DEVELOPMENT OF A FRAMEWORK FOR PROCESSING UNSTRUCTURED TEXT DATASET THROUGH NLP IN COST ESTIMATION AEC SECTOR

Chiara Gatto, Antonio Farina, Claudio Mirarchi, Alberto Pavan
Polytechnic University of Milan, Milan, Italy

Abstract

Cost estimation is one of the most critical steps in the building construction process. Currently, it requires humans to manually extract information from documents written in natural language, often resulting in human error. This paper aims to investigate an automated technique for extracting data from documents with the support of NLP techniques, in order to automatize the task of structuring information.

A framework for automatically classifying information from unstructured text was developed leveraging NER techniques. This research supports the cost estimation activity minimizing the loss of resources coming from human error when interpreting NL documents.

Introduction

Cost estimation for tendering constitutes a decisive phase in the construction process where cost, time and other resources have to be predicted, not only for effectively planning the investments necessary for the building work realization but also for successfully managing the construction phase (Choi et al., 2015).

This process can be summarized in some major stages consisting in: classifying all construction products that constitute a building project into assemblies or items, extracting all the quantities of the latters (QTO activity), collecting pricing information from construction specification, relating this data to assemblies and items and finally estimating the project cost (Ma et al., 2013).

Recent years have seen an increase in the digitalization within the architecture, engineering, and construction (AEC) industry, where Building Information Modeling (BIM) have played a key role. The advent of BIM resulted in a significant advancement of productivity, making it possible to perform specific tasks in an automated manner. The use of BIM has significantly reduced the time and manual effort involved in the building life cycle stages, thus improving the efficiency and accuracy of construction processes (Akanbi and Zhang, 2021).

Despite the widespread use of BIM approaches, information exchange in AEC industry is still mainly based on the production of documents. These documents are often written in natural language, conveying knowledge through unstructured or semi-structured data. Natural language is by nature unstructured and it is therefore difficult to be digitally managed. However, unstructured sources of information, such as text documents, are essential components of design and construction projects (Opitz et al., 2014).

In order to increase the level of efficiency during the information extraction process, data contained in textual documents needs to be structured, in order to reduce semantic ambiguity and to enhance the possibility for a machine to understand those without the support of human beings. This task can be performed with the support of NLP techniques, which are proving to be useful tools for supporting human activity.

State of the art revealed that the involvement of NLP techniques in the AEC sector brings concrete results and it is continuously expanding (Hong et al., 2020; Zabin et al., 2022). In the cost estimation field data acquisition from unstructured documents written in natural language is an intense labour activity prone to error. Several studies have been performed involving NLP for automatizing the process of classification and information extraction.

The main limitation found in existing studies regarding specification document processing is the scalability of the developed models and the classification level of detail.

In order to fix this gap and to cover the increasing need for public administrations and practitioners to migrate information from text to digital format, the intended research activity proposes a framework for classifying with a high level of detail information contained in item specification documents composing construction works. The research is mainly divided into two stages. The first concerns the definition of attributes that characterize the unit price of a cost entity. The second concerns the development of a classification model capable of autonomously recognizing attributes in a specification document.

This paper is organized as follows. The research background is outlined in the ”State of the Art” section. The subsequent section provides the sequence of steps performed to define the methodology. The ”Framework definition” section explains the design process of the framework. Subsequently, the practical implementation is presented in ”Framework testing”. Finally, conclusions are drawn in the last section.

State of the Art

Through the investigation of the currently available literature, it was possible to state that the application of AI methodologies is increasing over the years as the techniques developed are becoming more and more consolidated. Currently ML techniques are involved along the whole building life cycle stage, with a higher intensity in the design, operation, and control phase (Hong et al., 2020; Zabin et al., 2022).

NLP is a subset of AI defined as all those techniques which help machines understand human languages through analyzing structures of texts and meanings of words. NLP has been increasingly adopted in AEC sector mainly in four application scenarios: information extraction, document organization, expert systems, and automated compliance checking (Wu et al., 2022).

Concerning the documents organization Caldas and Soibelman (2003) describes a methodology for improving information organization and access in construction management information systems. The proposed approach is based on NLP and classification techniques for automatic hierarchical classification of construction project documents according to project components. The most satisfactory results were obtained with SVM classification system.

In a most recent study carried out form Qady and Kandil (2014), an unsupervised learning method to automatically cluster documents together based on textual similarities was developed.

The methodology supports the generation of classification models based on project information classification structures, such as construction information classification systems or project model objects.

Concerning the automated code checking, ML approaches have proven to be effective in performing this task. Techniques such as ANN, Case-Base Reasoning (CBR), NLP, and semantic logic-based information representations have been successfully applied to BIM with the purpose of speed up and automate the rule checking of a design code Fuchs and Amor (2021); Song et al. (2018). Zhang and El-Gohary (2017) developed one of the first building regulation compliance checking frameworks methodology that entirely relied on NLP. The developed framework was aimed at extracting design information from an IFC model into a semantic logic-based representation in order to matching the semantic logic-based representation of regulatory information. The limit of the search lies in the scalability of the approach, which was tested only with quantitative requirements of one International Building Code (IBC) chapter. The NLP techniques adopted fit well only for a particular text type and context (Zhang and El-Gohary, 2017; Fuchs and Amor, 2021).

In a further study, Zhang and El-Gohary (2023) used a deep learning technique for code compliant checking reaching satisfying results. The study leverages a transformer based model in order to provide additional contextual information and knowledge for better aligning and classifying the definitions of the concepts and an IFC knowledge. The proposed method was evaluated on IFC concepts and regulatory concepts from building codes and standards.

Concerning the cost analysis stage an interesting study is performed by Liu et al. (2022), who developed a Knowledge-based model in order to automatically extract geometry information from BIM models incorporating the standard method of measurement (SSM) rules. The result of this study is a BIM-based QTO automatised process which helps in reducing the inaccuracies, time, and errors of cost estimation.

One of the first studies where AI techniques were applied in the bidding process is carried out by Chua et al. (2001) who developed a case based reasoning (CBR) model for supporting contractors. The system retrieves similar cases assessing the possible level of competition and risk margin.Lee and Yi (2017) developed a methodology for predicting risk in the bidding process of construction projects by analyzing the uncertainty of the bidding document and using it as a factor to predict a project's bidding risk. In this study four representative classification algorithms for document classification have been explored: Artificial Neural Network (ANN), Support Vector Machine (SVM), k-nearest neighbors (KNN), and Naïve Bayes (NB). The best performance was achieved by the SVM classifier reaching an accuracy of 72.92%. Williams and Gong (2014) developed a model for predicting the level of cost overrun using text data mining classification algorithms. Different classification models have been tested: Ridor Rules; K-Star; Radial Basis Functions (with the highest prediction accuracy); Stanking.

Lee et al. (2019) proposed an information extraction method based on the lexicon. The model extracts the risk-related sentences in the contract and shows a warning message to the users to help them review those important clauses that should not be missed during the bidding and contract phases. As a result of the validation, the precision, recall, and F-measure of the model's performance against the experts' review of the contracts were all 81.8%.

Focusing on information extraction and classification data contained in cost estimation documents such as technical documents and specification documents, several studies have been performed.

Martínez-Rojas et al. (2013) describes a preliminary approach to automatically classify Work Descriptions in construction projects, coming from diverse sources. The methodology developed is able to classify work description independently from the linguistic framework and the original structure of its description. The just mentioned study was expanded and tested with a wider work descriptions dataset by Martínez-Rojas et al. (2018), who investigated six classification techniques for assigning work descriptions that come from very diverse projects under the common hierarchical warehouse structure of task groups. The six methods explored were the C4.5 decision tree, random forest, Naïve Bayes, neural networks, support vector machines, and k-nearest neighbors. Basic linguistic processing, such as cleaning and synonym replacement, was applied before applying these classification methods with the aim of reducing the vocabulary considered. It was experienced a high level of accuracy in the classification problem of a wide range of classes.In a further study performed by Moon et al. (2021), a name entity recognition (NER) model was developed with the aim of extracting information from construction specifications according to five information categories defined (Organization, action, element, standard, reference). A Bi-LSTM model was developed able in predicting the category of each word. This research constitutes one of the first successful attempts in applying this model in the construction industry. In a further study, an automated system for construction specification review using natural language processing was successfully developed. The objective of this research was to build an automated system for reviewing construction specifications by analyzing the varied semantic properties such as different vocabulary, different sentence structures, and differently organized provisions. Three types of semantic conflicts have been found in construction specifications that cause difficulties in automating the review process: different vocabulary, different sentence structures, and differently organized provisions. This difficulty is overcome through the use of NLP techniques based on Word2Vec embedding and Doc2Vec embedding Moon et al. (2022). One of the most recent studies shows a methodology for automatically processing work descriptions and laying a foundation for automated QTO and cost estimation through the NLP-based information extraction model integrating Hidden Markov Model (HMM) and formalized labeling rules. One limitation of this study is that it is based on RSMeans cost items, therefore it might not work effectively on cost items from other sources, furthermore training and testing data are limited Tang et al. (2022).

Wang et al. (2021) proposes a multi-scale information retrieval scheme for BIM both using the hierarchical structure of BIM and Natural Language Processing (NLP). The objective of this research is the on of developing a method for finding building components or attributes associated with the unified queries. The hierarchical BIM structure was represented through a BIM Hierarchy Tree (BIH-Tree) model. Subsequently, NLP and International Framework for Dictionaries (IFD) are employed to parse and unify the queries.

Akanbi and Zhang (2021) where the authors proposed a method that uses semantic modeling and NLP techniques for extracting required design information from CSI MasterFormat construction specifications documents, and automatically matches the extracted design information with unit prices of materials from a database. The cost database created is characterized by four main classes: identifier; building component; entity type (description of the entry); unit price.

Methodology

The methodology approach used during this study is explained in this section and synthesized as figure 1 shows.

Firstly, the development of the study stems from listening to the needs coming from the industry. Professionals and public administrations have been therefore interviewed in order to understand the shortcomings of the cost estimation process in tendering, especially when manually retrieving information from specification documents.

Starting from the industry needs, the subsequent step was the one of investigating the state of the art, with the aim of exploring tools and methodologies applied for automatizing the process of structuring data from textual documents in cost estimation AEC context.

Then, a framework for automatizing the process of structuring information from specification documents though have been developed leveraging NLP classification techniques. In order to build a tool that could structure data in a way that would satisfy the industry needs as much as possible, an iterative process was selected for performing this phase, where practitioners and public administrations have been involved through interviews. This strategy was adopted with the aim of identifying corrections and refining the framework before the test phase.

Finally, the framework was tested by its practical implementation on a first case study.

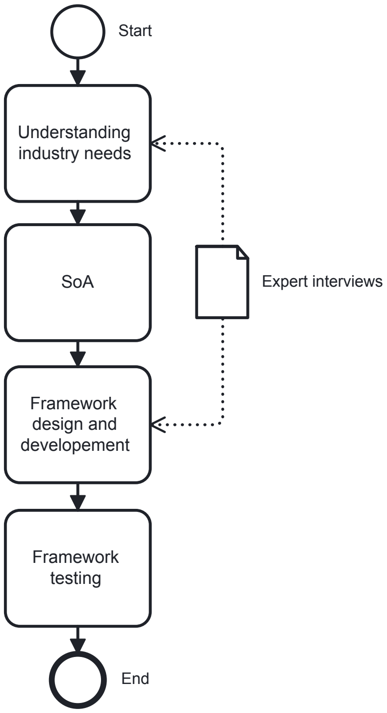

Start
Understanding industry needs
SoA
Expert interviews
Framework design and developement
Framework testing
End

Figure 1: Methodology chart.

Framework definition

In this section, the process of the framework design and development is explained and synthesized as figure 2 shows. The objective to be tent during this phase is the one of developing a procedure to convert textual data belonging to specification documents from unstructured to structured, through the support of NLP techniques, with the purpose of classifying information according to specific labels. This objective aims to respond to the increasing need for public administrations and practitioners to migrate in-

formation from text to digital format. The consequence of this task is the one of parameterizing the information allowing to manipulate data more easily and reducing the ambiguity that characterizes cost item descriptions.
This work phase mainly consists of two main stages as shown in the two lain of the chart in figure 2:

- labels definition;
- classifier definition.

The first step consists of the definition of specification documents ontology and semantic in the domain of cost estimation. Several Italian regional price list documents have been analyzed in order to identify the entities characterizing the construction cost analysis domain, their definition, and the relationships between them.
Those documents consist of a large amount of manufacturing items mainly defined by code, description, unit of measurement, and price information. Each item is built through a price analysis process, where elementary costs of materials, labour, transport, and rental involved in the manufacturing process compose the final price. The tool contains a first section related to construction work items, while the second is related to the elementary costs of products, transport, and labour involved in construction works. The cost entities defined are four: construction work item, product item, labour item, and equipment item.
Once the entities have been identified, the next step is the one of identifying common labels set according to which classifying the unstructured data inside technical specifications concerning the four different cost entities.
The label definition stage was performed in order to identify all attributes that characterize the price of a building cost item in the specification documents.
In order to achieve the goal, this step started with analyzing the item description inside the specification documents. Therefore cost items have been grouped according to their class in order to isolate a scattered sample of descriptions from different classes, so as to analyze a varied sample. The parameters that affect the price are extracted from the analysis of the item descriptions. Subsequently, the set of attributes is discussed with the practitioners through interviews to obtain the labels validation.

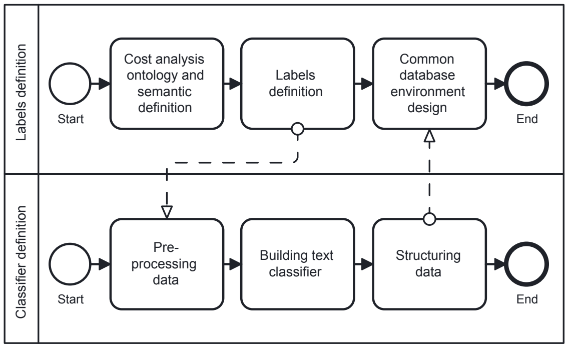
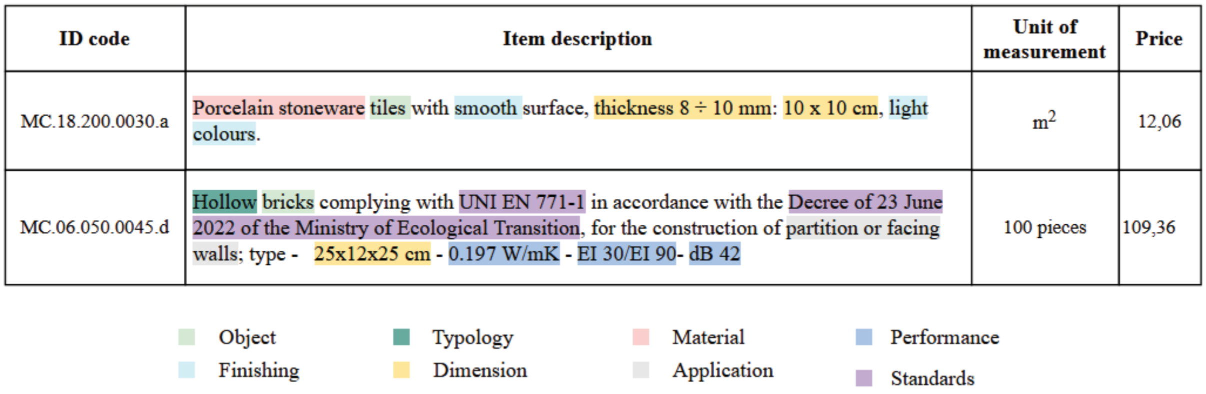


Figure 2: Framework design and development chart.

In order to further validate the definition of the labels, the just mentioned process is repeated on further cost items description, following an iterative process.
The labels found through this process are several and can be further grouped into four types of attributes:

- attributes related to the classification of the items (ID code, product class, typology);
- attributes related to the characteristics of the items (material, finishing, shape, size, function, physical performance, sub-components, standards, certificates, supply mode, yield);
- attribute related to miscellaneous information (included; excluded; notes, application);
- attribute related to unit price (unit of measurement, price).

The outcome of the first lane, as shown in figure 2, is the realization of a common environment, where all data involved in generating a new cost item are stored. On a practical level, the common environment consists of a relational database.
Once the labels are defined, the following step is the one of evaluating NLP techniques for structuring information. In particular, classification techniques are investigated since considered the most suitable for achieving the objective set in this research activity. Those are mainly divided between unsupervised and supervised techniques. Through the first approach, the machine learns without human supervision, while through the latter, the machine learns under human supervision.
Clustering could be one of the most suitable approaches among unsupervised NLP techniques, K-means model could be implemented in order to recognize common paths since often the descriptions that characterize the item belonging to the same item typology are similar (e.g. all those descriptions related to insulation panels). However, it would not be effective for easily identifying the predefined labels. Furthermore, it would not be effective in managing the heterogeneity existing among the description between different cost entity typologies and among items within the same entity class (e.g. all those descriptions related to insulation panels follow a common syntactic structure which is different from the masonry blocks).
The dataset that constitutes technical specification documents is considered heterogeneous since each item typology has its own characteristics set, which is different from one to the other affecting the way in which the composition of item descriptions is. Figure 3 highlights this aspect by showing two product item descriptions.
The supervised approach better fits this context as considered more effective for training a model in recognizing predefined labels in a heterogeneous dataset. In order to achieve this goal, the sequence labeling technique is considered the most suitable for the purpose of this research



ID code	Item description	Unit of measurement	Price
MC.18.200.0030.a	Porcelain stoneware tiles with smooth surface, thickness 8 ÷ 10 mm: 10 x 10 cm, light colours.	m^2	12,06
MC.06.050.0045.d	Hollow bricks complying with UNI EN 771-1 in accordance with the Decree of 23 June 2022 of the Ministry of Ecological Transition, for the construction of partition or facing walls; type - 25x12x25 cm - 0.197 W/mK - EI 30/EI 90- dB 42	100 pieces	109,36

Figure 3: Text annotation process with customized labels.

study. The most usual tasks performed with sequence labeling are mainly part of speech tagging (POS); name entity recognition (NER) and text chunking. Among those, name entity recognition (NER) is the one chosen for structuring the data coming from specification documents according to the predefined labels set. The goal of NER is to automatically extract entities from unstructured text.

The pre-trained NER models are capable of autonomously recognizing entities from text basing on generic labels such as location, person, and organization which are not useful for the purpose of this study. Therefore, for the development of this framework, a NER model is customized on the basis of the previously defined labels.

In order to do this, it is necessary to pre-process the text by means of a text annotation activity, which consists of manually labeling the cost entity description sample.

The classifier is trained, validated, and tested through a sample of labeled data. Once the classification model is built, it is then possible to obtain structured data and automatically populate the common database environment.

Framework testing

In this section, the process of testing the framework is shown with the aim of verifying the classifier's ability to automatically recognize labels in text.

The case study defined for building and evaluating the classifier model consists of product entity descriptions, which come from the Lombardy regional Price list document. It hosts about 6000 descriptions related to product items. The selected dataset consists of a 100 items description sample 50 of which are related to tiles products and the remaining 50 are related to masonry blocks products.

Figure 3 shows two examples of product descriptions that can be found within the sample. It is possible to see that those are not characterized by having a common precise rule in conveying information. In the first example, information related to the tile finishing is scattered throughout the text. In the second case, the characteristics of the block are provided in a noisy way through several concatenated sentences.

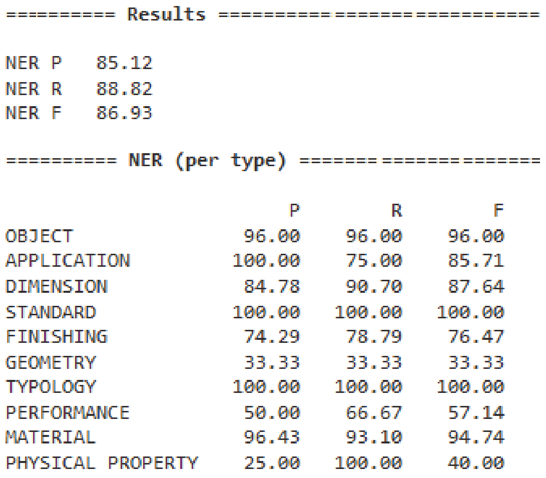
Once the case study has been defined, the following stage was the one of pre-processing the text through text annotation activity. Figure 3 shows an example of how this process is performed, consisting in manually selecting parts of the text and assessing their tag. Labels have been highlighted in the text sample with the support of an open-source text annotation tool, DOCCANO, which returned the processed text in a specific JSON format containing precise information of the labels position in the text.

It is important to note that not all labels defined during the first phase of the framework have been applied to this case study, as in this sample only 10 of the label set occur in the text.

Subsequently, the spaCy library was used to build the NER pipeline according to the customized labels. The configuration system has been set according to the Italian language.

The model has been trained with the 74% of the descriptions sample case study while the remaining was used for evaluating the model.

The results obtained from testing the developed framework are shown in figure 4 where the metrics are provided for both the entire model and individual labels.



```
========== Results ================================

NER P   85.12
NER R   88.82
NER F   86.93

========== NER (per type) =====================

                         P        R        F
OBJECT               96.00    96.00    96.00
APPLICATION         100.00    75.00    85.71
DIMENSION            84.78    90.70    87.64
STANDARD            100.00   100.00   100.00
FINISHING            74.29    78.79    76.47
GEOMETRY             33.33    33.33    33.33
TYPOLOGY            100.00   100.00   100.00
PERFORMANCE          50.00    66.67    57.14
MATERIAL             96.43    93.10    94.74
PHYSICAL PROPERTY    25.00   100.00    40.00
```

Figure 4: Classifier model score.

Conclusions

The research described in this article contributes to the goal of supporting public administration and practitioners need of migrating data from unstructured to structured format in the cost estimation field. Thus enriching the state of the art where a study on the structuring of information with numerous labels has not yet been carried out.
To automate the task of manually breaking down textual information according to specific parameters and populating a database, NLP techniques need to be leveraged.
In this study, we analyzed the labels needed to define a cost entity and subsequently we build a NER classifier. The approach used was a supervised one, leveraging the text annotation system for pre-process the data and train the classification model.
The framework was tested on two product typologies description (tiles and masonry blocks) coming from the Lombardy region price list case study, achieving an accuracy level of 86% in a classification problem with 10 labels. The lowest accuracy values have been recorded for those labels that occurred in the sample with lower frequency compared to those for which a higher level of accuracy was found.
In future developments, the case study analyzed will be extended in order to obtain results from a larger and more significant sample including entities other than product one. Furthermore, different classification techniques will be also evaluated in order to evaluate in order to assess the best solutions.

References

Akanbi, T. and Zhang, J. (2021). Design information extraction from construction specifications to support cost estimation. *Automation in Construction*, 131:103835.

Caldas, C. H. and Soibelman, L. (2003). Automating hierarchical document classification for construction management information systems. *Automation in Construction*, 12:395–406.

Choi, J., Kim, H., and Kim, I. (2015). Open bim-based quantity take-off system for schematic estimation of building frame in early design stage. *Journal of Computational Design and Engineering*, 2:16–25.

Chua, D. K. H., Li, D. Z., and Chan, W. T. (2001). Case-based reasoning approach in bid decision making. *Journal of Construction Engineering and Management*, 127:35–45.

Fuchs, S. and Amor, R. (2021). Natural language processing for building code interpretation: A systematic literature review.

Hong, T., Wang, Z., Luo, X., and Zhang, W. (2020). State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings*, 212:109831.

Lee, J., Yi, J.-S., and Son, J. (2019). Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based nlp. *Journal of Computing in Civil Engineering*, 33:04019003.

Lee, J. H. and Yi, J. S. (2017). Predicting project's uncertainty risk in the bidding process by integrating unstructured text data and structured numerical data using text mining. *Applied Sciences 2017, Vol. 7, Page 1141*, 7:1141.

Liu, H., Cheng, J. C., Gan, V. J., and Zhou, S. (2022). A knowledge model-based bim framework for automatic code-compliant quantity take-off. *Automation in Construction*, 133:104024.

Ma, Z., Wei, Z., and Zhang, X. (2013). Semi-automatic and specification-compliant cost estimation for tendering of building projects based on ifc data of design model. *Automation in Construction*, 30:126–135.

Martínez-Rojas, M., Marín, N., and Vila, M. A. (2013). A preliminary approach to classify work descriptions in construction projects. *Proceedings of the 2013 Joint IFSA World Congress and NAFIPS Annual Meeting, IFSA/NAFIPS 2013*, pages 1090–1095.

Martínez-Rojas, M., Soto-Hidalgo, J. M., Marín, N., and Vila, M. A. (2018). Using classification techniques for assigning work descriptions to task groups on the basis of construction vocabulary. *Computer-Aided Civil and Infrastructure Engineering*, 33:966–981.

Moon, S., Lee, G., and Chi, S. (2022). Automated system for construction specification review using natural language processing. *Advanced Engineering Informatics*, 51:101495.

Moon, S., Lee, G., Chi, S., and Oh, H. (2021). Automated construction specification review with named entity recognition using natural language processing. *Journal of Construction Engineering and Management*, 147.

Opitz, F., Windisch, R., and Scherer, R. J. (2014). Integration of document- and model-based building information for project management support. *Procedia Engineering*, 85:403–411.

Qady, M. A. and Kandil, A. (2014). Automatic clustering of construction project documents based on textual similarity. *Automation in Construction*, 42:36–49.

Song, J., Kim, J., and Lee, J. K. (2018). Nlp and deep learning-based analysis of building regulations to support automated rule checking system. *ISARC 2018 - 35th International Symposium on Automation and Robotics in Construction and International AEC/FM Hackathon: The Future of Building Things*.

Tang, S., Liu, H., Almatared, M., Abudayyeh, O., Lei, Z., and Fong, A. (2022). Towards automated construction quantity take-off: An integrated approach to information extraction from work descriptions. *Buildings*, 12.

Wang, J., Gao, X., Zhou, X., and Xie, Q. (2021). Multiscale information retrieval for bim using hierarchical structure modelling and natural language processing. *J. Inf. Technol. Constr.*, 26:409–426.

Williams, T. P. and Gong, J. (2014). Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Automation in Construction*, 43:23–29.

Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., and Yang, Z. (2022). Natural language processing for smart construction: Current status and future directions. *Automation in Construction*, 134:104059.

Zabin, A., González, V. A., Zou, Y., and Amor, R. (2022). Applications of machine learning to bim: A systematic literature review. *Advanced Engineering Informatics*, 51:101474.

Zhang, J. and El-Gohary, N. M. (2017). Integrating semantic nlp and logic reasoning into a unified system for fully-automated code checking. *Automation in Construction*, 73:45–57.

Zhang, R. and El-Gohary, N. (2023). Transformer-based approach for automated context-aware ifc-regulation semantic information alignment. *Automation in Construction*, 145:104540.