Management of construction safety and health plans based on automated content analysis

Abstract

Safety management in the construction sector continues to be of great concern due to the high accident rate. Enormous legislative efforts have been made to minimize this problem. In Europe, the Directive 92/57/EEC, which establishes a mandatory safety and health requirements, defines the Safety and Health Plan. This paper contains a methodology for automatically extracting structured relevant information from this textual document. Such plans are crucial to provide safety in this sector. However, current documents are often enlarged with nonessential information, whilst the most relevant information is missing or difficult to locate. The information extracted by our methodology makes it easier for Safety and Health Plans to be validated and approved at an early stage, as well as facilitating their later use and updating. The main steps of this methodology are: i) the definition of the relevant content to be extracted as seven items, in collaboration with experts in the area; ii) text preprocessing and enrichment using existing software resources for the natural language employed; and iii) automated content analysis for extracting information, on the basis of a collection of previously established rules for identifying the content defined in the first step. The output is a detailed and structured report containing the percentage of relevant information identified, as well as the desired items of information. To illustrate and validate our methodology, we have implemented and applied it to 50 real Safety and Health Plans. As a result, our system was able to identify most of the required information for five of the seven items in all of these plans. In addition, we have been able to detect the problems in the textual plan that caused the missing information in the other two items, which

will be corrected in future developments.

Keywords: Safety and Health Plans, Construction, Automated Content Analysis, Safety Management, KNIME

---

1. Introduction

The accident rate in the construction industry is among the highest compared to other industries all over the world. [1, 2, 3, 4]. Therefore, one of the most important tasks involves project management which requires proper data management is workplace safety [5, 6]. However, this is not an easy task due to both the context in which construction work is carried out and its specific characteristics. These include the traditional nature of the sector [7], the large number of documents throughout the project life cycle [8], the lack of standardization in documents and data [9], relative uniqueness and complexity of the projects [10, 11], legal requirements [12], and, the changing environment and the overlapping of multiple and diverse activities [13]. Focusing on safety, all these aspects, which are commonly reflected and collected in documents, make the effective management of safety difficult since information is usually scattered, repeated and fragmented in diverse data sources or even in the same document. Furthermore, the relevant information concerning safety can be difficult to find by people responsible for safety management.

Rigorous efforts have been made to provide a safe and healthy working environment for construction workers and to reduce casualties [14]. In this regard, strong governmental enforcement of safety laws and regulations has been established. These regulations involve a large amount of information through many very different documents. In Europe, Directive 92/57/EEC [15] establishes mandatory safety and health requirements at temporary or mobile construction sites. Article 3 of the European Directive states that the contractor in a worksite will have to create a Safety and Health Plan (hereinafter named S&H Plan). This plan should reflect the prevention management for that worksite adapted to the specific circumstances of the work and the execution process.

Concretely, this document should identify, plan, organize and control all activities to be performed from the perspective of prevention: work procedures, risks and preventive measures. An important aspect in this context is that this document must be updated whenever any modification occurs during the construction process, making it proactive and flexible. This living and dynamic document must be drafted and approved before work begins. In the case of private projects, the coordinator is responsible for approving it while in public projects it is the labor authority.

The S&H Plan should be considered a real instrument on the construction worksite to carry out the work safely. All the information contained in this document should be reliable, real and coherent considering the peculiarities of the project. Nevertheless, in most cases, the content of this document is far from ideal, being very long, heterogeneous and complex, with redundant and nonessential information. This means that it is not as useful as it should be for the required purposes. The main drawbacks identified in this document are: information overload (the more the better), unstructured information, generalization of risk assessment and preventive measures, and absence of relevant information. These aspects make it difficult to manage this document, both for the coordinator and the labor authority during review and validation. In addition, these issues make this document less practical than it should be during the execution process.

The Information and Communication Technologies (ICT) area can provide some interesting techniques for helping to deal with the aforementioned aspects in an efficient way. Nowadays, the construction research community is increasingly embracing the use of ICT to support data management [16, 17] and to achieve safe construction practices [18]. Safety research to date has tended to focus on several technologies, such as Virtual Reality (VR) [19, 20], Geographic Information System (GIS) [21], 4D CAD [22], Building Information Modeling (BIM) [23, 24], sensing technologies [25, 26], Internet of Things (IoT) [27], etc. These technologies are mainly designed to efficiently capture, store, update, manipulate, analyze, and display data on construction projects. However, not only

data management (in a strict sense) is needed but also the process which allows knowledge discovery in datasets, especially from textual data, is becoming more and more important for organizations [28, 29]. In this regard, Natural Language Processing (NLP) is a very active and rapidly evolving research area that deals with the comprehension and analysis of human-produced texts by computers through artificial intelligence, computer science and linguistics [30, 31]. One of the applications of NLP is automated content analysis, which is increasingly being used for a variety of applications even in the construction domain [32]. As mentioned previously, the S&H Plan is created by humans using natural language, which has irregularities, unstructured information, etc. This paper shows that these drawbacks can be addressed using automated content analysis.

The main objective of this paper is the integration of different elements and technologies into a methodological approach for dealing with S&H Plans to find relevant information to help people responsible for safety management. Our methodology has been implemented and tested in the specific case of constructing concrete-based structures. We have shown that it helps to detect at an early stage whether the S&H Plan meets specific requirements by analyzing the content and automatically extracting relevant information contained in the plan. Moreover, our approach can provide a significant time-saving advantage for coordinators and the labor authority when validating the S&H Plan and fulfilling its aim in construction sites: to execute work safely.

The first step of our proposal consists of determining what content should appear in the S&H Plan. Reference works in the literature and consultation with a group of experts in the field of construction and safety, with full knowledge of project management, were used to determine this. Most content refers to general aspects of the construction project. However, some requirements of the approach focus on the execution of structures (specifically, structures made of concrete) since it is one of the phases with the highest number of accidents in the construction industry [33]. Secondly, three phases are considered: preprocessing, enrichment and automated content analysis. In the preprocessing phase the methodology removes noise and inconsistent data to facilitate the extrac-

tion of the information. The enrichment phase adds useful information to the preprocessed text to locate the required information. Finally, the automated content analysis phase is composed of a set of rules that detect the information in the document. This proposal makes it possible to quickly verify at an early stage, whether the plan contains the required information that is relevant to the management of construction workplace safety.

The remainder of the paper is structured as follows. Section 2 is devoted to explain the current way of elaborating the S&H document. Section 3 presents previous works regarding NLP, while Section 4 describes the methodology based approach. Section 5 presents some results and finally, guidelines for future research (Section 6) and conclusions (Section 7) end the paper.

2. Safety and Health Plan

According to the European Directive 92/57/EEC [15], the S&H Plan is a document in which the contractor plans, organizes and controls each one of the activities from the point of view of workplace safety and health. In addition, this document constitutes the basic instrument for risk assessment and preventive planning. The S&H Plan is created before starting the activity since the coordinator or the labor authority must perform an evaluation of the plan. This process may involve considerable time since the document can exceed 100 pages in most cases. Unfortunately, the legislation does not establish a structure to organize the information in the document. This drawback along with those mentioned in the introduction makes the validation process very difficult.

In the construction industry, safety planning generally consists of identifying all potential hazards, assessing the risks, and choosing the corresponding safety measures for the worker, task, tools, and work environment [1]. It is thus necessary to carry out the risk assessment (tolerable, moderate, important, etc.) to define specific preventive and organizational measures and determine the necessary Individual Protection Equipment (IPE) and Collective Protection Equipment (CPE). Below is an example of the hazard “Objects falling while

being handled or after coming loose":

- Hazard type: Objects falling while being handled or after coming loose
- Hazard controls:
  - Do not stand under suspended loads
  - Always ensure hooks with their safety latch
  - Always check the state of cables, ropes, slings, etc.
  - Do not gather material on the edge of shutter work
  - Make sure materials are stored correctly
  - Do not leave tools, equipment or materials in the working platforms of the scaffolding
- Individual Protection Equipment (IPE): helmet
- Collective Protection Equipment (CPE): safety nets

However, this is not an easy task due to the construction characteristics, such as, changing environment and overlapping of multiple and diverse activities. If risks are not assessed the control measures cannot be developed and implemented since those involved are not aware of the hazard [34].

As mentioned before, each company must elaborate an initial assessment based on the activities that they develop, determining the preventive measures that will be applied to control the identified risks. But this document does not end at this point. The European Directive 92/57/EEC [15] establishes that this plan must be updated, expanded and modified whenever any modification occurs during the construction process. For example, these modifications can be produced as a result of changes in materials, processes, design, accidents, land characteristics or work methods, among others. Then, the S&H Plan must be a dynamic and flexible document.

Nevertheless, this document tends to be a mere requirement that must be prepared in case of a labor inspection. This poor perception is largely due to the quality of the document itself and, consequently, to its limited use in construction worksites [35]. The main drawbacks during its creation are lack

of knowledge of the execution processes, lack of data when the S&H Plan is being prepared, information overload (the idea that the more the better) and the generalization of risk assessment and preventive measures.

As such, the automated analysis of this information might represent an important challenge in this context. To the best of our knowledge, in the literature, there are no proposals to address this problem. With the approach presented in this paper, the S&H Plan will be analyzed automatically to check that it includes the information determined by experts.

3. Preliminaries

The construction is a greatly information-dependent industry since a lot of documents need to be transferred and exchanged in order to support the different important management tasks during the project life-cycle. For example, in a construction project, even a small one, a large amount of information is contained in textual and digital data coming from specifications, plans, process control, safety management, inventory, costs, and scheduling documents [8, 36, 7, 37]. However, traditional analytics can generate informative reports, but fail when it comes to content analysis. As a result, pattern recognition, data mining and Knowledge Discovery Data have received major attention, as they can provide reliable results and effectively assist in data analysis and knowledge extraction [32]. These techniques have been used to achieve diverse objectives, such as, to classify project information [10, 38], analyze construction contracts [8], costs [39] and, to structure and manage safety knowledge [40, 41, 42, 1].

In recent years, NLP is a very active and rapidly evolving area of research that deals with the comprehension and analysis of human-produced texts by computers [43]. It enables machines to derive meaning from human language inputs [44]. NLP lies at the confluence of artificial intelligence, linguistics, and computer science and aims at enabling computers to process natural language text in a human-like manner. Applications of NLP include speech recognition, machine translation, and automated content analysis [45]. Automated content

analysis is increasingly being used for a wide variety of applications. This is motivated by the necessity to make sense of the digital information and to take advantage of the fast-growing volume of it [30].

Some researchers have been trying to introduce NLP into the construction industry for several purposes, e.g. retrieval of computer-aided drawings [37], automatic analysis of injury reports [30], automatic clustering of construction project documents based on textual similarity [46], risk management [47] and construction site accident analysis [6]. Some of these proposals combine the use of NLP with other techniques with the aim of improving the efficiency and performance of particular applications.

However, some authors discussed that current applications of NLP in the construction industry represent a challenge [37, 41]. Generally, the main obstacles when creating a system that understands natural language are the difficulty to accurately model grammars and the disambiguation of words.

4. Methodology based approach

In this section, which is divided into two main parts, a step-by-step methodology based approach is detailed. Firstly, a reference structure which defines the content that the S&H Plans should contain is proposed. Subsequently, the methodology which analyzes the mentioned content of this document in an automatic way is detailed.

4.1. Required information

To evaluate the content of a given S&H Plan it is necessary to determine what content it should contain. As explained in Section 2, the S&H Plan is the document in which the contractor plans, organizes and controls each one of the activities from the point of view of workplace health and safety. Specifically, in this study we focus on projects using concrete structures.

There is no predefined regulation that determines the structure or the content of an S&H Plan [48]. This makes sense since, as mentioned previously,

this document should be adapted to the peculiarities of each project. The legislation gives contractors the freedom to choose the information that should be considered [15].

At this point, we should stress that the aim of our proposal is not to establish a model that designers can use as a reference for developing their plans, but rather to propose a methodology for testing the content that plans should contain. To do this, the content is structured in a hierarchical manner on two levels.

In the first level, reference works in this domain have been reviewed [48, 49] with the aim of defining an initial proposal regarding content requirements. Then, ten experts in the field of construction, with a deep knowledge of project management, were invited to review and validate this proposal. As a result of this process, Table 1 shows the items and subitems that should be contained in an S&H Plan in a structured way. In addition, each item has been evaluated with a percentage from 0% to 100% according to its importance in comparison to the rest of the items. Similar percentages, from 0% to 100%, were assigned to the subitems. The sum of the item level and subitem level is 100%.

It should be noted that some items are general and cover the entire construction process, but others are particular to the process of building concrete structures. This is the case of the item regarding "Risks, preventive measures and individual protection equipment", which will involve tasks that are needed for concrete formwork, placing rebars and pouring concrete. To identify automatically these issues in the S&H Plan, three lists containing all of the possible situations were created and reviewed by the panel of experts.

Table 1: Relevant items and subitems to consider in the S&H Plan

Items	**% items**	**Sub-items**	**% sub-items**
Construction ID	8%	Denomination	-
Information about professionals and companies	9%	Promoter	11
		Project designer	9
		Author of the S&H Study	16
		Project manager	16
		S&H Coordinator in the execution phase	27
		Contractor	21
Preventive structure of the main construction company	13%	Construction manager	25
		Foreman	16
		Foreman 2	11
		Organizational mode for preventive activity	24
		Preventive resource	24
Planning of the preventive action	23%	Access Control Plan	16
		Authorization plan for the use of machines	16
		Maintenance Plan	13
		Delivery Control Plan (IPE)	14
		Emergency and Self-Protection Plan	17
		Training and Information Plan	24
Execution process	29%	Risks	38
		Preventive measures	41
		Individual Protection Equipment (IPE)	21
Plans	9%		-
Measurement and budgets	9%		-

As can be observed, the items of greater importance are those that focus on the definition of preventive activities, such as structure and personnel, planning, and the execution process. The rest of the items have a similar value, around 9%. The next section explains the proposed methodology.

4.2. Proposed methodology

Once the features that our proposal will be able to evaluate have been defined and before explaining them in depth, Figure 1 details the flowchart of the proposed methodology based approach. Firstly, as can be observed, the S&H Plan is converted from a pdf file to an excel file. This step is necessary to

analyze the information contained in the original file. A preprocessing phase is carried out to remove unnecessary information from the document. Next, names and locations are identified in the text given that some items contain this kind of information. After that, the content is automatically analyzed to identify whether specific items and subitems are contained in the S&H Plan. Finally, a report with the results is obtained.

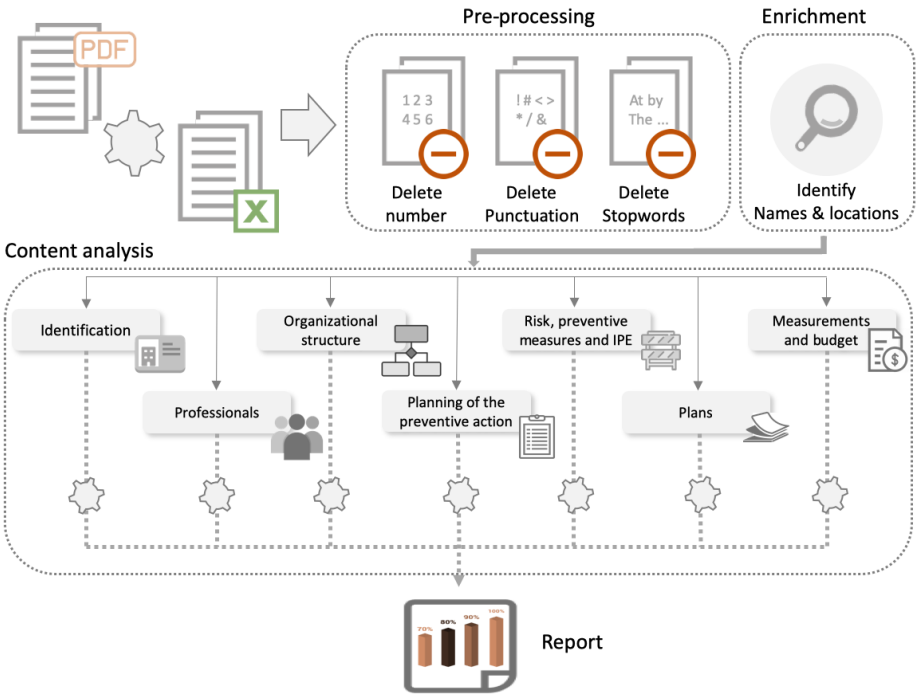

Figure 1: The flowchart of the proposed methodology based approach

The proposal has been implemented under the well-known KoNstanz Information MinEr (KNIME) software [50]. It is a freely available software, with a graphical interface where users can design workflows that allow to analyze and mine data in an intuitive and easy manner. Users can configure the workflows by dragging nodes from the repository and linking them through their output and input ports. Additionally, KNIME provides nodes for different stages of data mining techniques, such as, data sources, data pre-processing steps, model

building algorithms, as well as visualization tools [51].

Next subsections focus on explaining the design process of our proposal, which is divided in three phases: pre-processing, enrichment and content analysis phase.

4.2.1. Preprocessing phase

To carry out the preprocessing, the information must be extracted from the text of the S&H Plan. In our proposal, data collection is performed using the file reader node in Excel format.

Once the document is loaded, it should be preprocessed to remove noise and inconsistent data, and the format must be adapted to facilitate the extraction of the information. This phase could be considered one of the most important in terms of text analysis. Following, the steps that are involved in this phase are detailed:

- Lower case conversion. In this step the text is transformed into lower case which reduces variations of the same word, e.g., after transformation Employee and employee are indistinctly managed as the same word.
- Punctuation marks removal. They are removed given that they usually do not contribute to text analysis.
- Stopwords removal. Stopwords are frequent words that do not provide relevant information to text analysis, such as, conjunctions, prepositions, and pronouns. Therefore, they are removed in this step [52].
- Tokenization. Tokenization is the task of chopping a text sequence up into words, called tokens [53].
- Semantic pre-processing. This stage consists of the detection and grouping of synonyms. To do this, we have elaborated a domain specific list based on expert knowledge.

Following, an example of this preprocessing phase is shown:

– Using a sample sentence from the source S&H Plan:
"Safety and Health Plan for the execution of 3 isolated single-family houses,

which are located in the province of Málaga."

– The preprocessing process returns:
safety health plan execution 3 isolated single-family house are located province malaga

4.2.2. Enrichment phase

The objective of this phase is to add useful information to the preprocessed text. This means linking tags to words to provide information about the type of word they are, such as noun, verb, adjective, etc. In corpus linguistics, this is *part of speech tagging*.

Additionally, tags may also report the semantics of the word, namely, whether it corresponds to a person, location or organization. To do this, two complete lists are included with Spanish nouns and locations to identify automatically these words [1].

To continue with the previous example, the output of this enrichment phase would be: :
safety (noun), health (noun), plan (noun), execution (verb), 3 (number), isolated (adjective), single-family (adjective), house (noun), are (verb), located (verb), province (noun), málaga (location).

Once the text has been preprocessed, the automatic content analysis is conducted. The next section details this process.

4.2.3. Automated content analysis

As illustrated in Figure 1, the automated content analysis has been split into seven items: (i) identification, (ii) professionals, (iii) organizational structure, (iv) planning the preventive actions, (v) execution process (risk, preventive measures and IPE), (vi) plans and (vii) measurements.

The first step in the design of the automated content analysis process is the establishment of rules to detect variables based on combinations of generic

[1] Though the example uses terms in English for comprehension purposes, the current proposal works with Spanish terms.

keywords in the text. This is not an easy task when handling unstructured texts, due to the difficulty in examining similar structures. The second step is the development of a thesaurus, i.e. a keyword dictionary with specific terms in this context. Although there has been a great deal of construction safety research, to the best of our knowledge there is no lexicon related to S&H available at the time of this research.

The KNIME software allows to develop a fully automated content analysis system based on hand-coded rules and dictionaries of keywords. Following, we show how the seven aforementioned items are extracted from the text.

I. Construction site identification

The identification of the construction site is usually defined on the cover page of the plan or on the first page. Hence, an interval of the first 20 words of the document has been established to verify the identification of the construction site in two ways. Firstly, looking for words tagged as location and identifying construction typologies (hotel, isolated house, block of flats, etc.) in the previous words. Secondly, looking for the word "plan" and analyzing the subsequent words with the aim of identifying the construction typologies.

If either of these two ways provides a positive result, it is understood that the identification of the construction site is provided and the maximum score is reached in this item.

II. ID Professionals

In this item, the aim is to identify the different professionals involved in the construction process: promoter, project designer, author of the S&H study, project manager, safety and health coordinator in the execution phase and contractor. To do this, each professional is analyzed separately by means of a combination of words, including words tagged as "person" or "organization" and keywords for each professional. These keywords refer to the different denominations used for the professional. For example, the promoter may also be defined as owner.

Hence, the searching process is defined as follows: firstly, keywords are explored in the text concerning each professional. Secondly, the words that appear before and after each coincidence are analyzed to recognize words tagged as person or organization. This recognition process is possible thanks to the enrichment phase described above. Thirdly, if coincidences are found, it is considered that the professional is indicated in the plan and the corresponding percentage is assigned. Finally, the results of all the items are added up so as to estimate the total percentage of this second item.

III. Preventive structure

In this item, the aim is to identify the information concerning both the professionals with safety responsibilities and preventive structures: construction manager, foreman, foreman 2, organizational mode for preventive activity, and preventive resource. This is performed similarly to the previous one, except that only words tagged as "person" are analyzed. This is because there are no elements related to organizations in this item.

IV. Planning the preventive action

The objective of this item is to identify relevant information regarding the planning of preventive actions, such as access control plan, authorization plan for the use of machines, maintenance plan, delivery control plan (IPE), emergency and self-protection plan, training and information plan.

This item may be very difficult to examine because the information may be presented in different ways, using both unstructured and free text. To be able to analyze this information, the text is preprocessed in sentences, which are the input of this item. Items have to be detected based on combinations of keywords that should appear in a specific order in the text.

V. Execution process

In this item, the objective is to verify that the S&H Plan contains some

requirements that have been identified as essential for the execution of concrete structures with regard to risks, preventive measures and IPE. As explained previously, a group of experts in the construction domain completed and validated three lists containing these requirements.

On the one hand, 15 risks have been defined. For example: falling formwork, falls on different levels, cuts in general, stepping on sharp objects, etc. On the other hand, 25 sentences have been established for preventive measures. Examples include (i) formwork is forbidden without first having covered the risk of falling from high places by installing safety nets or handrails, (ii) the use of vertical nets in the perimeter of the structure is compulsory, (iii) when handling of rebars, the operator will protect his hands with gloves, etc. Finally, 8 individual protection devices have been defined as fundamental in the execution of concrete structures, such as helmets, safety boots, safety belts, etc. These three lists have to be preprocessed with the same process used previously with the text in the S&H Plan to be able to compare them.

Therefore, the implementation of this item has been divided into three parts, corresponding to each of the previously described objectives. Below a real example is shown to illustrate the process.

- Sentence from S&H Plan - "*The workplace must be ordered, free of obstacles and clear of waste*"
- Sentence from S&H Plan preprocessed - "*workplace must be ordered free obstacles clear waste*"
- Search sentence - "*A lot of effort must be put into order and cleanliness during the execution of the work*"
- Search sentence preprocessed - "*A lot of effort must be put in cleanliness during execution work*"

As can be seen in the previous examples, the phrases corresponding to risks, preventive measures or IPE that are searched for in the text might

be written in different ways, but they have the same meaning. In this case, a word-for-word search in the text would give the wrong result. To verify that this kind of information is contained in the text, a selection of representative words should be made for use as keywords.

First, the terms of the phrase are reduced to those strictly essential to maintain the meaning. For example, verbs, articles and prepositions are eliminated. Taking the previous example, the sentence would be: "*order cleanliness execution work*". Secondly, dictionaries of synonyms are used to avoid errors due to the use of different words with the same meaning. In the previous example, the word "cleanliness" will be replaced by the representative word "clean" given that it refers to the same concept. Then, the sentence would be: "*order clean execution work*". Once the sentence has been preprocessed, the methodology proceeds to verify in the S&H Plan those phrases in which all the selected terms appear. These terms might appear in any order, but the absence of one of them will result in a mismatch.

In the end, a percentage is assigned to the subitem based on the number of matching phrases in the text.

VI. Plans

The inclusion of the plans or blueprints is easy to verify at first glance, but this verification becomes complex when it is performed through text processing. In the previous analysis phase, it could be observed that all S&H H plans that contain blueprints include an entry in the index. Therefore, in this item, we search for a title or an entry in the index with the term "Blueprint". If a match is found, item 6 is assigned the highest rating.

VII. Measurement and budgets

This item is performed similarly to the previous one. The words "measurements" or "budget" are searched in the index and the maximum punctuation is achieved when an output is obtained.

5. Results

In this section, the results are presented with the aim of showing that the methodology can accurately identify the established requirements in the S&H Plan. Two different stages were used to achieve this objective: the learning phase and the testing phase. Firstly, twenty-five S&H Plans were analyzed to determine the differences between the output produced by the methodology and the actual content of the plans. Then, an additional twenty-five S&H Plans, different from those in the first phase, were analyzed.

Below, Table 2 shows the results from the learning phase:

- The *Output* column presents the results provided by the methodology. As can be observed, for the first item "construction site identification" 8 points have been obtained. This means that the methodology has correctly recognized the identification of the construction site in the plan. Notice that the total assignment for this item was 8, as shown in table 1.
- The second column *output** tries to verify that the automatic methodology works correctly. To do this, an expert reviews the outputs to detect possible errors. For example, it may be that the methodology is not able to correctly distinguish the different participants: contractor, promoter, etc. If the value in this column is the same as the first column, it would mean that the output was correct. Nevertheless, if the result is higher or lower, it would mean that the methodology has identified some error or has not identified any, respectively. The information in this column has been very useful in the learning phase to discover different ways to describe the items. The text in S&H Plans is unstructured, regarding both organization and language. Anomalies identified in this phase have been used to provide feedback for the methodology.
- The third column, *expert*, shows the results that the expert has previously assigned to each element after reviewing the information from the plans. This column is especially interesting in both in the learning and testing phase to verify the results.

Table 2: Data analyzed in the learning phase

	Output	Output*	Expert	Deviation
I. CONSTRUCTION SITE IDENTIFICATION	8	8	8	0
Promoter	11	0	11	0
Project designer	9	0	9	0
Author of the S&H Study	16	16	16	0
Project manager	0	0	0	0
Safety and Health Coordinator in the execution phase	27	27	27	0
Contractor	21	21	21	0
II. ID PROFESSIONALS	7,56	5,76	7,56	0
Construction manager	25	25	25	0
Foreman	0	0	0	0
Foreman 2	0	0	0	0
Organizational mode for preventive activity	0	0	0	0
Preventive resource	24	24	24	0
III. PREVENTIVE STRUCTURE	6,37	6,37	6,37	0
Access Control Plan	0	0	16	-16
Authorization Plan for the use of machines	0	0	0	0
Maintenance Plan	13	0	13	0
Delivery Control Plan (IPE)	0	0	0	0
Emergency and Self-Protection Plan	0	0	0	0
Training and Information Plan	24	24	24	0
IV. PLANNING THE PREVENTIVE ACTION	8,51	5,52	12,19	-3,68
Risks	28,15	27,87	25,333333	2,5333334
Preventive measures	11,9	11,48	16,4	-4,92
Individual Protection Equipment (IPE)	17,06	21	18,375	2,625
V. EXECUTION PROCESS	16,5619	17,5015	17,431417	-0,8695167
VI. PLANS	9	**9**	9	0
VII. MEASUREMENT AND BUDGETS	9	9	9	0
TOTAL	65,0019	61,1515	69,551417	**-4,5495167**

- Finally, the last column shows the deviation between the automatic output column and the expert column.

These data are obtained for each of the S&H Plans considered in the learn-

ing phase and are then used to improve the content analysis methodology. A subsequent testing phase is carried out with another twenty-five plans that were not used in the learning phase.

Table 3 provides a summary of the results for each of the seven proposed items and each column of the table is detailed. Notice that the deviation is calculated by the difference between the output of the methodology and the expert's assessment:

- N°S&H Plan >0 : This column represents the number of S&H plans where the output is greater than 0. This means that the methodology has identified something wrongly.
- % N°S&H Plan >0 : The percentage of plans where the deviation is greater than 0 is represented in this column.
- Avg Deviation >0 : This column represents the error average percentage. Notice that in most items, this value corresponds to the punctuation assigned to this item, since it is an absolute value. However, in point 5, the mean is relative.
- N°S&H Plan <0 : Contrary to the first column, this column represents the number of S&H plans where the output is less than 0. This means that the methodology has not totally identified the information that is described in the plan.
- % N°S&H Plan <0 : The percentage of plans where the deviation is less than 0 is represented in this column.
- Avg Deviation <0 : Similarly to the third column, the error average percentage is represented in this column.

Table 3 allows us to analyze the results both in a particular and general way. Firstly, from a deeper perspective, the table provides the results for each subitem. As can be observed, items 4 and 5 show greater differences. This is due to the freedom allowed when providing this kind of information, as we will explain below. However, although the number of plans is higher, the average deviation is lower than in other items, such as item 2 or 3.

From the results in column 1, we can observe that in some items (concretely, 1, 2, 6 and 7) the methodology works and has correctly identified the information. However, in some plans, not all of the required information has been identified, as it is apparent from the values in column 4.

As can be seen from the table, the methodology currently works reasonably well and correctly identifies the information. It can be deduced from the data in the first and fourth column given that the number of times where the result is greater than 0 is considerably less considering all items. Nevertheless, the number of times with a result lower than 0 is more frequent in the different items. In both columns, the number of plans is very reduced in most items (between 2 and 3), except in item 5 due to its own peculiarities.

Table 3: Overall results

	N°S&H Plan >0	% N°S&H Plan >0	Avg Deviation >0	N°S&H Plan <0	% N°S&H Plan <0	Avg Deviation <0
I. CONSTRUCTION SITE IDENTIFICATION	0	0%	-	3	12%	-8,00
Promoter	0	0,00%	-	2	8,00%	-11,00
Project designer	0	0,00%	-	5	20,00%	-9,00
Author of the S&H Study	0	0,00%	-	4	16,00%	-16,00
Project manager	2	8,00%	16,00	0	0,00%	-
Safety and Health Coordinator in the execution phase	0	0,00%	-	2	8,00%	-27,00
Contractor	0	0,00%	-	2	8,00%	-21,00
II. ID PROFESSIONALS						
Construction manager	0	0,00%	-	0	0,00%	-
Foreman	0	0,00%	-	0	0,00%	-
Foreman 2	0	0,00%	-	0	0,00%	-
Organizational mode for preventive activity	1	4,00%	24,00	1	4,00%	-24,00
Preventive resource	1	4,00%	24,00	1	4,00%	-24,00
III. PREVENTIVE STRUCTURE						
Access Control Plan	0	0,00%	-	2	8,00%	-16,00
Authorization Plan for the use of machines	2	8,00%	16,00	0	0,00%	-
Maintenance Plan	3	12,00%	13,00	0	0,00%	-
Delivery Control Plan (IPE)	1	4,00%	14,00	2	8,00%	-7,50
Emergency and Self-Protection Plan	1	4,00%	17,00	13	52,00%	-17,15
Training and Information Plan	5	20,00%	24,00	1	4,00%	-24,00
IV. PLANNING THE PREVENTIVE ACTION						
Risks	8	32,00%	2,85	11	44,00%	-4,90
Preventive measures	3	12,00%	6,01	20	80,00%	-9,48
Individual Protection Equipment (IPE)	20	80,00%	3,28	0	0,00%	-
V. EXECUTION PROCESS						
VI. PLANS	0	0	-	2	0,08	-9,00
VII. MEASUREMENT AND BUDGETS	0	0	-	0	0	-
TOTAL	5	20,00	2,97	20	80,00	-9,07

Additionally, a total result is shown in the last row. It is an average value that is calculated by adding the results for each item in each plan. 520 As a general overview, in 20% of the plans additional information has been identified and there was a deviation of only 2.97 points. On the other hand, in 80% of the plans, part of the information has not been identified, which represents a total of 9.07 points.

The output results from our approach are shown in a detailed report for each S&H Plan. This report allows the coordinator and labor authority to explore the scores that our approach has assigned to each item and their corresponding subitems as well as to examine the information retrieved. The first part of the report shows a table with the overall output results. The first column describes each of the items while the second column shows the percentages represented in two values. The first corresponds to the percentage achieved by our approach while the second one corresponds to the total percentage assigned by the expert panel. The rest of the report shows detailed information about the items and subitems.

As an example and for the sake of simplicity, Figure 2 shows an extract of the report for a real S&H Plan which contains a total of 15 pages. In this example, items I, VI and VII have obtained the maximum score hence our approach has been able to retrieve the information from these items in the S&H Plan. The rest of the items present a lower percentage in relation with that assigned by the expert panel. The second part of the first page of the report shows the information retrieved for items I, VI and VII as well as the percentages for subitems II, III, IV and V. As can be observed, our approach has been able to correctly retrieve the identification of the project corresponding to item I: "Isolated house", "PLOT U.A. 1.26","CAPANES SUR 1 URB", "BENAHAVIS (MALAGA)". Meanwhile, the second item contains the percentages obtained for each subitem. Some professionals corresponding to item II have also been identified due to the maximum score achieved (i.e. promoter, the author of the S&H study, S&H coordinator and the constructor).

Evaluation of Safety and Health Plan number: E_044

OVERALL RESULTS FOR THE DIFFERENT ITEMS:

Item	percentage	
	System	Expert
I. CONSTUCTION SITE ID	8,00 %	8,00 %
II. PROFESSIONAL ID	6,75 %	9,00 %
III. PREVENTIVE STRUCTURE	3,12 %	13,00 %
IV. PLANNING PREVENTION	15,87 %	23,00 %
V. EXECUTION PROCESS	19,14 %	29,00 %
VI. PLANS	9,00 %	9,00 %
VII. MEASUREMENTS AND BUDGETS	9,00 %	9,00 %

Detailed results for each item/subitem:

ITEM I. CONSTRUCTION SITE ID
"ISOLATED HOUSE", "PLOT U.A. 1.26", "CAPANES SUR 1 URB.", "BENAHAVÍS (MÁLAGA)"

ITEM II. PROFESSIONALS AND COMPANIES	Percentage	
	System	Expert
Promoter	11 %	11 %
Project designer	0 %	9 %
Author of the S&H Study	16 %	16 %
Project manager	0 %	16 %
S&H Coordinator in the execution phase	27 %	27 %
Contractor	21 %	21 %

ITEM III. PREVENTIVE STRUCTURE OF THE MAIN CONSTRUCTION COMPANY	Percentage	
	System	Expert
Construction manager	0 %	25 %
Foreman	0 %	16 %
Foreman 2	0 %	11 %
Organization mode for prevention	24 %	24 %
Preventive resource	0 %	24 %

ITEM IV. PLANNING OF THE PREVENTIVE ACTION	Percentage	
	System	Expert
Access control Plan	16 %	16 %
Authorization plan for the use of machines	16 %	16 %
Maintenance plan	13 %	14 %
Delivery control plan (IPE)	0 %	14 %
Emergency and Self-Protection plan	0 %	17 %
Training and information plan	24 %	24 %

ITEM V. RISKS, PREVENTIVE MEASURES AND IPE	Percentage	
	System	Expert
Risks	25,3 %	38 %
Preventive Measures	19,6 %	41 %
IPE	21 %	21 %

ITEM VI. PLANS
"PLANS", "TECHNICAL SHEETS"

ITEM VI. MEASUREMENT AND BUDGETS
"MEASUREMENT", "BUDGETS"

Figure 2: Page 1 of the report as output where general punctuations are shown

The rest of the report provides the information retrieved for each item but, for the sake of simplicity, an extract is provided in 3. As can be seen, the name of the identified professionals is shown in the second column of the table. Thanks to this information, both the coordinator and the labor authority can review the plan and request from the constructor the missing or incomplete information to validate the S&H Plan with the complete data.

Retrieved information for the item II:

ITEM II. PROFESSIONALS AND COMPANIES	
Promoter	*COPROANSA S.L [ORGANIZATION (NE)]*
Project designer	
Author of the S&H Study	*MIGUEL [PERSON (NE)]* *JESUS [PERSON (NE)]*
Project manager	
S&H Coordinator in the exec. phase	*JAVIER [PERSON (NE)]*
Contractor	*COPROANSA S.L [ORGANIZATION (NE)]*

Figure 3: Report of information regarding promoter subitem identified

The main drawbacks found for each of the items are listed below:

1. Item 1 regarding *construction site identification* is quite reliable, giving erroneous results only in plans where there is no cover. In these cases, the location of the work is expressed throughout the document. In none of the 25 plans has incorrect information been located for this item, as can be seen in Table 3.
2. The second item concerning *ID professionals* works correctly, but problems arise when professionals are described in great detail in the text. This situation is not very frequent but, in this case, the methodology provides more than one output or an incorrect output if one of them is not described. In the first option, there is no real problem because the methodology is able to detect the different professionals. However, if the professional is not described, but there are others described in great detail, the output may be incorrect.
3. For the item concerning the *preventive structure* the behavior is similar to the previous item. For example, the person designated by [company] as

the construction manager, with the aim of supervising the work executed, is [name].

4. The main problem in the *planning prevention* item is that the fact that one of these plans (access control, maintenance, ...) appears in the text does not mean that it is defined. For example, in a given S&H Plan the following appears: "Contractor will be required to make a plan for training and information for workers". In this case, the training and information plan is not defined but the methodology will provide an output. Specifically, item 4 is where fewer lower results are obtained due to its characteristics. The person in charge of writing the S&H Plan is allowed considerable freedom when explaining these items and they depend on the characteristics of the construction process.
5. The main drawbacks in the fifth item concerning the *execution process* occur with the preventive measures due to the freedom given when explaining them. Two sentences written with different vocabulary might have the same meaning. This issue may be easily solved by means of a more complete dictionary of synonyms from the construction domain. This dictionary should be flexible with the aim of being able to include any new terms that may appear. The next section details a proposal for future work in which the appearance of some terms can be more relevant than others. In this way, the search would not be so strict and would allow the identification of more similar sentences.
6. Regarding the *plans*, only two plans were not identified due to the fact that they did not have a cover and index. However, this is not very common.
7. Finally, regarding the *measurement and budget* item, our approach works correctly. In all plans, this item has been perfectly identified and inaccurate information has not been detected.

6. Limitations and recommendations for future research

Designing and implementing a methodology that can understand natural language is a very challenging task. Noteworthy to this context, the difficulty to accurately model grammars and the importance of word sense disambiguation. In this section, some limitations and reflections of the methodology are detailed as well as recommendations for future research.

First, the methodology is limited by the use of solid detection rules: it is not so robust to unknown and erroneous inputs, such as misspelled or missing words. Although a complete thesaurus of this domain has been developed, it should be flexible with the aim of incorporating new terms and concepts. It would be very interesting to apply different metrics to measure the distance or similarity between two text chains, that is, to measure the differences between them. This type of metrics is widely used in the solution of diverse problems, especially in word processing or natural language. The most known metrics that can be considered in future work is the "Hamming Distance" [54] and the "Levenshtein Distance" [55].

Secondly, the item 5, regarding the execution process, focuses on the execution of structures made of concrete because it is one of the activities that presents a higher accident rate. However, the proposal can be easily extended in order to cover the entire execution process. To do this, the panel of experts of the construction domain should decide the risks, measures and IPE that are essential for each construction process. Additionally, the specific vocabulary of these activities should be reviewed in order to include it in the thesaurus so as to achieve a better performance of the methodology.

Concretely, in the item 4 is where little lower results are obtained due to the characteristics of this item. The person in charge to write the S&H Plan is allowed much freedom when explaining these items and, they depend on the characteristics of the construction process. Finding similar text strings in this context is not an easy task, as mentioned before. For this purpose, we propose for future work, the use of Ordered Weighted Averaging (OWA) operator, in-

troduced by Ronald R. Yager [56] that provides a parameterized aggregation. This operator should allow to assign weights to the words in the search phrase, depending on their importance to the meaning of the phrase. Based on the words that appear in the analyzed plan, a level of coincidence is established, which will be greater if the most relevant terms are found. A threshold might be established, above which, the search phrase would be considered identified in the S&H Plan.

7. Conclusions

Construction is one of the most hazardous industries, resulting in a huge number of work accidents. For this purpose, safety management is one of the priorities during the execution of construction projects. According to the current legislation, during the execution phase, each contractor must create a S&H Plan which sets out the provisions for guaranteeing safety and health for everyone on the construction site and all others who may be affected. In most cases, this is a general document that does not specify the characteristics of a given construction site. In addition, it should be a proactive, living and dynamic document that is updated whenever any modifications are made during the construction process. Because of this, it is sometimes ignored and becomes a mere documentary requirement, making it difficult to manage and not very practical in the real world.

In this paper, we have thus proposed a methodology-based approach that promises to be very beneficial, especially for the labor authority and health and safety coordinators in terms of management safety in construction projects. Our approach facilitates the verification process and detects at an early stage whether the document meets specific requirements through automatic content analysis. As we have mentioned, this is not a trivial task given that the S&H Plans are unstructured and contain natural language since there is no defined structure or instructions on how to create them. To verify the content of the S&H Plan, some requirements have been proposed along with a percentage

representing its importance with respect to others. These requirements have been organized into seven items depending on the type of information to which they refer. For each of these items a set of rules has been designed so as to detect variables based on combinations of generic keywords in the text.

To validate our proposal, fifty S&H Plans have been used, twenty-five for learning and the other twenty-five for testing. The learning phase has been very useful to refine the proposal as well as to enhance the thesaurus. On the other hand, the results obtained in the testing phase are very promising and demonstrate that the proposed approach is able to identify the required information in an easy and intuitive way for the coordinator and labor authority. Thanks to the report provided by the proposal, the validation process is faster and more thorough. In this way, the coordinator and the labor authority can quickly identify aspects contained in, or missing from, the S&H Plan so that the constructor can update or correct them before work begins. Therefore, the document will be validated when it contains reliable and coherent information on the construction project. Additionally, this approach can help when reviewing updates and project changes, placing in context the flexible and dynamic nature of the S&H Plan. Finally, we have exposed the weaknesses of the methodology as well as future proposals to improve its performance.

References

[1] S. Zhang, F. Boukamp, J. Teizer, Ontology-based semantic modeling of construction safety knowledge: Towards automated safety planning for job hazard analysis (JHA), Automation in Construction 52 (2015) pp. 29–41. doi:https://doi.org/10.1016/j.autcon.2015.02.005.

[2] M. Gunduz, M. T. Birgonul, M. Ozdemir, Fuzzy structural equation model to assess construction site safety performance, Journal of Construction Engineering and Management 143 (4) (2016) pp. 04016112. doi:10.1061/(ASCE)CO.1943-7862.0001259.

[3] C. Q. Poh, C. U. Ubeynarayana, Y. M. Goh, Safety leading indicators for construction sites: A machine learning approach, Automation in Construction 93 (2018) pp. 375–386. doi:https://doi.org/10.1016/j.autcon.2018.03.022.

[4] M. T. Newaz, P. Davis, M. Jefferies, M. Pillay, Using a psychological contract of safety to predict safety climate on construction sites, Journal of Safety Research 68 (2019) pp. 9–19. doi:https://doi.org/10.1016/j.jsr.2018.10.012.

[5] R. M. Choudhry, D. Fang, S. M. Ahmed, Safety management in construction: Best practices in hong kong, Journal of Professional Issues in Engineering Education and Practice 134 (1) (2008) pp. 20–32. doi:10.1061/(ASCE)1052-3928(2008)134:1(20).

[6] F. Zhang, H. Fleyeh, X. Wang, M. Lu, Construction site accident analysis using text mining and natural language processing techniques, Automation in Construction 99 (2019) pp. 238–248. doi:https://doi.org/10.1016/j.autcon.2018.12.016.

[7] M. Martínez-Rojas, N. Marín, M. A. Vila, The role of information technologies to address data handling in construction project management, Journal of Computing in Civil Engineering 30 (4) (2016) pp. 04015064. doi:10.1061/(ASCE)CP.1943-5487.0000538.

[8] M. Marzouk, M. Enaba, Text analytics to analyze and monitor construction project contract and correspondence, Automation in Construction 98 (2019) pp. 265–274. doi:https://doi.org/10.1016/j.autcon.2018.11.018.

[9] L. Soibelman, J. Wu, C. Caldas, I. Brilakis, K.-Y. Lin, Management and analysis of unstructured construction data types, Advanced Engineering Informatics 22 (1) (2008) pp. 15–27. doi:https://doi.org/10.1016/j.aei.2007.08.011.

[10] C. H. Caldas, L. Soibelman, Automating hierarchical document classification for construction management information systems, Automation in Construction 12 (4) (2003) pp. 395–406. doi:https://doi.org/10.1016/S0926-5805(03)00004-9.

[11] Y. Lu, Y. Li, M. Skibniewski, Z. Wu, R. Wang, Y. Le, Information and communication technology applications in architecture, engineering, and construction organizations: A 15-year review, Journal of Management in Engineering 31 (1) (2015) pp. A4014010. doi:10.1061/(ASCE)ME.1943-5479.0000319.

[12] F. Salguero-Caparrós, M. Pardo-Ferreira, M. Martínez-Rojas, J. Rubio-Romero, Management of legal compliance in occupational health and safety. a literature review, Safety Science 121 (2020) pp. 111–118. doi:https://doi.org/10.1016/j.ssci.2019.08.033.

[13] H. Moon, V. R. Kamat, L. Kang, Grid cell-based algorithm for workspace overlapping analysis considering multiple allocations of construction resources, Journal of Asian Architecture and Building Engineering 13 (2) (2014) pp. 341–348. doi:10.3130/jaabe.13.341.

[14] P. Swuste, "you will only see it, if you understand it" or occupational risk prevention from a management perspective, Human Factors and Ergonomics in Manufacturing & Service Industries 18 (4) (2008) pp. 438–453. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/hfm.20101, doi:10.1002/hfm.20101.

[15] Council of the European Union, Council directive 92/57/EEC of 24 june 1992 on the implementation of minimum safety and health requirements at temporary or mobile construction sites. (Accessed: Jun. 20, 2020).
URL https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A31992L0057

[16] M. Martínez-Rojas, N. Marín, M. A. V. Miranda, An intelligent system for the acquisition and management of information from bill of quantities in

building projects, Expert Systems with Applications 63 (2016) pp. 284–294. doi:https://doi.org/10.1016/j.eswa.2016.07.011.

[17] S. Alsafouri, S. K. Ayer, Review of ICT implementations for facilitating information flow between virtual models and construction project sites, Automation in Construction 86 (2018) pp. 176–189. doi:https://doi.org/10.1016/j.autcon.2017.10.005.

[18] W. Zhou, J. Whyte, R. Sacks, Construction safety and digital design: A review, Automation in Construction 22 (2012) pp. 102–111. doi:https://doi.org/10.1016/j.autcon.2011.07.005.

[19] X. Li, W. Yi, H.-L. Chi, X. Wang, A. P. Chan, A critical review of virtual and augmented reality (VR/AR) applications in construction safety, Automation in Construction 86 (2018) pp. 150–162. doi:https://doi.org/10.1016/j.autcon.2017.11.003.

[20] Y. Shi, J. Du, C. R. Ahn, E. Ragan, Impact assessment of reinforced learning methods on construction workers' fall risk behavior using virtual reality, Automation in Construction 104 (2019) pp. 197–214. doi:https://doi.org/10.1016/j.autcon.2019.04.015.

[21] V. Bansal, Application of geographic information systems in construction safety planning, International Journal of Project Management 29 (1) (2011) pp. 66–77. doi:https://doi.org/10.1016/j.ijproman.2010.01.007.

[22] S. Choe, F. Leite, Construction safety planning: Site-specific temporal and spatial information integration, Automation in Construction 84 (2017) pp. 335–344. doi:https://doi.org/10.1016/j.autcon.2017.09.007.

[23] M. Li, H. Yu, P. Liu, An automated safety risk recognition mechanism for underground construction at the pre-construction stage based on BIM, Automation in Construction 91 (2018) pp. 284–292. doi:https://doi.org/10.1016/j.autcon.2018.03.013.

[24] M. D. Martínez-Aires, M. López-Alonso, M. Martínez-Rojas, Building information modeling and safety management: A systematic review, Safety Science 101 (2018) pp. 11–18. doi:https://doi.org/10.1016/j.ssci.2017.08.015.

[25] I. Awolusi, E. D. Marks, Active work zone safety: Preventing accidents using intrusion sensing technologies, Frontiers in Built Environment 5 (2019) pp. 21. doi:10.3389/fbuil.2019.00021.

[26] J. M. G. de Gabriel, J. A. Fernández-Madrigal, A. López-Arquillos, J. C. Rubio-Romero, Monitoring harness use in construction with BLE beacons, Measurement 131 pp. 329–340. doi:https://doi.org/10.1016/j.measurement.2018.07.093.

[27] C. Zhou, L. Ding, Safety barrier warning system for underground construction sites using internet-of-things technologies, Automation in Construction 83 (2017) pp. 372–389. doi:https://doi.org/10.1016/j.autcon.2017.07.005.

[28] M. Martínez-Rojas, N. Marín, C. Molina, M. Vila, Cost analysis in construction projects using fuzzy OLAP cubes, in: 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2015, pp. 1–8. doi:10.1109/FUZZ-IEEE.2015.7338048.

[29] N. M. Ruiz, M. Martínez-Rojas, C. M. Fernández, J. M. Soto-Hidalgo, J. C. Rubio-Romero, M. A. V. Miranda, Flexible management of essential construction tasks using fuzzy OLAP cubes, Emerald. ISBN:9781787438699, 2018. doi:10.1108/978-1-78743-868-220181010.

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

[31] J. M. Soto-Hidalgo, J. M. Alonso, G. Acampora, J. Alcalá-Fdez, JFML: A java library to design fuzzy logic systems according to the IEEE std 1855-2016, IEEE Access 6 (2018) pp. 54952–54964. doi:10.1109/ACCESS.2018.2872777.

[32] E. Petrova, P. Pauwels, K. Svidt, R. L. Jensen, In search of sustainable design patterns: Combining data mining and semantic data modelling on disparate building data, in: I. Mutis, T. Hartmann (Eds.), Advances in Informatics and Computing in Civil and Construction Engineering, Springer International Publishing, 2019, pp. 19–26. doi:https://doi.org/10.1007/978-3-030-00220-6_3.

[33] A. López-Arquillos, J. C. Rubio-Romero, A. Gibb, Accident data study of concrete construction companies' similarities and differences between qualified and non-qualified workers in spain, International Journal of Occupational Safety and Ergonomics 21 (4) (2015) pp. 486–492. doi:10.1080/10803548.2015.1085750.

[34] G. Carter, S. D. Smith, Safety hazard identification on construction projects, Journal of Construction Engineering and Management 132 (2) (2006) pp. 197–205. doi:10.1061/(ASCE)0733-9364(2006)132:2(197).

[35] T. Baxendale, O. Jones, Construction design and management safety regulations in practice—progress on implementation, International Journal of Project Management 18 (1) (2000) pp. 33–40. doi:https://doi.org/10.1016/S0263-7863(98)00066-0.

[36] M. Martínez-Rojas, N. Marín, M. A. Vila, A preliminary approach to classify work descriptions in construction projects, in: 2013 Joint International Fuzzy Systems Association World Congress and North American Fuzzy Information Processing Society Annual Meeting (IFSA/NAFIPS), 2013, pp. 1090–1095. doi:10.1109/IFSA-NAFIPS.2013.6608552.

[37] W. der Yu, J. yang Hsu, Content-based text mining technique for retrieval

of cad documents, Automation in Construction 31 (2013) pp. 65–74. doi:https://doi.org/10.1016/j.autcon.2012.11.037.

[38] M. Martínez-Rojas, J. M. Soto-Hidalgo, N. Marín, M. A. Vila, Using classification techniques for assigning work descriptions to task groups on the basis of construction vocabulary, Computer-Aided Civil and Infrastructure Engineering 33 (11) (2018) pp. 966–981. doi:https://doi.org/10.1111/mice.12382.

[39] T. P. Williams, J. Gong, Predicting construction cost overruns using text mining, numerical data and ensemble classifiers, Automation in Construction 43 (2014) pp. 23–29. doi:https://doi.org/10.1016/j.autcon.2014.02.014.

[40] H.-H. Wang, F. Boukamp, Ontology-based representation and reasoning framework for supporting job hazard analysis, Journal of Computing in Civil Engineering 25 (6) (2011) pp. 442–456. doi:10.1061/(ASCE)CP.1943-5487.0000125.

[41] H. Fan, H. Li, Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques, Automation in Construction 34 (2013) pp. 85–91. doi:https://doi.org/10.1016/j.autcon.2012.10.014.

[42] N.-W. Chi, K.-Y. Lin, S.-H. Hsieh, Using ontology-based text classification to assist job hazard analysis, Advanced Engineering Informatics 28 (4) (2014) pp. 381–394. doi:https://doi.org/10.1016/j.aei.2014.05.001.

[43] A. J.-P. Tixier, M. R. Hallowell, B. Rajagopalan, D. Bowman, Construction safety clash detection: Identifying safety incompatibilities among fundamental attributes using data mining, Automation in Construction 74 (2017) pp. 39–54. doi:https://doi.org/10.1016/j.autcon.2016.11.001.

[44] G. G. Chowdhury, Natural language processing, Annual Review of Information Science and Technology 37 (1) (2003) pp. 51–89. doi:10.1002/aris.1440370103.

[45] C. D. Manning, H. Schütze, Foundations of statistical natural language processing, MIT Press. ISBN:0262133601, 1999.

[46] M. A. Qady, A. Kandil, Automatic clustering of construction project documents based on textual similarity, Automation in Construction 42 (2014) pp. 36–49. doi:https://doi.org/10.1016/j.autcon.2014.02.006.

[47] Y. Zou, A. Kiviniemi, S. W. Jones, Retrieving similar cases for construction project risk management using natural language processing techniques, Automation in Construction 80 (2017) pp. 66–76. doi:https://doi.org/10.1016/j.autcon.2017.04.003.

[48] J. C. Rubio Romero, M. d. C. Rubio Gámez, Manual de coordinación de seguridad y salud en las obras de construcción (Safety and health coordination manual in construction works), Ediciones Díaz de Santos. ISBN:8479786752, 2000.

[49] INSHT, Guía Técnica para la evaluación de prevención de los riesgos relativos a las obras de construcción (Technical Guide for the prevention evaluation of risks related to construction works), Tech. rep. (Accessed: Jun. 20, 2020).
URL https://www.insst.es/

[50] Konstanz Information Miner (KNIME) (Accessed: Jun. 20, 2020).
URL https://www.knime.com/

[51] K. Thiel, The KNIME Text Processing Plugin (Accessed: Jun. 20, 2020).
URL https://www.knime.com/sites/default/files/KNIME-TextProcessing-HowTo.pdf

[52] nlp.stanford.edu, Introduction to information retrieval, dropping common terms:stop words. (Accessed: Jun. 2, 2020).

URL https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html

[53] nlp.stanford.edu, Introduction to information retrieval, tokenization (Accessed: Jun. 2, 2020).
URL https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html

[54] R. W. Hamming, Error detecting and error correcting codes, The Bell System Technical Journal 29 (2) (1950) pp. 147–160. doi:https://10.1002/j.1538-7305.1950.tb00463.x.

[55] L. Yujian, L. Bo, A normalized levenshtein distance metric, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (6) (2007) pp. 1091–1095. doi:10.1109/TPAMI.2007.1078.

[56] R. R. Yager, On ordered weighted averaging aggregation operators in multicriteria decisionmaking, IEEE Transactions on Systems, Man, and Cybernetics 18 (1) (1988) pp. 183–190. doi:10.1109/21.87068.