www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

A MULTI-LABEL TEXT CLASSIFIER: APPLICATION ON AN ITALIAN PUBLIC TENDER PROCEDURE, PROJECT ISCOL@

SUBMITTED: December 2023
REVISED: June 2024
PUBLISHED: October 2024
EDITOR: Robert Amor
DOI: 10.36680/j.itcon.2024.038

Mirko Locatelli, PhD,
Department of Management "Valter Cantino", Università degli Studi di Torino, 10134 Turin, Italy
ORCID: https://orcid.org/0000-0003-0100-3169
mirko.locatelli@unito.it

Lavinia Chiara Tagliabue, Associate professor,
Department of Computer Science, Università degli Studi di Torino, 10149 Turin, Italy
ORCID: https://orcid.org/0000-0002-3059-4204
laviniachiara.tagliabue@unito.it

Giuseppe M. Di Giuda, Full Professor,
Department of Management "Valter Cantino", Università degli Studi di Torino, 10134 Turin, Italy
ORCID: https://orcid.org/0000-0002-2294-0402
giuseppemartino.digiuda@unito.it

SUMMARY*: The main means of communication during the pre-design phase is natural language. Effective communication during the pre-design phase through text documents and reports is essential to the success of a design and construction project. The study develops and evaluates a Natural Language Processing (NLP) tool called ArchiBERTo to process textual data related to design tender documents in the Italian public procurement process. Documenti di Indirizzo alla Progettazione (DIPs) are key documents, as they outline the demands, needs, and objectives of the public appointing party. ArchiBERTo is developed to process and convert DIP quality objective sentences into a hierarchy of objectives and criteria. The performances are evaluated by comparing the tool's rankings with those provided by a group of domain experts. The results demonstrate ArchiBERTo's capability to reflect the collective ability of a panel of experts and to properly reflect the different contents of the DIP in the objectives hierarchy. The proposed system aims to address the issue of information asymmetry and potential misunderstandings, or varying interpretations, among the key actors of the Italian tendering procedure, namely the public appointing party, the design teams, and the external committee, regarding the relative importance of quality objectives and evaluation criteria. The utilization of the NLP systems ArchiBERTo to establish a shared hierarchy of objectives is expected to enhance communication and promote consensus during the pre-design phase. The minimization of the consensus issue among the actors can have a positive impact on the overall quality of the design proposals and facilitate the evaluation process conducted by the external committee, bridging the gap between expected and actual quality, ensuring that design proposals align with the quality objectives and demands of the public actor.*

KEYWORDS*: Natural Language Processing (NLP), Bidirectional Encoder Representations from Transformers (BERT), Large Language Model, Deep Learning, Consensus Issue, Collective Intelligence, NLP-enhanced Communication Flow.*

REFERENCE*: Mirko Locatelli, Lavinia Chiara Tagliabue, Giuseppe M. Di Giuda (2024). A multi-label text classifier: application on an Italian public tender procedure, project ISCOL@. Journal of Information Technology in Construction (ITcon), Vol. 29, pg. 864-893, DOI: 10.36680/j.itcon.2024.038*

1 INTRODUCTION

1.1 Natural language in Pre-design phase

Digital Transformation significantly impacts both the economy and society, influencing people's lifestyles. The European Union (EU) has identified the Architecture, Engineering, Construction, and Operation (AECO) industry as a strategic sector in its digitalization strategy, with the aim of enhancing productivity and the quality of building production (European Commission, 2019). To maximize digitalization's benefits in the AECO sector, the EU Commission suggests integrating Building Information Modeling (BIM) with other digitalization techniques and technologies. The public actor plays a pivotal role in the AECO digital transition being the design project and building owner. A public actor is any public contracting authority, such as a local government authority or public institution. Public actors have the possibility to directly influence and guide the outcomes of design and construction projects by defining quality objectives, needs, and requirements. However, the AECO sector faces challenges related to objective and requirement definition and management, such as a lack of identification, management, and traceability. AECO sector is document-centric, with various complex pieces of information exchanged and modified using documents, often relying on human natural language, which can be inconsistent and prone to misinterpretations mainly during the Pre-design phase. Pre-design is the initial phase of an architectural design and construction process, and it has a significant impact on the project's value (Senescu et al., 2014). The main purpose of this phase is to define project goals and objectives which must satisfy stakeholders' needs and demands reaching a consensus between the stakeholders' needs and design teams' proposals. Thus, effective communication and understanding of the expressed requirements and demands are crucial for the project's success (Taleb et al., 2017). Effective communication means that the intended messages and information have been properly processed and understood by all the actors involved (Norouzi et al., 2015). Typically, during the Pre-design phase, communication is mainly based on natural language: verbal expressions, written or spoken, are collected in text documents (Di Giuda et al., 2020). Natural language is pervasive and one of the richest forms of knowledge representation and communication. As stated, natural language is also prone to ambiguity due to its complex form and can lead to misinterpretations, or at least have different interpretations (Sun and Li, 2022).

1.2 Actors, procedure, and main documents: the consensus issue

To clarify the research objective, key actors and documentation involved in a public procurement procedure for design and construction services in Italy are explained. There are three primary actors: the Appointing Party, which defines project goals, objectives, and specifications, which are shared in a document called *Documento di Indirizzo alla Progettazione* (DIP) at the start of the call for tenders. The DIP in the Italian tender procedure acts as a guiding document of the design phase; the multiple Design teams, which compete aiming to secure the tender by submitting design proposals that align with the public actor's demands; an External Committee, which evaluates the design proposals to determine the best project proposal that meets the specified requirements. In the context of the Italian public design call for tenders procedure, issues arise in the communication and interpretation of information among the involved parties. Typically, design teams analyze the contents of the DIP to understand the quality requirements and needs, establish design objectives, and develop proposals that align with the defined design objectives. However, these design objectives are often shaped by the subjective perspectives and experiences of the designers, leading to variations in objectives and priorities among different design teams. When an external committee evaluates design proposals, they collectively or individually define the key evaluation criteria to evaluate the design proposals, and their relative importance. Each committee member may use different evaluation criteria, considering some design aspects more important than others (e.g., aesthetics, context integration, energy performances), according to their subjective experience and expertise. In addition, the criteria selected by the external committees may differ from the design objectives defined by the design teams and from the genuine intentions of the public client. This misalignment can lead to design proposals that fail to meet the public client's needs and demands, because the design teams may have defined their proposal according to design objectives that may be different from the real intentions that the public client tried to communicate via the DIP, as well as different from the evaluation criteria considered by the evaluation committee. The primary cause of these misunderstandings and discrepancies is the use of natural language expressions in the DIP, which can be subject to various interpretations based on personal biases and experiences. Even when the design teams and the external committee share the same objectives, their hierarchy of priorities may differ from those of the public client. This discrepancy can be identified as a consensus issue, where the involved parties lack a shared set of objectives and their relative

importance. The situation is further complicated by the mandatory steps in the Italian public tender procedure, which restrict additional communication among the parties, thus increasing the risk of misinterpreting the genuine objectives and priorities.

1.3 Research objective: NLP for the mitigation of the consensus issue

To address the identified consensus issue, Natural Language Processing (NLP) technology can be employed in the design call for tenders process. NLP technology is employed to process the quality needs and demands section of a DIP, which are expressed via natural language, automatically identify and convert expressions of quality objectives and needs into a machine-readable format, and generate a hierarchy of objectives. The hierarchy of objectives reflects the contents of the quality needs and demands section of the DIP and the public client's real intentions. The objectives and related hierarchy are shared with the design teams and the external committee along with the DIP. Therefore, the objectives and their hierarchy are not subjectively defined by design teams and committee members. On the contrary, by providing a common basis for defining and evaluating design proposals, the shared hierarchy of objectives minimizes the consensus issues among the involved parties. Consequently, it reduces the gap between the expected and actual quality of the design proposals, aligning them more closely with the needs and demands of the public actor.

The research project aims to provide public clients, design teams and external committee with a NLP-based tool to process and translate the quality objectives, which are stated via natural language in a DIP, into a hierarchical list of objectives shared with the involved parties, improving communication among them (Figure 1) and addressing the previously mentioned issues of the Italian procurement process. The proposed system aims to prevent the generation of the consensus issues, reducing the information asymmetry, possible misunderstandings and different interpretations of the relative importance of the quality objectives and evaluation criteria by the public appointing party, design teams, and external committee. Defining the hierarchy of objectives by using NLP systems, and sharing it among the involved parties, is expected to enhance the communication and foster the consensus among the actors during the Pre-design phase. This generates a positive impact on the overall quality of the competing design offers, which are more aligned with the needs and demands of the public actor, and also facilitates the evaluation activity of the external committee, that are given a hierarchy of objectives as a reference to evaluate the design proposals. All the above will help reducing the gap between the expected and actual quality of design proposals and of the output of the overall procedure, ensuring that the design proposals meet the public actor quality objectives and demands.

Additionally, the digitization of the preliminary quality objectives and needs by means of the proposed NLP system is expected to help overcoming the document-based nature of the construction industry and improving the digital management of objectives and needs. The possibility of digitally managing unstructured natural language data and information represents a key step in the digitalization of the design and construction industry.

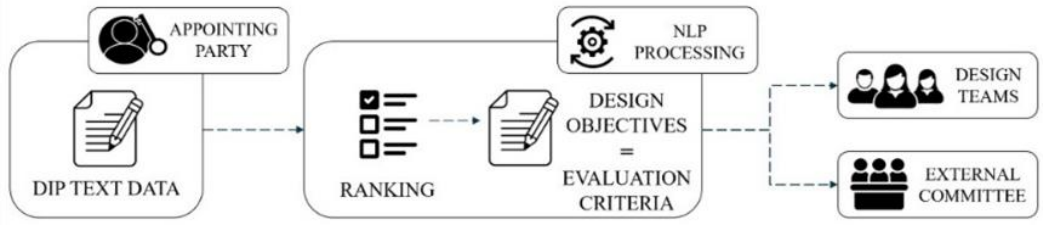

Figure 1: NLP enhanced information flow and communication process.

2 LITERATURE REVIEW

2.1 Latest NLP approaches: Pre-Trained Bidirectional Encoder Representation from Transformers

NLP techniques are generally used to process and analyze textual information and text corpora, and the latest NLP approaches are based on Deep Learning (DL) algorithms. Among the latest NLP approaches there are Pre-Trained Language Models, which are black box models trained with a large amount of unannotated data, allowing the language models to process and understand a general language (e.g., English). The language models can be fine-

tuned via supervised approach by feeding the model with smaller task-specific datasets to train it on specific NLP tasks. The availability of language models for specific languages (e.g., Italian) can be limiting, however, pre-trained models on several different languages including the Italian language have been developed (e.g., the 'dbmdz/bert-base-italian-uncased' deployed by Hugging Face). The Bidirectional Encoder Representation from Transformers (BERT) model is a widely used Pre-Trained Large Language Model (LLM) for large corpora of texts (Devlin et al., 2019). Pre-trained BERT models can be seen as the digital replica of the hidden knowledge contained in large-scale datasets, and they can be fine-tuned to solve NLP downstream tasks in specific knowledge domains (Fang et al., 2020). BERT-based models perform better than traditional word embedding models, which are context-free models using static word embedding where each word has a single vector, regardless of context. On the contrary, BERT-based models can consider the rich context-related information hidden in the text creating contextualized word representations, where word vectors are sensitive to the context in which they appear (Ethayarajh, 2019). In other words, Pre-Trained context-aware language models like BERT models can capture the nuances of word usage in different contexts, improving the accuracy of text processing. Examples of the latest variations of Transformers models are Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019) and Generative Pre-trained Transformer (e.g., GPT-3 (Brown et al., 2020)).

2.2 NLP applications in AECO sector

Studies on the application of NLP techniques and tools in AECO sector are collected and categorized according to the application field, namely: Construction Safety Management and Accident Prevention, Compliance and Code Checking, Procurement-Risk and Dispute Management, Construction and Project Management, and Facility-Life Cycle Management.

2.2.1 Construction Safety Management and Accident Prevention

Early applications of NLP in the Construction Safety Management and Accident Prevention field have aimed at automatically organizing and categorizing documents like accident and injury reports (Kim and Chi, 2019). More recent applications regard the use of DL algorithms to solve the classification problem (Qiao et al., 2022). The latest applications aim to build a knowledge base useful for predicting possible risks at the worksite analyzing the structured and unstructured sources of information, like text documents or even workers' conversations (Xu et al., 2021a). The overall goal of these applications is to extract and make the most of the value from the knowledge contained in textual documents (i.e., injury reports) to predict and reduce or avoid accidents or dangerous situations on construction sites (Baker et al., 2020). In fact, by using NLP techniques to process sources of unstructured information and knowledge related to the Safety Management field (Li et al., 2021), it is possible to create knowledge bases that are useful for predicting potential future risks for workers and occupants (Koc et al., 2022). Studies investigating the processing of languages different from the English language (i.e., Korean and Chinese) are also present in the literature (Kim et al., 2022b). A Question and Answering system based on the fine-tuning of the BERT language model in the construction incident reports domain is also developed (Mohamed Hassan et al., 2022).

2.2.2 Compliance and Code Checking

NLP-enabled Automated Compliance Checking (ACC) is one of the most active and investigated topics in the literature (Salama and El-Gohary, 2011). The traditional compliance checking procedure involves several time-consuming steps: I) regulatory documents are analyzed by experts and rules are extracted from texts; II) the rules identified are translated into machine-readable logic; III) information to be checked is extracted from an informative source (e.g., a document or a BIM model); IV) compliance checking activity is performed filling the to be checked information into the rules previously extracted and translated in machine-language (Wu et al., 2022a). The application of NLP aims to optimize and reduce time by automatizing one or more of the previously described tasks. In particular, studies are listed regarding the automatization of the following ACC tasks:

1. Rules extraction and analysis of regulatory documents (Zheng et al., 2022);
2. Rules translation into machine-readable format (Xue et al., 2022);
3. Extraction of information to be checked from documents or BIM models, and IV) Compliance Checking activity (Peng and El-Gohary, 2018);
4. The entire ACC workflow and BIM model code validation (Wu et al., 2022c).

It is possible to state that ACC is one of the fields where BIM, ontologies, and NLP can be applied in a virtuous circle that leverages the possible optimizations of the methodologies and technologies (Zhang and El-Gohary, 2022b).

2.2.3 Procurement-Risk and Dispute Management

Similar to the applications in the field of Safety Management, clauses or contract documents are processed to create a useful knowledge base for predicting contractual and dispute risk in the field of Procurement-Risk and Dispute Management. The main studies in the field of Procurement-Risk and Dispute Management focus on:

- Knowledge discovery, acquisition, and modeling from claim documents (Lee et al., 2021), and automated content analysis and Information Extraction from legal documents (claim, contract, and dispute documents) (Chalkidis et al., 2017). The studies rely on the use of metadata (Zhu et al., 2010), ontologies (Guévremont and Hammad, 2021), and Knowledge Graphs (Zhang et al., 2021);
- Similar case retrieval and pattern identification for dispute resolution and defect litigation cases (Jallan et al., 2019);
- Project-oriented contract risk mining (Yang et al., 2022), bidding and contract uncertainty evaluation (Park et al., 2021), contract and clauses vagueness prediction and extraction (Candaş and Tokdemir, 2022), requirements detection (Hassan et al., 2020), and automated benchmarking profitability (Bilal and Oyedele, 2020);
- Automated contract standard selection (Elkhayat and Marzouk, 2022) and contract drafting (Hassan and Le, 2021).

The procurement-legal Artificial Intelligence field has recently attracted attention. However, applications have not been successfully developed and deployed due to the huge effort and time required to produce large-annotated corpora (Zhang et al., 2022). Despite that, the recent possibilities given by Pre-trained language models fostered a series of research studies on legal AI, also in the construction sector (e.g., (Moon et al. 2022a) apply the pre-training and fine-tuning approach to reaching high accuracy and efficiency levels in processing legal documents (Moon et al., 2022a)). The author also highlights the presence of a study that applies in a combined way a Chatbot developed to classify and extract relevant information, with permissioned Blockchain technologies for providing features such as document search, history tracking, automatic extraction of related documents, and authenticity verification for document management to support claims and dispute tasks (Kim et al., 2022a). A study investigating the processing of the Chinese language is also present in the literature (Zhang et al., 2021).

2.2.4 Construction and Project Management

During the design, construction, and maintenance of a building, an overwhelming quantity of data and information is produced. Moreover, 80% of data in the AECO sector is unstructured (Gharehchopogh and Khalifelu, 2011), and most of them are textual data. However, most Project Management information system focuses and relies on producing and managing structured data. For that reason, recent studies focus on addressing the described gap of previous systems, developing NLP systems to support the management of unstructured data during the Construction and Project Management activity (Botha, 2018). The application of NLP in such a context can play a key role in the digitalization of the Project Management activity bringing unstructured data into the equation. In particular, studies found in the literature focus on:

- Construction activities extraction and scheduling evaluation (Hong et al., 2022);
- Project risk and time-cost evaluation and prediction (Tang et al., 2022);
- Construction progress monitoring (Ren and Zhang, 2021);
- Value analysis and value for money assessment (Ren et al., 2021; Zhang and El-Gohary, 2022a);
- Design and construction specifications review (Moon et al., 2022b);
- Constrain modeling and management (Wu et al., 2022b);
- Quality control and management (Zhong et al., 2022).

In Adel et al., 2022, the authors highlight an innovative approach in the Construction Project Management field which introduces a novel Information Exchange and Management system for construction firms based on Blockchain technology and Chatbots. The proposed system leverages the characteristics of Blockchain technology in terms of peer-to-peer operation mode, data integrity, structuring, and privacy, and the chatbot merits regarding

ease of use and degree of automation. The study proposes the use of a private Blockchain network configured for data distribution and storage. A smart contract is coded for regulating data writing and reading operations, and a Chatbot is developed for data collection and retrieval through textual conversations. Furthermore, a serverless cloud function and database are configured to enable the linkage between the Blockchain network and chatbot.

2.2.5 Facility-Life Cycle Management

Similarly, to the applications in the field of Project Management, unstructured data associated with Operation and Maintenance activities are processed to create a useful knowledge base bringing the unstructured data into the digitalization process of the Life-Cycle Management of a building (Ng et al., 2006). Asset management data like document databases (Williams and Halling, 2014), maintenance records data (Stenström et al., 2015), maintenance staff indication activities (Mo et al., 2020), building defect reports (Jeon et al., 2022), maintenance requests (D'Orazio et al., 2022), and occupants' feedback (Alhaj et al., 2021) are processed via Text Mining and NLP techniques to extract and retrieve valuable information for the Facility Management activity. Potential uses of NLP to match real-world facilities with BIM elements (Xie et al., 2019), and smart building control operations using NLP for voice recognition (i.e., Chatbot assistants) are also present in the literature (Alexakis et al., 2019).

2.3 Literature review findings

Recent applications at several stages of the design and construction process, shown above, highlight a positive trend for the combined use of BIM and NLP methodology in semantic and knowledge modeling to manage the construction process. Research in the field of Information Technology (IT) in the AECO sector seems to have taken the strategy of modeling and managing semantic information and knowledge. Moreover, the use of State-Of-The-Art pre-trained language models based on the Transformers mechanism is increasingly common due to the possibility of fine-tuning the model using a limited data sample compared to systems based on traditional DL algorithms (Mohamed Hassan et al., 2022; Moon et al., 2022a). NLP techniques are predominantly used to develop applications that process the English language, but applications in different languages are also present (e.g., Chinese (Xu et al., 2021b)). Several studies propose the development of real chatbots for querying and extracting information from BIM models. Frontier research topics see the combined application of Information Modeling methodology and NLP techniques for developing Chatbots integrated with Blockchain systems (Adel et al., 2022). However, no cases of systematic application to design and construction projects consider the early Planning and Preliminary Design stages. Applications and studies of NLP technologies in the AECO sector typically follow the Pre-design and Planning phase.

3 METHODOLOGY

3.1 NLP tool-enhanced information flow

The research methodology proposes the development of an NLP tool, named ArchiBERTo, to analyze and translate the needs and quality demands section of a DIP to enhance communication among actors. All the documents used in the development, assessment, and evaluation of the tool are written in Italian language belonging to the case study of Project Iscol@ (Seghezzi et al., 2020). The tool outputs are meant to be shared with the design teams participating in the call for tenders as a support for a better identification and understanding of the design goals. Sharing the hierarchical scale of objectives (i.e., the NLP outputs) with the bidding party (design teams) and the evaluating party (external committee), it is possible to reduce the possible differences in the interpretation of the relative importance of quality objectives and demands. The proposed process can also make the appointing party more aware and conscious of the relative importance of the design quality objectives and demands, which are expressed via natural language in the DIP. A comparison between the traditional communication method based on the sharing of a DIP and the NLP tool-enhanced information flow is illustrated in Figure 2. In the traditional communication method, the DIP, which is written via natural language, is shared with the design teams and the evaluation committee. The design teams analyze the contents of the DIP to understand the quality requirements and needs, establish design objectives, and develop proposals that align with the public actor's expectations. Likewise, the external committee members collectively or individually define the key evaluation criteria and their relative importance by analyzing the DIP, and use the evaluation criteria to evaluate the design proposals. However, both the design objectives and the evaluation criteria are often shaped by the subjective perspectives, experiences, and biases of the subjects, leading to variations in objectives and related priorities among different actors and with

respect to the real intentions, needs, and demands of the public client. This misalignment can result in design proposals that fail to meet the public client's needs and demands. The primary cause of these misunderstandings and discrepancies is the use of natural language expressions in the DIP, which can be subject to various interpretations based on personal biases and experiences. This phenomenon that often occurs in the traditional communication method can be identified as a consensus issue. On the other hand, in the proposed NLP tool-enhanced information flow, an additional step is introduced after the completion of the DIP. The DIP is processed by the NLP tool to extract a list of prioritized objectives according to the contents of the DIP, i.e, reflecting the real intentions of the public actor. The prioritized list of objectives is then shared with design teams and evaluation committee. As a consequence, the prioritized list of objectives provides both the design teams with the design objectives and the external committee with the evaluation criteria, minimizing the possible subjective misunderstanding and different interpretations. Therefore, the proposed information flow has the potential to minimize the consensus issue in the Italian tender procedure.

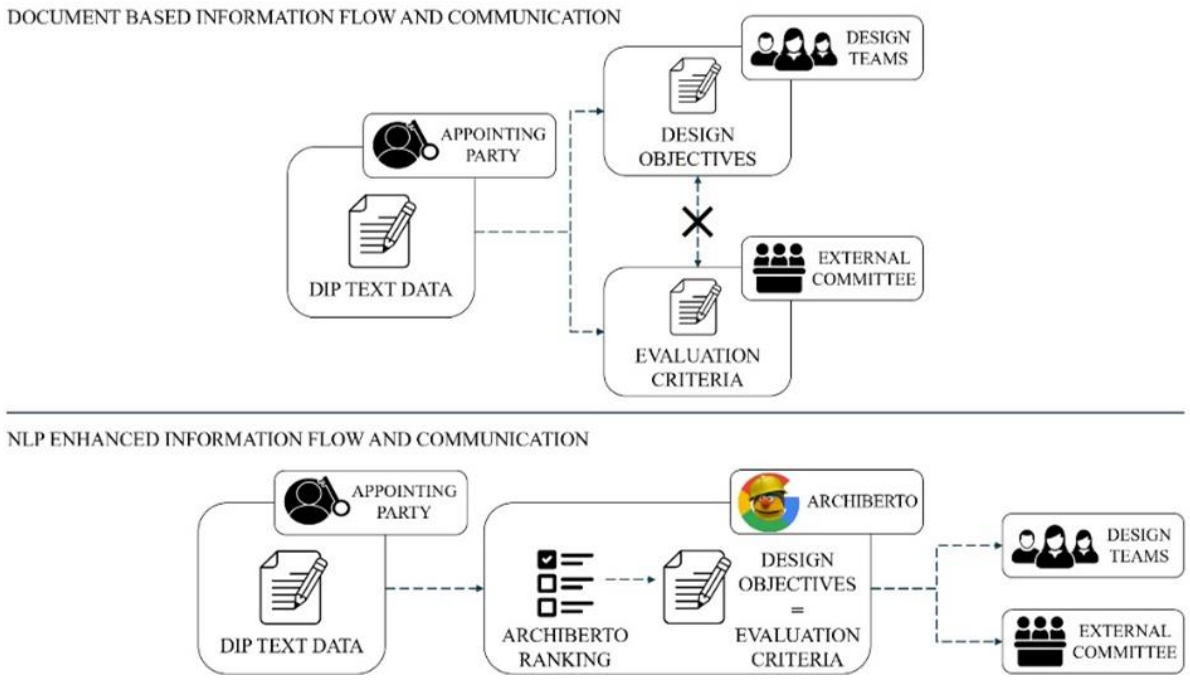

Figure 2: Traditional and NLP tool-enhanced information flow comparison.

3.2 ArchiBERTo: a BERT-based Multi-label Text classifier

Pre-trained context-aware LLMs, like BERT, can capture the nuances of word usage and how a term is employed in different contexts, improving the accuracy of text processing, by considering the rich context-related information hidden in the text and creating contextualized word representations, where word vectors are sensitive to the context in which they appear (Ethayarajh, 2019). In the AECO sector, (Zheng et al. 2022) and (Erfani and Cui 2021) demonstrated that BERT-based models' performances in sentence-level tasks, like the Text Classification, are significantly better if compared to traditional word embedding methods.

Consequently, to develop the NLP tool to process and translate the quality objectives expressed in a DIP into a ranking of criteria and objectives, which serves as the computational equivalent of the natural language information, a Pre-trained context aware BERT-based LLM is fine-tuned to solve the Text Classification task. Among the Text Classification problems, Multi-label Text Classification (MTC) regards the problem of automatically applying one or more predefined classification labels to a single piece of text. MTC is different from regular classification tasks as it involves assigning multiple, non-exclusive labels to a piece of text, rather than just one exclusive label (Venkatesan and Er, 2014). Consequently, the main feature of ArchiBERTo is the capability to solve an MTC problem. In the case of ArchiBERTo, the tool is applied to automatically assign these predefined labels or objectives to each sentence in a DIP, determining a weight for each label based on the sentence's correlation with the label and objective. Once the entire DIP quality section is processed the tool generates a

priority ranking of the DIP quality objectives based on the total weights of each objective. Consequently, a ranking for evaluating the proposals is established, which can be shared with the design teams and the evaluation committee to reach a consensus between the main parties involved in an Italian public tendering process. Being the case study the Project Iscol@, an Italian public tender procedure for the design and construction of school buildings, all the DIPs and the labels representing the appointing party quality objectives and demands are in the Italian language.

3.3 ArchiBERTo development and assessment steps

The main activities to develop and assess the tool are the following (Figure 3):

- Labels (quality objectives and needs) definition;
- Production of the Training and Validation datasets;
- Model fine-tuning and setting;
- Model performance evaluation;
- Testing the fine-tuned and evaluated model

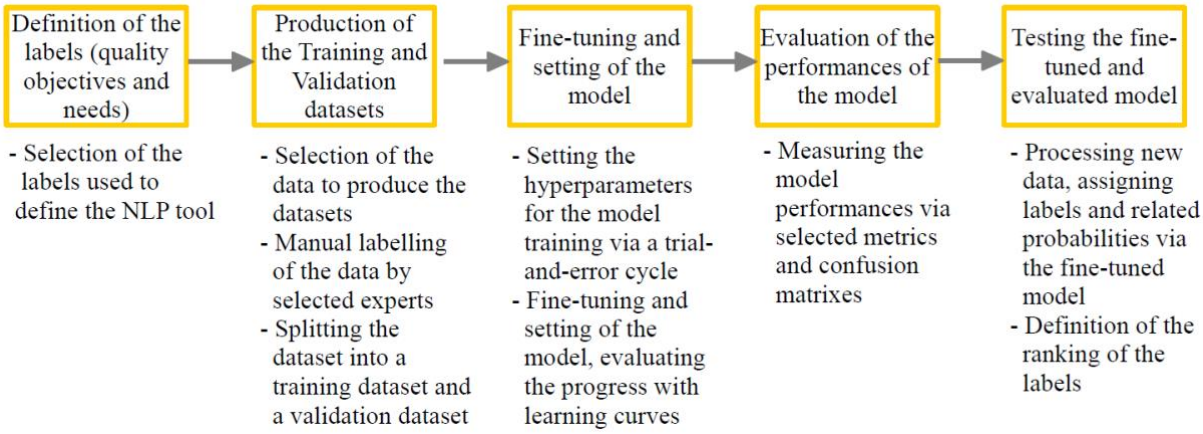

Figure 3: Steps to develop and assess the NLP-tool.

3.3.1 Labels (quality objectives and needs) definition

The NLP tool needs to be trained to sort sentences based on a set of predefined labels that reflect the quality objectives and demands of the appointing party. To establish consensus and agreement on the labels that represent the interests, objectives, and demands of the appointing party and end-users, the set of labels must be established in collaboration with the appointing party, end-users (when possible), and domain experts such as architects, building engineers, and designers. In the proposed case study (Project Iscol@), a list of predefined labels, as defined by the appointing party, was already available. These labels are the outcome of cooperation among various field experts and end-users. For clarity, it must be stated that there is a single predefined list of labels for all the Iscol@ DIPs, and the experts don't agree on labels on an individual document basis. The development and integration of ArchiBERTo in the Project Iscol@ call for tenders aims for each DIP to provide a ranking of the same predefined labels according to the specific DIP content on a project-by-project basis.

3.3.2 Production of the Training and Validation datasets

The Pre-trained BERT Large Language Model, at the very base of the proposed tool, needs to be fine-tuned to solve the MTC problem in the architecture and design knowledge domain of the case study of Project Iscol@. Consequently, to fine-tune the BERT language model for the specific case study of Project Iscol@, a sufficient amount of training and validation data is required. The data is collected into a general dataset, which is then randomly divided into two distinct datasets, i.e., a training dataset and a validation dataset, with a proportion of 0.8:0.2 respectively. In other words, the division of the data into the two datasets is performed randomly, but a total of 80% of the general dataset constitutes the training dataset, while a total of 20% of the general dataset constitutes the validation dataset. The model is fine-tuned using a dataset referred to as the training dataset. This dataset is used to train the model and improve its performance. Another dataset, called the validation dataset, is

utilized to assess the model's capabilities without any bias. This dataset is not used in the training process and serves to evaluate the model's performance. The dataset used for the model training and validation is created by selecting sentences belonging to several DIPs and manually assigning labels to them. The task of manual labeling is crucial for the accuracy and effectiveness of the NLP tool and thus is done through collaboration between experts with knowledge in architecture, design, and construction. Since the NLP system is tested on a specific case study (Project Iscol@), the experts involved have been instructed on the Iscol@ objectives, guidelines, and context in order to correctly label the training sentences. To avoid biases in the dataset production, each expert independently proposes labels for each sentence, and any disagreements are discussed and resolved through a consensus process. The goal is to have a model able to represent and utilize the collective knowledge of the experts. In particular, three annotators (i.e., the knowledge domain experts) are involved in the sentence labeling. As stated, they independently proposed the labeling of each sentence and then shared their labeling hypothesis. In case of disagreement on the labels to be attached to the sentences, the annotators follow the instructions explained in Table 1 and listed as follows:

- Agreement on the label by all three annotators: in the example of Label_01, all the annotators propose to tag the sentence with the same label, and Label_01 is automatically attached without a discussion between the three experts;
- Partial agreement on the label by two out of three annotators: in the example of Label_02, two out of three annotators propose to tag the sentence with the same label, and Label_02 is automatically attached without a discussion between the three experts;
- Partial agreement on the label by one out of three annotators: in the example of Label_03, one out of three annotators proposes to tag the sentence with the label. After a discussion about the possibility of attaching the label between the two remaining annotators, if one or both of them agree with the proposal of Annotator 03, Label_03 is attached;
- Partial agreement on the label by one out of three annotators: in the example of Label_04, one out of three annotators proposes to tag the sentence with a certain label. After a discussion about the possibility of attaching the label between the two remaining annotators, if none of them agrees with the proposal of Annotator 02, Label_04 is not attached;
- Agreement on not attaching the label by all three annotators: in the example of Label_05, all the annotators propose not to tag the sentence with the label, and Label_05 is automatically not attached without a discussion between the three experts.

Table 1: Possible labeling scenarios and final labeling according to the annotators' inter-agreement.

	Labeling proposals				
Annotator	Label_01	Label_02	Label_03	Label_04	Label_05
Expert 01	x	x	-	-	-
Expert 02	x	x	-	x	-
Expert 03	x	-	x	-	-
	Final labeling				
	x	x	x	-	-

Ultimately, the NLP tool, ArchiBERTo, aims to reduce subjectivity in the analysis of textual information by utilizing the collective intelligence of the group of experts. This method is expected to surpass the capability of single experts in managing the complexity of analyzing multiple sentences, as it incorporates the combined collective knowledge of the group.

3.3.3 Model fine-tuning and setting

After the dataset is produced, the model is fine-tuned using parameters, known as hyperparameters, which are typically used to train and fine-tune a BERT model (Devlin et al., 2019). The list of hyperparameters used for the NLP tool training is here provided (Devlin et al., 2019):

- MaximumLength. The maximum number of tokens (words) that will be considered during the training process;

- TrainingBatchSize. The number of training examples utilized in a single iteration. For instance, if the batch size is set to 16, then it means 16 samples from the training dataset will be used to calculate the error gradient before the model's weights are updated;
- ValidationBatchSize. The number of examples utilized for validation in a single iteration. For instance, if the batch size is set to 8, then it means 8 samples from the validation dataset will be used to validate the model;
- EpochsNumber. The number of full passes through the entire training dataset. During each epoch, the internal parameters of the model are updated. One epoch ends when the learning algorithm completes one pass through the subsets of the training dataset. The size of these subsets is determined by the TrainingBatchSize;
- LearningRate. It refers to the rate at which the weights of the neural network are adjusted based on the loss gradient descent. It determines the velocity at which the model moves toward the optimal weights.

The values of the hyperparameters cannot be determined or estimated from the data, and must be identified and set via a trial-and-error cycle: the model is run and tested several times while different values of the hyperparameters are set within predefined ranges, defined by the developers of the BERT model (Devlin et al., 2019). Via the trial-and-error cycle, it is possible to identify the configuration of hyperparameter values that allows the model to perform best. That configuration is selected for the model fine-tuning.

3.3.4 Model performance evaluation

To estimate the precision of the NLP tool, a comparison is made between the output of the model and the human-annotated validation dataset. The F1-score is chosen as the metric to measure the tool's performance (Sokolova and Lapalme, 2009). To calculate the F1-score for each label False Positives (FP), False Negatives (FN), True Positives (TP), and True Negatives (TN) are calculated for each prediction provided by ArchiBERTo processing the sentence of the validation dataset. All the possible combinations of TP or FP and FN or TN are shown in Table 2.

Table 2: Combinations of True and False positives and True and False negatives (Confusion matrix).

Actual label	Predicted label	
		Positive (Pp)
	Positive (Pa)	True Positive (TP)
	Negative (Na)	False Positive (FP)

Being the F1-score (Blair, 1979) classification accuracy metric that combines Precision (P) and Recall (R), consequently both P and R for each label are calculated according to Formula 1 and Formula 2.

$$Precision\ (P) = TP/(FP+TP) \quad (1)$$

$$Recall\ (R) = TP/(FN+TP) \quad (2)$$

Once calculated P and R the F1-score for each label is computed according to Formula 3, for a classifier to have a high F1-score, it needs to have a high P and a high R, in fact, the F1-score is defined as the harmonic mean of P and R.

$$F1\text{-}score\ (F1) = 2PR/(P+R) \quad (3)$$

Micro, macro, weighted, and samples F1-score are calculated for each label using the *sklearn.metrics Python module*, specifically:

- Micro average F1-score computes the F1 by considering the total amount of TP, FN, and FP (no matter the prediction for each label in the dataset);
- Macro average F1-score computes the F1 for each label and returns the average without considering the proportion for each label in the dataset;
- Weighted average F1-score computes the F1 for each label and returns the average considering the proportion for each label in the dataset;
- Samples average F1-score computes the F1 for each instance (i.e., each sentence) and returns the average.

Samples average F1-score is selected to identify the best model (i.e., to define the setting of the hyperparameters that produces the model with the highest performances).

3.3.5 Confusion matrix

Furthermore, TP, FP, FN, and TN can be visualized in a confusion matrix. A confusion matrix is a performance method for measurement in Machine Learning (ML) classification problems like the MTC problem. A confusion matrix can show what the model classifies correctly and what it does not and what types of errors it is making; in other terms, a confusion matrix provides a comparison between actual and predicted values. For an MTC problem, a confusion matrix is produced for each label. Each confusion matrix is a 2x2 matrix (like the one in Table 2) showing in the matrix cells the number of TP, FP, TN, and FN calculated for each label.

3.3.6 Learning curves

During the training of the LLM BERT-based model the learning curve shows the progress over time (number of epochs) of a specific learning metric. Learning curves are the mathematical representation of the algorithm learning process that can provide insight into learning behavior by plotting generalization performance against the number of training examples. In the ML field, the attention is focused on the generalization performance, in other words, the algorithm performance on new and unseen data (Viering and Loog, 2021). In literature, learning curves show the progress of the model's training by plotting the training error against the number of iterations in the optimization process. This allows for monitoring of the model's performance during the learning phase and can aid in identifying any issues and improving predictions (Osborne, 1975). One of the most widely used metrics is the training loss and validation loss comparison over iterations or epochs. In particular:

- Training loss curve indicates how well the model fits the training data. In this specific case, the Binary Cross Entropy Loss function is used to calculate the loss curve;
- Validation loss curve indicates how well the model fits new and unseen data.

By analyzing the training and validation curves, it is possible to monitor the model's progress and identify any underfitting or overfitting behaviors, in fact: if a model is underfitted, the training and validation loss will not decrease with increasing iterations or epochs. This is because the model has a high bias and fails to consider data and information. Conversely, if a model is overfitted, the training loss will decrease with low error values per iteration or epoch, while the validation loss will initially decrease but then increase after reaching a minimum point. This signifies the onset of overfitting and the model's poor performance on new data. The model may memorize the training data, but its generalization performance will be poor. Training and validation loss curves are therefore fundamental in finding the right set of training hyperparameters, the amount of data and iterations needed, and monitoring underfitting or overfitting model behavior. It is crucial to halt the training process at the global minimum, where the validation error transitions from decreasing to increasing, to properly train a model. If the training stops prior to the global minimum, the model is underfitted, whereas if it stops after the global minimum, the model becomes overfitted, as demonstrated in Figure 4.

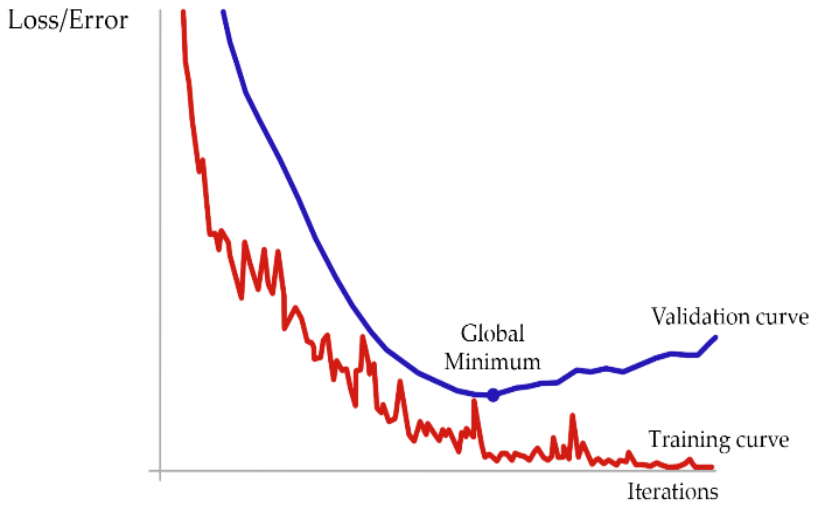

Figure 4: Possible trend of training and validation curves.

3.3.7 ArchiBERTo outputs and objectives ranking

Once ArchiBERTo has been fine-tuned and its performance evaluated, it can be tested by processing new sentences and assigning labels along with their predicted probabilities. These probabilities determine the priority of each label and quality objective for a given sentence. The probabilities of each label for all sentences in a document are then summed, normalized (relative to the sum of all label probabilities), and used to calculate the label weight for the entire document, as described in Formula number 4.

$$Label\ Weight\ i = (\sum Li/(\sum L1+\sum L2+\sum L3+\cdots+\sum Ln)*100 \qquad (4)$$

where Li denotes the probability value of the i-th label, with i ranging from 1 to n, and n being the total number of labels. Label 1, 2, 3, …, n = (A.1, B.1, C.1, …, P.1)

The sum of the probability values of each label considering the entire document $(\sum Li)$ divided by the sum of all probability values of all labels $(\sum L1+\sum L2+\sum L3+\ldots+\sum Ln)$, is considered as the weight of each label. The label weight reflects the relative significance of a specific quality objective. Once the label weights for all labels are calculated, a ranking of the labels, or quality objectives, is created for the processed document. The capability of the NLP tool to automatically assign labels and probabilities is the foundation for generating the prioritized ranking of quality objectives.

3.4 ArchiBERTo evaluation

3.4.1 Sentence splitting procedure

The sentences in the DIPs used to evaluate ArchiBERTo to produce the objectives rankings are obtained by manually splitting the different text sections of the document. No automatic sentence splitting tool is used; splitting is done manually to better control the procedure. In fact, text data can be inconsistent, sentences may be broken in the middle of the line or may have punctuation marks in the wrong positions leading automatic sentence splitting tools to fail. In addition, the splitting activity is not excessively time-consuming given the size of the documents required for the training of the model.

3.4.2 Evaluation metrics

To assess the NLP tool's capability to reflect the collective expertise and sensitivity of a group of experts in the objective hierarchization task and to determine the level of subjectivity involved, the outputs from the tool are compared to those provided by a group of experts in architecture and construction. Three experts were engaged in the process, with comprehensive knowledge of Project Iscol@ and expertise in the fields of architecture, construction, and tender procedures. Neither of them received any specialized training to participate in the research project. The comparison between the outputs of the tool and the assessments of the experts aims to measure the tool's capability to hierarchize and prioritize objectives and criteria. The tool's capabilities are measured by means of two metrics that are empoloyed to evaluate the following aspects:

- The level of the subjectivity of ArchiBERTo, gauging its ability to reflect the collective ability of the expert group in the quality objective hierarchization task. The more ArchiBERTo aligns with the experts' collective ability, the less subjective its ranking will be;
- The customization capability of ArchiBERTo, adapting its objective ranking to reflect different DIP contents. Hence, ArchiBERTo's rankings for different DIPs should be as diverse as the contents of the analyzed DIPs.

The two metrics are described in the following sections 875 and 876.

3.4.3 ArchiBERTo subjectivity degree

The performance of the NLP tool ArchiBERTo is evaluated by comparing its outputs to the rankings provided by a group of experts in the architecture and construction field. The experts individually analyze different DIPs and hierarchize the quality objectives, then the same DIPs are collectively analyzed by the group, providing a benchmark ranking. The rankings generated by ArchiBERTo are compared to both the individual expert rankings and the collective expert ranking. A score from 1 to 21 is assigned, a score equal to 1 represents the most important objective and a score equal to 21 represents the least important. The proposed evaluation process is illustrated in Figure 5.

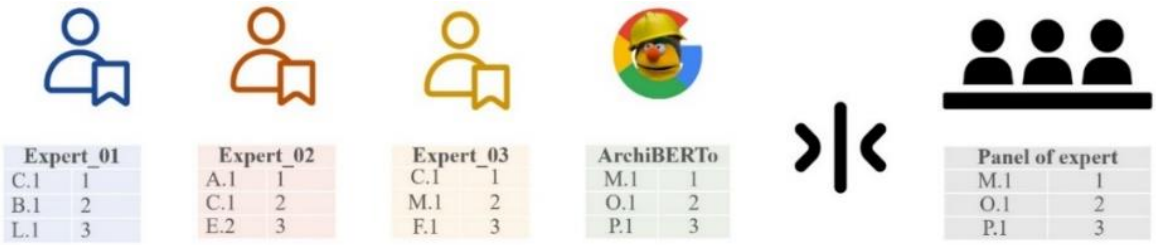

Figure 5: Subjectivity measurement schema.

To investigate the similarity of the rankings provided by the single experts and of ArchiBERTo in relation to the collective group ranking, the Kendall Tau (τ) coefficient is calculated for each ranking. Kendall τ is a nonparametric measure of the degree of correlation, introduced by Maurice Kendall in 1938 (Kendall, 1938). It allows solving the problem of comparing two different rankings of the same set of individuals defining if the two orders are sufficiently alike to indicate similarity of judgment in the individuals who provide the two rankings. The aim is to measure by means of the Kendall τ metric which ranking (among the individual rankings and ArchiBERTo ranking) is more similar to the ranking produced collectively by the group of experts. Formula 5 is used to calculate the Kendall τ.

$$Kendall\ \tau=((C-D))/((C+D)) \qquad (5)$$

where C is the number of Concordant pairs and D is the number of Discordant pairs

In particular, the objectives are listed from the most to the least important (from 1 to 21) for the group of experts, while concerning the individual rankings and ArchiBERTo only the rank related to each objective is provided. The Concordant (C) pairs identify for each rank of the individual expert and of ArchiBERTo the number of higher ranks that are positioned below the selected rank. On the other hand, the Discordant (D) pairs identify for each rank of the individual expert and of ArchiBERTo the number of lower ranks that are positioned below the selected rank. The Kendall τ coefficient returns a value between 0 and 1, where:

- 0 means no relationship between the compared rankings;
- 1 means a perfect relationship.

When comparing each ranking produced by the individual experts and by ArchiBERTo with the collective one by calculating the related Kendall τ coefficients, two possible outcomes can occur:

- ArchiBERTo Kendall τ < individual experts Kendal τ: the ranking provided by ArchiBERTo has a lower relationship with the collective one than the single experts' rankings. Thus, the tool has a higher level of subjectivity compared to individual experts, meaning it does not effectively reflect the collective capability of translating quality objectives stated in natural language into a ranked list of objectives;
- ArchiBERTo Kendall τ > individual experts Kendal τ: the ranking provided by ArchiBERTo has a higher relationship with the collective one than the single experts' rankings. Thus, the tool has a lower degree of subjectivity, accurately reflecting the collective capability of a group of experts in translating natural language expressions into objective rankings.

The second option is the most preferable one since it demonstrates the lower subjectivity of ArchiBERTo ranking compared with the individual expert rankings. The ranking of ArchiBERTo is in fact more similar to the collective one, which being a group effort is less subjective and less biased among all the rankings.

3.4.4 ArchiBERTo customization capability

Once calculated the ArchiBERTo subjectivity, to measure its customization capability to adapt the objectives ranking mirroring the DIP contents, different DIPs are processed, and different objectives rankings are produced using the NLP tool. In addition, a unique ranking is produced by the group of experts from the analysis of a whole different DIP document. The DIPs processed by ArchiBERTo, and the DIP analyzed by the group of experts are related to different school types and involve different contents, needs, and objectives. As explained for the subjectivity degree measurement a score from 1 to 21 is assigned, a score equal to 1 represents the most important objective and a score equal to 21 represents the least important, as shown in Figure 6. The final aim is to measure the capability of ArchiBERTo to generate a ranking that is customized to the contents of the different documents.

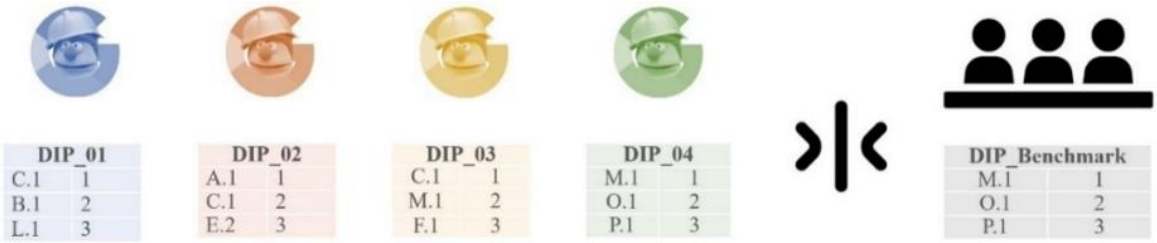

Figure 6: Customization capability measurement schema.

The processed DIPs and the benchmark DIP are selected to have substantial differences in the hierarchy of objectives:

- DIP_03 concerns the design and construction of a Secondary school;
- DIP_04 concerns the design and construction of a Primary and kindergarten school;
- DIP_05 concerns the design and construction of a Secondary school;
- DIP_06 concerns the design and construction of a Primary and kindergarten school;
- the DIP used as a benchmark concerns the design and construction of a Primary and kindergarten school.

The DIPs analyzed being related to different types of schools will present a different degree of similarity with the benchmark DIP. In particular, DIP 03 and DIP 05 will present a lower similarity (i.e., the lowest Kendal τ values, closer to zero). DIP 04 and DIP 06 will likely be the most similar to the benchmark DIP (i.e., the highest Kendal τ, closer but not equal to 1), being Primary and kindergarten schools like the benchmark DIP but with different objectives and needs and socio-economic context. To confirm the above assumptions, the Kendall τ values for the four DIPs are calculated and compared with the benchmark DIP.

4 CASE STUDY

4.1 Application on DIP of school buildings, Project Iscol@

The case study aims to test the developed NLP tool, ArchiBERTo, and the proposed methodology on documents related to the design and construction of school buildings. School projects are considered an appropriate building typology for the assessment as they have a high heterogeneity of quality objectives, needs, and demands, and a significant impact on the social and urban context. The NLP tool will be evaluated in the Italian AECO context of Project Iscol@. Project Iscol@ aims to address the problem of the backwardness of the pedagogical and educational regional system in the Sardinia Region by renovating and expanding the regional school building stock. The main goal of Iscol@ is to create a school system focused on architectural quality and social and environmental sustainability of the interventions.

4.2 Project Iscol@ tender procedure

The tender procedures of Project Iscol@ are single-stage design tenders and the project delivery method adopted is the Design-Build (DB). The DB involves a single operator for design and construction, and it is increasingly being adopted to replace the Design-Bid-Build where design and construction are totally separated phases and are managed by two different operators. The DB allows the public actor (i.e., the appointing party) to dialogue with a single operator, increasing the efficiency of Information Management and Exchange of the construction process in a virtuous cycle with the application of Information Modeling methods (e.g., BIM). In view of the positive characteristics of the Iscol@ calls for tenders, regarding the application of Information Management in the design and construction process, the pilot study would concern the processing of DIPs, i.e., the basis of the DB tenders. Moreover, the heterogeneity of objectives, needs, and demands and the social impact that an educational building can have on the urban context make Iscol@ a valid pilot study application.

4.3 Project Iscol@ DIPs guidelines and evaluation grid, positive and negative impacts

In the early stage of Project Iscol@, the Sardinia Region provided general guidelines for the drafting of DIPs by local municipalities. These guidelines ensured consistency in following regional strategies and homogenizing the objectives for school building interventions across the region. The Iscol@ team also established a standardized evaluation grid for design proposals and shared it with municipalities. After the first round of design calls in 2021,

it was analyzed how the use of this standardized grid with fixed priorities and objectives affected the projects. Specifically, the implementation of guidelines resulted in all DIPs adhering to regional directives and standardizing the quality objectives of school building renovations on the Island. Summarizing, the first round of tenders showed that using a fixed list and ranking of objectives was helpful in aligning projects with Iscol@ strategic goals, but also too rigid in accommodating differing building projects based on their unique geographical-environmental and socio-cultural context. In fact, each building design and construction project has unique qualities due to its correlation and influence by the surrounding context and specific socioeconomic and territorial requirements. In fact, buildings can be considered "prototypes of themselves" and are strictly correlated and influenced by the context. The use of fixed priorities and weights for objectives, however, may lead to uniform outcomes and limit the ability to tailor to the individuality of each project. The use of ArchiBERTo, which adapts the ranking of objectives for each call based on the specific content of each DIP, aims to restore flexibility and accommodate unique needs and requirements on a project-by-project basis. The Iscol@ team has validated and published 26 DIPs, with 21 serving to build and test ArchiBERTo, and 5 used to gauge the tool's level of subjectivity and customization.

5 RESULTS AND DISCUSSION

5.1 Labels (quality objectives and needs) definition

As introduced in the methodology section, a consensus about the interests and quality objectives of the appointing party and end-user is established by defining a set of labels in collaboration with the appointing party and experts in the field, such as architects, building engineers, pedagogues, and agronomists. The full list of labels for the case study is presented in Table 3.

Table 3: Labels code and description.

Label	Description
A.1)	Capability of the school building to be used as a Civic Center.
B.1)	Visibility and integration of sustainable design choices (educational medium) and integration of the intervention into nature and application of landscape enhancement strategies.
C.1)	Possibility of personalization of spaces and equipment to prevent vandalism creating a feeling of belonging in users.
D.1)	Spatial and volumetric integration of the intervention in the context and with existing buildings (shape, materials, colors, connections, etc.) and proper mediation with the demand for visibility and architectural quality of the intervention as a building containing public functions.
E.1)	Articulation of spaces and accesses with a focus on simple and clear identification of the various functions, including using colors and signages.
E.2)	Presence of green spaces as an integral part of the design.
F.1)	Perceptual quality (natural and artificial light) and psychophysical comfort (visual, thermo-hygrometric, acoustic, etc.) to promote comfort and learning.
F.2)	Indoor air quality and healthiness.
G.1)	Cleanability, durability, maintainability, and replaceability of landscaping, materials, and greenery to reduce operating and maintenance costs.
I.1)	Integration of the intervention with the road system and distinction between driveways, bicycle, and pedestrian paths; provision of areas and equipment to encourage slow and non-motorized mobility.
I.2)	Ensuring accessibility and usability for people with disabilities.
L.1)	Fostering interactions between students and teachers, group work and peer learning (collaborative learning and peer tutoring) by supporting innovative and inclusive teaching. Architecture should support the idea of space as a "third teacher".
L.2)	Visual and spatial continuity between outdoor (green and non-green) and indoor environments to encourage outdoor educational activities and enhance contact with the natural environment (outdoor space can be used as a second classroom). Connection between classroom and circulation spaces. The architecture should support the concept of openness of the traditional classroom and the concept of the learning landscape.
M.1)	Use of renewable, natural (non-harmful), local materials or materials with recycled content.
M.2)	Minimization of the impact of the building on the surrounding environment (noise, light, water pollution, heat island effect, minimization of land consumption and use of soil defense strategies, etc.).

Label	Description
M.3)	Integration between design and renewable energy production systems and exploitation and management of solar, light, and natural cooling and heating inputs.
M.4)	Requests regarding energy standards and minimization of consumption (energy, water, etc.) including using monitoring systems.
N.1)	Ensuring safety during school activities and separation between activities conducted by people not belonging to the school staff, maintenance activities (spaces and paths). Adequate delimitation of the school perimeter and need for control and supervision.
O.1)	Spatial flexibility (furniture, facilities, etc.).
O.2)	Temporal flexibility, possibility of use during curricular and extracurricular hours by citizens and long-term temporal flexibility, adaptability of spaces (readiness for change, adaptability).
P.1)	Usability of technological devices and integration with learning theories. Integration of space and technology; widespread presence of ICT technologies.

5.2 Production of the Training and Validation datasets

The selected BERT-based language model to be fine-tuned is a BERT model pre-trained on the Italian language and available in the Hugging Face repository. All the details related to the language model fine-tuned to develop ArchiBERTo are available at the following link: https://huggingface.co/dbmdz/bert-base-italian-uncased. The model is fine-tuned by defining a dataset through a manual process of selecting, gathering, and labeling sentences from the qualitative sections of the DIPs, following the procedure outlined in the methodology section. The labeling and dataset definition process was completed as a group effort, reflecting the collective expertise and sensitivity of the panel of experts. The jointly produced dataset is expected to be less biased and subjective in the automatic labeling of new sentences and objective ranking definition compared to individual expert assessments. The training and evaluation datasets represent the group of experts' collective and shared ability to identify and prioritize quality objectives and needs. The general dataset is composed of 1268 sentences labeled with 21 tags (i.e., the quality objectives and needs). The labeled sentences composing the general dataset are then randomly split into the training dataset and the validation dataset. Out of the total quantity of sentences of the general dataset, the 80% constitutes the training dataset that is used to train the model (for a total of 1014 sentences), while the 20% constitutes the validation dataset that is used to provide an unbiased evaluation of the model (for a total of 254 sentences). The validation dataset is never used in the training of the model, and viceversa the training dataset is never used in the evaluation of the model. Both the training and validation datasets have approximately the same percentage of samples per label as the general dataset. As stated in the methodology section, the general dataset of labeled sentences is produced by the collaboration among several experts in the architecture and construction field with deep knowledge about the strategic objectives related to the specific case study (Project Iscol@). Consequently, the model is trained using a dataset produced by the collaboration of different experts. Therefore, the NLP tool is expected to be able to represent the collective ability of the group and outperform the ability of the single expert to judge and classify the sentences related to the quality objectives. This can help avoid subjectivity in the hierarchization activity and better manage the complexity of analyzing and interpreting a huge number of sentences. The NLP tool can in fact be considered the numerical counterpart of the group of experts' knowledge.

5.3 Model fine-tuning and setting

As described in the methodology, section 872, the second step of the NLP tool development is the hyperparameters definition and setting. Hyperparameters, which are a variable configuration external to the model and whose values cannot be estimated from the data, are defined via a trial-and-error cycle: the model is run and tested several times while different values of the hyperparameters are set within predefined ranges. The configuration of hyperparameter values that allows the model to perform best is selected. Via the trial-and-error cycle, the following values of the hyperparameters are selected to be used for the NLP fine-tuning:

- MaximumLength = 128;
- TrainingBatchSize = 2;
- ValidationBatchSize = 32;
- EpochsNumber = 20;
- LearningRate = 2 E-05.

The listed values allow for obtaining the best fine-tuned model considering the training dataset.

5.4 Model performance evaluation

5.4.1 Precision, Recall, and F1-score

The performance of ArchiBERTo is evaluated using Precision (P), Recall (R), and F1-score (F1) metrics for each label (as shown in Table 4), to gauge the model's effectiveness, as outlined in the methodology section.

Table 4: Model Precision, Recall, and F1-score per label.

Label	Precision (P)	Recall (R)	F1-score (F1)
A.1)	0.86	0.86	0.86
B.1)	0.86	0.75	0.80
C.1)	0.75	0.50	0.60
D.1)	0.62	0.56	0.59
E.1)	0.67	0.57	0.62
E.2)	0.67	0.67	0.67
F.1)	0.63	0.73	0.68
F.2)	1.00	0.33	0.50
G.1)	1.00	1.00	1.00
I.1)	1.00	0.75	0.86
I.2)	0.86	0.75	0.80
L.1)	0.72	0.81	0.76
L.2)	1.00	0.52	0.69
M.1)	1.00	1.00	1.00
M.2)	0.90	0.75	0.82
M.3)	0.89	0.73	0.80
M.4)	0.78	0.93	0.85
N.1)	0.67	0.33	0.44
O.1)	0.86	0.84	0.85
O.2)	0.52	0.73	0.61
P.1)	0.88	0.94	0.91
Metric	Precision (P)	Recall (R)	F1-score (F1)
micro avg	0.79	0.75	0.77
macro avg	0.82	0.72	0.75
weighted avg	0.81	0.75	0.77
samples avg	0.77	0.76	0.75

With only three values of the F1-score lower than 0.6 (i.e., labels D.1, F.2, and N.1) and the samples average F1-score value higher than 0.75, the NLP model can be considered properly fine-tuned.

5.4.2 Confusion matrix

Confusion matrixes are produced for each label, and the values of TP, FP, FN, and TN for each label are summarized in Table 5. The number of correct predictions (TP and TN) and incorrect predictions (FP and FN) broken down by labels (Table 5) provides insight into the type and number of errors made by ArchiBERTo during the training phase.

Some labels obtained better results than others in the training of the model, considering all three values of Precision, Recall and F1-score (Table 4). When that occurred, the same labels also showed good results in the confusion matrix (Table 5). For example, label M.1 is among the labels that obtained the best results. A reason for this can be that the label or objective is extremely specific and, therefore, with a sufficient number of sentences in the dataset associated with the label, it is easier to fine-tune the model on that label or objective. In fact, it is easier

to train a model on a specific and in detail objective rather than training it on a wide and general concept. As for the label M.1, it refers to the "Use of renewable, natural (non-harmful), local materials or materials with recycled content", which is a very detailed concept and objective. The variability of the sentences related to this concept is limited, and the concept can be easily identified in a new sentence. On the contrary, some labels, like N.1, obtained less good results in the training, and this can be caused by an insufficient number of sentences associated with the label in the training dataset or because the objective is wide and more difficult to define. In the case of N.1, the objective is described as "Ensuring safety during school activities and separation between activities conducted by people not belonging to the school staff, maintenance activities (spaces and paths). Adequate delimitation of the school perimeter and need for control and supervision". This concept can involve a greater number of variations, examples, and different sentences and aspects in comparison to label M.1. In addition, N.1 is the label with the lower number of sentences in the training dataset, i.e., only 15 sentences, while 36 sentences refer to label M.1, and the highest quantity, i.e., 161 sentences, refers to label O.1, which also obtained good results in the training of the model.

Table 5: True Positive, False Positive, True Negative, and False Negative values per label.

Labels	TP	FP	FN	TN
A.1)	24	4	4	222
B.1)	6	1	2	245
C.1)	3	1	3	247
D.1)	5	3	4	242
E.1)	8	4	6	236
E.2)	4	2	2	246
F.1)	19	11	7	217
F.2)	2	0	4	248
G.1)	7	0	0	247
I.1)	3	0	1	250
I.2)	6	1	2	245
L.1)	29	11	7	207
L.2)	12	0	11	231
M.1)	10	0	0	244
M.2)	18	2	6	228
M.3)	8	1	3	242
M.4)	14	4	1	235
N.1)	2	1	4	247
O.1)	31	5	6	212
O.2)	11	10	4	229
P.1)	15	2	1	236

5.4.3 Learning curves: training and validation loss

The model's overfitting or underfitting behavior is determined by plotting the training and validation loss learning curves (represented in Figure 7 and Figure 8). The training loss shows the model's performance in fitting the training data, while the validation loss illustrates its ability to fit new data.

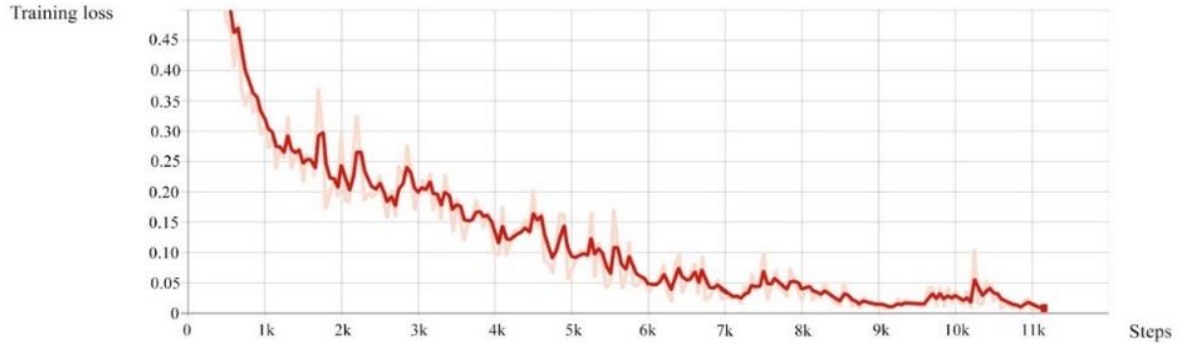

Figure 7: Training loss chart.

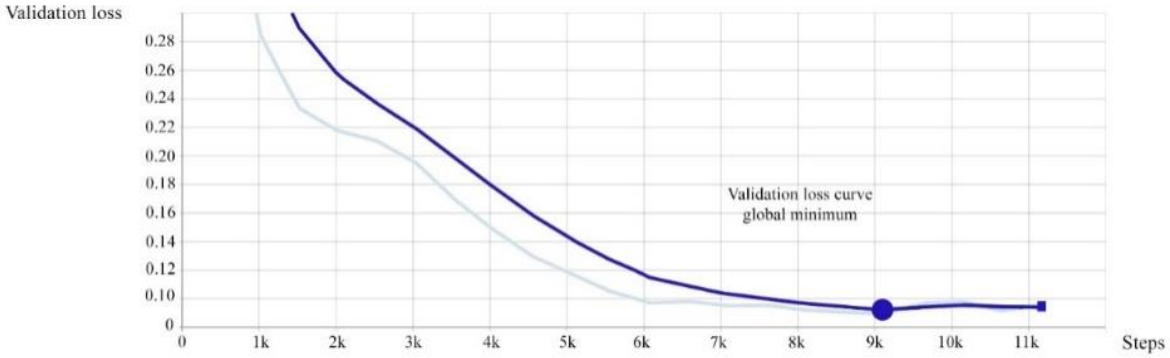

Figure 8: Validation loss chart.

As depicted in the graphs, the training and validation loss curves steadily decrease and level off, with both curves reaching similar values as reported below:

- Training_loss = 0.01842
- Validation_loss = 0.08986

The trends of the two curves show the absence of overfitting and underfitting phenomena showing an optimal fit of the two learning curves. Consequently, the results confirm that the model is properly fine-tuned.

5.5 ArchiBERTo document level outputs: DIP rankings

The capability of ArchiBERTo to process an entire DIP is evaluated as described. In particular, the subjectivity degree and the customization capability of ArchiBERTo are evaluated in sections 882 and 884. All the sentences of the DIPs used to evaluate the subjectivity degree and the customization capability of ArchiBERTo are split as previously described.

5.6 ArchiBERTo evaluation: subjectivity degree

The subjectivity degree of the NLP system is measured as follows: the rankings generated by the NLP model and the rankings provided by the single experts are compared with the rankings provided collectively by the group of three experts considered as the benchmark. Based on the rating, a score was assigned between 1 (representing the first and most important objective in the ranking) and 21 (representing the least important goal in the ranking). The two DIPs used to evaluate the subjectivity degree and the DIP used as a benchmark are selected among the five ones available and were not used to produce the training and validation datasets. For both DIPs, the Kendall τ coefficient calculation for each expert and the NLP tool is shown in Figure 9 and Figure 10. The Kendall τ coefficient calculation is performed as described in the methodology section.

5.6.1 DIP number 1: Kendall τ comparison

The similarity between the rankings provided by the single experts and ArchiBERTo with the collective one is estimated by calculating the Kendall τ coefficient for each ranking. The ranking with the higher coefficient, the nearest to the value 1, is the most similar to the benchmark, i.e., the collective ranking. Consequently, the ranking with the higher Kendall τ is the one less affected by subjectivity and better mirrors the collective capability of the group of experts to translate the natural language expressions. Kendall τ calculation and values for the first analyzed DIP are provided in Table 6.

The ArchiBERTo ranking reaches the highest Kendall τ values among the rankings. Consequently, it is the most similar to the benchmark ranking, the one produced collectively by the group of experts, considering the DIP_01. Consequently, ArchiBERTo ranking can be considered the least affected by subjectivity being the most similar to the collective judgment and demonstrating its capability to mirror the collective intelligence of the group of experts in the objectives ranking task, as shown in Figure 9.

Table 6: Concordant and Discordant pairs of ArchiBERTo and single experts' rankings compared with the collective rank.

Labels	Group	BERT	C	D	Exp03	C	D	Exp01	C	D	Exp02	C	D
A.1)	1	2	19	1	2	19	1	1	20	0	1	20	0
L.1)	2	1	19	0	1	19	0	6	15	4	2	19	0
F.1)	3	4	17	1	5	16	2	3	17	1	4	17	1
O.1)	4	3	17	0	10	11	6	2	17	0	8	13	4
M.4)	5	7	14	2	7	13	3	7	14	2	9	12	4
P.1)	6	6	14	1	12	9	6	5	14	1	10	11	4
D.1)	7	13	8	6	19	2	12	4	14	0	5	13	1
L.2)	8	9	11	2	3	13	0	20	1	12	7	11	2
O.2)	9	5	12	0	11	8	4	14	6	6	6	11	1
E.1)	10	14	7	4	14	6	5	15	5	6	3	11	0
I.1)	11	10	9	1	6	9	1	17	3	7	11	10	0
M.3)	12	11	8	1	9	7	2	10	7	2	16	5	4
M.1)	13	12	7	1	8	7	1	9	7	1	20	1	7
N.1)	14	16	5	2	13	6	1	12	5	2	12	7	0
C.1)	15	15	5	1	20	1	5	8	6	0	13	6	0
G.1)	16	8	5	0	4	5	0	19	1	4	19	1	4
M.2)	17	18	3	1	15	4	0	11	4	0	21	0	4
E.2)	18	21	0	3	21	0	3	13	3	0	15	2	1
F.2)	19	19	1	1	17	1	1	16	2	0	18	0	2
I.2)	20	20	0	1	16	1	0	18	1	0	17	0	1
B.1)	21	17	0	0	18	0	0	21	0	0	14	0	0
Sum			181	29		157	53		162	48		170	40
Kendall τ			0.72			0.49			0.54			0.62	

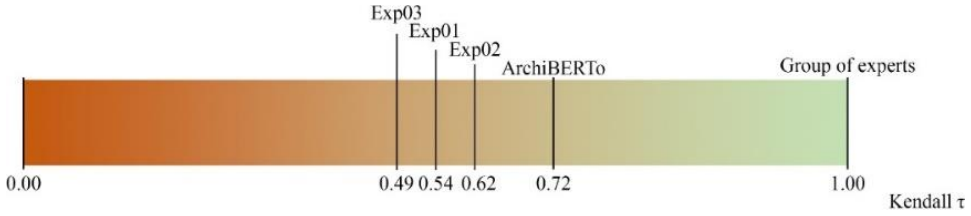

Figure 9: Single experts and NLP tool Kendall τ values, DIP_01.

5.6.2 DIP number 2: Kendall τ comparison

Kendall τ calculations for the second analyzed DIP are provided in Table 7. Kendall τ values are visualized in Figure 10. The ArchiBERTo ranking reaches the highest Kendall τ values among the rankings. Consequently, it is the most similar to the benchmark ranking, the one produced collectively by the group of experts, considering the DIP_02.

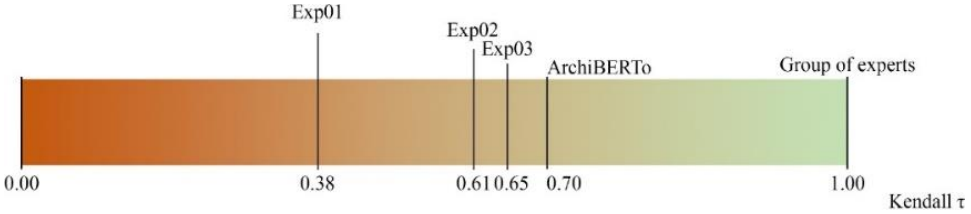

Figure 10: Single experts and NLP tool Kendall τ values, DIP_02.

Table 7: Concordant and Discordant pairs of ArchiBERTo and single experts' rankings compared with the collective rank, DIP 02.

Labels	Group	BERT	C	D	Exp03	C	D	Exp01	C	D	Exp02	C	D
F.1)	1	3	18	2	5	16	4	1	20	0	4	17	3
O.1)	2	2	18	1	6	15	4	2	19	0	3	17	2
L.2)	3	5	16	2	1	18	0	4	17	1	7	14	4
L.1)	4	1	17	0	7	14	3	8	13	4	1	17	0
A.1)	5	6	15	1	2	16	0	14	7	9	2	16	0
P.1)	6	4	15	0	3	15	0	11	9	6	12	9	6
M.4)	7	8	13	1	11	10	4	10	9	5	9	11	3
E.1)	8	15	6	7	9	11	2	20	1	12	5	13	0
M.3)	9	10	10	2	15	6	6	9	8	4	10	10	2
B.1)	10	12	8	3	13	7	4	6	9	2	15	6	5
D.1)	11	11	8	2	4	10	0	17	3	7	14	6	4
G.1)	12	9	8	1	10	8	1	12	6	3	13	6	3
O.2)	13	7	8	0	8	8	0	19	1	7	8	7	1
I.2)	14	17	4	3	18	3	4	7	5	2	11	6	1
M.2)	15	20	1	5	12	6	0	5	5	1	19	2	4
C.1)	16	13	5	0	16	4	1	15	3	2	6	5	0
E.2)	17	16	3	1	20	1	3	3	4	0	18	2	2
M.1)	18	14	3	0	14	3	0	13	3	0	17	2	1
N.1)	19	18	2	0	19	1	1	16	2	0	16	2	0
F.2)	20	19	1	0	17	1	0	18	1	0	20	1	0
I.1)	21	21	0	0	21	0	0	21	0	0	21	0	0
Sum			179	31		173	37		145	65		169	41
Kendall τ			0.70			0.65			0.38			0.61	

ArchiBERTo ranking can be considered less affected by subjectivity being the most similar to the collective judgment (Figure 10) demonstrating its capability to mirror the collective intelligence of the group of experts in the objectives ranking task. The outcomes of the second DIP validate the reduced subjectivity of the NLP tool and its capability to reflect the collective knowledge of the panel of experts.

5.7 ArchiBERTo evaluation: customization capability

ArchiBERTo's capability to generate a ranking tailored to the DIP content is evaluated by comparing the rankings generated from processing four different DIPs with a fixed ranking determined collectively by three experts. The four DIPs used to evaluate the customization capability are selected among the five ones available and were not used to produce the training and validation datasets. To calculate the variation between the evaluations of the NLP tool and the collective evaluation, a score is assigned from 1 (representing the first and most important objective in the ranking) and 21 (representing the least important goal in the ranking). The aim is to measure the capability of ArchiBERTo to generate a ranking that is customized to the contents of the different DIPs. The variations of

ArchiBERTo rankings of the processed DIPs are verified with respect to the ranking collectively produced by the experts, considered as the benchmark for the analysis, and which is selected among the five ones available and was not used to produce the training and validation datasets.

5.7.1 DIP number 3-4-5-6: results and discussion

The DIP used as a benchmark concerns the design and construction of a primary and kindergarten school building. The panel of experts is asked to analyze and translate the content of a document into a fixed ranking that is used as a comparison for the outputs of ArchiBERTo related to four different DIPs. The typology of the school building, location, and the number of sentences processed by ArchiBERTo to produce the ranking of the four DIPs are listed below:

- DIP 03 concerns the design and construction of a secondary school building in the municipality of Sassari;
- DIP 04 concerns the design and construction of a primary and kindergarten school building in the municipality of Nuoro;
- DIP 05 concerns the design and construction of a secondary school building in the municipality of Monte Attu, province of Nuoro;
- DIP 06 concerns the design and construction of a primary and kindergarten school building in the municipality of Abbasanta province of Oristano.

As stated in the methodology section the analyzed DIPs being related to different types of schools will present different degrees of similarity with the benchmark DIP. In particular, DIP 03 and DIP 05 are expected to present a lower similarity (i.e., the lowest Kendal τ values, closer to zero). DIP 04 and DIP 06 will likely be the most similar to the benchmark DIP (i.e., the highest Kendal τ, closer but not equal to 1), being Primary and kindergarten schools like the benchmark DIP but with different objectives and needs and socio-cultural and economic context. To demonstrate the customization capability of the NLP tool, the Kendall τ values for the four DIPs are calculated and compared with the benchmark DIP, as presented in Table 8.

As shown in Figure 11, the ranking related to the DIP 04 and DIP 06 have the highest Kendall τ values. Both documents are related to the design and construction of primary and kindergarten school buildings like the benchmark DIP. DIP 03 and DIP 05 are positioned more distant from the DIP benchmark, considering the Kendall τ similarity values, being both documents related to the design and construction of secondary schools. None of the rankings reached a Kendall τ value of 1. This can be explained by the inner differences and the specificity of each DIP. In fact, each document reflects the individual specificities in terms of socio-economic, cultural, and territorial characteristics of the place where the school buildings will be constructed. The phenomenon can also be explained by the fact that the appointing party and the community of students, teachers, and citizens involved in the use of the buildings can have different interests, demands, and needs concerning the topic of each label. Looking at the label list, it is clear that different projects can require different degrees of accessibility and different necessities for the intervention to be integrated with the surrounding natural or built environment context according to the specificities of each location. Moreover, the appointing party can be more or less interested in objectives like the creation of ecological awareness, preferring objectives related to the sociocultural value and impact of the intervention or the development of a sense of respect in the users (e.g., in communities that are in particular situations of economic, social, and environmental distress). Conversely, objectives like the quality of the layout plan and indoor space, durability and maintainability of furniture and assets, or pedagogical objectives, shared among all the municipalities involved in the Project Iscol@, are all present and with equal relative importance.

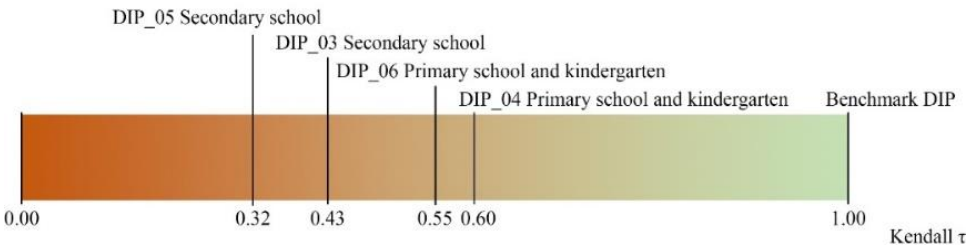

Figure 11: DIPs of different school (primary and secondary) Kendall τ values comparison.

Table 8. Concordant and Discordant pairs of ArchiBERTo different DIPs ranking compared with the benchmark DIP.

Labels	DIP_03	C	D	DIP_04	C	D	DIP_05	C	D	DIP_06	C	D
A.1)	8	13	7	5	16	4	6	15	5	6	15	5
L.1)	1	19	0	1	19	0	1	12	0	1	19	0
F.1)	2	18	0	3	17	1	5	8	3	3	17	1
O.1)	3	17	0	4	16	1	4	8	2	2	17	0
M.4)	5	15	1	7	14	2	8	6	3	8	13	3
P.1)	14	7	8	2	15	0	3	9	1	4	15	0
D.1)	13	7	7	13	8	6	10	3	3	11	10	4
L.2)	11	8	5	6	13	0	7	8	1	5	13	0
O.2)	7	10	2	8	12	0	2	8	0	7	12	0
E.1)	12	7	4	18	3	8	17	1	7	15	6	5
I.1)	18	3	7	15	5	5	15	2	5	21	0	10
M.3)	9	7	2	9	9	0	12	6	2	10	8	1
M.1)	10	6	2	11	7	1	19	4	6	14	5	3
N.1)	19	2	5	21	0	7	14	1	3	18	2	5
C.1)	16	3	3	17	2	4	21	1	6	13	4	2
G.1)	6	4	1	10	5	0	9	3	0	9	5	0
M.2)	4	4	0	12	4	0	16	3	2	20	0	4
E.2)	20	1	2	16	2	1	18	2	2	16	2	1
F.2)	21	0	2	20	0	2	20	1	2	19	0	2
I.2)	17	0	1	14	1	0	11	1	0	17	0	1
B.1)	15	0	0	19	0	0	13	1	0	12	0	0
Sum		151	59		168	42		103	53		163	47
Kendall τ		0.43			0.60			0.32			0.55	

The results support the hypothesis made that ArchiBERTo can adjust the objectives ranking based on the distinct DIP content. Additionally, none of the Kendall τ values are equal to 1, indicating the system's flexibility and not adhering to a fixed evaluation grid. Thus, the evaluation of the tool customization capability shows ArchiBERTo's capability in providing a tailored prioritization of objectives for different DIPs, reflecting the semantic content of each document, and maintaining a suitable degree of flexibility and compliance with the specific requirements of different designs and construction projects.

6 CONCLUSIONS

6.1 Recalling the outputs and results of the research project application

6.1.1 NLP-enhanced procurement model and ArchiBERTo performances

The study stands as one of the first applications of NLP methods and tools to documents belonging to the Pre-design phase in the Italian construction sector. ArchiBERTo exhibits high Precision, Recall, and F1-score values during the fine-tuning stage and shows promising results when processing text from a DIP of Project Iscol@ that was not part of the training and validation datasets, which are therefore unknown to the system. The objective of the research project was to evaluate the capability of ArchiBERTo to reflect and outperform the capability of single experts in the evaluation of a DIP and in the related hierarchization of the objectives. Therefore, the aim is to assess the capability of ArchiBERTo to reflect via the hierarchized list of objectives (labels) the real intentions that a public client originally intended when defining the DIP contents. On the contrary, evaluating and optimizing the performances of the algorithm on a single sentence is out of the scope of the research. The intended result of the

NLP system is the hierarchized list of labels or objectives that represents the whole DIP.

6.1.2 ArchiBERTo performances on DIPs

According to the results, ArchiBERTo's capability of mirroring the collective ability and sensitivity of a group of experts in interpreting, judging, and ranking the DIP sentences related to quality objectives in the architecture and construction knowledge domain is demonstrated. ArchiBERTo shows a lower subjectivity in the interpretation, judgment, and ranking process than the individual experts. ArchiBERTo, as a representation of a group of experts' knowledge, demonstrated better performance in analyzing multiple sentences in a DIP than a single expert. It also has the capability to generate tailored rankings based on DIP content with a good level of customization. This highlights the system's flexibility and capability to prioritize objectives based on the unique requirements of each project on a project-by-project basis, unlike the fixed evaluation grid currently used by Project Iscol@ to assess the design proposals.

6.1.3 Replicability and generalizability of the proposed methodology

The application of the proposed methodology to the pilot study is replicable with some limitations described as follows. The specific results of the application of the methodology to the case study depend on the subjective assessments of domain experts, to measure the subjectivity degree and the customization capability of ArchiBERTo, which are key aspects for the successful application of the proposed methodology. Consequently, it is not possible to directly reproduce the specific metrics of the application, which depend on the judgments and interpretations of the specific panel of experts in relation to which the ArchiBERTo performances are compared. However, it would be possible to reproduce the experiment with a different set of experts and achieve comparable results since the methodology and the application are extensively described, and the code for the ArchiBERTo development is provided at the following link: https://github.com/Mrk624/ArchiBERTo-dataset-and-code/blob/main/ArchiBERTo_github.ipynb. Therefore, the specific rankings produced on any given DIP document will be easily reproducible, with slightly different outputs due to the different panel of experts involved in the evaluation of the tool. In addition, the proposed methodology is generalizable. It can in fact be applied to other case studies of educational buildings belonging to the Project Iscol@ without changes to the methodology. In the case of application to other case studies of educational buildings not belonging to the Project Iscol@ or other building types, the methodology can be applied with minor changes. In particular, the NLP tool should be re-trained according to the specificities of the DIP documents (and related quality objectives corresponding to the labels of the MTC), procedures, and building types.

6.1.4 Advantages of the investigation

In Italian public tender procedures such as Iscol@, the NLP tool-generated prioritization ranking, which represents the numerical counterpart of the DIP contents, can be shared with the design teams involved in the tender process to enhance communication and provide them with a clear understanding of the appointing party's needs and quality objectives. Furthermore, this process assists the evaluation committee in comparing design proposals more objectively, thereby reducing the potential biases and subjectivity of the committee members (Figure 12). Consequently, the consensus issue that affects the traditional Italian public tender procedure is mitigated by the proposed methodology, which fosters and improves communication and consensus among the actors concerning the most important quality objectives and the relative hierarchy to define and evaluate the design proposals. Better communication and shared knowledge during the Pre-design phase can enable increasing the compliance between the design proposals and the public actor's quality objectives, needs, and demands, therefore minimizing the gap between expected and actual quality.

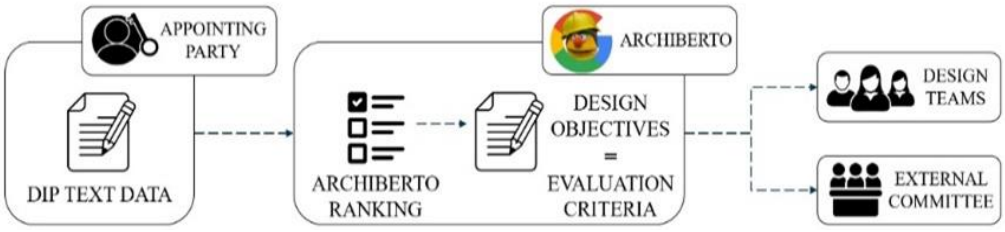

Figure 12: ArchiBERTo-enhanced evaluation system in the Italian Design-Bid public tender procedure.

The current experience of Project Iscol@ resulted in overly constrained projects as a result of a set of fixed, global objectives being applied to every project as opposed to acknowledging that distinct projects have their unique characteristics and needs. The proposed research methodology and the NLP tool ArchiBERTo ensure a flexible assessment of the prioritization of objectives on a project-by-project basis. The proposed methodology aims to expand the digitalization of the design and construction process to unstructured natural language data, addressing the limitations of the BIM approach in managing such information as the information in the qualitative section of a DIP document. The integration of BIM and NLP methodologies and tools can enable architects and engineers to manage both structured (alpha-numeric) and unstructured (natural language) data necessary for the design and construction process, advancing the digitalization of the industry. The proposed methodology, based on the NLP tool ArchiBERTo, allows the public appointing party to communicate their quality requirements using natural language, by acting exclusively on the subsequent translation phase. The digitalization of the preliminary quality objectives and needs by means of NLP systems could enhance the delivery of quality buildings and foster the introduction of digital methods and technologies into the construction process, both crucial steps for the future of the construction sector. The definition of the hierarchy of objectives improved by the NLP tool can enhance the communication between the actors during the Pre-design phase generating a positive impact on the overall quality of the competing design offers, facilitating at the same time the evaluation activity.

6.1.5 Delimitations of the research project

The delimitations of the research, i.e., the decisions that influenced the direction and parameters of the research, are the following:

- The research project is applied to the Project Iscol@ that focuses on educational buildings call for tenders' procedures, consequently, the application is only focused on educational buildings;
- Project Iscol@ currently uses a fixed evaluation grid to evaluate design proposals. The fixed evaluation grid includes quality objectives that are the result of the work and cooperation among different experts and end-users. Therefore, the fixed quality objectives were employed as, at the same time, labels and objectives for the NLP tool. The NLP tool is then applied to prioritize the labels according to the specificities of each DIP document, resulting in a hiererchized list of labels and objectives. Consequently, the research project did not focus on the definition of labels andobjectives;
- The research project involved the use of a LLM BERT-based model relying on the results of a literature review. Hence, different language models, e.g., BERT, RoBERTa, or GPT-3, are not compared in the methodology and application to find the most suitable model to develop ArchiBERTo, which could be considered a possible further development of the research.

6.1.6 Limitations of the investigation

One limitation of using the developed NLP system and, more generally, DL systems is their characteristic of being black boxes. In fact, despite the many advantages of DL approaches over statistical and rule-based methods, the DL approach offers little, or no explanation of the relationships modeled between the data. This phenomenon is called the black-box effect, which makes it impossible to understand how and what is learned by the DL algorithm. Moreover, the developers and designers of ArchiBERTo have far more control over the system than the end-users who are impacted by the system itself. Another limitation of the proposed application is the participation of only 3 experts in the development of ArchiBERTo, for reasons of practicality. The involvement of a larger number of experts would probably improve the performances of ArchiBERTo decreasing, even more, the subjectivity degree of the tool. Furthermore, the NLP tool is trained and evaluated with a limited dataset since the AECO field does not produce the large quantities of structured or labeled data that DL models typically thrive on. Therefore, the tool has necessarily been trained using a relatively limited dataset in terms of size not influencing the ArchiBERTo's performances.

7 FURTHER DEVELOPMENTS

7.1 Technological further development: from Text to Knowledge Graph

A possible further step of the research can be the design of an improved NLP system for the semi-automatic translation of natural language into digital formal entities. The formal entities suitable to represent needs and requirements are identified in the concept of Knowledge Graph (KG). A KG is a knowledge representation method

that belongs to the class of semantic networks. A KG represents semantic relationships between concepts (or entities). Each node represents a concept, and each arc or link represents a semantic relationship. Textual information, once processed through NLP systems, can be integrated into a queryable KG. Consequently, NLP tools based on BERT-like algorithms can act as a "bridge" that connects the world of documents and texts with the world of digital entities. NLP services built on the BERT language model or similar algorithms can process text documents and return digital entities (i.e., Knowledge Graphs). Therefore, the research could be improved by testing the feasibility and efficiency of NLP-based systems for generating a KG of a DIP document. The KG produced by the NLP tool processing the DIP content could be queried by the different actors as a unique and shared source of information.

FUNDING

This research received no external funding.

DATA AVAILABILITY STATEMENT

Publicly available datasets were produced in this study. Data and code can be found at the following link: https://github.com/Mrk624/ArchiBERTo-dataset-and-code.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Adel K., Elhakeem A. and Marzouk M. (2022). Chatbot for construction firms using scalable blockchain network. *Automation in Construction,* Vol. 141, n. 104390. https://doi.org/10.1016/j.autcon.2022.104390.

Alexakis G., Panagiotakis S., Fragkakis A., Markakis E. and Vassilakis K. (2019). Control of Smart Home Operations Using Natural Language Processing, Voice Recognition and IoT Technologies in a Multi-Tier Architecture. *Designs (Basel),* Vol. 3, n. 32. https://doi.org/10.3390/designs3030032.

Alhaj M.B., Liu H. and Sulaiman M. (2021). Towards Occupant-Centric Facility Maintenance Management: Automated Classification of Occupant Feedback Using NLP, in Walbridge, S., Nik-Bakht, M., Ng, K.T.W., Shome, M., Alam, M.S., el Damatty, A., Lovegrove, G. (Eds.), *Proceedings of the Canadian Society of Civil Engineering Annual Conference 2021,* Springer Nature, Online, 297–307. https://doi.org/10.1007/978-981-19-0968-9_24.

Baker H., Hallowell M.R. and Tixier A.J.P. (2020). AI-based prediction of independent construction safety outcomes from universal attributes. *Automation in Construction,* Vol. 118, n. 103146. https://doi.org/10.1016/j.autcon.2020.103146.

Bilal M. and Oyedele L.O. (2020). Big Data with deep learning for benchmarking profitability performance in project tendering. *Expert Systems With Applications* 147, 1–19. https://doi.org/10.1016/j.eswa.2020.113194.

Blair D.C. (1979). Information Retrieval. *Journal of the American Society for Information Science,* Vol. 30, 374–375. https://doi.org/10.1002/asi.4630300621.

Botha L.J. (2018). Data Mining Construction Project Information To Aid Construction Project Management.

Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D.M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I. and Amodei D. (2020). Language models are few-shot learners, *Advances in Neural Information Processing System,* Vol. 33, n. (NeurIPS 2020), Online, 1–75. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Candaş A.B. and Tokdemir O.B. (2022). Automated Identification of Vagueness in the FIDIC Silver Book Conditions of Contract, *Journal of Construction Engineering and Management*, Vol. 148.

https://doi.org/10.1061/(asce)co.1943-7862.0002254.

Chalkidis I., Androutsopoulos I. and Michos A. (2017). Extracting contract elements, *Proceedings of the International Conference on Artificial Intelligence and Law,* London, United Kingdom, 19–28. https://doi.org/10.1145/3086512.3086515.

Devlin J., Chang M.-W., Lee K. and Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Di Giuda G.M., Locatelli M. and Seghezzi E. (2020). Natural Language Processing and BIM In AECO Sector: A State Of The Art, *Proceedings of the Fifth Australasia and South-East Asia Structural Engineering and Construction Conference,* ISEC Press, Christchurch, New Zealand, 1–6. https://doi.org/10.14455/ISEC.2020.7(2).CON-22.

D'Orazio M., Di Giuseppe E. and Bernardini G. (2022). Automatic detection of maintenance requests: Comparison of Human Manual Annotation and Sentiment Analysis techniques, *Automation in Construction*, Vol. 134, n. 104068. https://doi.org/10.1016/j.autcon.2021.104068.

Elkhayat Y. and Marzouk M. (2022). Selecting feasible standard form of construction contracts using text analysis, *Advanced Engineering Informatics*, Vol. 52, n. 101569. https://doi.org/10.1016/j.aei.2022.101569.

Erfani A. and Cui Q. (2021). Natural Language Processing Application in Construction Domain: An Integrative Review and Algorithms Comparison, *Computing in Civil Engineering 2021 - Selected Papers from the ASCE International Conference on Computing in Civil Engineering 2021,* ASCE, Orlando, Florida, 26–33. https://doi.org/10.1061/9780784483893.004.

Ethayarajh K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMO, and GPT-2 embeddings, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, 55–65. https://doi.org/10.18653/v1/d19-1006.

European Commission (2019). Supporting digitalisation of the construction sector and SMEs: Including Building Information Modelling. https://doi.org/10.2826/422658.

Fang W., Luo H., Xu S., Love P.E.D., Lu Z. and Ye C. (2020). Automated text classification of near-misses from safety reports: An improved deep learning approach, *Advanced Engineering Informatics,* Vol. 44 (2020), n. 101060. https://doi.org/10.1016/j.aei.2020.101060.

Gharehchopogh F.S. and Khalifelu Z.A. (2011). Analysis and evaluation of unstructured data: Text mining versus natural language processing, *2011 5th International Conference on Application of Information and Communication Technologies AICT 2011*, IEEE, Baku, Azerbaijan, 44–47. https://doi.org/10.1109/ICAICT.2011.6111017.

Guévremont M. and Hammad A. (2021). Ontology for Linking Delay Claims with 4D Simulation to Analyze Effects-Causes and Responsibilities, *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction*, Vol. 13. https://doi.org/10.1061/(asce)la.1943-4170.0000489.

Hassan F. U. and Le T. (2021). Computer-assisted separation of design-build contract requirements to support subcontract drafting. *Automation in Construction*, Vol. 122. https://doi.org/10.1016/j.autcon.2020.103479.

Hassan F. U., Le T. and Tran D.H. (2020). Multi-Class Categorization of Design-Build Contract Requirements Using Text Mining and Natural Language Processing Techniques, *Construction Research Congress 2020: Project Management and Controls, Materials, and Contracts - Selected Papers from the Construction Research Congress 2020,* Tempe, Arizona, 1266–1274. https://doi.org/10.1061/9780784482889.135.

Hong Y., Xie H., Hovhannisyan V. and Brilakis I. (2022). A graph-based approach for unpacking construction sequence analysis to evaluate schedules, *Advanced Engineering Informatics*, Vol. 52, n. 101625.

https://doi.org/10.1016/j.aei.2022.101625.

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, 1–6. https://doi.org/10.1061/(ASCE)LA.1943-4170.0000308.

Jeon K., Lee G., Yang S. and Jeong H.D. (2022). Named entity recognition of building construction defect information from text with linguistic noise, *Automation in Construction*, Vol. 143, n. 104543. https://doi.org/10.1016/j.autcon.2022.104543.

Kendall M.G. (1938). A New Measure of Rank Correlation, *Biometrika*, Vol. 30, 81–93. https://doi.org/10.2307/2332226.

Kim E.W., Park M.S., Kim K. and Kim K.J. (2022a). Blockchain-Based Automatic Tracking and Extracting Construction Document for Claim and Dispute Support, *KSCE Journal of Civil Engineering*, Vol. 26, 3707–3724. https://doi.org/10.1007/s12205-022-2181-z.

Kim J.M., Lim K.K., Yum S.G. and Son S. (2022b). A Deep Learning Model Development to Predict Safety Accidents for Sustainable Construction: A Case Study of Fall Accidents in South Korea, *Sustainability (Switzerland)*, Vol. 14. https://doi.org/10.3390/su14031583.

Kim T. and Chi S. (2019). Accident Case Retrieval and Analyses: Using Natural Language Processing in the Construction Industry, *Journal of Construction Engineering and Management*, Vol. 145, 1–13. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001625.

Koc K., Ekmekcioğlu Ö. and Gurgun A.P. (2022). Prediction of construction accident outcomes based on an imbalanced dataset through integrated resampling techniques and machine learning methods, *Engineering, Construction and Architectural Management*. https://doi.org/10.1108/ECAM-04-2022-0305.

Lee J., Ham Y. and Yi J.S. (2021). Construction disputes and associated contractual knowledge discovery using unstructured text-heavy data: Legal cases in the United Kingdom, *Sustainability (Switzerland)*, Vol. 13. https://doi.org/10.3390/su13169403.

Li X., Zhu R., Ye H., Jiang C. and Benslimane A. (2021). MetaInjury: Meta-learning framework for reusing the risk knowledge of different construction accidents, *Safety Science*, Vol. 140, n. 105315. https://doi.org/10.1016/j.ssci.2021.105315.

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L. and Stoyanov V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach, *ArXiv*. http://arxiv.org/abs/1907.11692.

Mo Y., Zhao D., Du J., Syal M., Aziz A. and Li H. (2020). Automated staff assignment for building maintenance using natural language processing, *Automation in Construction*, Vol. 113, n. 103150. https://doi.org/10.1016/j.autcon.2020.103150.

Mohamed Hassan H.A., Marengo E. and Nutt W. (2022). A BERT-Based Model for Question Answering on Construction Incident Reports, Rosso, P., Basile, V., Martínez, R., Métais, E., Meziane, F. (Eds.), *Natural Language Processing and Information Systems*, Springer International Publishing, Cham, 215–223.

Moon S., Chi S. and Im S.B. (2022a). Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (BERT), *Automation in Construction*, Vol. 142, n. 104465. https://doi.org/10.1016/j.autcon.2022.104465.

Moon S., Lee G. and Chi S. (2022b). Automated system for construction specification review using natural language processing, *Advanced Engineering Informatics*, Vol. 51, 1–16. https://doi.org/10.1016/j.aei.2021.101495.

Ng H.S., Toukourou A. and Soibelman L. (2006). Knowledge Discovery in a Facility Condition Assessment Database Using Text Clustering, *Journal of Infrastructure Systems*, Vol. 12, 50–59. https://doi.org/10.1061/(asce)1076-0342(2006)12:1(50).

Norouzi N., Shabak M., Embi M.R.B. and Khan T.H. (2015). The Architect, the Client and Effective Communication in Architectural Design Practice, *Procedia – Social and Behavioral Sciences*, Vol. 172, 635–642. https://doi.org/10.1016/j.sbspro.2015.01.413.

Osborne M.L. (1975). A Modification of Veto Logic for a Committee of Threshold Logic Units and the Use of 2-Class Classifiers for Function Estimation, Oregon State University, USA.

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 (Basel)*, Vol. 14. https://doi.org/10.3390/en14185901.

Peng Z. and El-Gohary N. (2018). Automated Matching of Design Information in BIM to Regulatory Information in Energy Codes, *Construction Research Congress 2018, Proceedings,* New Orleans, Louisiana, pp. 75–85. https://doi.org/10.1061/9780784481264.008.

Qiao J., Wang C., Guan S. and Shuran L. (2022). Construction-Accident Narrative Classification Using Shallow and Deep Learning, *Journal of Construction Engineering and Management*, Vol. 148, 1–13. https://doi.org/10.1061/(asce)co.1943-7862.0002354.

Ren G., Li H., Liu S., Goonetillake J., Khudhair A. and Arthur S. (2021). Aligning BIM and ontology for information retrieve and reasoning in value for money assessment, *Automation in Construction*, Vol. 124, n. 103565. https://doi.org/10.1016/j.autcon.2021.103565.

Ren R. and Zhang J. (2021). An Integrated Framework to Support Construction Monitoring Automation Using Natural Language Processing and Sensing Technologies, *Computing in Civil Engineering 2021 - Selected Papers from the ASCE International Conference on Computing in Civil Engineering 2021,* Orlando, Florida, 1101–1109. https://doi.org/10.1061/9780784483893.135.

Salama D.M. and El-Gohary N.M. (2011). Semantic Modeling for Automated Compliance Checking, *Computing in Civil Engineering, Proceedings*, American Society of Civil Engineers (ASCE), Miami, Florida, United States, 641–648. https://doi.org/10.1061/41182(416)79.

Seghezzi E., Locatelli M. and Di Giuda G.M. (2020). Iscol@ Sardegna, *Arketipo*.

Senescu R.R., Haymaker J.R., Meža S. and Fischer M.A. (2014). Design Process Communication Methodology: Improving the Effectiveness and Efficiency of Collaboration, Sharing, and Understanding, *Journal of Architectural Engineering*, Vol. 20, 1–14. https://doi.org/10.1061/(ASCE)AE.1943-5568.0000122.

Sokolova M. and Lapalme G. (2009). A systematic analysis of performance measures for classification tasks, *Information Processing & Management*, Vol. 45, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.

Stenström C., Aljumaili M. and Parida A. (2015). Natural language processing of maintenance records data, *International Journal of COMADEM*, Vol. 18, 33–37.

Sun S. and Li L. (2022). Application of Deep Learning Model Based on Big Data in Semantic Sentiment Analysis, *The 2021 International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy. SPIoT 2021. Lecture Notes on Data Engineering and Communications Technologies*, Springer Science and Business Media Deutschland GmbH, Shanghai, China, 590–597. https://doi.org/10.1007/978-3-030-89508-2_76.

Taleb H., Ismail S., Wahab M.H. and Rani W.N.M.W.M. (2017). Communication management between architects and clients, *AIP Conference Proceedings,* Kedah, Malaysia, 1–6. https://doi.org/10.1063/1.5005469.

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

Viering T. and Loog M. (2021). The Shape of Learning Curves: a Review, *ArXiv*, 1–46. https://doi.org/10.48550/arXiv.2103.10948.

Williams T. and Halling M. (2014). Analyzing Asset Management Data Using Data and Text Mining.

Wu C., Li X., Guo Y., Wang J., Ren Z., Wang M. and Yang Z. (2022a). Natural language processing for smart

construction: Current status and future directions, *Automation in Construction*, Vol. 134. https://doi.org/10.1016/j.autcon.2021.104059.

Wu C., Xiao L., Rui J., Yuanjun G., Jun W. and Zhile Y. (2022b). Graph-based deep learning model for knowledge base completion in constraint management of construction projects, *Computer-Aided Civil and Infrastructure Engineering*, 1–18. https://doi.org/10.1111/mice.12904.

Wu J., Zhang J., Jin W. and Jiansong Z. (2022c). Model Validation Using Invariant Signatures and Logic-Based Inference for Automated Building Code Compliance Checking, *Journal of Computing in Civil Engineering*, Vol. 36, n. 4022002. https://doi.org/10.1061/(asce)cp.1943-5487.0001002.

Xie Q., Zhou X., Wang J., Gao X., Chen X. and Chun L. (2019). Matching Real-World Facilities to Building Information Modeling Data Using Natural Language Processing, *IEEE Access*, Vol. 7, 119465–119475. https://doi.org/10.1109/access.2019.2937219.

Xu N., Ma L., Liu Q., Wang L. and Deng Y. (2021a). An improved text mining approach to extract safety risk factors from construction accident reports, *Safety Science*, Vol. 138. https://doi.org/10.1016/j.ssci.2021.105216.

Xu N., Ma L., Wang L., Deng Y., Ni G. (2021b). Extracting Domain Knowledge Elements of Construction Safety Management: Rule-Based Approach Using Chinese Natural Language Processing, *Journal of Management in Engineering*, Vol. 37, 1–11. https://doi.org/10.1061/(asce)me.1943-5479.0000870.

Xue X., Hou Y. and Zhang J. (2022). Automated Construction Contract Summarization Using Natural Language Processing and Deep Learning, *Proceedings of the 39th International Symposium on Automation and Robotics in Construction,* Waterloo, Belgium, 459–466. https://doi.org/10.22260/isarc2022/0063.

Yang J., Chen Y., Yao H. and Zhang B. (2022). Machine Learning–Driven Model to Analyze Particular Conditions of Contracts: A Multifunctional and Risk Perspective, *Journal of Management in Engineering*, Vol. 38, 1–16. https://doi.org/10.1061/(asce)me.1943-5479.0001068.

Zhang G., Nulty P. and Lillis D. (2022). Enhancing Legal Argument Mining with Domain Pre-training and Neural Networks, *Journal of Data Mining & Digital Humanities NLP4DH*. https://doi.org/10.46298/jdmdh.9147.

Zhang L. and El-Gohary N. (2022a). Human-centred and BIM-integrated automated value analysis of buildings, *International Journal of Construction Management*, 1–13. https://doi.org/10.1080/15623599.2022.2025555.

Zhang Q., Hong Z. and Su X. (2021). Content Analysis Based on Knowledge Graph: A Practice on Chinese Construction Contracts, Ye, G., Yuan, H., Zuo, J. (Eds.), *Proceedings of the 24th International Symposium on Advancement of Construction Management and Real Estate*, Springer, Chongqing, China, 823–837. https://doi.org/10.1007/978-981-15-8892-1_59.

Zhang R. and El-Gohary N. (2022b). Building information modeling, natural language processing, and artificial intelligence for automated compliance checking, Lu, W., Anumba, C.J. (Eds.), *Research Companion to Building Information Modeling. Edward Elgar Publishing Limited*, 248–267. https://doi.org/10.4337/9781839105524.00022.

Zheng Z., Lu X.Z., Chen K.Y., Zhou Y.C. and Lin J.R. (2022). Pretrained domain-specific language model for natural language processing tasks in the AEC domain. *Comput* , Vol. 142. https://doi.org/10.1016/j.compind.2022.103733.

Zhong B., Wu H., Xiang R. and Guo J. (2022). Automatic Information Extraction from Construction Quality Inspection Regulations: A Knowledge Pattern–Based Ontological Method, *Journal of Construction Engineering and Management*, Vol. 148, 1–15. https://doi.org/10.1061/(asce)co.1943-7862.0002240.

Zhu Y., Emre Bayraktar M. and Chen S.C. (2010). Application of metadata modeling to dispute review report management, *Journal of Civil Engineering and Management*, Vol. 16, 491–498. https://doi.org/10.3846/jcem.2010.55.