembled with the competitive voting method. After implementing the competitive voting method, accuracy increased to 89% for sentence type classification and 83% for risk ownership classification. Table 1 presents the individual performance results of each ML and competitive voting results for sentence type and risk ownership classification.

Table 1. Classification Performance of Individual Models and Competitive Voting

			Sentence Type		**Risk Ownership**	
ML Model No	**Text Vectorization**	**ML Algorithm**	**f1-score**	**Accuracy**	**f1-score**	**Accuracy**
1	BAG OF WORDS	Logistic Regression	0.77	0.79	0.67	0.71
2	BAG OF WORDS	Support Vector Machine	0.75	0.77	0.64	0.69
3	BAG OF WORDS	Decision Tree	0.72	0.68	0.67	0.72
4	TFIDF	Logistic Regression	0.78	0.77	0.66	0.72
5	TFIDF	Support Vector Machine	0.83	0.81	0.69	0.77
6	TFIDF	Decision Tree	0.70	0.66	0.45	0.70
7	Spacy	Logistic Regression	0.70	0.71	0.59	0.66
8	Spacy	Support Vector Machine	0.75	0.72	0.61	0.70
9	Spacy	Decision Tree	0.61	0.52	0.49	0.61
10	Keras Embedding	RNN	0.80	0.79	0.65	0.72
11	Glove Embedding	RNN	0.80	0.78	0.67	0.73
12	Word Embedding	BERT	0.83	0.85	0.73	0.80
13	Competitive voting (combination of Model 5, 11 and 12)		0.86	0.89	0.76	0.83

Discussion of Findings and Conclusions:
This study explores the potential of using NLP and ML for automated contract review in the construction industry. Manual contract analysis is currently time-consuming, costly and error-prone. The study uses FIDIC books to create datasets for ML models and compares their classification performance. With the proposed method, sentences in terms and conditions can be classified as type and ownership to identify parties' risks, rights and obligations. The results obtained, with an accuracy of 0.89 and an f1 score of 0.86, are promising, especially considering the relatively small training dataset. The study highlights the importance of using pre-trained models based on large datasets to improve classification performance, which is particularly useful when there is a limited amount of input in a domain. This approach can provide a way to combine domain-free information from large datasets with domain-specific information to solve problems. The automated construction contract review model, while may not be ideal as a stand-alone method at the bidding stage, can provide valuable information to reduce time and errors due to overlooking. The proposed approach can reduce staff workload and increase the quality of work in risk assessment at the bid stage, which can be helpful for contractors when deciding on risk premiums. Overall, this study provides a new and promising approach for contractors to review construction contracts using automated methods, which can improve efficiency and reduce errors.

Future work and limitations:
Although the results are promising about utilization of ML and NLP for automated contract review, there are limitations and further research is needed in this area. The dataset used in this study is limited to FIDIC books and the classification model was only tested on construction contracts based on FIDIC. To build a more general classification model, the dataset needs to be extended to include different types of standard contracts. While 12 ML models have been trained based on 5 algorithms and 6 vectorisation methods, further research needs to evaluate other alternatives for both the algorithm and vectorisation sides. In addition, the integration of a rule-based approach and the consideration of ambiguity in the natural language are also important factors which may increase the classification performance. Finally, the usefulness of the classification model depends on the appropriateness of the labels in the training dataset and needs to be verified according to the company's risk perception.

Keywords: Construction Contract Review, Machine Learning, NLP, Text Classification, Deep Learning