AN AUTOMATED FRAMEWORK FOR ENSURING INFORMATION CONSISTENCY IN PRICE LIST TENDERING DOCUMENT

Chiara Gatto, Maryam Gholamzadehmir, Marta Zampogna, Claudio Mirarchi & Alberto Pavan
Polytechnic of Milan, Italy

ABSTRACT: *Effective cost estimation for tendering plays a critical role in the building construction process, enabling efficient investment management and ensuring successful execution of the construction phase. Traditional cost estimation procedure involves manual information processing to extract and match technical data from textual description construction resources. This activity requires practitioner deep experience and manual effort, often resulting in errors and, in the worst scenario, judicial disputes.*

In response to the increasing demand for structured information and automated processes, this study addresses the need for Public Administrations to achieve better control over the data contained in public tendering documents provided to practitioners. To fulfill this objective, a framework is proposed to automatically retrieve information from these documents, serving as a support tool to map items within the documents, highlight missing data, and critical semantic ambiguity.

The designed framework aims to develop a tool for automatically identifying similarities between work items and their corresponding elementary resource items in Price List tendering documents. By leveraging the information retrieval NLP technique of cosine similarity through TF-IDF, a methodology was developed to support and facilitate practitioners' activities. Finally, the framework was tested on four case studies extracted from Lombardy Regional Italian price list documents showing that the resulting support tool is able to automate the analysis process and efficiently reveal inconsistency. The model successfully extracted and correctly matched the elementary resource to the corresponding work query in 75% of the cases where the elementary resource was present in the list. Additionally, the model proved to be a valuable tool in helping practitioners identify missing resources.

KEYWORDS: *Automated cost estimation, Information retrieval, Text similarity, NLP, Tendering document, Public Administrations*

1. INTRODUCTION

Cost estimation plays a pivotal role in effective decision-making within construction project management. Numerous studies underscore its significance (M.E.Sepasgozar et al., 2021). However, traditional construction cost estimation often involves multiple manual processes, with limited automation, resulting in time-consuming efforts and susceptibility to human errors (Akanbi & Zhang, 2021). Despite the increasing use of BIM approaches, information exchange in AEC industry is still mainly based on the production of paper-based 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. Manually data extracting can lead to discrepancies and inaccuracies in information, thereby posing financial risks, project delays, potential failure, and in the worst case scenario to judicial disputes (Jafari et al., 2021a). These issues impact the effectiveness of the projects along with the credibility/reputation of the stakeholders. Moreover, the gap between traditional document-based (i.e., semi-structured and unstructured) and model-based information can lead to information loss and inconsistency. (Opitz et al., 2014). Thus, effective data management turns out to be essential to the overarching project strategy.

Public administrations play a key role in the construction process, especially in the context of public procurement where they assume the role of contracting authorities. These entities encounter a large daily influx of data, much of which needs to be made accessible to external stakeholders. Among these diverse datasets, a significant portion consists of unstructured textual information. Consequently, there is a need for these administrations to enhance their data management capabilities.

Effectively addressing these challenges necessitates the adoption of a methodology capable of efficiently handling and structuring the considerable volume of semi-structured and unstructured data for tasks such as cost and time estimation. Within this framework, data pre-processing emerges as a fundamental phase, acknowledged for its role as the most time-intensive aspect of text classification.

Referee List (DOI: 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Chiara Gatto, Maryam Gholamzadehmir, Marta Zampogna, Claudio Mirarchi, Alberto Pavan, *An Automated Framework for Ensuring Information Consistency in Price List Tendering Document,* pp. 837-847, © 2023 Author(s), CC BY NC 4.0, DOI 10.36253/979-12-215-0289-3.83

To address the aforementioned issues, and to meet the growing demand for public administrations and practitioners to convert textual information into digital formats, this research proposes a methodology to develop a procedure for checking information consistency within price list tendering. This is achieved through the application of Natural Language Processing (NLP) techniques, ensuring the coherence of information within the document. The methodology focuses on confirming the alignment between work descriptions and the employed elemental resources. The proposed research activity follows the prior study focused on automating the process of structuring data from textual documents (Gatto et al., 2023). Specifically, this research shifts the focus on responding to the public administration's need to assess the consistency of information within the regional price list before structuring and subsequently providing it to the user, by verifying the correspondence between the textual information related to the construction works and the textual information of the respective elementary resources involved.

To minimize semantic ambiguity and enhance machine comprehension without human intervention, data in textual documents can be handled using NLP, which has demonstrated its efficacy in supporting human activity (Zabin et al., 2022). In this direction, (Tang et al., 2022a) developed NLP and rule-based algorithms to automate the information extraction from work descriptions in building construction. They integrated different algorithms such as Hidden Markov model and improved the accuracy by 89% compared to other common named entity recognition algorithms. However, despite the wide application of NLP in different construction fields, the application of these techniques in the pre-design phase is still a research gap in the literature (Locatelli et al., 2022). Data pre-processing of unstructured data is known as the most time-consuming phase of text classification in the whole process (Munková et al., 2013).

This research is organized as follows: The initial section, "State of the Art," presents the research background. Following that, the "Research Methodology" and "Framework Development" sections detail the study's approach and implementation. The subsequent segments, "Testing Framework" and "Results and Discussion," demonstrate the practical application of the framework and its evaluation. Ultimately, the "Conclusion" section encapsulates the key findings from the study.

2. STATE OF THE ART

Manual extraction of reporting requirements from extensive construction documents can lead to time and cost underestimations. In this direction, the application of NLP techniques has been increasingly adopted in the AEC sector to manage the information contained in documents ((Jafari et al., 2021b); (J. Zhang et al., 2020)). NLP is mainly applied in four scenarios of information extraction, document organization, expert systems, and automated compliance checking (Wu et al., 2022).

Recently, NLP has been used in the construction industry to facilitate cost estimation through document management (Tang et al., 2022b). To automate extracting information from construction regulatory documents, a study has been developed by applying a semantic rule-based NLP approach for a text recognition algorithm based on semantic analysis (J. Zhang & El-Gohary, 2016). In a later study, an automated framework was developed using NLP and machine learning techniques to automatically recognize and prioritize important contract terms, enabling managers to quickly and fully understand contract agreements (Hassan & Le, 2020). Furthermore, a model that automatically identifies the most relevant pairs of provisions from various specifications using semantic text similarity was developed (Moon et al., 2021). This assists practitioners by reducing the effort to complete tasks that involve written documents, enhancing the objectivity of outcomes, and minimizing human errors.

In the construction industry, dealing with inconsistent information, semi-structured and unstructured data in price list documents is an ongoing challenge. The causes of these issues often include human errors during data entry, outdated price lists, and issues with the software tools used for creating BIM models and price lists (Cha & Lee, 2018). These issues can cause inaccurate cost estimates, budget overruns, delays in project timelines, and disagreements between project stakeholders.

Concerning the inconsistency and ambiguity checking, a recent research activity has been performed using the support vector machine (SVM) supervised learning model methodology, leading to an automated method for detecting ambiguity in building requirements, which can then be reviewed and interpreted by domain experts to support the automated compliance checking process (Z. Zhang & Ma, 2023). In another work, the authors proposed an uncertain knowledge graph-based method to eliminate potential conflicts and acquire the ‘most likely scenarios’ by integrating multiple representations of building information (Xie et al., 2023).

Document classification is crucial in the process of digitizing and structuring information. Text data pre-processing serves as a foundational step for this process (Lee & Yi, 2017). Text classification as part of NLP, involves the automated categorization of text. This task can be accomplished using two main approaches: rule-based techniques and Machine Learning (ML) algorithms.

The conversion of textual information into digital formats is necessary for the building tendering process. The digitalization process not only enhances the accessibility and usability of the information but also opens up new scenarios for data analysis and decision-making. Application of advanced technologies such as NLP and machine learning, can automate the structure of price lists and show similarities between products and their corresponding resource items. This can help practitioners improve the consistency of information in documents.

2.1 Research gaps and challenges

As the construction industry continues to embrace digital transformation, the application of NLP has emerged as a promising area of research. NLP has the potential to revolutionize various aspects of construction, from project management to BIM. However, the integration of NLP into the construction sector faces some challenges that are explained in this section to highlight future research interests in this field.

(Ding et al., 2022) provided a review of the NLP-related research articles in the construction field and they pointed out related challenges such as data accessibility/monopoly to develop the intelligent agent and data diversity from various devices, such as text, images, sensors, and audio, presenting a challenge in developing comprehensive models with higher performance. Additionally, they mentioned that achieving full automation and high-level reasoning requires advanced extraction and understanding of models because NLP models can struggle with understanding the complex technical context in construction.

Another challenge is achieving semantic interoperability between BIM and NLP in the construction industry. The use of ontology as a bridging tool is a potential solution to address this gap, yet this area remains underexplored (Locatelli et al., 2021).

2.2 Tendering documents

Tender documents serve as a communication tool between the project owner and potential contractors, outlining project specifications, execution conditions, and the rights and obligations of all parties. The clarity of these documents is important to avoid financial disputes. The type of tender documents depends on the procurement method and contract type, and typically include drawings, specifications, and bills of quantities (Cunningham, 2015).

The quality of tender documentation can significantly impact on the procedure, alongside other factors like contract content and tender management. Despite the challenges, contractors must carefully prepare their bids to increase their chances of securing the contract (Leśniak & Janowiec, 2020).

Construction projects, seen as transient businesses, require careful project management, particularly during the tendering process. This process, which involves numerous variables and substantial resources, is influenced by factors such as the financial stability of contractors, offered price, delivery timeline, experience, environmental considerations, and personnel qualifications (Naji et al., 2022).

The Public Italian Contracts Code, art.23 D.lgs 18 Aprile 2016, n.50, imposes on each Italian region to annually provide a price list that contracting authorities have to use for setting the project cost base for tenders. Therefore, each region provides practitioners with a price list containing work items and their respective cost (Sdino & Rosasco, 2021). The tool mainly stores data associated with construction activities, including their unit prices. This resource assists practitioners in generating estimated metric calculations. Additionally, to ensure more transparency in the composition of the price of construction works, the price list provides a catalog of elemental resources involved in the latter. Therefore, there must be a full correspondence between works and elemental resources, otherwise inconsistency arises. Considering the need for annual updates, the price list is subjected to periodical revisions, consisting in the unit prices update, the addition of new work and elementary resource items or removal of outdated entries.

Information is conveyed by the tool in verbal form: sentences composed by words and syntax delivering knowledge. Since each item is written in natural language and because the document doesn’t follow a standard in providing information, a lack of homogeneity has been recorded between each item phrase structure and information typology transmitted. Public Administrations therefore are looking for tools to help them structure

high amount of data.

3. RESEARCH METHODOLOGY

The methodological approach employed in the presented study is explained in this section and summarized as depicted in Figure 1Figure 1 Methodology chapter. Firstly, the development of the study starts by listening to the needs of public administrations in the AEC sector, who have been engaged to comprehend the challenges they face in managing information during public tendering processes. Following the prior study focused on automating the process of structuring data from textual documents (Gatto et al., 2023), a need emerged from public administrations to incorporate a preliminary step to assess the consistency of information within the regional price list before structuring and subsequently providing it to the user. The subsequent step was to explore the state of the art, aiming to delve into tools and methodologies applied for automating the management of unstructured data. This was followed by the development of a framework, leveraging information retrieval NLP techniques. Specifically textual similarity recognition based on cosine similarity using TF-IDF, was selected to build a tool designed to aid public administrations in establishing associations among essential resources within a construction project, thereby highlighting any gaps that may exist. This technique has proved to be valuable and effective within the domain of text similarity and in this study it is applied in the specific field of cost estimation, focusing on price list documents. Finally, the framework was tested in a practical case study to assess its effectiveness.

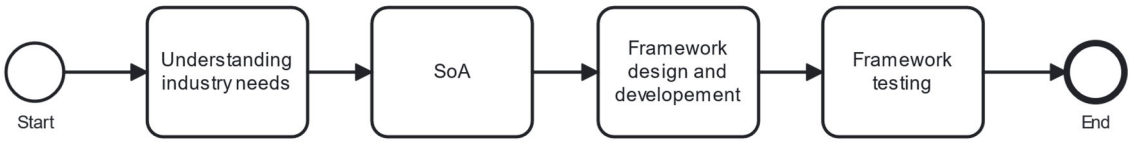

Figure 1 Methodology chapter.

4. FRAMEWORK DEVELOPMENT

This section presents the design and development process of the framework, as synthesized in Figure 2. The objective of this phase is to develop a comprehensive procedure for checking information consistency within price list tendering documents using Natural Language Processing (NLP) techniques, specifically aimed at verifying the correspondence between the textual information related to the construction works and the textual information of the respective elementary resources involved. The primary goal of public administrations is to provide users with a tool that ensures utmost clarity, free from semantic ambiguities and inconsistencies during the crucial phase of estimating construction costs.

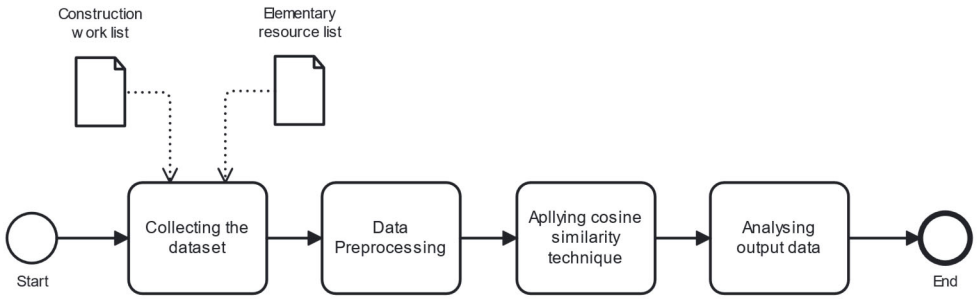

Figure 2 Framework flowchart.

By leveraging NLP technique, this framework allows the public administration to speed up the process of addressing elementary resource item to the work items where it is involved, thus helping with the detection of missing information and inconsistencies, ensuring the accuracy and reliability of how the data is presented to the user. The parameterization of information through NLP techniques empowers public administrations with greater control over the textual dataset, facilitating easier data manipulation and analysis. Furthermore, this process seeks to address the longstanding issue of ambiguity that often arises from cost item descriptions, ensuring a higher level

of accuracy and precision in subsequent analyses and decision-making processes.

The developed framework mainly consists of four stages.

The first step in the process involves collecting the dataset, which comprises both the list of completed works and the corresponding list of elementary resources involved. The work's textual descriptions typically explicitly state the elementary resources utilized in the activity, allowing practitioners to identify them accurately. Once the dataset is assembled, it undergoes a pre-processing phase to ensure the correct execution of the subsequent steps.

Since the description of the elementary resources is expected to be contained in the description of the works, the cosine similarity technique using TF-IDF is chosen for the development of this framework. This measures the similarity between two vectors, A and B, by calculating their dot product and dividing it by the product of their magnitudes as:

$$(|A||B| * \cos(\alpha))/(|A||B|)$$

The resulting value ranges between 0 and 1, where 0 indicates no match (completely dissimilar vectors), and 1 represents complete similarity (vectors pointing in the same direction). This metric is widely used in text analysis to measure document similarity. TF-IDF is primarily concerned with determining the importance of words within individual documents and in literature is commonly used for tasks like information retrieval and document ranking, and in text. Alternatively, the utilization of cosine similarity through Word2Vec is centered around capturing semantic meanings. This is achieved by representing each word as a dense vector within a continuous space. Notably, certain studies have suggested that for text similarity, TF-IDF often outperforms the other method, highlighting its effectiveness (Sitikhu et al., 2019).

Once the dataset has been collected and the NLP technique to be used identified, the next step is to vectorize the list of elementary resources dataset and query them with the respective work descriptions, with the aim of deriving the elementary resource item associations. The cosine similarity technique allows to rank the list of elementary resources based on their similarity to the work descriptions, with the most similar resources being assigned with higher score similarity value.

After obtaining a ranked list of elementary resources associated with each work description, building construction cost estimation practitioners were involved to validate the accuracy and effectiveness of the output results. They carefully reviewed the output and assessed whether the correctness of the output with a higher score rate. The expert validation process not only serves as a critical quality control measure but also provides valuable insights and feedback, enhancing the overall robustness and practical applicability of the framework.

5. TESTING FRAMEWORK

In this section, the testing process of the framework is presented, with the aim of assessing its capability to accurately determine the appropriate elementary resource for each work item. The evaluation was conducted using four sample case studies extracted from the Lombardy Region Price List document (January 2023 version). The objective was to verify the framework's ability to assign the correct elementary resource to four specific types of works: masonry clay block, masonry concrete block, tile floor, and thermal insulation work.

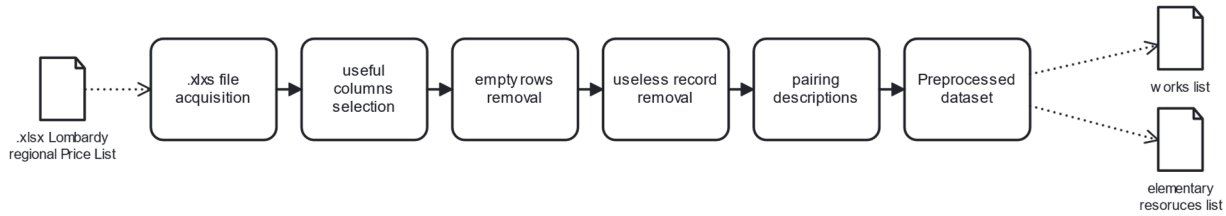

Figure 3 Preprocessing flowchart.

To ensure effective vectorization and handling of relevant information, dataset pre-processing plays a crucial role. The steps undertaken during this phase are depicted in Figure 3. As explained later, the pre-processing is focused on isolating and extracting the pertinent textual data; however, it does not involve further cleaning, such as removing stop words or those that appear with high or limited frequency in the text. The objective is to retain the essential context and meaningful information while preparing the data for vectorization and subsequent analysis.

Data was acquired from a single spreadsheet. where the knowledge organization follows a semi-structured format as shown in Figure 3, consisting of seven columns in the following order: item ID code, textual description, unit of measurement, unit price, percentage of labor incidence, percentage of material incidence, and percentage of equipment incidence.

CODICE	DESCRIZIONE	U.M.	P.U.	% Inc. M.O.	% Inc. MAT	% Inc. ATT
1C.06.100	MURATURE FACCIA A VISTA	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN
1C.06.100.0050	Muratura faccia a vista con mattoni pieni tipo...	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN
1C.06.100.0050.a	- con mattoni 25 x 12 x 5.5 cm, spessore 12 cm	m²	98,31	33,90	41,38	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN
1C.06.100.0050.b	- con mattoni 25 x 5.5 x 5.5 cm (bastonetto), ...	m²	88,99	36,52	40,52	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN
1C.06.100.0100	Muratura faccia a vista con mattoni semipieni ...	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 4 Raw dataset.

The raw dataset is characterized by many rows with null records, therefore, after the acquisition of the Lombardy Regional Price List document, the first step performed for cleaning the dataset is the removal of empty rows. Subsequently, only the useful columns have been selected, containing the ID code information and textual description of items.

Moreover, the hierarchy knowledge of information, such as chapters and sub-chapters, is provided by the price list tool by code length. The shorter the code, the higher is the hierarchical level of information; conversely, the longer the code, the deeper is the hierarchical level of information transmitted. In Figure 4, the first record characterized by the code "1C.06.100" represents the chapter referring to the exposed brick wall works, while the records with longer codes are the specific work activity within the masonry chapter. For the achievement of the objectives set within this paper, the data type to work with belongs to the last hierarchical level, where data are conveyed with the highest level of detail through textual items description. Work and elementary resource descriptions which a unit price is associated with are characterized by a code length higher than 13 digits.

Since elementary resources could be described at both single code and parent-child code levels, the last pre-processing step involved pairing the child entries with their respective parent entries, ensuring the framework's accuracy in resource assessment, as shown in Figure 5, where the process of pairing items description is shown.

By following these steps, we ensured that only relevant and properly organized data were used in the testing process, further validating the framework's effectiveness in elementary resource allocation for construction works.

The following stage consists of the cosine similarity technique application. Scikit-learn library have been used for

```
('MC.06.050.0025', 'Exposed bricks 6 x 11 x 23 cm sandblasted'),
 ('MC.06.050.0030', 'Semisolid exposed bricks:'),
 ('MC.06.050.0030.a', '- brick 25 x 10 x 5.5 cm'),
 ('MC.06.050.0030.b', '- brick 25 x 10 x 10 cm'),
 ('MC.06.050.0030.c', '- brick 25 x 12 x 5.5 cm'),
 ('MC.06.050.0030.d', '- brick 25 x 12 x 7 cm'),
 ('MC.06.050.0030.e', '- brick 25 x 12 x 10 cm'),
 ('MC.06.050.0030.f', '- double UNI 25 x 12 x 12 cm')
```

```
['MC.06.050.0025', 'Exposed bricks 6 x 11 x 23 cm sandblasted'],
 ['MC.06.050.0030.a', 'Semisolid exposed bricks:- brick 25 x 10 x 5.5 cm'],
 ['MC.06.050.0030.b', ' Semisolid exposed bricks:- brick 25 x 10 x 10 cm'],
 ['MC.06.050.0030.c', ' Semisolid exposed bricks:- brick 25 x 12 x 5.5 cm'],
 ['MC.06.050.0030.d', 'Mattoni semipieni faccia a vista:- brick 25 x 12 x 7 cm'],
 ['MC.06.050.0030.e', 'Semisolid exposed bricks:- brick 25 x 12 x 10 cm'],
 ['MC.06.050.0030.f', 'Semisolid exposed bricks:- double UNI 25 x 12 x 12 cm']
```

Figure 5 From left to right, the process of pairing child entries with their respective parent entries.

this purpose importing in a Google Colab notebook the *TfidfVectorizer* class and using the function *cosine_similarity*.

Descriptions from the elementary resource list are converted from textual to numerical representation through TF-IDF feature extraction action, obtaining a sparse matrix. This process effectively maps the vocabulary of the dataset's domain knowledge. As a result, each phrase in the dataset is transformed into a vector within the TF-IDF feature space, representing its unique characteristics in relation to the entire corpus of documents.

Later, the same process is repeated for a single description query, which comes from the works list. This query is transformed into a TF-IDF feature vector using the *TfidfVectorizer* that was fitted on the product descriptions.

Finally, the *cosine_similarity* function calculates the similarity between the query and all records (product

descriptions) in the product list. This provides a similarity score for each elementary resource within the list with respect to the query. The products characterized by the most similar description to the query are retrieved by sorting the similarity scores in descending order, from the highest to the lowest value. For this study, it was decided to recall only the first 5 products, leaving out the later ones.

In the table below it is shown an example of the framework, by querying the list of products with the "1C.06.050.0100" work, whose description is "Semi solid masonry wall, 8 x 12 x 24 cm, with cement mortar, including the charge for the formation of shoulders, vaults, corners, pilasters, internal worktops". The first elementary resource returned by the proposed methodology is the correct characterizing product of the enquired work. The overall proposed framework exploits cosine similarity through TF-IDF techniques for retrieving products in the Price List document. Those are ranked based on their similarity scores, allowing the user to identify the most relevant matches for the query.

The framework is tested on four samples. Each sample represents a different domain (masonry clay blocks, masonry concrete blocks, tile floors, and thermal insulations) with unique terminologies and sentence structures. By testing the developed methodology on a diverse dataset, it is possible to assess its scalability on different knowledge subdomain.

The last step of the framework requires the evaluation of the output, performed by practitioners, who assesses the correctness of the first output, characterized by the highest score rate.

Table 1 Framework output, queried with "1C.06.050.0100" work.

Score rate	ID code	description
0.50	MC.06.050.0040.a	Semi-solid bricks:- semisolid brick 8 x 12 x 24 cm
0.43	MC.06.050.0040.b	Semi-solid bricks:- semisolid brick 8 x 24 x 24 cm
0.38	MC.06.050.0040.e	Semi-solid bricks:- double UNI semi-solid brick 24 x 12 x 12 cm
0.34	MC.06.050.0045.c	Semi-solid bricks complying with UNI EN 771-1 and the Minimum Environmental Criteria set forth in the Decree of 23 June 2022 of the Ministry of Ecological Transition, for the construction of partitions or counterwalls; type - dimensions (length x width x height) in cm - perforation (%<) - thermal conductivity (λ) according to UNI 1745 of dry brick - fire resistance with normal and fireproof plaster* - soundproofing power:- block with horizontal holes 30x4.5x15 cm - dB 39
0.33	MC.06.050.0015	Solid bricks 25 x 12 x 5.5 cm complying with UNI EN 771-1 and the Minimum Environmental Criteria set forth in the Decree of June 23, 2022 of the Ministry of Ecological Transition, for the construction of load-bearing masonry according to NTC 2018, thermal conductivity (λ) according to UNI 1745 of dry brick 0.431 W/mK

6. RESULTS AND DISCUSSION

In this section, a preliminary phase of analysis and discussion of the analyzed dataset is presented. This approach allows us to have a more comprehensive view of the results obtained before delving into further discussions.

6.1 Dataset preliminary analysis

A preliminary analysis was performed in order to provide a better description and visualization of the four selected dataset. Table *1* collects some significant data on which the evaluations are performed. It provides the size of the tested sample, both for works and elementary resources, for the subdomain knowledge analyzed (Clay brick wall, concrete brick wall, tiles, and thermal insulation). Furthermore, it provides the description with major, minor, and average number of words per work and elemental resource. A delta is also given to highlight the differences between works and elemental resources.

As it is possible to see from the table below, a homogeneous number of 40 work items have been analyzed for each test campaign. The related number of elementary resources varies according to the type of knowledge subdomain, ranging from a minimum of 33 to a maximum of 64 items.

Concerning the analysis of textual descriptions, the number of words in them was investigated. The clay Brick Wall campaign is characterized by 50.8 average words per work item. and 30.7 average words per elementary resource item, registering a delta of 20.1 words. Among the test campaigns, the Concrete Brick Wall campaign stood out with the longest descriptions, averaging 123.9 words per work item. In contrast, the Elementary Resources campaign displayed a mean word size of 59.9, exhibiting the largest word delta between the two.

The Thermal Insulation test campaign also featured lengthy descriptions. In this case, the average word count for both elementary resources and jobs was quite similar. On the other hand, the Tile Floors knowledge subdomain had the shortest textual descriptions for both elementary works and resources; moreover, the average length of the descriptions for works and elementary resources nearly matched, resulting in an almost zero delta.

Based on this overview, it is evident that there are differences among the various knowledge subdomains. Specifically, the Concrete Brick Wall subdomain requires a greater number of words to convey information. Additionally, the descriptions of works within this subdomain offer more extensive information compared to their corresponding elementary resources. Conversely, the Tile Floor subdomain necessitates fewer details in its descriptions, with both works and elementary resources providing relatively concise information. These differences highlight the varying informational needs and content richness across the different knowledge subdomains.

Table 2 Preliminary dataset analysis. W: work items; ER: elementary resource items.

	Clay brick wall			**Concrete brick wall**			**Tile floor**			**Thermal insulation**		
	W	ER	Δ	W	ER	Δ	W	ER	Δ	W	ER	Δ
.n° of items	*40*	*33*	*7*	*40*	*50*	*-10*	*40*	*59*	*-19*	*40*	*64*	*-24*
Max length	*105*	*89*	*16*	*140*	*121*	*19*	*82*	*94*	*-12*	*150*	*121*	*29*
Mean length	*50.8*	*30.7*	*20.1*	*123.9*	*59.9*	*64*	*46.3*	*46.5*	*-0.2*	*85.6*	*98.1*	*12.5*
Min length	*13*	*10*	*3*	*54*	*28*	*26*	*10*	*10*	*0*	*49*	*33*	*16*

6.2 Discussing framework result

Table 4 in this paragraph presents the results obtained, by displaying in the first row the number of times the model successfully assigned the elementary resource to the work query. Conversely, in the second row, the table shows the number of times when the model was unable to assess the correct output. The last row provides the number of missing elementary resource items. According to the domain of knowledge, different outcomes were achieved.

The poorest results were recorded for the Concrete Brick Wall and Clay Brick Wall domains. Also, as previously shown in table 3, these domains exhibited a substantial delta between the number of words between works and elementary resources, indicating that the works conveyed more information than the descriptions of elementary resources. For the Clay Brick Wall sample, 19 tests over 40 provided incorrect output, being 47.5% of the total tests; however, practitioners verified that in the 57.8% of incorrect outputs, the elementary resource is missing in the price list. Concerning the Concrete Brick Wall sample, 26 tests over 40 provided incorrect output. being 65% of the total tests.

Furthermore, it was observed that in certain instances, standardizing works and elementary resources resulted in a shift from negative to positive outcomes. Table 3 shows the output of the work query “Load-bearing masonry made of hollow core brick blocks, thermo-acoustic, with cement mortar, including formation of vaults, pilasters, corners; with:- simple blocks 13 x 30 x 19 cm, thickness 13”. The model does not return the correct elementary resource

as the first output, but rather as the third ranked output. Just by equalizing the lexicon from “block” to “blocks”, like the other items are, the similarity score of the correct output turns from 0.37 to 0.44, with a higher score.

Table 3 Framework output, queried with “1C.06.050.0300.b” work.

Score rate	ID code	description
0.39	MC.06.100.0010.b	Thermal insulation blocks, 45% drilling:- interlocking blocks, 30 x 25 x 19 cm
0.39	MC.06.100.0010.a	Thermal insulation blocks, 45% drilling:- interlocking blocks, 25 x 30 x 19 cm
0.37	MC.06.100.0010.c	Thermal insulation blocks, 45% drilling:- simple block, 13 x 30 x 19 cm

Conversely, Tile Floor and Thermal Insulation domains exhibited a larger number of positive outcomes, registering respectively 35 and 37 correct output tests out of 40, 87.5% and 92.5% respectively. In these cases, unlike previous tests, the descriptions of works and elementary resources had a smaller word delta. For Thermal Insulation, practitioners verified that in 67% of the negative outcomes, the correct data was indeed missing from the elementary resources list.

Table 4 Testing framework results.

	Clay brick wall	Concrete brick wall	Tile floor	Thermal insulation
correct	*21*	*17*	*35*	*37*
incorrect	*19*	*26*	*5*	*3*
Missing ER	*11*	*0*	*0*	*2*

7. CONCLUSION

The research presented in this study contributes to the empowerment of public administrations in effectively managing data from Price List Documents, aligning with the growing necessity to transition textual information into structured and machine-readable formats. Building upon the investigation conducted in the research titled (Gatto et al., 2023), this study introduces a methodology capable of extracting elementary resource information from a price list document and linking it to the corresponding construction work using cosine similarity NLP techniques.

The model successfully extracted and correctly matched the elementary resource to the corresponding work query in 75% of the cases where the elementary resource was present in the list. Additionally, the model proved to be a valuable tool in supporting practitioners in identifying missing resources. A limitation associated with this approach is its reliance on retrieving explicitly mentioned information from the text. Indeed, it cannot derive implicit information.

The study also highlighted that due to the lack of standardization in conveying information, the machine encountered ambiguity in interpreting the text. It emphasized the importance of adopting a more standardized approach to delivering information, making it more understandable not only to machines but also to humans.

It is important to note that the developed framework does not replace human activity but rather acts as a supporting tool. Human verification and validation of the model's outputs are essential.

Given the success of this framework, future developments aim to extend its application to broader contexts, with a focus on extracting cost information from various textual documents, including technical specifications and price list documents, and linking them to verify the consistency of cost information with the data contained in BIM models.

REFERENCES

Akanbi, T., & Zhang, J. (2021). Design information extraction from construction specifications to support cost estimation. *Automation in Construction*, *131*(April 2020). https://doi.org/10.1016/j.autcon.2021.103835

Cha, H. S., & Lee, D. G. (2018). Framework Based on Building Information Modelling for Information Management by Linking Construction Documents to Design Objects. *Https://Doi.Org/10.3130/Jaabe.17.329*, *17*(2), 329–336. https://doi.org/10.3130/JAABE.17.329

Cunningham, T. (2015). *Tender Documentation for Construction Projects - An Overview*.

Ding, Y., Ma, J., & Luo, X. (2022). Applications of natural language processing in construction. *Automation in Construction*, *136*, 104169. https://doi.org/10.1016/J.AUTCON.2022.104169

Gatto, C., Farina, A., Mirarchi, C., & Pavan, A. (2023). *Development of a framework for processing unstructured text dataset through NLP in cost estimation AEC sector*. *4*, 0–0. https://doi.org/10.35490/EC3.2023.232

Hassan, F. ul, & Le, T. (2020). Automated Requirements Identification from Construction Contract Documents Using Natural Language Processing. *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, *12*(2), 04520009. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000379

Jafari, P., Al Hattab, M., Mohamed, E., & Abourizk, S. (2021a). Automated extraction and time-cost prediction of contractual reporting requirements in construction using natural language processing and simulation. *Applied Sciences (Switzerland)*, *11*(13). https://doi.org/10.3390/app11136188

Jafari, P., Al Hattab, M., Mohamed, E., & Abourizk, S. (2021b). Automated Extraction and Time-Cost Prediction of Contractual Reporting Requirements in Construction Using Natural Language Processing and Simulation. *Applied Sciences 2021, Vol. 11, Page 6188*, *11*(13), 6188. https://doi.org/10.3390/APP11136188

Lee, J. H., & 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*(11), 1141. https://doi.org/10.3390/APP7111141

Leśniak, A., & Janowiec, F. (2020). Analysis of tender procedure phases parameters for railroad construction works. *Open Engineering*, *10*(1), 846–853. https://doi.org/10.1515/eng-2020-0095

Locatelli, M., Pattini, G., Seghezzi, E., Tagliabue, L. C., & Di Giuda, G. M. (2022). NLP-based system for automatic processing of quality demands in italian public procedure: a system engineering formalization. *Proceedings of the 2022 European Conference on Computing in Construction*, *3*. https://doi.org/10.35490/ec3.2022.176

Locatelli, M., Seghezzi, E., Pellegrini, L., Tagliabue, L. C., & Di Giuda, G. M. (2021). Exploring Natural Language Processing in Construction and Integration with Building Information Modeling: A Scientometric Analysis. *Buildings 2021, Vol. 11, Page 583*, *11*(12), 583. https://doi.org/10.3390/BUILDINGS11120583

M.E.Sepasgozar, S., Costin, A. M., Reyhaneh, K., Sara, S., & Abbasian, Ezatollah Li, J. (2021). BIM and Digital Tools for State-of-the-Art Construction Cost Management. *Buildings*. https://doi.org/10.1142/9789814447935_0007

Moon, S., Lee, G., & Chi, S. (2021). Semantic text-pairing for relevant provision identification in construction specification reviews. *Automation in Construction*, *128*, 103780. https://doi.org/10.1016/J.AUTCON.2021.103780

Munková, D., Munk, M., & Vozár, M. (2013). Data pre-processing evaluation for text mining: Transaction/sequence model. *Procedia Computer Science*, *18*, 1198–1207. https://doi.org/10.1016/j.procs.2013.05.286

Naji, K. K., Gunduz, M., & Falamarzi, M. H. (2022). Assessment of Construction Project Contractor Selection Success Factors considering Their Interconnections. *KSCE Journal of Civil Engineering*, *26*(9), 3677–3690. https://doi.org/10.1007/s12205-022-1377-6

Opitz, F., Windisch, R., & Scherer, R. J. (2014). Integration of document- and model-based building information

for project management support. *Procedia Engineering*, *85*, 403–411. https://doi.org/10.1016/j.proeng.2014.10.566

Sdino, L., & Rosasco, P. (2021). The Regional Price Lists for Estimating the Costs of Construction. *GREEN ENERGY AND TECHNOLOGY*, 213–229. https://doi.org/10.1007/978-3-030-49579-4

Sitikhu, P., Pahi, K., Thapa, P., & Shakya, S. (2019). A Comparison of Semantic Similarity Methods for Maximum Human Interpretability. *International Conference on Artificial Intelligence for Transforming Business and Society, AITB 2019*. https://doi.org/10.1109/AITB48515.2019.8947433

Tang, S., Liu, H., Almatared, M., Abudayyeh, O., Lei, Z., & Fong, A. (2022a). Towards Automated Construction Quantity Take-Off: An Integrated Approach to Information Extraction from Work Descriptions. *Buildings*, *12*(3). https://doi.org/10.3390/buildings12030354

Tang, S., Liu, H., Almatared, M., Abudayyeh, O., Lei, Z., & Fong, A. (2022b). Towards Automated Construction Quantity Take-Off: An Integrated Approach to Information Extraction from Work Descriptions. *Buildings 2022, Vol. 12, Page 354*, *12*(3), 354. https://doi.org/10.3390/BUILDINGS12030354

Xie, X., Chang, J., Kassem, M., & Parlikad, A. (2023). *Resolving inconsistency in building information using uncertain knowledge graphs: a case of building space management*. *4*, 0–0. https://doi.org/10.35490/EC3.2023.267

Zabin, A., González, V. A., Zou, Y., & Amor, R. (2022). Applications of machine learning to BIM: A systematic literature review. *Advanced Engineering Informatics*, *51*(April 2021). https://doi.org/10.1016/j.aei.2021.101474

Zhang, J., & El-Gohary, N. M. (2016). Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking. *Journal of Computing in Civil Engineering*, *30*(2). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000346

Zhang, J., Zi, L., Hou, Y., Deng, D., Jiang, W., & Wang, M. (2020). A C-BiLSTM Approach to Classify Construction Accident Reports. *Applied Sciences 2020, Vol. 10, Page 5754*, *10*(17), 5754. https://doi.org/10.3390/APP10175754

Zhang, Z., & Ma, L. (2023). *Using machine learning for automated detection of ambiguity in building requirements*. *4*, 0–0. https://doi.org/10.35490/EC3.2023.211