

Automation of text document classification in the budgeting phase of the Construction process: a Systematic Literature Review

Luís Jacques de Sousa and João Poças Martins

CONSTRUCT-Gequaltec, Faculty of Engineering, University of Porto, Porto, Portugal

Luís Sanhudo

BUILT CoLAB – Collaborative Laboratory for the Future Built Environment, Porto, Portugal, and

João Santos Baptista

Associated Laboratory for Energy, Transports and Aeronautics (LAETA/PROA), Faculty of Engineering, University of Porto, Porto, Portugal

Abstract

Purpose – This study aims to review recent advances towards the implementation of ANN and NLP applications during the budgeting phase of the construction process. During this phase, construction companies must assess the scope of each task and map the client's expectations to an internal database of tasks, resources and costs. Quantity surveyors carry out this assessment manually with little to no computer aid, within very austere time constraints, even though these results determine the company's bid quality and are contractually binding.

Design/methodology/approach – This paper seeks to compile applications of machine learning (ML) and natural language processing in the architectural engineering and construction sector to find which methodologies can assist this assessment. The paper carries out a systematic literature review, following the preferred reporting items for systematic reviews and meta-analyses guidelines, to survey the main scientific contributions within the topic of text classification (TC) for budgeting in construction.

Findings – This work concludes that it is necessary to develop data sets that represent the variety of tasks in construction, achieve higher accuracy algorithms, widen the scope of their application and reduce the need for expert validation of the results. Although full automation is not within reach in the short term, TC algorithms can provide helpful support tools.

Originality/value – Given the increasing interest in ML for construction and recent developments, the findings disclosed in this paper contribute to the body of knowledge, provide a more automated perspective



on budgeting in construction and break ground for further implementation of text-based ML in budgeting for construction.

Keywords Artificial neural networks, Budgeting for construction, Construction management, Natural language processing, PRISMA review

Paper type Literature review

1. Introduction

The tendering process is a crucial phase in the construction project for companies. The quality of bids can determine whether companies are awarded construction contracts, as well as their loss or profit from the project (Aman and Azeanita, 2021; Moon *et al.*, 2022). Additional costs can lead to severe problems for all stakeholders, including possible work stoppages due to the lack of resources (Alaka *et al.*, 2019; Pessoa *et al.*, 2021; Suneja *et al.*, 2021). One solution to mitigate these difficulties is to predict the real costs of the work accurately, i.e. good cost predictions can promote better planning and avoid later setbacks (Wang *et al.*, 2021a).

However, achieving a good cost prediction is challenging, as the tendering process demands a careful examination of the Bills of Quantities (BoQ), project specifications, project requirements and the choice of contract, all of which can influence the final cost. These challenges are aggravated by the austere time constraints in which this task must be performed and the multiple variables that must be considered simultaneously to achieve an accurate prediction (e.g. project duration, resource availability and construction quality or on-site safety, among others) (Jafari *et al.*, 2021). As each project is prepared considering different assumptions with distinct degrees of certainty, these deliverables become susceptible to errors and omissions (Martins, 2009).

Since traditional workflows for cost estimation in the early stages of construction can result in constraints for stakeholders (e.g. work stoppages due to the lack of resources and budget sheets (Elhegazy *et al.*, 2021), there is a need for decision support tools that mitigate these problems, avoiding the perpetuation and accumulation of errors throughout the construction process (Jacques de Sousa *et al.*, 2022a).

To this end, the architectural engineering and construction (AEC) industry can rely on recent artificial intelligence (AI) developments, using machine learning (ML) and natural language processing (NLP) methods to perform budget probabilistic analyses and predictions based on large amounts of data that would be impossible for humans to leverage properly otherwise (Juszczuk, 2018a, 2018b; Elmousalami, 2020a; Xue and Zhang, 2020).

As such, the present paper performs a systematic literature review using the PRISMA guidelines to find the primary AI methodologies and research trends applied to this problem (Page *et al.*, 2021). Furthermore, this paper seeks to compile applications of ML and NLP in the AEC sector while recognising that, despite their widespread use in analogous engineering domains, these techniques are only beginning to be implemented in the AEC sector. Indeed, the growing adoption of ML tools for addressing construction challenges has not yet significantly impacted the budgeting process. Budgeting in the construction sector continues to be a manual process, even though NLP and artificial neural network (ANN) have demonstrated the potential for automating some of construction's budgeting procedures (Mukanov *et al.*, 2020).

As a result, while substantial theoretical knowledge exists for developing these tools and algorithms, built mainly upon their successful implementation in other fields, there remains a pressing need to experiment with how these tools adapt to the unique realities of the construction sector. This review sheds light on the necessary adaptations in terms of these tools and the industry itself to identify critical barriers and best practices for successfully integrating NLP and ML technologies within the AEC domain.

In addition, this systematic literature review seeks to identify research gaps in the current literature and to find which methodologies can assist in construction budgeting, enabling the automatic classification of construction tasks. This review provides essential support for future studies that aim to develop software to automate the budgeting process in construction.

The document is organised as follows: Section 2 explains the research methodology and strategy; in Section 3, the findings of the selected bibliography are revealed; in Section 4, these findings are examined and their implications are summarised; and finally, Section 5 presents the conclusions and answers to the research questions.

1.1 Research questions

This paper performs a systematic literature review on applying AI methods for construction budgeting. The following questions structure the objectives of the review:

- Q1. What are the main approaches for implementing AI methods in construction project budgeting?
- Q2. What are the main techniques applied in those approaches?
- Q3. What methodologies are used by the authors in developing these tools and their respective algorithms?
- Q4. To what types of tasks/projects were these algorithms applied?
- Q5. What are the most used programming languages and code libraries?
- Q6. How was the data used during the development of the algorithms obtained?
- Q7. What are the results and performance obtained by these algorithms?
- Q8. What were the primary indicators used to calculate the performance of the algorithms?
- Q9. What is the relationship between these algorithms and their performance?
- Q10. What are the main limitations of the selected literature?
- Q11. What are the research gaps to be explored in the future?

2. Methods

This systematic literature review was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) Statement ([Page et al., 2021](#)), the last updated guidelines of the PRISMA statement ([Moher et al., 2009](#)). PRISMA 2020 checklist is an evidence-based minimum set of items for reporting in systematic reviews and meta-analyses.

While alternative approaches, such as scoping reviews or narrative reviews, may be better suited for specific situations, the PRISMA method for systematic reviews is implemented in this study since this methodology's explicit inclusion and exclusion processes filter a large corpus of literature to select the most relevant publications for a given purpose. Moreover, the documentation of the inclusion/exclusion criteria allows for the replicability of research and, ultimately, for transparency of the review.

In addition, to enhance the value of the study, the gathered metadata from selected articles was subjected to a scientometric analysis using VOSviewer ([van Eck and Waltman, 2010](#)). This analysis facilitated three key components ([Zhong et al., 2019](#)):

- (1) *Co-author analysis*: This analysis revealed the co-occurrence network of authors and countries, illustrating collaboration patterns;
- (2) *Keyword co-occurrence*: This encompassed both keyword co-occurrence networks and keyword evolution networks; and
- (3) *Abstract and title term co-occurrence*: This analysis aimed to categorise the selected documents into labelled clusters.

The keyword and abstract-title co-occurrence analysis helped identify thematic trends and their evolution over time, revealing potential future research directions.

The following sub-chapters will illustrate the application of the PRISMA methodology to the current research.

2.1 Eligibility criteria

A study must address at least one of the topics mentioned in the list of research questions to be considered eligible. Studies considered include case studies and controlled trials, among other research methodologies. Any record with data to measure effectiveness, identify techniques, methods of implementation and the significance of the outcome was considered.

All literature reviews and conference proceedings were excluded. Including conference proceedings and abstracts in systematic reviews is challenging (Scherer and Saldanha, 2019). Leading AEC conferences are often absent from well-known citation databases like Scopus or WoS. In contrast, databases like Google Scholar or BASE, offering broader coverage, include a substantial portion of non-journal sources and conference papers with varying levels of peer review, some of which might be classified as grey literature (Gusenbauer, 2022). This trade-off between quantity and quality becomes crucial when evaluating citation databases for systematic reviews, as it can introduce unwanted bias during article selection. This paper also excludes literature reviews because it aims to cover original approaches towards the automation of text document classification in the budgeting phase.

Focussing on the latest developments, only articles from the last five years were included (2018–2022). Only papers that applied ML and NLP techniques in the context of civil engineering during the budgeting phase were eligible to reflect the application of these technologies in the AEC sector. Lastly, only articles written in English were included, and the most relevant electronic databases in the engineering field (SCOPUS, Web of Science, Dimensions, IEEE and Journal Storage) were used to meet the research questions. While there is no authority governing database selection, these databases are well-established and recognised resources within the academic community. As such, this study restricted its selection to these databases, as they have been identified as comprehensive sources covering impactful publications (Gusenbauer, 2022).

2.2 Search strategy

The keywords considered for this bibliographical research were: “machine learning”, “natural language processing” and “artificial intelligence” to define the techniques that are intended to be applied; “BIM” and “Construction” to address the area where one plans to use the preferred methods; and “Cost” and “Tendering” to specify the part of the construction process to be addressed. Other keywords were added to cover the most comprehensive number of records, resulting in the following search string:

“Machine Learning” OR “Text Classification” OR “NLP” OR “Natural Language Processing” OR “AI” OR “Artificial Intelligence”) AND (“BIM” OR “Building Information Modelling” OR “Construction” OR “Civil Engineering”) AND (“Budget*” OR “Tender*” OR “Specification” OR “Cost”).

Due to the limited number of operators in the Journal Storage database, the following search string was used for this specific case:

“Machine Learning” OR “Text Classification” OR “Natural Language Processing”) AND “Construction” AND (“Budget” OR “Tender” OR “Cost”).

The total number of records was registered for sample statistical analysis. After these initial records were screened and accessed for eligibility, a second search was performed. This second search focused on the references to the selected records, a process usually referred to as backward snowballing (Wohlin, 2014). Each article cited in the initially chosen papers was reviewed to identify additional relevant studies that could be incorporated into the review.

2.3 Data collection process

Qualitative data was extracted from the selected records and stored in an Excel file. The goal was to collect data that accurately answered the research questions. The file was populated with the combined results from the different databases and records. All selected papers were exported and analysed to remove duplicates. All references were managed using the Endnote software.

2.4 Selection process

Figure 1 illustrates the three-phased selection process. In the investigation stage, the selected keywords and boolean operators identified 5,961 records in the electronic databases.

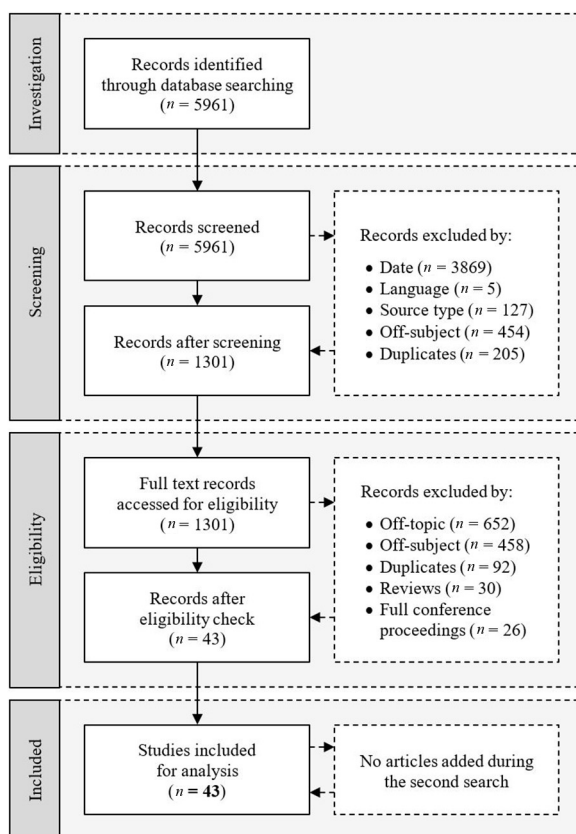
In the screening stage, records were eliminated using the search tools provided by electronic databases: 3,869 records were cleared as they did not fit the date restrictions; five records were removed as they were not written in English; 127 records were eliminated due to the type of source; and, in the particular case of the JSTOR platform, 454 further records were eliminated for being off-subject (e.g. medicine, history, arts). The remaining records were inputted into the software EndNote 20, where the duplicates were automatically eliminated, resulting in 1,301 records.

In the eligibility stage, the entire records were analysed, with the following being removed: 652 records were deemed off-topic; 458 were out of the subject area; 92 were duplicates that had to be manually withdrawn because EndNote did not identify them as such; 30 were reviews; and 26 were full conference proceedings, resulting in a total of 43 included records.

No further relevant articles were identified during the second search, which focused on the included records' references, leading to the final count remaining at 43 records.

2.5 Risk of bias

Two independent authors analysed the selected articles according to an adaptation of the Cochrane Collaboration tool to assess the risk of bias (Montgomery). In research, methodology bias is a systematic error that prevents the impartial consideration of an aspect of the study. Bias is not a binary variable; an author cannot claim that it does not exist (Gerhard, 2008). Authors must take measures to mitigate the level of bias. To do this, one must consider the level of bias and determine its inclusion in the selected sample (Pannucci and Wilkins, 2010). In this review, one can consider the existence of “selection bias” in including some articles. Selection bias is one type of bias that occurs during the selection of the study population. It manifests itself as an impediment to selecting the sample adequately, possibly affecting the study's results (Pannucci and Wilkins, 2010). For instance, in Wang *et al.* (2021b), all the information comes from a BIM model, which can affect the results in terms of accuracy and error. Also, (Akanbi and Zhang, 2020; Dimitriou *et al.*, 2018) use only eight projects of a single scope (wood construction and



Source: Created by authors

Figure 1.
PRISMA workflow

road projects, respectively). Given the significant difference in characteristics, specificities and diversity between construction project specifications over a small sample, a slightly negative impact on the conclusions is expected. Nevertheless, the existence of selection bias does not dispute all the findings of the article and its contribution to this research. Since the results of the implemented methodologies, functionalities, solutions and scope of the studies on which this research stands remain highly relevant, it was determined not to exclude these records.

3. Results

3.1 Record characteristics

The quantitative information collected included: the authors' name, country or region, year of publication, source of the record, number of citations and the project phase concerning the data gathering process. Supplementary Table 1 summarises the characteristics of the selected papers.

Regarding the records' release dates, they span from 2018 to 2022. Over half of the papers were published in the last two years, proving the prominence of the subject area in recent years. Since only records from the previous five years were eligible, a low citation score is

expected. The final sample averaged 3.88 citations per record. The most cited authors in the selected literature were M. Juszczuk, with 41 and 22 citations in two of his articles, and Y. Jallan, with 16 citations. The most prominent researcher is M. Juszczuk, with six published articles, followed by J. S. Zhang, with three publications, and S. Moon in third, with two records. The co-authorship analysis conducted using VOSViewer revealed the existence of two recurring work groups that have made significant contributions to this research area. Specifically, M. Juszczuk collaborated twice with A. Agnieszka, while S. Moon collaborated twice with both G. Lee and S. Chi.

The main sources were: IEEE Access with three records, followed by *Journal of Construction Engineering and Management*, *Journal of Computing in Civil Engineering* and *Automation in Construction*, with two papers in the reviewed literature.

The countries/regions that produced the most research in this area were the USA, with seven records, and Poland, South Korea and China, all with six papers in the subject area.

Finally, the project phase where ML and NLP techniques were used the most were the design phase with nineteen articles, followed by the execution and planning phases with six articles each. The remaining phases, procurement and maintenance, appeared in five and two articles, respectively. Different project phases have different levels of detail. This granularity is essential to understanding the refinement of the data inserted into the models and their purpose.

3.2 Type of task application

The construction industry is an extensive sector with a vast diversity of projects. This diversity adds to the complexity of the ML tools to be developed since it requires them to contain information about multiple variables and documents (Elhegazy *et al.*, 2021). There has been a tendency among researchers to apply their models to specific types of construction projects. In most cases, the authors have tested their algorithms on subsections of a construction project or a particular type of construction project.

Regarding the type of project to which ML and NLP techniques have been most applied, “highway construction” leads the way with six publications (Cao and Ashuri, 2020; Gaussmann *et al.*, 2020; Xue *et al.*, 2020; Moon *et al.*, 2021a; Moon *et al.*, 2021b; Suneja *et al.*, 2021), followed by “composite slabs” (Elhegazy *et al.*, 2021; Juszczuk, 2018a, 2018b) and “sports buildings” with two each (Juszczuk *et al.*, 2018; Juszczuk *et al.*, 2019). The remaining types of projects include: “wood construction” (Akanbi and Zhang, 2020), “stairs” (Zhang *et al.*, 2018), “building renovation” (Cho *et al.*, 2019), “educational buildings” (Yaqubi and Salhotra, 2019), “bridges and piers” (Dimitriou *et al.*, 2018), “electrical installations” (Ronghui and Liangrong, 2021) and “aggregate pavement and bases” (Jeon *et al.*, 2021a).

This concentration of models on a single type of project is mainly due to two reasons: (1) the need to test on a smaller range of variables to obtain acceptable results and, by doing so, improve the efficiency of algorithms; and (2) the difficulty in obtaining information. Data about projects stored in electronic formats is essential for analysis when taking a big data perspective (Kim *et al.*, 2020; Alaka *et al.*, 2019). Only in the last couple of years has the importance of storing data for probabilistic analysis been recognised by the AEC sector companies. Furthermore, in most cases, within the selected bibliography, information is provided by private companies. A public, reliable and open-source repository of this information would be essential for computational learning (Jacques de Sousa *et al.*, 2023).

3.3 Main techniques

When focusing on ML and NLP techniques, there are several approaches to controlling and forecasting costs in the AEC sector. These are discussed in the following sections.

3.3.1 Machine learning. AI can help automate early estimation of the parameters that affect project costs, reducing human error and enhancing prediction accuracy (Sharma *et al.*, 2021). ML is a subset of AI focused on approaches that allow machines to learn from data without being explicitly programmed (Das *et al.*, 2015; Alpaydin, 2010). The “learning” aspect relates to ML algorithms attempting to minimise their associated error while maximising their likelihood of success. There are several types of ML, such as supervised, unsupervised, semi-supervised and reinforcement learning, that differ according to the nature of the training data (Bishop, 2006).

Within the literature reviewed, artificial neural networks (ANNs) were the most commonly applied ML technique. ANNs are mathematical models inspired by how neurons communicate and function in living beings’ brains. These models are usually implemented in computer applications (Juszczyk *et al.*, 2018). Neurons have three main structures: dendrites, cell bodies (or soma) and axons. Likewise, artificial neurons work with three fundamental parts: the synaptic weights, sum and activation functions, emulating the structures identified in real neurons (Pessoa *et al.*, 2021).

In neural networks, neurons may be organised to form layers, creating the structure that characterises the ANN architecture (Chowdhury *et al.*, 2019). In each layer, neurons perform feature extraction or selection (Kumar, 2014; Tajziyehchi *et al.*, 2020). The ANN workflow introduces data from various neurons in the input layer, conveying the accumulated data to subsequent neurons in further layers of the model’s blueprint (Wang and Gibson, 2008; Sharma *et al.*, 2021). These layers are dubbed hidden layers. The number of layers in a model is usually predefined and adapted to a specific problem by the authors to extract the best performance. The data is then processed at these layers and passed to the output layer, yielding the results (Sharma *et al.*, 2021).

End-users do not need expert knowledge to benefit from neural networks’ capabilities. Furthermore, large supporting libraries such as PyTorch and TensorFlow facilitate the creation of these models (Jeon *et al.*, 2021a).

ANNs can be applied to different ends, such as prediction, approximation, association, classification, pattern recognition and data analysis (Zhang *et al.*, 2022). For the specific case of cost estimation in construction, their ability to acquire knowledge during the training process, as well as build and store the gained knowledge, makes them valuable. This allows for generalising and collecting useful patterns in regression problems, making it possible to find dependencies between different variables (Jacques de Sousa *et al.*, 2022a; Juszczyk *et al.*, 2018).

In the reviewed literature, authors apply ANNs to predict project costs using databases, primarily based on construction projects, defined by a set of variables that can correlate with project price variations. The algorithms can determine which variables affect the price the most through the training process. In addition, these algorithms were implemented in different phases of the project. Table 1 presents the techniques mentioned in the selected literature.

3.3.2 Natural language processing. Another solution to increase budget accuracy in the proposal phase is correctly classifying tasks according to the client’s specifications. This can be done with decision-support tools based on NLP. As with ML, NLP is also a subdivision of AI. It encapsulates methods that enable computers to understand natural language information within text files (Wang *et al.*, 2022). The main goal in most NLP models is to convert unstructured text data into a structured representation, namely, through text extraction, TC and topic modelling (Kessler *et al.*, 2019; Zhao *et al.*, 2020; Ren and Zhang, 2021; Shen *et al.*, 2022).

Table 1.
ANN techniques
mentioned in the
reviewed literature

Reference	Technique	Brief description	Phase of application	Input
<i>Dimitriou et al. (2018)</i>	Feed-forward neural networks (FNN)	FNN is one of the simplest forms of ANN. Communication between neurons is only processed in one direction in this neural network. The information can be passed to the subsequent layer of neurons, but never backwards	Planning	68 Bill-of-quantities from road bridge projects
<i>Juszczyk et al. (2018), Pessoa et al. (2021)</i>	Multilayer perceptron (MLP)	MLP is a specific case of FNN, in which every layer is a fully connected layer (FCL). A perceptron is a computational unit used for learning binary classifiers. Perceptrons have weighted input signals and produce output signals based on an activation function. The association of these units in layers creates an MLP network	Design, Planning	129 construction projects defined by ten different variables; 1094 construction projects defined by up to four variables.
<i>Moon et al. (2021b)</i>	Recurrent neural networks (RNN)	Derived from FFNNs, RNNs contain loops that allow for information to be stored and accessed by the network in the future. This information works as valuable experience in forthcoming decision-making, enabling a better performance when facing sequence-based problems (e.g., action classification over time)	Design	4,659 sentences labelled according to five categories of information
<i>Xue et al. (2020)</i>	Convolutional neural networks (CNN)	A convolutional theoretically is an operation on two functions that produces a third. The computational model emulates this operation by stacking convolutional layers, each capable of recognising more sophisticated and complex features within the same data (usually images)	Execution	415 expressways projects defined by five factors
<i>Cao and Ashuri (2020), Jeon et al. (2021b), Cheng et al. (2020)</i>	Long short-term memory (LSTM)	LSTM networks are improvements over an RNN. LSTM includes units that can maintain information in memory for long periods. It is possible to control when information enters the memory, when it is outputted and when it is forgotten. This is possible by using three gates: "Input", "output" and "forget" gates. The input gate decides how much information from the last sample is stored in memory, the output gate determines how much data is transmitted to the subsequent layer and forget gates control the rate at which memory is eliminated. This structure allows for longer-term dependence analysis	Execution	13 projects, defined by five different variables; 11,060 sentences manually labelled to rule-based classification; Cost indexes collect over 20 years

(continued)

Reference	Technique	Brief description	Phase of application	Input
Juszczyk <i>et al.</i> (2018)	Radial basis function (RBF)	RBF networks consist of an input vector followed by a layer of RBF neurons and an output layer comprised of a set of neurons. This algorithm classifies the similarity between the input points and points from the training set (prototypes) that each neuron stores in memory. Each neuron computes the Euclidean distance between the input and its prototype. From this comparison, a similarity measure of 0 to 1 is produced. When the input is equal to the prototype, the value is 1; when they are not similar, the value drops exponentially to 0	Design	115 construction projects defined with ten variables

Source: Created by authors

Table 1.

In the specific case of TC, this task supports quantitative and qualitative analysis of the information collected from the text file (Baker *et al.*, 2020). There are two main approaches to completing this task: (1) rule-based, relying on predefined and hand-coded rules to classify the text; and (2) AI-based, using supervised ML algorithms to classify the text into predefined categories (Zhou *et al.*, 2022). Within construction, the second approach is more advantageous, given that construction budgets vary from project to project following different assumptions and that there is a lack of standardisation of deliverables in the tender phase. These approaches usually follow the steps described below:

- *Data preparation and pre-processing*: Transform raw data into a labelled data set to develop and train the TC models;
- *Feature selection*: Identify discriminating features in the data set, with most applications utilising techniques like tokenisation, lemmatisation, parts-of-speech tagging and stemming. The data set consists of text and words that ML algorithms cannot directly identify. These terms are converted into feature vectors;
- *Algorithm selection and training*: Code different NLP algorithms and train them with a large part of the pre-prepared data set;
- *Model testing*: Test the different NLP algorithms with the remaining part of the data set; and
- *Model evaluation*: Evaluation of the performance of the algorithms in terms of accuracy and errors via metrics (Ul Hassan *et al.*, 2020; Akanbi and Zhang, 2021).

When focusing on the specific case of cost estimation in construction, NLP algorithms can be used to analyse and classify BoQ and project specifications using linguistics and data science concepts to assist technicians' decision-making (Baker *et al.*, 2020). The efficiency and accuracy of the NLP model in tasks such as text retrieval and semantic analysis prove to be an asset for the decision-making process in the budgeting phase of a construction project (Li *et al.*, 2020; Schönfelder *et al.*, 2022).

Like ANNs, NLP is also made up of various techniques and approaches. Table 2 presents those mentioned in the reviewed literature according to three categories: TC, text processing and text vectorization.

3.4 Programming languages and machine learning packages

There are several ways to write code for ML and NLP applications. In the reviewed literature, the primary language used was Python with six entries, followed by MatLab with three, and, finally, R and SPARQL, both used in one of the studies. Large open-source ML libraries facilitate the implementation and development of these algorithms. Python seems to have a slight advantage, with most identified libraries working with this language. However, some libraries can serve several languages, such as TensorFlow. One may separate these libraries into categories: TensorFlow, PyTorch and Keras are ML libraries, with the latter specialising in ANN, while Spacy, NLTK and FastTextAI are NLP-support repositories.

3.5 Data collection

Data access represents the greatest conceptual challenge to the development of ML algorithms. Data access is essential in the learning process of ML models and, therefore, fundamental to the goal of cost estimation (Elmousalami, 2020b). Since implementing NLP and ANN techniques requires a large amount of data to support their training phases, the lack of a reliable database can be prohibitive to implementing these techniques. A high-quality data set is crucial for

Reference	Technique	Brief description
<i>Text classification</i> Jafari <i>et al.</i> (2021)	Naïve Bayes (NB)	NB is a probabilistic algorithm that uses the Naive Bayes equation to calculate the most likely classification. According to the literature sample, it is one of the most widely used algorithms for classifying text documents
Alaka <i>et al.</i> (2019), Baker <i>et al.</i> (2020), Juszcyk (2018b)	Support vector machine (SVM)	An SVM algorithm can classify an example set into two categories. In other words, this method is a binary linear classifier. SVM puts the training points on a plane and separates them into two intervals. The test points are then mapped into that same space and classified according to which side of the interval they fall into
Tajzیهchi <i>et al.</i> (2020), Ul Hassan <i>et al.</i> (2020)	K-Nearest neighbours (KNN)	KNN is based on the premise that similar data is found close to each other. KNN captures the idea of similarity using mathematical equations. Often, this similarity is calculated by the distance between points using simple equations like the Euclidean distance, although there are many other ways to calculate this distance
Bloch and Sacks (2020)	K-Means clustering (KMC)	K-means clustering is an unsupervised learning algorithm. Although it also has the letter k in its name, it is a different method than KNN. This method uses an iterative process where k is the number of clusters to find in the database, and this number is defined as a priori. Each data point is assigned to the closest k. After all objects are assigned, the positions of the k centroids are recalculated. This process is repeated until the k centroids do not change position
Jallan <i>et al.</i> (2019), Hong <i>et al.</i> (2021)	Latent dirichlet allocation (LDA)	LSA is a good algorithm for topic building, a subproblem of NLP. For this purpose, the algorithm in question takes a geometric approach. In this geometric approach, a plane is created (Dirichlet distribution), where each vertex is a classification category and each point inside this plane is a document. The number of classes is defined previously. The number of categories will determine the number of dimensions of the plan. A second Dirichlet distribution is formed where the vertices of the plan are terms within the documents, and the points within that plan are the topics. These terms within the documents constitute another geometric space. These distributions are associated with multinomial distributions. From the first distribution, we get topics, and from the second one, combinations of terms. The association of these two distributions forms new classified documents that try to replicate the initial input ones. N documents are created, corresponding to the N input documents (corpus). By comparing this corpus with the original one, we obtain the precision of the results
Hong <i>et al.</i> (2021)	Latent semantic analysis (LSA)	Latent semantic analysis is an unsupervised algorithm for classifying topics in documents or text. This technique is used to find hidden topics within the text. Hidden topics are then used to group similar documents ("clustering"). The LSA returns concepts instead of topics; concepts are combinations of words that describe the document. LSA works by performing a matrix decomposition on the document-term matrix using singular value decomposition (SVD) to reduce the computational complexity and increase the algorithmic efficiency. SVD decomposes the term co-occurrence matrix into three different matrixes: orthogonal

(continued)

Table 2. NLP Techniques mentioned in the reviewed literature

Reference	Technique	Brief description
		column matrix, orthogonal row matrix and one singular matrix. The product of these matrixes represents the term co-occurrence matrix
Tajziyehchi <i>et al.</i> (2020), Ul Hassan <i>et al.</i> (2020), Yaqubi and Salhotra (2019)	Random forest (RF)	As its name implies, RF consists of a broad group of singular decision trees that run as an ensemble. RF can be used for classification and regression tasks. Every individual tree in the RF yields a class prediction, and the most recurrent class becomes the model's predictions. It is advantageous because it creates an uncorrelated prediction in every individual tree through bagging and feature randomness
Pessoa <i>et al.</i> (2021), Tajziyehchi <i>et al.</i> (2020), Yaqubi and Salhotra (2019)	Gradient boosting regression	GB is an ML algorithm for structured data sets. It is an ensemble method that combines multiple weak models and combines them to achieve better performance as a combined entity. It is capable of finding nonlinear correlations between the model target and features. Similar to RF, it has greater usability as it can deal with missing values and outliers
<i>Text processing</i> Baker <i>et al.</i> (2020)	Bag of words (BOW)	BOW simplifies the representation of text in NLP applications. The BOW method takes all unique words from a corpus of text and stores the frequency of occurrence of these unique terms. This frequency metric represents the text or documents and can help algorithms select features in the training phase to enable later text classification
Kessler <i>et al.</i> (2019)	N-gram analysis	N-grams are combinations of adjacent words or letters of length n. An n-gram is a phrase made of n-words: a 1-gram is a single word, a 2-gram is a phrase made of two words and so forth. The most advantageous length of the n-grams depends on the type of utilisation.
Moon <i>et al.</i> (2021b)	Named entity recognition (NER)	NER is a subtask of information extraction that aims to identify and classify rigid designator members (named entities) from data such as organisations, people and places, among others (Goyal <i>et al.</i> , 2018)
Kim <i>et al.</i> (2020), Guo <i>et al.</i> (2021)	Part-of-speech tagging (POS)	POS tagging is to mark words in a sentence to a POS. POS includes nouns, verbs, articles, adjectives, prepositions, pronouns and many other categories. POS tags are used to indicate lexical and functional categories of words
<i>Text vectorization</i> Hong <i>et al.</i> (2021), Jeon <i>et al.</i> (2021a)	Word2Vec	These algorithms use neural network models to learn the association between words in a text with a large corpus. These models, once trained, can detect synonyms or suggest similar words. Word2Vec represents each word as a vector. These vectors are an optimised way of representing words in NPL applications which, when examined by functions such as cosine similarity, can determine the level of resemblance between vectors
Moon <i>et al.</i> (2021a), Moon <i>et al.</i> (2021b)	Doc2Vec	Like Word2Vec, this method represents documents in vector form, as the name implies. It uses the same word-vector representation as Word2Vec and adds a new vector specific to each document (paragraph vector). In the word vector training phase, the document vector is also trained and holds the numerical representation of a document. This representation is helpful in

(continued)

Table 2.

Table 2.

Reference	Technique	Brief description
Jeon <i>et al.</i> (2021a)	GloVe	NLP applications as it allows training for future classification of topics in documents GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Words are mapped in space, and their distance is related to their semantic similarity. It is an open-source project released by Stanford University. It bases itself on a log-bilinear regression model capable of word analogy, word similarity and NER tasks (Pennington <i>et al.</i> , 2014)

Source: Created by authors

computational models to gain experience in correlations and tendencies, all while dealing with the subjectivity of human communication (Somntag, 2004). There was a significant discrepancy in the number of projects used to deliver the model data set in the reviewed literature. While Yaqubi and Salhotra used only ten educational projects developed in India, Juszczuk *et al.* used 129 sports complex projects. In another study, Baker *et al.* gathered close to 2,345 safety reports from a United Kingdom-based construction company and an undetermined number of projects, while Jeon, Xu *et al.* relied on 2,736 sentences extracted from just one specification document. Naturally, this discrepancy raises questions about the quality of the data and the type of information the authors can extract from them. The authors must process the data to create a meaningful and compatible sample with their models.

The most widely used method in the reviewed literature is a repository of specifications containing documents such as BoQs and planning schemes dedicated to construction tasks. Most of the work studied started with digital text files, which were then classified. However, there is also a more ambitious approach starting from BIM models (Guo *et al.*, 2021; Juszczuk, 2018a, 2018b; Wang *et al.*, 2021b).

In the literature reviewed, there was no correlation between the number of projects and the performance of the models. Despite this, one can conclude that more extensive and higher-quality data sets result in better algorithm performances. This is due to the efficiency of these processes resulting from experience acquired during the training phase.

3.6 Evaluation metrics and performance

To determine the algorithms' effectiveness and accuracy, the authors of the reviewed literature applied several evaluation metrics. These metrics are essential in deciding which algorithms perform the best, ensuring a proper model selection for each situation. In addition, they represent the algorithm's quality and its predictions' accuracy and illustrate the models' ability to achieve their objectives realistically. These metrics are assessed in the following section according to the types of problems they measure: classification problems and regression problems.

3.6.1 Reported evaluation metrics. In classification problems, a confusion matrix organises the number of correct and incorrect model predictions into rows and columns, where each row corresponds to a predicted class and each column represents an actual class. In two-class classification problems, it is possible to organise the results into four categories: true positives (TP), true negatives (TN), false positives and false negatives. Table 3 showcases evaluation metrics for classification that rely on these four categories. A perfect classifier has precision and recall equal to 1, with both metrics being dependent on each other, as one can improve recall at the expense of precision and vice versa. Because of this,

these metrics should be reported together, as they allow further measurements to be made, such as the F1 Score.

In regression problems, several metrics can be used to evaluate an algorithm's performance. Table 4 presents the metrics identified in the reviewed literature.

3.6.2 *Reported efficiency and accuracy.* The results reported by the authors of the reviewed literature can give an idea of the idealised tool's impact. These results range from highly effective, reasonable or less effective than humans.

Jeon *et al.* showed that their ANN-based long short-term memory (LSTM) and convolutional neural network (CNN) could accurately extract and classify information from official documents, achieving an accuracy of 92% (Jeon *et al.*, 2021b). Suneja *et al.* reached a mean absolute percentage error (MAPE) of 70% in their regression model, while Hong *et al.* achieved an F1 score of 88% in their text clustering problem, showing satisfactory results. Jallan *et al.* proved that, in a pilot implementation, their AI algorithm could not provide similar results to manual experts regarding content analysis. By combining LSTM and named entity recognition (NER), Moon *et al.* reached a 91.9% precision and 91.4% recall, while Jeon *et al.* reached similar values by combining GloVe and CNNs, achieving 91.9% accuracy (Jeon *et al.*, 2021a). The experiment comparisons conducted by Zhao *et al.*, who tested different types of TC algorithms, helped conclude that the tested CNN algorithms outperformed the remaining options, such as K-Nearest neighbours (KNN) and SVM, achieving an F1 score of 87.7%. It is concluded that the industry standard concerning the level of accuracy must be upwards of 90% to be satisfactory. Moreover, the algorithms that stand out as the most successful are CNNs coupled with techniques that enhance them.

Juszczyk *et al.* tested the predictions of two ANNs (multilayer perceptron [MLP] and radial basis function [RBF]). Using the MLP type of ANN, the authors achieved an average MAPE error of less than 15%, translating it into acceptable results (Juszczyk *et al.*, 2018). The same cannot be said for the RBF variant, which the authors discarded as unsuitable for cost predictions. In the same line, Pessoa *et al.* obtained a MAPE error below 6% in the forecasts of their MLP algorithm. Also, using the MAPE indicator, Xue *et al.* obtained a 17% prediction error with their CNN-based model (Xue *et al.*, 2020). In another work, M. Juszczyk applied an SVM algorithm for cost prediction and calculated an average RSME error of 15 and MAPE of 5%. This latter author confirms that the trend for the various cost prediction algorithms regarding MAPE error is around 20% (Juszczyk, 2020).

Algorithm results may not always be more accurate than those performed by humans, but they are often similar. The great advantage is that they are performed in a fraction of the

Reference	Indicator	Definition
Akanbi and Zhang (2021), Hong <i>et al.</i> (2021), Jeon <i>et al.</i> (2021b), Zhao <i>et al.</i> (2020), Moon <i>et al.</i> (2022)	Precision	Precision is the fraction of relevant instances among all retrieved instances
	Recall	Recall is the fraction of retrieved instances among all relevant instances
	Accuracy	Accuracy is the fraction of the sum of TP and TN divided by the total of measurements. It states how often the model is correct
	F1 Score	F1-score is a measure of a model's accuracy on a dataset. It evaluates binary classification systems, classifying examples into "positive" or "negative". The F-score is a harmonic mean of Recall and Precision

Table 3.
Supervised learning
evaluation metrics
utilised in the
reviewed literature

Source: Created by authors

Reference	Indicator	Definition
Cheng <i>et al.</i> (2020), Ronghui and Liangrong (2021), Sumeja <i>et al.</i> (2021)	RMSE	Root mean square error (RMSE) is the square root of the average of the squared differences between the estimated and the actual value of the variable/feature. RMSE is the standard deviation of the residuals, which are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it measures how concentrated the data is around the line of best fit
Juszczyk and Leśniak (2019), Juszczyk <i>et al.</i> (2019), Juszczyk (2020), Cheng <i>et al.</i> (2020), Sumeja <i>et al.</i> (2021)	MAPE	Mean absolute percentage error (MAPE) measures the accuracy as a percentage and can be calculated as the average absolute percentage error for each time period minus actual values divided by actual values
Cheng <i>et al.</i> (2020), Sumeja <i>et al.</i> (2021)	MAE	Mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. Examples of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time and one technique of measurement versus an alternative technique of measurement
Cheng <i>et al.</i> (2020)	R2	R2 or coefficient of determination is a regression score function. The results span from 1 to -1 (1 being the best possible). R2 measures the proportion of the change in the dependent variable from the independent variables

Source: Created by authors

Table 4.
Regression problems
evaluation metrics
used in the reviewed
literature

time compared to a technician, proving to be an excellent tool to assist during the budgeting process (Park *et al.*, 2021). The best practice for testing and building models is to test various algorithms for the problem. The algorithms' accuracy results allow the authors to know which algorithms best fit the specific problem. A model with high accuracy for one situation may not perform well in another. These results depend largely on the base data and the models' objectives.

3.7 Keyword co-occurrence analysis

The keyword co-occurrence analysis was done using VOSViewer software on the 43 selected research articles. This analysis employed full counting and set a minimum threshold of two keyword co-occurrences. This approach was chosen to provide a comprehensive overview of the relationships between the selected articles, as increasing the minimum threshold would yield overly restricted results.

As illustrated in Figure 2, the 16 keywords that met the analysis criteria created 91 links and were organised into three distinct clusters. The first cluster revolved around "artificial intelligence", the second cluster centred on "natural language processing" and the third cluster was focused on "machine learning". Cluster formation in the analysis reveals distinct trends: AI is commonly utilised for construction management, often involving regression analysis. ML applications typically employ ANNs and SVM, requiring substantial data sets for effective model training. Finally, NLP is applied across various facets of construction, including management, cost estimation and information retrieval. Furthermore, NLP techniques find application in enhancing BIM for improved management and information retrieval.

The top five most frequently occurring keywords were as follows: "artificial intelligence" (frequency, $f = 10$); "machine learning" ($f = 8$); "natural language processing" ($f = 8$); "artificial neural network" ($f = 7$); "cost estimation" ($f = 5$). The presence of the top five keywords suggests that there is active experimentation with the applications of AI, ML, ANN and NLP in the domain of cost estimation within the construction field. The top five keywords with the

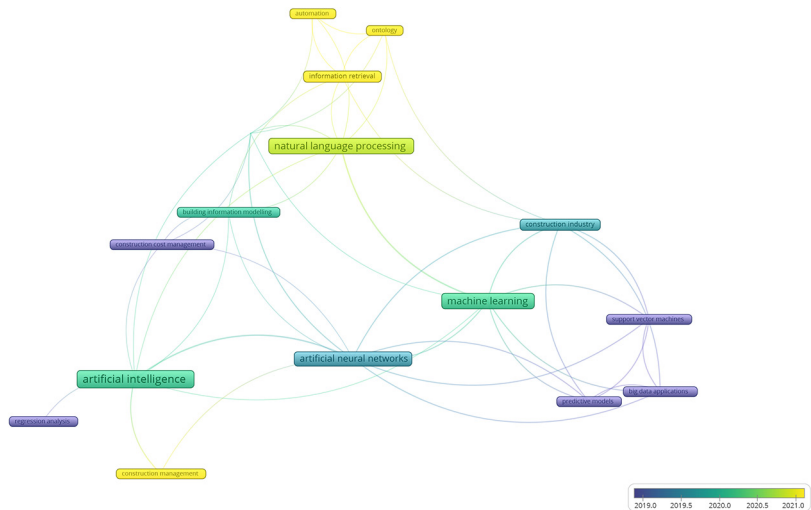


Figure 2.
Keyword co-occurrence overlay visualisation

Source: Created by author using VOS Viewer Software (van Eck and Waltman, 2010)

most links, along with their respective frequencies (f), are as follows: “artificial neural networks” ($f = 10$), “artificial intelligence” ($f = 8$), “machine learning” ($f = 8$), “construction industry” ($f = 7$), “natural language processing” ($f = 7$) and “cost estimation” ($f = 7$). Based on the provided frequencies, it suggests that ANN is the most frequently used technique for cost prediction in the AEC sector and is more commonly used than NLP.

From Figure 2, it is evident that three distinct temporal groups can be observed. SVM and regression analysis research dates back several years, while ML and ANN occupy a middle position along the timeline. Notably, NLP research has only recently begun. This observation suggests a research gap concerning the study of NLP for construction cost estimation and automation. The emerging interest in NLP indicates a potential new area for exploration, highlighting the need for further investigation in this field as it is still in its early stages.

3.8 Abstract-title term co-occurrence analysis

Similarly to the keyword analysis in the previous section, an abstract-title co-occurrence analysis was conducted using VOSViewer software on the selected records. This analysis involved binary counting, which considered the presence of a term in the abstract without considering the frequency of its occurrence within the same abstract. The study also used a 60% threshold for selecting the most relevant terms, per the standard recommendations from VOSViewer, and a minimum occurrence threshold of 3. This particular configuration was the most suitable after multiple attempts, as it provided a comprehensive overview of the key terms within the corpus without generating excessive noise or too many irrelevant terms.

These analysis criteria resulted in the identification of 56 terms organised into four distinct clusters. Figure 3 illustrates that the red cluster is centred around the term “artificial neural network”, the green cluster is associated with “natural language processing”, the blue cluster is related to the word “requirement” and the yellow cluster pertains to “prediction”.

Within the red cluster, critical terms associated with ANN include “cost estimation”, “forecasting” and “construction management”. This observation suggests that ANN is predominantly employed for developing cost prediction models in the context of construction management. In the green cluster, NLP is associated with terms such as “document”, “specification”, “classification” and “information extraction”. This pattern suggests that NLP is utilised in the context of construction documents, particularly construction specifications. Moreover, NLP is being applied in developing classification models for information extraction tools. The blue cluster contains words such as “requirement”, “maintenance”, “contractor” and “execution”. This highlights additional information sources for ANN and NLP models, especially the association between term “requirements” and NLP since it is close to the green node. In the yellow cluster, terms such as “text mining”, “logistics regression”, “support vector machine” and “prediction” are grouped. This cluster emphasises various methods used for prediction and underscores the importance of data for these models. Notably, the proximity of the term “prediction” to the ANN node suggests that ANN models are primarily used for prediction tasks. Furthermore, the presence of two links from “text mining” to both NLP and ANN nodes indicates that both techniques are capable of text mining activities.

Consistent with the keyword analysis, the temporal dispersion of terms in the title and abstract reveals that technologies like ANN, SVM and prediction are more prominent in research conducted around 2019 or earlier. On the other hand, terms such as NLP, requirement and specification are more prevalent in research closer to 2022 or the present time. This temporal pattern indicates evolving research trends and areas of focus within the AEC sector.

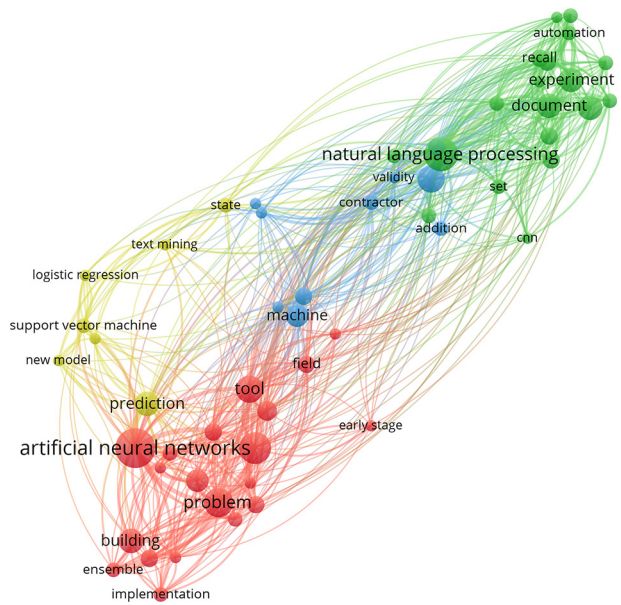
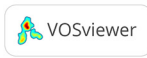


Figure 3.
Abstract-title co-
occurrence network
visualisation



Source: Created by author using VOS Viewer Software (van Eck and Waltman, 2010)

4. Discussion and future study directions

4.1 Summary of evidence

This review presents the findings of 43 selected papers. The reviewed bibliography is classified and studied, considering essential themes for the question defined in Section 1.

Highway construction was the primary type of project where ML (most notably ANNs) and NLP techniques were applied for budgeting purposes. The most common ANN algorithm was LSTM. In the case of NLP, various algorithms were used with the same frequency. Most of the code mentioned in the literature uses Python language, while Spacy is the most common code library. Most authors obtained data from private companies that provided information from past projects. F1-score, MAPE and root mean square error (RMSE) are the most applied indicators to measure results. The results reported by the authors range from highly effective to less effective than humans. A model with high accuracy in one situation may not be suitable in another. Different algorithms must be tested to find the best technique for each case. The main challenges reported by the authors were the difficulties in finding large, high-quality data sets for training. Sometimes, because of the limited diversity of this information, the accuracy of the algorithms may be overestimated and may not reflect reality. Experts need to validate the predictions in some models due to the significant subjectivity of natural language. In conclusion, the temporal variations in keywords suggest that the use of NLP applications in cost estimation is in a relatively early phase of development compared to the application of ANN.

4.2 Limitations of the analysed studies and future work

The highlighted limitations in the reviewed literature may define future research gaps and pave research paths. Indeed, Jallan *et al.* found that the topics their algorithm identified were too generic and did not have sufficient detail to see trends that might be expected between different types of construction projects. Wording defects related to technical details of construction are complex and misleading. Because of this, additional analysis by human agents was still required for meaningful interpretation of the topic selection (Jallan *et al.*, 2019).

Access to data is one of the main barriers to disseminating ML applications in construction. Jafari *et al.* extraction model's accuracy was influenced by its small training sample and the effort required to collect pertinent information. Tajziyehchi *et al.* and Hassan *et al.* found that their studies had limited data sets, hinting at the data sourcing problem described in Section 3.5 of this paper. Moreover, in the study by Bloch and Sacks, the sample size was deemed small, and the results obtained cannot be generalised as they are not representative of all codes in the AEC industry. In addition, the number of classes identified in the regulations is repetitive, preventing more classes from being found.

The authors conclude that contract documentation remains an immature area of practice, and there is a need to find more reliable and efficient approaches. To this end, Hassan *et al.* propose that authors develop models on larger data sets comprising project requirements of many construction projects and evaluate the different feature extraction methods to examine their effect on classification accuracy. In addition, future works should explore how to perform semantic enrichment of the classes based on the research already conducted (Bloch and Sacks, 2020) and be able to retain previous results to gain experience, producing better results that consider this experience when making new predictions (Wang *et al.*, 2021a).

Companies still have reservations about data sharing, even if the projects have been closed according to company rules. In fact, Kim *et al.* point out that in other research areas, ML algorithms are at a more advanced stage of development, and the AEC sector can take advantage by mimicking them. Ji *et al.* and Alaka *et al.* suggested that future studies should find a way to automate the download of large amounts of construction information from firms to develop higher-reliability algorithms.

Park *et al.* state that models should be tested on several projects rather than a single type in their study. Although obtaining data is difficult, applying it to a single type of construction does not reflect the algorithm's real applicability. For example, Akabi and Zhang only applied their algorithm to wood elements observed in the development data and construction specifications that followed a specific format (Akanbi and Zhang, 2021). This trend is seen in more studies as authors reduce the scope of the models to focus on a specific part of a construction project to obtain better results from the information provided. Analogously, Ren and Zhang pointed out that the model developed only used data from the execution stage and construction procedural documents. Future works should introduce a greater variety and typology of documents in the models (Ren and Zhang, 2021). Still, on this subject, Guo *et al.* highlighted that creating different training data sets and ML algorithms for each construction regulation is not feasible. There is the need to verify if a training data set or an algorithm of ML can be used for different regulations checking.

The literature shows that ANNs are predominantly employed for developing cost prediction models and predicting the budget of projects according to different project features. ANNs are commonly used for construction management, often involving regression analysis. Conversely, NLP is used for indirect approaches to construction budgeting, such as measurement rule extraction, enhancing BIM models for improved management, information retrieval and contract collusion detection. Furthermore, the research on budget prediction algorithms has received more attention than NLP applications

for budgeting. This distinction emphasises a broader research gap within NLP applications, indicating the pressing need for further investigation in this field, as it is still in its nascent stages and holds significant unexplored potential.

In summary, these limitations identify four main challenges:

- (1) reducing the need for final confirmation of classifications by human agents;
- (2) the need to test the applicability of the developed algorithms to different tasks rather than to an exclusive type of task;
- (3) the creation of a data set transversal to all tasks in the construction industry that can be used openly by the scientific community; and
- (4) developing efficient and effective algorithms, that save time for technicians while displaying good accuracy, thus, not giving up on one competence to acquire the other.

4.3 Limitations of the study

The main limitations of this study come from the deliberately implied restrictions on the scope through the inclusion criteria. Therefore, the quality of the included studies may vary, and good-quality studies may have been excluded based on compliance with the prior-defined criteria. Moreover, NLP and ANN techniques have found success and are more commonly used in industries such as industrial or mechanical (Dogan and Birant, 2021). Since the research has been restricted to applications in the construction industry, valuable computational methods may have been excluded.

5. Conclusions

This paper presents a systematic literature review of 43 articles. These papers encompass the most recent advances in using ML for the construction budgeting phase. The present work follows the last update of the PRISMA methodology and applies the general principles of scientific methods in the review processes, namely the reproducibility and transparency of the procedures. One of the main differences between PRISMA compared to state-of-the-art reviews is the focus only on a very specific area of knowledge. This systematic literature review obtained in-depth knowledge of the research area, answering the questions outlined in Section 1.1 of the present article.

A1. There are two main approaches to implementing AI in construction project budgeting. The first method relies on ML, most prominently ANNs, to predict the variables affecting construction's budgeting process. The second one uses NLP to categorise project specifications and assist in producing more accurate budgets;

A2. In the ANN approach, LSTM was the most used algorithm by researchers. In NLP, no type of algorithm stood out above all others;

A3. The methodology generally implemented by the authors was to obtain data from private companies, followed by developing algorithms considering this data. A training phase took place next, followed by a testing phase. For the specific case of the NLP approach, the authors tended to develop several algorithms and select the one that obtained better performance in the studied case, focusing the remaining work on this algorithm;

A4. Algorithms were applied to different types of projects or specific tasks within a project. Highway construction was the most frequent type of project;

A5. Python was the most used programming language. Spacy was the most common support library;

A6. Companies provided the data mainly from past projects. Some authors also found information through government institutions;

A7. The results range from highly effective, reasonable or less effective than humans. The accuracy results let the authors know which tested algorithms best fit the problem. Most of the authors obtained good results. However, it is essential to understand that the results largely depend on the base data and the final goal. A model with high accuracy for one situation may not be suitable for another;

A8. F1-Score, RMSE and MAPE were the most used indicators to calculate algorithm performance;

A9. No technique stood out as bringing indisputably better results than all others. As reported in this paper, the results depend largely on the initial data and, therefore, can have very different performances for distinct situations. However, CNN algorithms, enhanced with other techniques, set a 90% accuracy benchmark for this type of application;

A10. The difficulty in finding a complete, high-quality database limited the authors' works. Contract documentation remains an immature area of practice. The authors reduced the scope to only one type of project or one specific task within that project. Creating different training data sets and machine learning algorithms for each construction regulation is not feasible. Specialist knowledge and manual operations are still required for final evaluations; and

A11. The research gaps identified in this area which outline future directions in text document classification are:

- the need to generalise the algorithms to different tasks and documents used in construction;
- to provide a holistic solution such as using standard formatting for contracts, although the authors of this work recognise the logistical difficulty in achieving this solution;
- to develop or obtain more extensive databases, ideally open-source, allowing for a set of multiple project types and a more accurate evaluation of the models; and
- future applications should be able to perform continuous learning to produce results more consistent with the previously predicted results.

The answers presented above identify the main barriers to the development and application of these technologies in the Construction industry, as well as the main techniques applied and the expected results regarding the effectiveness level and accuracy of these algorithms. Thus, the conclusions drawn in this work can support future initiatives to develop automated solutions for construction budgeting based on text documents such as BoQs or technical specifications.

The significant implication for the development of future ANN and NLP applications in the AEC sector is the fundamental importance of accessing data before developing the tools because, as seen, different algorithmic architectures can achieve acceptable results. If AI techniques are to be implemented in the construction industry, there is a need for a cultural change in how participants treat and share data.

Moreover, there is the need to reduce reliance on human agents for final classification confirmation, broaden the applicability of developed algorithms across diverse tasks in the construction industry and create effective algorithms that enhance technicians' productivity while maintaining high accuracy without trade-offs between these aspects.

Lastly, the literature shows that ANN models are implemented primarily for cost prediction in construction management. At the same time, NLP aids in indirect budgeting approaches such as measurement rule extraction, information retrieval and detecting contract collusion. Finally, the application of ANN for construction budget prediction has

been more thoroughly researched than NLP applications. This distinction underscores a substantial research gap within NLP applications in construction budgeting, stressing the imperative for more comprehensive research.

References

- Akanbi, T. and Zhang, J. (2020), "Automated design information extraction from construction specifications to support wood construction cost estimation", *Construction Research Congress 2020*, pp. 658-666.
- Akanbi, T. and Zhang, J.S. (2021), "Design information extraction from construction specifications to support cost estimation", *Automation in Construction*, Vol. 131.
- Alaka, H., Oyedele, L., Owolabi, H., Akinade, O., Bilal, M. and Ajayi, S. (2019), "A big data analytics approach for construction firms failure prediction models", *IEEE Transactions on Engineering Management*, Vol. 66 No. 4, pp. 689-698.
- Alpaydin, E. (2010), *Introduction to Machine Learning*, The MIT Press.
- Aman, M.S.A. and Azeanita, S. (2021), "Building information modelling for project cost estimation", *Recent Trends in Civil Engineering and Built Environment*, Vol. 3 No. 1, pp. 621-630.
- Baker, H., Smith, S., Masterton, G. and Hewlett, B. (2020), "Data-led learning: using natural language processing (NLP) and machine learning to learn from construction site safety failures", pp. 356-365.
- Bishop, C.M. (2006), *Pattern Recognition and Machine Learning*, Springer, New York, NY.
- Bloch, T. and Sacks, R. (2020), "Clustering information types for semantic enrichment of building information models to support automated code compliance checking", *Journal of Computing in Civil Engineering*, Vol. 34 No. 6.
- Cao, Y. and Ashuri, B. (2020), "Predicting the volatility of highway construction cost index using long short-term memory", *Journal of Management in Engineering*, Vol. 36 No. 4, p. 4020020.
- Cheng, M.-Y., Cao, M.-T. and Herianto, J.G. (2020), "Symbiotic organisms search-optimised deep learning technique for mapping construction cash flow considering complexity of project", *Chaos Solitons and Fractals*, Vol. 138, p. 109869.
- Cho, K., Kim, J. and Kim, T. (2019), "Decision support method for estimating monetary value of post-renovation office buildings", *Canadian Journal of Civil Engineering*, Vol. 46 No. 12, pp. 1103-1113.
- Chowdhury, S., Dong, X. and Li, X. (2019), "Recurrent neural network based feature selection for high dimensional and low sample size micro-array data", *2019 IEEE International Conference on Big Data (Big Data)*, pp. 4823-4828.
- Das, S., Dey, A., Pal, A. and Roy, N. (2015), "Applications of artificial intelligence in machine learning: review and prospect", *International Journal of Computer Applications*, Vol. 115 No. 9, pp. 31-41.
- Dimitriou, L., Marinelli, M. and Fraggkakis, N. (2018), "Early bill-of-quantities estimation of concrete road bridges: an artificial intelligence-based application", *Public Works Management and Policy*, Vol. 23 No. 2, pp. 127-149.
- Dogan, A. and Birant, D. (2021), "Machine learning and data mining in manufacturing", *Expert Systems with Applications*, Vol. 166.
- Elhegazy, H., Chakraborty, D., Elzarka, H., Ebid, A.M., Mahdi, I.M., Haggag, S.Y.A. and Rashid, I.A. (2021), "Artificial intelligence for developing accurate preliminary cost estimates for composite flooring systems of multi-storey buildings", *Journal of Asian Architecture and Building Engineering*, Vol. 21 No. 1.
- Elmoussalimi, H.H. (2020a), "Artificial intelligence and parametric construction cost estimate modeling: state-of-the-art review", *Journal of Construction Engineering and Management*, Vol. 146 No. 1, p. 3119008.

- Elmoussalimi, H.H. (2020b), "Data on field canals improvement projects for cost prediction using artificial intelligence", *Data in Brief*, Vol. 31, p. 105688.
- Gaussmann, R., Coelho, D., Fernandes, A.M.R., Crocker, P. and Leithardt, V.R.Q. (2020), "Using machine learning for road maintenance cost estimates in Brazil: a case study in the federal district", *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, pp. 1-7.
- Gerhard, T. (2008), "Bias: considerations for research practice", *American Journal of Health-System Pharmacy*, Vol. 65 No. 22, pp. 2159-2168.
- Goyal, A., Gupta, V. and Kumar, M. (2018), "Recent named entity recognition and classification techniques: a systematic review", *Computer Science Review*, Vol. 29, pp. 21-43.
- Guo, D., Onstein, E. and Rosa, A.D.L. (2021), "A semantic approach for automated rule compliance checking in construction industry", *IEEE Access*, Vol. 9, pp. 129648-129660.
- Gusenbauer, M. (2022), "Search where you will find most: comparing the disciplinary coverage of 56 bibliographic databases", *Scientometrics*, Vol. 127 No. 5, pp. 2683-2745.
- Hong, Y., Xie, H.Y., Bhumbra, G. and Brilakis, I. (2021), "Comparing natural language processing methods to cluster construction schedules", *Journal of Construction Engineering and Management*, Vol. 147 No. 10.
- Jacques de Sousa, L., Martins, J. and Sanhudo, L. (2023), "Portuguese public procurement data for construction (2015-2022)", *Data in Brief*, doi: [10.1016/j.dib.2023.109063](https://doi.org/10.1016/j.dib.2023.109063).
- Jacques de Sousa, L., Martins, J., Baptista, J., Sanhudo, L. and Mêda, P. (2022a), "Algoritmos de classificação de texto na automatização dos processos orçamentação".
- Jafari, P., Al Hattab, M., Mohamed, E. and AbouRizk, S. (2021), "Automated extraction and time-cost prediction of contractual reporting requirements in construction using natural language processing and simulation", *Applied Sciences*, Vol. 11 No. 13, p. 6188.
- Jallan, Y., Brogan, E., Ashuri, B. and Clevenger, C.M. (2019), "Application of natural language processing and text mining to identify patterns in construction-defect litigation cases", *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, Vol. 11 No. 4, p. 4519024.
- Jeon, J., Xu, X., Zhang, Y., Yang, L. and Cai, H. (2021a), "Extraction of construction quality requirements from textual specifications via natural language processing", *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2675 No. 9, pp. 222-237.
- Jeon, K., Lee, G. and Jeong, H.D. (2021b), "Classification of the requirement sentences of the US DOT standard specification using deep learning algorithms", *Lecture Notes in Civil Engineering*, pp. 89-97.
- Juszczuk, M. (2018a), "Implementation of the ANNs ensembles in macro-BIM cost estimates of buildings' floor structural frames", p. 20014.
- Juszczuk, M. (2018b), "Residential buildings conceptual cost estimates with the use of support vector regression", Vol. 196.
- Juszczuk, M. (2020), "Development of cost estimation models based on ANN ensembles and the SVM method", *Civil and Environmental Engineering Reports*, Vol. 30 No. 3, pp. 48-67.
- Juszczuk, M. and Leśniak, A. (2019), "Modelling construction site cost index based on neural network ensembles", *Symmetry*, Vol. 11 No. 3, p. 411.
- Juszczuk, M., Leśniak, A. and Zima, K. (2018), "ANN based approach for estimation of construction costs of sports fields", *Complexity*, Vol. 2018, pp. 1-11.
- Juszczuk, M., Zima, K. and Lelek, W. (2019), "Forecasting of sports fields construction costs aided by ensembles of neural networks", *Journal Of Civil Engineering and Management*, Vol. 25 No. 7, pp. 715-729.
- Kessler, R., Béchet, N. and Berio, G. (2019), "Extraction of terminology in the field of construction", *2019 First International Conference on Digital Data Processing (DDP)*, pp. 22-26.
- Kim, Y., Lee, J., Lee, E.-B. and Lee, J.-H. (2020), "Application of natural language processing (NLP) and text-mining of big-data to engineering-procurement-construction (EPC) bid and contract documents".

- Kumar, V. (2014), "Feature selection: a literature review", *The Smart Computing Review*, Vol. 4 No. 3.
- Li, R.Y.M., Li, H.C.Y., Tang, B. and Au, W.C. (2020), "Fast AI classification for analysing construction accidents claims", pp. 1-4.
- Martins, J. P. D. S. P. (2009), "Modelação do fluxo de informação no processo de construção: aplicação ao licenciamento automático de projectos", No. Porto.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G. and The, P.G. (2009), "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement", *PLoS Medicine*, Vol. 6 No. 7, p. e1000097.
- Moon, S., Chi, S. and Im, S.-B. (2022), "Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (BERT)", *Automation in Construction*, Vol. 142, p. 104465.
- Moon, S., Lee, G. and Chi, S. (2021a), "Semantic text-pairing for relevant provision identification in construction specification reviews", *Automation in Construction*, Vol. 128.
- Moon, S., Lee, G., Chi, S. and Oh, H. (2021b), "Automated construction specification review with named entity recognition using natural language processing", *Journal of Construction Engineering and Management*, Vol. 147 No. 1, p. 4020147.
- Mukanov, A., Saginbayev, A. and Salikhov, R. (2020), "From field operations to economics: breaking the barriers. Next level of integration", *Society of Petroleum Engineers – SPE Annual Caspian Technical Conference 2020, CTC 2020*.
- Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P. and Moher, D. (2021), "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews", *BMJ*, Vol. 372, p. n71.
- Pannucci, C.J. and Wilkins, E.G. (2010), "Identifying and avoiding bias in research", *Plastic and Reconstructive Surgery*, Vol. 126 No. 2, pp. 619-625.
- Park, M.J., Lee, E.B., Lee, S.Y. and Kim, J.H. (2021), "A digitalized design risk analysis tool with machine-learning algorithm for EPC contractor's technical specifications assessment on bidding", *Energies*, Vol. 14 No. 18.
- Pennington, J., Socher, R. and Manning, C. (2014), "Glove: global vectors for word representation".
- Pessoa, A., Sousa, G., Maues, L.M.F., Alvarenga, F.C. and Santos, D.D. (2021), "Cost forecasting of public construction projects using multilayer perceptron artificial neural networks: a case study", *Ingenieria E Investigacion*, Vol. 41 No. 3.
- Ren, R. and Zhang, J. (2021), "Semantic rule-based construction procedural information extraction to guide jobsite sensing and monitoring", *Journal of Computing in Civil Engineering*, Vol. 35 No. 6.
- Ronghui, S. and Liangrong, N. (2021), "An intelligent fuzzy-based hybrid metaheuristic algorithm for analysis the strength, energy and cost optimisation of building material in construction management", *Engineering with Computers*, Vol. 38 No. S4.
- Scherer, R.W. and Saldanha, I.J. (2019), "How should systematic reviewers handle conference abstracts? A view from the trenches", *Systematic Reviews*, Vol. 8 No. 1, p. 264.
- Schönfelder, P., Al-Wesabi, T., Bach, A. and König, M. (2022), "Information extraction from text documents for the semantic enrichment of building information models of bridges".
- Sharma, S., Ahmed, S., Naseem, M., Alnumay, W.S., Singh, S. and Cho, G.H. (2021), "A survey on applications of artificial intelligence for pre-parametric project cost and soil shear-strength estimation in construction and geotechnical engineering", *Sensors*, Vol. 21 No. 2, p. 463.
- Shen, Q., Wu, S., Deng, Y., Deng, H. and Cheng, J.C.P. (2022), "BIM-based dynamic construction safety rule checking using ontology and natural language processing", *Buildings*, Vol. 12 No. 5.

- Sonntag, D. (2004), "Assessing the quality of natural language text data".
- Suneja, N., Shah, J.P., Shah, Z.H. and Holia, M.S. (2021), "A neural network approach to design reality oriented cost estimate model for infrastructure projects", *Reliability: Theory and Applications*, Vol. 16, pp. 254-263.
- Tajziyehchi, N., Moshirpour, M., Jergeas, G. and Sadeghpour, F. (2020), "A predictive model of cost growth in construction projects using feature selection", *2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pp. 142-147.
- Ul Hassan, F., Le, T. and Tran, D.H. (2020), "Multi-class categorization of design-build contract requirements using text mining and natural language processing techniques", pp. 1266-1274.
- van Eck, N.J. and Waltman, L. (2010), "Software survey: VOSviewer, a computer program for bibliometric mapping", *Scientometrics*, Vol. 84 No. 2, pp. 523-538.
- Wang, B., Yuan, J.J. and Ghafoor, K.Z. (2021a), "Research on construction cost estimation based on artificial intelligence technology", *Scalable Computing: Practice and Experience*, Vol. 22 No. 2, pp. 93-104.
- Wang, J., Gao, X., Zhou, X. and Xie, Q. (2021b), "Multi-scale information retrieval for BIM using hierarchical structure modelling and natural language processing", *Journal of Information Technology in Construction*, Vol. 26, pp. 409-426.
- Wang, N., Issa Raja, R.A. and Anumba Chimay, J. (2022), "NLP-based query-answering system for information extraction from building information models", *Journal of Computing in Civil Engineering*, Vol. 36 No. 3, p. 4022004.
- Wang, Y.-R. and Gibson, G. Jr (2008), "A study of preproject planning and project success using ANN and regression models".
- Wohlin, C. (2014), "Guidelines for snowballing in systematic literature studies and a replication in software engineering", *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, London, England, United Kingdom, Association for Computing Machinery, p. 38.
- Xue, X.R. and Zhang, J.S. (2020), "Evaluation of seven part-of-speech taggers in tagging building codes: identifying the best performing tagger and common sources of errors", *Construction Research Congress (CRC) on Construction Research and Innovation to Transform Society, AZ State Univ, Del E Webb Sch Construct, Tempe, AZ*, pp. 498-507.
- Xue, X., Jia, Y. and Tang, Y. (2020), "Expressway project cost estimation with a convolutional neural network model", *IEEE Access*, Vol. 8, pp. 217848-217866.
- Yaqubi, M.K. and Salhotra, S. (2019), "The automated cost estimation in construction", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 8 No. 7, pp. 845-849.
- Zhang, F., Chan, A.P.C., Darko, A., Chen, Z. and Li, D. (2022), "Integrated applications of building information modeling and artificial intelligence techniques in the AEC/FM industry", *Automation in Construction*, Vol. 139, p. 104289.
- Zhang, J., Chen, Y., Hei, X., Zhu, L., Zhao, Q. and Wang, Y. (2018), "A RMM based word segmentation method for Chinese design specifications of building stairs", 2018 14th International Conference on Computational Intelligence and Security (CIS), pp. 277-280.
- Zhao, H., Pan, Y. and Yang, F. (2020), "Research on information extraction of technical documents and construction of domain knowledge graph", *IEEE Access*, Vol. 8, pp. 168087-168098.
- Zhong, B., Wu, H., Li, H., Sepasgozar, S., Luo, H. and He, L. (2019), "A scientometric analysis and critical review of construction related ontology research", *Automation in Construction*, Vol. 101, pp. 17-31.
- Zhou, Y.-C., Zheng, Z., Lin, J.-R. and Lu, X.-Z. (2022), "Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking", *Computers in Industry*, Vol. 142, p. 103746.

CI
24,7

Further reading

Ji, W. and Abourizk, S.M. (2018), "Data-driven simulation model for quality-induced rework cost estimation and control using absorbing markov chains", *Journal of Construction Engineering and Management*, Vol. 144 No. 8.

318

Supplementary material

The supplementary material for this article can be found online.

Corresponding author

Luís Jacques de Sousa can be contacted at: luisjsousa@fe.up.pt

For instructions on how to order reprints of this article, please visit our website:

www.emeraldgroupublishing.com/licensing/reprints.htm

Or contact us for further details: permissions@emeraldinsight.com