FOSTERING THE CONSENSUS: A BERT-BASED MULTI-LABEL TEXT CLASSIFIER TO SUPPORT AGREEMENT IN PUBLIC DESIGN CALL FOR TENDERS

Mirko Locatelli, Giulia Pattini, Laura Pellegrini, Silvia Meschini, Daniele Accardo

TEMA Technologies Engineering Materials Architecture
e-ISSN 2421-4574
Vol. 9, No. 1 - (2023)

DOI: 10.30682/tema0901f

Abstract

Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) concerned with allowing computers to process natural human language. NLP is applied to solve several tasks in the design and construction process. However, in scientific literature, no applications are related to the pre-design phase and the processing of quality objectives and needs. The pre-design phase aims to reach a consensus between the stakeholders' quality demands and the design solution, relying on written natural language. Human language is the most pervasive and richest form of human knowledge representation and communication; however, at the same time, it is ambiguous, prone to misinterpretations, and hardly machine computable. Moreover, the mandatory procedural steps of the public tender procedure exacerbate the risk of misinterpretations inherent in using natural language. The study provides a methodology to design, assess, and evaluate an NLP tool based on the latest language model (i.e., BERT) to translate quality demands sentences into an evaluation grid in the Italian public tender context. The methodology is validated against a case study of an educational facility tender. The first results show good accuracy and capability of the NLP system to translate natural language into a numerical grid to support communication and foster consensus among actors, clarifying the appointing party and end-users' objectives to be reached via the design proposals.

Keywords

Deep Learning, Natural Language Processing, Knowledge mirroring, Knowledge representation, Educational facilities tender.

Mirko Locatelli*
DABC - Dipartimento di Architettura, Ingegneria delle Costruzioni e Ambiente Costruito, Politecnico di Milano, Milano (Italy)

Giulia Pattini
DABC - Dipartimento di Architettura, Ingegneria delle Costruzioni e Ambiente Costruito, Politecnico di Milano, Milano (Italy)

Laura Pellegrini
DABC - Dipartimento di Architettura, Ingegneria delle Costruzioni e Ambiente Costruito, Politecnico di Milano, Milano (Italy)

Silvia Meschini
DABC - Dipartimento di Architettura, Ingegneria delle Costruzioni e Ambiente Costruito, Politecnico di Milano, Milano (Italy)

Daniele Accardo
DM - Dipartmento di Management, Università degli Studi di Torino, Torino (Italy)

** Corresponding author:*
e-mail: mirko.locatelli@polimi.it

1. INTRODUCTION

1.1. TEXT SOURCES AND NATURAL LANGUAGE PROCESSING (NLP) IN THE CONSTRUCTION SECTOR

The combination of the latest technological advances and the increasing number of different typologies of available data sources has fostered data-driven research in the construction industry. The design and construction process deals with different and complex forms of information that are mainly captured and exchanged using text documentation: in the construction sector, «documents are interfaces, used to access and navigate through collections of information» [2]. Among the recent technologies, various text-based knowledge discovery techniques (i.e., Text Mining, Text processing, and Natural Language Processing) have emerged as a rapidly

growing set of literature, aiming at processing text data and information as one of the leading sources of analysis [1]. In particular, existing studies assessed and applied Natural Language Processing (NLP) technology to process different document types in the construction sector [8], such as contracts and legal agreements [3], design requirements specification [4], risk and safety reports [5, 6], building legislation [7].

NLP refers to the branch of Artificial Intelligence (AI) concerned with giving computers the ability to process natural human language in written or verbal form. Project, Safety, and Risk Management are the areas with the highest number of applications. Moreover, cases of combined applications of the Building Information Modelling (BIM) method and text processing to streamline Automated Compliance Checking tasks are currently being investigated, and NLP-based systems to convert regulatory information represent an active field of research [9]. However, in the scientific literature, no applications are related to the pre-design phase and specifically to the processing of quality objectives and needs expressed in written form via text documents [10]. The pre-design phase's main objective is to foster a consensus between the stakeholders' and end-users' quality demands and the design solution. The definition and sharing of quality demands primarily rely on natural language, which is the most pervasive and valuable form of human knowledge. However, natural language is subject to ambiguity and misinterpretation and is hardly machine-processable [11].

NLP technology can still be considered an innovative topic, especially for NLP applications in the design and construction research field. Consequently, the following section proposes an overview of the latest developments in NLP algorithms in the computer science research field.

1.2. NLP LATEST DEVELOPMENTS

Recently, NLP research has significantly improved with novel approaches that emphasize semantic meaning and context awareness, using the generalization capability of modern Deep Learning (DL) algorithms that enable semantic meaning processing and streamline NLP tasks.

Leading global companies such as Google and Facebook have helped popularize the use of pre-trained language models for word embedding (i.e., the conversion of text into numbers). In fact, for computers to comprehend human language, text must be converted into numbers (i.e., matrixes and vectors). The first pre-trained word embedding models, such as Word2vec, released by the Google team in 2013 [12], and Glove, published by the Stanford NLP research [13], are defined as non-contextual models. These models could not differentiate the different meanings of a word according to the context, failing to capture the impact of surrounding words on the meaning of individual words.

The limitations of non-contextual models prompted the development of language models able to provide contextualized embedding. The first example is ELMo (Embeddings from Language Models), which provides a deeply contextualized word representation that directly addresses the challenges of modeling complex characteristics of word use and how words use varies according to the context (i.e., polysemy phenomenon) [14]. The latest example is BERT (Bidirectional Encoder Representation from Transformers), released by the Google research team, which overcomes the main limitation of previous standard language models, which are unidirectional and limit the choice of architectures that can be used during the pre-training phase [15]. Both ELMo and BERT-based models can learn contextual relationships between words in a text by emphasizing semantic meaning and the impact of the context and, as a result, they are defined as context-aware models. A further key contribution of Google's BERT was the use of the pre-training using unlabeled text by conditioning jointly on both left and right context in all layers and the possibility to fine-tune the model [15, 16]. In fact, the BERT model is pre-trained on a massive amount of text data in an unsupervised manner to learn general linguistic patterns. The BERT model can be easily fine-tuned by adding a single output layer to create state-of-the-art models for a wide range of NLP applications on specific knowledge domains without an excessive array of new data [15]. From this point of view, BERT can be defined as a basic framework and a starting point for producing BERT-like versions fine-tuned on specific knowledge domains.

1.3. NLP APPLICATION IN ITALIAN PUBLIC DESIGN CALL FOR TENDERS

Considering the Italian public design call for tenders procedure, three main actors are involved: the appointing party, the design teams competing for the tender, and an external commission in charge of evaluating the design proposals. Each actor in the tender procedure has well-defined goals and roles:

1. the appointing party defines needs, objectives, and requirements, representing the end-users' needs to be satisfied through the design project. The Italian regulation requires the appointing party to communicate the needs in text form via a design guidance document called *Documento di Indirizzo alla Progettazione* (DIP);
2. the design teams participating in the call for tenders aim to deliver a design proposal to satisfy the appointing party and end-users' demands and win the call for tenders;
3. the external committee, composed of experts appointed by the appointing party, evaluates the various design proposals to identify the best design project, i.e., the most compliant with the DIP.

The DIP represents the appointing party and end-users' expectations about the design and, ultimately, about the building in a text form. The DIP structure is regulated by national law, and it has different mandatory contents which can be grouped into two main sections. A quantitative section provides technical requirements (e.g., minimum square meters per student (m^2/student)), economic and legal constraints (e.g., construction cost (€/m^2)), and regulations (e.g., minimum dimensions required by regulations: minimum height) that can be defined through alphanumeric parameters. A qualitative section describes quality objectives, needs, and demands defined and shared through verbal and natural language expressions (e.g., space flexibility or spatial and volumetric integration within the context).

During the design call for tenders procedure, the different actors are prohibited from communicating with each other. Moreover, each design team and the external committee must refer to the DIP to understand the main appointing party objectives during the design proposals' definition and evaluation, respectively. Consequently, a hierarchical organization of the appointing party's objectives and needs is typically manually implemented based on the actors' education, knowledge background, and experience and, accordingly, can be strongly subjective. This can lead to different interpretations of the relative importance of each quality objective due to the subjectivity of the DIP interpretation caused by the impossibility of confrontation and the absence of consensus among the involved actors. This, in turn, can ultimately cause a quality gap between the DIP quality objectives and needs and the design proposals. In fact, design proposals are typically defined and evaluated relying on the subjective view of individuals, thus increasing the risk of misinterpretations inherent in the use of natural language.

In such a context, the study proposes a methodology to design, assess, and evaluate an NLP tool based on the latest language model (i.e., Google BERT) to translate quality demands sentences included in a DIP into numerical values to support the definition of an evaluation grid, aiming at:

- reducing misinterpretation, or at least minimizing the different interpretations, of the relative hierarchy of quality objectives expressed by the appointing party, by defining a common and shared evaluation system;
- supporting the design teams to clearly identify the relative importance of the quality objectives and demands to be pursued by the design proposals;
- supporting the external committee in evaluating the design proposals according to the importance of the quality objectives and demands expressed in the DIP.

Summarizing, the evaluation system set by processing the DIP qualitative section via the proposed NLP tool aims to create a consensus about the relative hierarchy of quality needs and objectives among the three main actors involved. Consequently, the possible interpretations of the hierarchy and weights of the appointing party objectives by the design teams and the external committee are

minimized by providing them with an evaluation grid of the hierarchized objectives resulting from applying the NLP tool to the DIP quality objectives section. A case study is selected to assess the proposed methodology on a real DIP document for designing and constructing an educational facility.

2. METHODOLOGY

In order to develop an NLP-based tool to process and translate the quality objectives expressed in a DIP into an evaluation grid, which must represent the computational counterpart of the natural language information, the automatic labeling task was identified as the most suitable to achieve the goal.

Among the NLP techniques, the Multi-label Text Classification (MTC) was selected, which is a text analysis technique that automatically applies one or more predefined classification labels to a single text or sentence. Unlike common classification tasks, in which class labels are mutually exclusive, multi-label classification allows predicting and assigning multiple mutually non-exclusive classes, i.e., the predefined labels.

As stated, the research study aims to develop an NLP-based system to perform a multi-label classification to automatically process and translate the needs expressed in a DIP document into an evaluation grid. The predefined labels for the MTC are the general objectives guiding the design processes and evaluating design proposals. The MTC is performed to automatically assign labels, i.e., the predefined objectives, to each DIP sentence, also assigning a weight to each label depending on the correlation of the sentence with each objective/label. Once applied the NLP tool to all DIP quality section sentences, a priority ranking of the DIP quality objectives is generated according to their total weights. Consequently, an evaluation grid is defined, which can eventually be revised by the appointing party and then shared with design teams and the evaluation committee to achieve a convergence of consensus among the three main actors (Fig. 1).

2.1. NLP TOOL DEVELOPMENT AND EVALUATION

As stated, the NLP tool is trained to classify sentences by assigning multiple labels, taken from a predefined list, to the natural language expressions. The main activities to develop the tool, explained in detail in the following paragraphs, are listed as follows: labels definition, training and validation dataset production, model fine-tuning, and performance evaluation.

2.1.1. LABELS DEFINITION

The NLP tool must be trained to classify sentences according to a set of predefined labels. To create a consensus on labels, which represent the interests and quality objectives of appointing parties and end-users, the set of labels must be jointly defined by them, when possible, and by domain experts (architects, building engineers, and designers).

2.1.2. TRAINING AND VALIDATION DATASETS DEFINITION

The proposed NLP tool is based on the BERT language model, which is based on a neural network architecture optimized for language processing. The pre-trained BERT language model must be fine-tuned to solve the

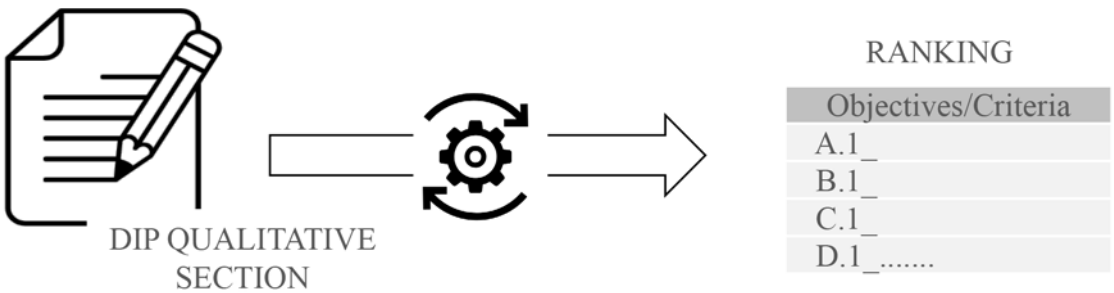

Fig. 1. *Schema of the proposed methodology.*

multi-label classification problem in the architecture and design knowledge domain. Consequently, a certain amount of training and validation data is needed. In particular, two datasets are defined:

- a training dataset for the first fine-tuning activity, which is used to further train the BERT model: the model learns from this dataset;
- a validation dataset to evaluate the performances of the trained model: the validation dataset is the set of data used to provide an unbiased evaluation of the model.

The general dataset is defined and then randomly split into the training and validation datasets at a 0.8:0.2 ratio.

The general dataset is defined by selecting DIP sentences and manually assigning labels, which is a critical task influencing the tool's accuracy and capability to automatically process and properly label the needs and quality objectives.

The production of the general dataset must result from a collaboration among experts with knowledge in the architectural, design, and construction fields. In addition, experts of specific knowledge domains, according to the specificity of each practical application of the methodology, must be involved in producing the general dataset. The involvement of domain experts allows for a proper representation of the knowledge domain and the avoidance of bias in the production of the general dataset. Consequently, the pre-trained BERT language model can be properly fine-tuned, representing a less biased capability of the group of experts to represent the knowledge domain.

Moreover, to further avoid any bias in the production of the dataset, each expert must be asked to independently propose a hypothesis for labeling each sentence. Then the experts would share their hypothesis and, in case of disagreement on some labels, they would be asked to share the motivation for their label choices and converge on a single common proposal. In fact, the construction of the dataset by different experts allows the model to represent and use their collective knowledge in the labeling activity.

By representing a collective intelligence of a group of people, larger than the ability of a single expert to judge and classify quality objectives related sentences, the NLP system aims to avoid subjectivity in interpreting textual information. Furthermore, the NLP tool will likely outperform the capability of a single expert to manage the complexity of analyzing several sentences, representing the group of experts' knowledge.

2.1.3. MODEL FINE-TUNING PARAMETERS

Once defined the dataset to properly train the BERT model, a set of hyperparameters must be defined. A hyperparameter is a variable configuration external to the model whose value cannot be estimated from the data. The list of hyperparameters used for the BERT NLP tool training follows:

- MAX_LEN: maximum number of tokens (words) processed during the training;
- TRAIN_BATCH_SIZE: refers to the number of training examples used in one iteration. A batch size of 32 means that 32 samples from the training dataset will be used to estimate the error gradient before the model weights are updated;
- VALID_BATCH_SIZE: refers to the number of examples used to validate in one iteration. A batch size of 16 means that 16 samples from the validation dataset will be used to validate the model;
- EPOCHS: an epoch is an entire transit of the training data through the algorithm. At each epoch, the internal model parameters of the dataset are updated. A training epoch ends when the learning algorithm has made one pass through the subgroups of the training dataset. The dimension of the subgroups is defined by the training batch size;
- LEARNING_RATE: it defines the adjustment in the neural network weights with respect to the loss gradient descent, determining how fast or slow the model will move towards the optimal weights.

2.1.4. PERFORMANCE EVALUATION METRICS

In order to measure the system's accuracy, the model predictions are compared with the human annotation, considered the best standard, and the F1-score metric is selected to measure the accuracy [17]. An explanation

of the metrics utilized for the evaluation of the model is provided as follows:

- False Positives (FP) occur when a classifier predicts a label that does not match the input sentence. Considering the sentence "Spaces must be accessible", if the model assigns the labels "Space flexibility" and "Space maintainability", both errors are false positives;
- False Negatives (FN) occur when a classifier misses a label that matches the input sentence.

Considering the previous example, if the classifier does not predict "Space accessibility", this is an example of a false negative.

Similarly, there are two ways classifier predictions can be corrected: True Positives (TP) and True Negatives (TN), described as follows:

- True Positives occur when a classifier correctly predicts the existence of a label;
- True Negatives occur when a classifier correctly predicts the in-existence of a label.

All the combinations are shown in Table 1.

		Predicted label	
		Positive (Pp)	Negative (Np)
Actual label	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

Tab. 1. *Possible combination of True/False positives and True/False negatives.*

Consequently, performance metrics can be calculated, i.e., Precision, Recall, and F1-score [18]:

- Precision is the ratio of correct predictions among all predictions of a certain class, i.e., the proportion of True Positives among all positive predictions (1);
- Recall is the ratio of examples of a certain class that the model has predicted as belonging to that class, i.e., the proportion of True Positives among all true examples (2);
- F1-score is the harmonic mean of a certain class Precision and Recall; it can be considered as an overall measure of the quality of the classifier predictions (3).

$$Precision\ (P) = {}^{TP}/_{FP + TP} \quad (1)$$

$$Recall\ (R) = {}^{TP}/_{FN + TP} \quad (2)$$

$$F_1 = {}^{2PR}/_{(P + R)} \quad (3)$$

F1-score values can range from 0 to 1. F1-score values equal to 1 represent a model that perfectly matches each observation with the correct label, and F1-score values equal to 0 represent a random classifier, i.e., a model that cannot match any observation with the corresponding label. Consequently, to evaluate the model performances, for each label, the authors agreed on a threshold value of the F1-score equal to 0.5, which is used to evaluate the tool predictions.

2.1.5. NLP TOOL OUTPUTS

Once the model is fine-tuned and the tool performances are evaluated, the NLP tool can be used to process new sentences and assign the corresponding labels (e.g., O.1. Spatial flexibility) and the accuracy degree with which the labels are associated to the new sentences (Fig. 2). The accuracy degree values of each processed sentence represent the weights of the labels and thus the relative priority of the labels/quality objectives for the single sentence.

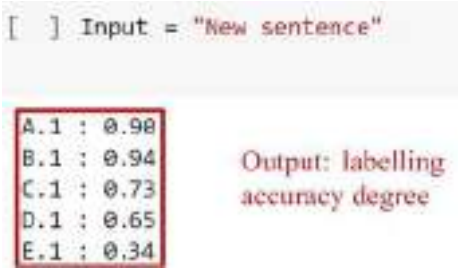

Fig. 2. *NLP tool outputs format.*

The accuracy values/weights of the labels obtained by processing all the sentences can then be summed and normalized to define the total weight of each label for the entire DIP (4).

$$Label\ Weight_i = \Sigma\ (L_i) / \Sigma\ (L_1) + \Sigma\ (L_2) + \Sigma\ (L_3) + \cdots + \Sigma\ (L_n) \quad (4)$$

where Li denotes the confidence of the i-th label, $i = (A.1, B.1, C.1..., P.1)$

The total weights of the labels represent the relative importance of the quality objectives to be pursued by the design teams in the definition of the design proposals and, at the same time, the evaluation criteria to be used by the external committee in the evaluation of the design proposals. The use by the design teams and the external committee of the same set of objectives, prioritized according to the appointing party's needs as expressed in the DIP, allows for an increase in the consensus among the three main actors about the most important objectives for the specific design.

3. CASE STUDY

3.1. CASE STUDY: *PROGETTO ISCOL@*

As introduced in Section 1, the case study aims to apply and assess the proposed methodology on a DIP document to design and construct a new school building. Since a school project has a high heterogeneity of quality objectives, needs, and requirements, and a high impact on the social and urban context, it was considered an appropriate building typology to assess the methodology. Specifically, the NLP tool is assessed on an Italian regional project, *Progetto Iscol@*, started in 2014 to realize several new school buildings in Sardinia. The Sardinian Regional Council introduced the *Progetto Iscol@* to address the problem of the backwardness of the regional educational system. The *Iscol@* team aims to renovate and expand the regional school building stock, improving the educational offer. The public investment of 265 million euros also has the economic objective of reactivating the Sardinian construction industry, setting a school system focused on architectural quality and social and environmental sustainability of the interventions.

At the beginning of *Progetto Iscol@*, the Sardinia Region shared general indications and guidelines for the drafting and the main contents of the DIPs to be produced by the involved local municipalities. Using common guidelines ensures that all DIPs follow the regional strategies, homogenizing the objectives of the interventions on the building school regional portfolio. In addition, a standardized evaluation grid for the design proposals evaluation process was established by the *Iscol@* team and shared with the local municipalities.

After completing the first set of calls for tenders and school designs of *Progetto Iscol@* in 2021, it was possible to analyze the impact of using a standardized evaluation grid (with fixed priority and weights of the predefined objectives) for all projects. On one hand, it was a useful tool to define and keep the focus on shared criteria and objectives that are in line with *Iscol@* strategic goals. On the other hand, a standardized evaluation grid with fixed priority and weights of the objectives is an overly rigid method to evaluate projects located in different contexts and with divergent specificities. In fact, there is an inner specificity of design and construction projects: buildings are considered "prototypes of themselves" and are strictly correlated and influenced by the context and the specific socio-economical and territorial needs. The use of fixed objectives priority and weights ultimately tends to flatten results, preventing a focus on the specificity and needs of each project.

Customizing the order priority and weight of the objectives for each call for tender, which is the output of the proposed NLP tool, can introduce the proper flexibility and specificity in the procedures.

Two phases of the *Progetto Iscol@* have been performed as of now: the first phase included ten project design competitions, and the second phase fifteen calls for tenders. All ten DIPs from the first phase and ten DIPs from the second phase were collected and used to produce the dataset (training and validation datasets), while the remaining five DIPs of the second phase were used to test the fine-tuned NLP system.

3.2. TOOL DEVELOPMENT AND EVALUATION

In this sub-section, the NLP tool fine-tuning is presented. The tool is trained and evaluated through the following steps described in the methodology section.

3.2.1. LABELS LIST

A list of predefined labels, defined within *Progetto Iscol@*, was already available for the proposed case study. The labels result from cooperation among experts and end-users (i.e., architects, designers, pedagogues, agronomists, and citizens). The list of predefined labels/objectives and the number of related sentences for the definition of the dataset (training and validation datasets) is provided in Table 2.

Labels/Objectives list	Number of Sentences
A Sociocultural value:	
A.1 Capability of the school building to be used as a Civic Center	130
B Creation of ecological awareness in users:	
B.1 Visibility and integration of sustainable design choices (educational medium) and integration of the intervention into nature and application of landscape enhancement strategies	42
C Development of a sense of belonging and respect for the common thing:	
C.1 Possibility of personalization of spaces and equipment to prevent vandalism creating a feeling of belonging	24
D Architectural and landscape quality of the intervention and integration with the pre-existing context:	
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	128
E Quality of layout plan:	
E.1 Articulation of spaces and accesses with a focus on simple and clear identification of the various functions, including using colors and signages	63
E.2 Presence of green spaces as an integral part of the design	32
F Indoor space quality:	
F.1 Perceptual quality (natural and artificial light) and psychophysical comfort (visual, thermo-hygrometric, acoustic etc.) to promote comfort and learning	128
F.2 Indoor air quality and healthiness	26
G Durability and maintainability:	
G.1 Cleanability, durability, maintainability, and replaceability of landscaping, materials, and greenery to reduce operating and maintenance costs	46
I Accessibility to the area:	
I.1 Integration of the intervention with the road system and distinction between driveways and bicycle and pedestrian paths; provision of areas and equipment to encourage slow and non-motorized mobility	36
I.2 Ensuring accessibility and usability for people with disabilities	45
L Integration between architecture and innovative pedagogical method:	
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	201
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 learning landscape	107
M Minimization of environmental impact with a view to ameliorative change:	
M.1 Use of renewable, natural (non-harmful), local materials or materials with recycled content	44
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.)	90
M.3 Integration between design and renewable energy production systems and exploitation/management of solar, light, and natural cooling and heating inputs	48
M.4 Requests regarding energy standards and minimization of consumption (energy, water etc.) including using monitoring systems	102
N Safety:	
N.1 Ensuring safety during school activities and separation between activity conducted by people not belonging to the school staff, maintenance activities (spaces and paths), the adequate delimitation of the school perimeter, and need for control/supervision	34
O Flexibility and adaptability of spaces:	
O.1 Spatial flexibility (furniture, facilities etc.)	204
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)	118
P Fostering the use of multimedia technologies in education:	
P.1 Usability of technological devices and integration with learning theories. Integration of space and technology; widespread presence of ICT technologies	106

Tab. 2. *Labels list and description.*

3.2.2. TRAINING AND VALIDATION DATASETS DEFINITION

In order to fine-tune the model, the dataset was defined by manually identifying, collecting, and labeling the sentences from the qualitative sections of the DIPs, according to the procedure described in the methodology section. The production of the general dataset resulted from a collaboration between three experts with knowledge in the architectural, design, and construction fields. In addition, since the proposed NLP system is applied to a specific case study (*Progetto Iscol@*), a deep knowledge of the strategic objectives of *Progetto Iscol@* was needed to correctly label the training sentences. Since *Iscol@* members could not be directly involved in the project, a preliminary study of the overall goals, guidelines, and context of *Iscol@* was conducted by the three selected experts before labeling the training sentences.

In addition, the authors want to highlight that, as the labeling and dataset definition is a group activity, the dataset should represent the collective ability and sensitivity of the group of people and experts. The dataset and, consequently, the NLP tool should be less biased and with a lower grade of subjectivity in automatically labeling new sentences and defining the evaluation grid. The tool, in fact, represents the numerical counterpart of the ability of the group to prioritize quality objectives and needs, representing their collective and common knowledge.

3.2.3. MODEL FINE-TUNING PARAMETERS

The values of the hyperparameters were defined according to the dataset characteristics and after a cycle of trial and error (Tab. 3).

Consequently, these are the values of the hyperparameters that allow the obtainment of the best fine-tuned model considering the available dataset.

3.2.4. PERFORMANCE EVALUATION METRICS

Precision, Recall, and F1-score are calculated for each label (Tab. 4).

The performances of the NLP tool in the sentence labeling task seem to be good, showing only two labels with an F1-score value below the threshold of 0.5 and nineteen labels with an F1-score above the threshold. Consequently, the model can be considered properly fine-tuned.

MAX_LEN	TRAIN_BATCH_SIZE	VALID_BATCH_SIZE	EPOCHS	LEARNING_RATE
85	2	32	20	2 E-05

Tab. 3. *Hyperparameters values adopted for the model fine-tuning.*

Metrics	A.1	B.1	C.1	D.1	E.1	E.2	F.1	F.2	G.1	I.1	I.2
Precision	0.71	0.75	0.75	0.56	0.88	0.67	0.68	1.00	1.00	0.67	0.55
Recall	0.92	0.75	1.00	0.42	0.64	0.33	0.77	1.00	0.90	0.67	0.86
F1-score	0.80	0.75	0.86	0.48	0.74	0.44	0.72	1.00	0.95	0.67	0.67
Metrics	**L.1**	**L.2**	**M.1**	**M.2**	**M.3**	**M.4**	**N.1**	**O.1**	**O.2**	**P.1**	
Precision	0.67	0.85	0.67	0.82	0.67	0.64	1.00	0.88	0.44	0.65	
Recall	0.80	0.39	1.00	0.90	0.86	0.60	0.50	0.71	0.78	1.00	
F1-score	0.73	0.54	0.80	0.86	0.75	0.62	0.67	0.78	0.56	0.79	

Tab. 4. *Precision, Recall, and F1-score values per each label.*

4. RESULTS AND DISCUSSION

In the following paragraphs, the fine-tuned NLP tool is applied to new sentences from one of the five available DIPs of *Progetto Iscol@* second phase to evaluate the results of the application of the proposed methodology.

4.1. NLP TOOL SENTENCE LEVEL EVALUATION

In order to demonstrate the tool's ability to translate sentences related to qualitative aspects into numerical values in the context of *Progetto Iscol@*, the developed algorithm was tested on a sample of sentences belonging to a new DIP among the five available. Three examples of the labeling and numerical translation of quality-related sentences are provided in Figure 3.

```
[27] test_commenti = "Spazi delle aule\
     devono essere aperti, modulabili, facilmente riconfigurabili. L'aula deve diventare uno spazio\
     flessibile e deve adattarsi alle nuove esigenze della didattica che prevede che in questi spazi\
     siano sviluppati lavori di gruppo, interazioni continue con l'insegnante che in questo spazio\
     prepara e verifica la programmazione didattica complessiva"

     _, test_prediction = trained_model(encoding["input_ids"], encoding["attention_mask"])
     test_prediction = test_prediction.flatten().numpy()

     for label, prediction in zip(LABEL_COLUMNS, test_prediction):
       if prediction < THRESHOLD:
         continue
       print(f"{label}: {prediction}")

     i.1 : 0.98
     0.1 : 0.97
```

```
[52] test_comment7 = "Lo spazio esterno è parte integrante del progetto e come detto si caratterizza per la\
     presenza di un oliveto: inteso come continuazione ed estensione dello spazio interno, dovrà\
     prevedere un'alternanza tra percorsi e spazi per lo svolgimento di attività all'aperto, essere\
     progettato come luogo privilegiato per il gioco in spazi in ombra e spazi più soleggiati, il\
     movimento, l'apprendimento attivo, l'incontro tra pari e tra i due ordini di scuola, dove\
     sviluppare la consapevolezza dell'importanza di crescere in un ambiente sostenibile e\
     salubre e incrementare comportamenti e stili di vita rispettosi dell'ambiente"

     _, test_prediction = trained_model(encoding["input_ids"], encoding["attention_mask"])
     test_prediction = test_prediction.flatten().numpy()

     for label, prediction in zip(LABEL_COLUMNS, test_prediction):
       if prediction < THRESHOLD:
         continue
       print(f"{label}: {prediction}")

     B.1 : 0.85
     E.2 : 0.72
     L.2 : 0.72
```

```
[34] test_comment8 = "Rappresenterà il luogo d'incontro privilegiato tra la scuole dell'infanzia e la scuola primaria\
     per le attività legate al Progetto continuità e costituirà lo spazio di mediazione tra il\
     quartiere e la scuola, ambito di possibile socializzazione dei genitori, spazio adeguato per\
     attività extrascolastiche. Sarà uno spazio funzionale ad ospitare manifestazioni, mettere in\
     scena rappresentazioni e organizzare attività laboratoriali, anche sotto forma di grandi tavoli\
     di lavoro, proiezioni di contenuti multimediali."

     _, test_prediction = trained_model(encoding["input_ids"], encoding["attention_mask"])
     test_prediction = test_prediction.flatten().numpy()

     for label, prediction in zip(LABEL_COLUMNS, test_prediction):
       if prediction < THRESHOLD:
         continue
       print(f"{label}: {prediction}")

     A.1 : 0.48
     L.1 : 0.67
     O.2 : 0.91
     P.1 : 0.10
```

Fig. 3. *Example of the programming code developed to process and label three new sentences.*

Labels	A.1	B.1	E.2	L.1	L.2	O.1	O.2	P.1
Confidence sum	0.40	0.05	0.72	1.66	0.72	0.98	0.91	0.10
Criteria weights	7.18	0.95	12.98	29.96	13.00	17.63	16.46	1.84

Tab. 5. *Weights values calculated per each label/criterion according to the NLP outputs.*

Considering only the results of the three sentences from the selected DIP, it was possible to define the evaluation grid customized to the DIP textual content, as shown in Table 5.

The grid obtained through processing the three sentences (Fig. 3) represents the numerical counterpart of the objectives expressed in natural language. The highest weight is obtained by objective L.1, related to integrating space and innovative teaching based on group work and peer learning. Objectives O.1 and O.2, respectively related to space flexibility and extracurricular use (temporal flexibility), get the second and third highest weights. According to the grid, the remaining objectives are sorted as follows: L.2, the openness of indoor space to outdoor space; E.2, the presence of greenery; A.1, the capability of the school building to be used as a Civic Center; P.1, the integration of space and technology and widespread presence of ICT technologies; and, finally, B.1, the visibility and integration of sustainable design choices as an educational medium. The identified objectives and the related weights assigned by the NLP tool seem to reflect the meaning of the sentences used as a sample, confirming the system's ability to identify and translate the objectives, expressed in textual form in the DIP, into a numerical scale of objectives/criteria.

5. CONCLUSIONS AND FURTHER DEVELOPMENTS

The study stands as the first application of NLP methods and tools to documents belonging to the pre-design phase in the Italian construction sector. It demonstrates good levels of precision and recall during the fine-tuning process and useful results in processing samples of textual information derived from a DIP of *Progetto Iscol@* previously excluded from the training and validation datasets of the NLP tool.

The ability of the NLP tool (i.e., a Multi-label Classifier based on the latest language model BERT) to translate DIP quality demands sentences into numerical values to support the definition of an evaluation grid will reduce the possible misinterpretations about the relative hierarchy of the appointing party quality objectives by the design teams and by the external committee. Consequently, the proposed system enables the definition of a common and shared evaluation system fostering a consensus among the involved actors.

Moreover, having been the training dataset definition a group activity, the tool seems to be able to mirror the collective ability, sensitivity, and knowledge of the group of experts involved in the dataset definition. Consequently, the NLP tool can be considered less biased and with a lower grade of subjectivity on the definition of the evaluation grid, being the tool the numerical counterpart of the ability of a group to prioritize quality objectives and needs. These aspects are discussed in more detail in the following chapter.

5.1. FURTHER DEVELOPMENTS

The next step of the research involves measuring the developed system's ability to assign labels/objectives relying on the collective capability of the expert panel and to produce customized evaluation grids for each DIP and, consequently, for each project. For this purpose, two scenarios A and B are proposed.

Scenario A aims to measure the NLP tool's subjectivity degree and capability to represent collective knowledge and intelligence. The NLP tool will process sentences related to the quality objectives, and a weight will be provided for each objective. Firstly, the experts will manually analyze the same sentences, individually providing a weight for each objective. Then the group of experts will perform the same activity collectively. The scores assigned individually, collectively, and by the NLP tool will be compared.

Scenario B intends to evaluate the customization capability of the NLP tool to customize the outputs, mirroring the contents of different DIPs. DIPs of different school buildings (primary and secondary schools) will be processed, and different evaluation grids will be produced using the NLP system. The customized evaluation grids will then be compared with the standardized evaluation grid of *Progetto Iscol@*.

The proposed further developments of the research will further improve the NLP tool, verifying the subjectivity degree and capability of customization employing the two proposed scenarios.

6. REFERENCES

[1] Erfani A, Cui Q (2021) Natural Language Processing Application in Construction Domain: An Integrative Review and Algorithms Comparison. In: Computing in Civil Engineering 2021 - Selected Papers from the ASCE International Conference on Computing in Civil Engineering 2021. ASCE, Orlando, Florida, pp 26–33

[2] Haimes R (1994) Document Interface. Interactions 1(4):15–18. https://doi.org/10.1145/194283.194296

[3] Kim Y, Lee JHJ-H, Lee E-BB, Lee JHJ-H (2020) Application of Natural Language Processing (NLP) and Text-Mining of Big-Data to Engineering-Procurement-Construction (EPC) Bid and Contract Documents. In: Proceedings - 2020 6th Conference on Data Science and Machine Learning Applications, CDMA 2020. IEEE, Riyadh, Saudi Arabia, pp 123–128

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

[5] Tixier AJP, Hallowell MR, Rajagopalan B, Bowman D (2016) Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. Automation in Construction 62:45–56. https://doi.org/10.1016/j.autcon.2015.11.001

[6] Zou Y, Kiviniemi A, Jones SW (2017) Retrieving similar cases for construction project risk management using Natural Language Processing techniques. Automation in Construction 80:66–76. https://doi.org/10.1016/j.autcon.2017.04.003

[7] Lee H, Lee J-K, Park S, Kim I (2016) Translating building legislation into a computer-executable format for evaluating building permit requirements. Automation in Construction 71:49–61. https://doi.org/10.1016/j.autcon.2016.04.008

[8] Yan H, Yang N, Peng Y, Ren Y (2020) Data mining in the construction industry: Present status, opportunities, and future trends. Automation in Construction 119:1–16. https://doi.org/10.1016/j.autcon.2020.103331

[9] Di Giuda GM, Locatelli M, Seghezzi E (2020) Natural Language Processing and BIM In AECO Sector: A State Of The Art. In: Proceedings of International Structural Engineering and Construction. Emerging Technologies and Sustainability Principles in Structural Engineering and Construction. ISEC Press, Christchurch, New Zealand, pp 1–6

[10] Locatelli M, Seghezzi E, Pellegrini L, Tagliabue LC, Di Giuda GM (2021) Exploring Natural Language Processing in Construction and Integration with Building Information Modeling: A Scientometric Analysis. Buildings 11:33. https://doi.org/10.3390/buildings11120583

[11] Sun S, Li L (2022) Application of Deep Learning Model Based on Big Data in Semantic Sentiment Analysis. In: 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, pp 590–597

[12] Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv: 1–12. https://doi.org/10.48550/arXiv.1301.3781

[13] Pennington J, Socher R, D. Manning C (2014) GloVe: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp 1532–1543

[14] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, New Orleans, Louisiana, pp 2227–2237

[15] Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 16. https://doi.org/10.48550/arXiv.1810.04805

[16] Malte A, Ratadiya P (2019) Evolution of Transfer Learning in Natural Language Processing. arXiv. https://doi.org/10.48550/arXiv.1910.07370

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

[18] Blair DC (1979) Information Retrieval. Journal of the American Society for Information Science 30:374–375. https://doi.org/10.1002/asi.4630300621