ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/356068248

UNDERSTANDING CONSTRUCTION SITE SAFETY HAZARDS THROUGH OPEN DATA: TEXT MINING APPROACH

Article · October 2021

DOI: 10.11113/aej.v11.17871

CITATIONS
0

READS
102

2 authors:

Heshani Rupasinghe
Sirindhorn International Institute of Technology (SIIT)
3 PUBLICATIONS 2 CITATIONS
SEE PROFILE

Kriengsak Panuwatwanich
Sirindhorn International Institute of Technology, Thammasat University, Thailand
128 PUBLICATIONS 2,319 CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Enhancing innovation in the Australian Public Service: Strategy to foster high performance engineering workforce View project

Project Mixed Reality Applications in the Construction Industry View project

All content following this page was uploaded by Kriengsak Panuwatwanich on 09 November 2021.

The user has requested enhancement of the downloaded file.

UNDERSTANDING CONSTRUCTION SITE SAFETY HAZARDS THROUGH OPEN DATA: TEXT MINING APPROACH

Neththi Kumara Appuhamilage Heshani Rupasinghe[1] and Kriengsak Panuwatwanich[2]

[1]Department of Civil Engineering, Faculty of Engineering, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka, Tel: +94771126059, e-mail: heshani.r@sliit.lk, heshanirupasinghe@gmail.com

[2]School of Civil Engineering and Technology, Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand, Tel: +6629869009 Ext: 1909, e-mail: kriengsak@siit.tu.ac.th

Received Date: June 29, 2020; Revised Date: April 26, 2021; Acceptance Date: September 15, 2021

Abstract

Construction is an industry well known for its very high rate of injuries and accidents around the world. Even though many researchers are engaged in analysing the risks of this industry using various techniques, construction accidents still require much attention in safety science. According to existing literature, it has been found that hazards related to workers, technology, natural factors, surrounding activities and organisational factors are primary causes of accidents. Yet, there has been limited research aimed to ascertain the extent of these hazards based on the actual reported accidents. Therefore, the study presented in this paper was conducted with the purpose of devising an approach to extract sources of hazards from publicly available injury reports by using Text Mining (TM) and Natural Language Processing (NLP) techniques. This paper presents a methodology to develop a rule-based extraction tool by providing full details of lexicon building, devising extraction rules and the iterative process of testing and validation. In addition, the developed rule-based classifier was compared with, and found to outperform, the existing statistical classifiers such as Support Vector Machine (SVM), Kernel SVM, K-nearest neighbours, Naïve Bayesian classifier and Random Forest classifier. The finding using the developed tool identified the worker factor as the highest contributor to construction site accidents followed by technological factor, surrounding activities, organisational factor, and natural factor (1%). The developed tool could be used to quickly extract the sources of hazards by converting largely available unstructured digital accident data to structured attributes allowing better data-driven safety management.

Keywords: Construction, Hazards, Natural language processing, Safety, Text mining

Introduction

Despite the technological improvements and vastly available safety management techniques, occupational accidents have become a vital phenomenon in construction [1], making the industry one of the most hazardous industries [2]. Therefore, construction health and safety related researchers have a vested interest in investigating effective health and safety management systems [3], constructing frameworks and models on site safety climate [4, 5] as well as predicting safety performance [6, 7]. According to the International Labour Organisation (ILO) [8], over 2.3 million fatalities occur due to work related accidents or ill health around the world in every year. Further, there have been more than 374 million victims of non-fatal workplace injuries annually [9]. However, the factors which influence the construction safety cannot be completely eliminated for both technical and economic reasons [10]. Therefore, construction site safety has been constantly scrutinised by researchers [11-13].

Conventionally in construction safety research, data collection is mostly conducted using questionnaire surveys, semi structured interviews and structured surveys which are criticised as unreliable [14]. This is because the survey questionnaires are often filled with inaccurate answers, skipped questions, interpretation issues, lack of nuance, and accessibility issues. The interviews also rely on the respondent's ability to accurately and honestly answer the questions without bias or fear of legal implications [15]. Hence, root causes of the occupational injuries may not be reliably alarmed whenever the organisational factors such as management and safety issues forefront the accidents.

With the adoption of new technologies such as robotics, sensory machines and ICT in construction [16], research studies have found means to automate construction planning [17], enhance the productivity and quality [18], material handling [19], defect identification in concrete [20], and building information modelling (BIM) [20-24]. In recent years, construction automation has been effectively brought out through Artificial Intelligence (AI). AI has been employed for planning [25], safety management [26, 27], construction management [28], automatic analysis of injury reports [29], automatic clustering of construction project documents based on textual similarity [30] and retrieval of Computer Aided Design (CAD) drawings [31]. Text mining (TM) is a widespread AI technology that uses Natural Language Processing (NLP) to transform unstructured documents into normalised, structured data for information retrieval, data mining, machine learning, statistics, and computational linguistics [32]. Therefore, it has been used recently by construction-related research to analyse construction management processes, including safety management. There are many topics which show that the TM-based classifiers have an immense potential for future project improvement, avoiding mistakes, and making aware of previously unknown facts [4].

With an increased awareness of open data, construction industries around the world have started to make construction accident reports publicly available [33]. A number of research studies have utilised Support Vector Machine (SVM) [34], Random Forest (RF) [35], Naïve Bayesian (NB) Classifier [36], K Nearest Neighbours (k-NN) [37, 38], Decision Tree (DT), Vector Space Model (VSM) [39], Non-negative matrix factorization (NNMF) based classifier [40], and Neural Networks (NN) [41, 42] for automatic classification of accident data. Moreover, Symbiotic Gated Recurrent Unit (SGRU) model has been created to address the drawbacks of Gated Recurrent Unit (GRU) using the accident data from Occupational Safety and Health Administration (OSHA), Department of Labour, USA during 1998 and 2016, which extracted accident causing events such as traffic, collapse of object, fall in, caught in between objects, struck by moving objects, exposure to chemical substances, fires and explosion, electrocution, struck by falling objects, and exposure to extreme temperatures [43]. The Sequential Quadratic Programming (SQP) algorithm is utilised to optimise the weight of each classifier involved in ensembled model and rule based chunking [33], using OSHA accident data to extract the accident-causing events as similar to above. RF Classifier has been utilised to predict the type of construction accidents in Korea [35]. The falls from heights, collision by objects, rollover, and falling objects are the main accidents causing factors that have been identified [35].

Although the adoption of TM techniques has become more popular, the utilisation of open data to identify the nature of the factors leading to hazards causing construction accidents is limited. This study thus focuses on applying an automatic extraction method by applying TM and NLP techniques for classification of final narrative data into sources of hazards, utilising OSHA accident data. Additionally, the main objectives of the study presented in this paper are to identify the source of the accident-causing events (sources of hazards) through TM and to evaluate the performance of existing classifiers compared to the model presented in the paper.

Theoretical Background

Construction Safety Hazards

In occupational health and safety, the term ‘hazard’ possesses several definitions. Health and safety commission defines hazard as ‘potential cause for harm’ while the International Labour Organisation [8] defines hazard as the inherent potential to cause injury or damage people’s health. However, the existence of hazard in construction workplace is an outturn of the actions such as planning and preparation, site environment, project management and safety culture [44].

Sources of hazard were defined as a condition or a factor which would lead to a hazard in the construction site environment. According to the demonstration of accident causation model [44], accidents are appeared from failures of the worker interactions with their work place, material and equipment [45, 46]. Thus, accident root causes can be identified as originating influences (client requirements, economical constrains and construction education), shaping factors (design and specification), worker factors (errors and violations by worker), site factors (layout, lighting, cite constraints, scheduling and housekeeping), material and equipment factors [44], distal factors (project conditions and management decisions), proximal factors (inappropriate site conditions or actions) [47], work technology (tool, material and actions required to perform a specific task), physical conditions (working environment), surrounding activities (such as falling objects from upper elevations and vibrations due to piling), human factors (all environmental, organisational and human characteristics which affects health and safety) [48], organisational factors (poor safety and management) [49, 50] and natural factors (natural processes or phenomenon) [51]. In particular, existing literature discusses more on the human factor involvement towards occupational accidents. The Health and Safety Executive (HSE) [52] defines human factors as ‘environmental, organisational and job factors and human and individual characteristics which influence behaviour at work in a way which can affect health and safety’. Human factors also concern the interaction between people, their characteristics, abilities, organisation, management and technology [53]. Therefore, in this study, the human factors were divided into worker factors, technological factors, organisational factors and surrounding activity. Thus, it can be concluded that above factors are mainly derived from identified sources of hazard factors shown in Table 1.

Text Analysis and Natural Language Processing (NLP)

Text analysis is a tool/automated process employed to extract and classify information from a document in textual format such as emails, customer review reports, survey responses, tweets etc. NLP is a branch of AI which allows computer to understand natural language. Moreover, these alters the textual data into numerical data which can be further utilised in data mining algorithms [54] such as k-NN, SVM, NB algorithm, k-means, etc. Also, it requires certain platforms, libraries and packages for data processing. Recently, research studies have adopted deep learning-based NLP models to extract information. The Hierarchical Attention Network (HAN) is used to separately encode the sentences to identify the important words in each sentence [55]. It was proposed with two basic insights on the document structure where words form sentences and sentences form documents [56]. Thus, both sentence level and word level attention models were developed. However, such algorithms perform poorly when the positive sample for training is limited for each category [57]. For such instances, hand-coded rules and keyword dictionaries were used to integrate human judgment and knowledge into the TM system by increasing the accuracy [29].

Table 1. Categories of Sources of Hazards

Source of Hazard	Description
Worker factor	Every possible error, violation, mental and health issue, lack of skills and behaviour of the workers inside the site environment [49, 58-60].
Technological factor	Design, methodology, tool, material, machinery or equipment breakdown and technical faults and errors occur while utilising tools and machinery in work place [32, 48, 61].
Natural factor	Any natural phenomenon which negatively affects the site condition and cause harm [51, 62].
Organisational factor	Project conditions, management decisions, health and safety issues and controlling which are beyond the level of workers [63].
Surrounding activity factor	Activities in progress inside or outside the site environment other than the activity in which the victim is engaged [48].

Research Methodology

The methodology adopted in this paper consists of six main steps. They are data collection, selection of extraction tool, lexicon building, composing extraction rules, data pre-processing and model validation. Figure 1 illustrates these steps and the following subsections provide detailed illustration of the methodology.

Data Collection

Data has been downloaded from OSHA, Department of Labour, USA [64]. There were 50,032 accident data records including both construction and non-construction from January 2015 to September 2019. This study focuses only on construction accident data which has been extracted from the dataset using primary North American Industry Classification System (NAICS) and Microsoft Excel Macro. According to the NAICS, an industry code beginning with ‘23’ belongs to the construction industry and 8,940 such injury reports were obtained.

Selection of Extraction Tool

In the process of selecting an extraction tool, existing data mining algorithms plays a major role. However, the statistical classifiers required thousands of data records to achieve a significant accuracy. Thus, this study developed a rule-based classification tool using initial 1,500 data records from the original data set. Thus, classifiers such as DT, RF, SVM, k-NN and NB were trained manually only for the initial 1,500 data records from the dataset and its performance were evaluated using F1 score (see model validation part for more details) to justify the selection of rule-based classifier. The results of F1 score showed that SVM has the highest F1 score of 0.68 while DT, RF, NB, kNN and kernel SVM has F1 scores of 0.62, 0.62, 0.33, 0.32 and 0.28, respectively. This low performance is mainly due the scarcity of the training sample and linguistic diversity of the accident data reports. Attributes related to each source of hazards co-occur only a limited number of times resulting the poor performance of the classifier. As the main concern of this study is to develop a solution for disinclination of manual analysis of large unstructured data reports, rule-based extraction tool was selected without further manual training.

Rule based extraction tool is a classifier which allows to classify data into any classification scheme depending on the ranking of the rules written using IF, THEN and ELSE IF clauses. The developing tool utilised in this study was Spyder 3.2.3 which promotes and facilitates the use of Python 3.5 programming language for scientific and engineering software development. Anaconda Navigator 1.3.1 created by Continuum Analytics was used as the desktop graphical user interface which supports to launch Spyder and easily manage Conda packages (a compressed '.conda' file that consists of Python packages, modules, meta data and system level libraries), environments and channels without the use of command line prompt. Moreover, TM was carried out on a computer running macOS Catalina version 10.15.3 with 2.3 GHz intel dual-core i5 processor with 8GB memory.

Lexicon Building

Lexicon is normally a vocabulary of a language or subject. In this study, a word lexicon refers to a stock of words related to construction accidents. In building the lexicon for this study, 1,500 accident records were arbitrarily selected, representing about 17% of the entire dataset (8,940 records). Previous studies such as Desvignes [65] used a similar proportion of 1,280 out of 7,000 (around 18%). The 1,500 records were used as a reference to initially create the lexicon by extracting key phrases which describe the cause of accident (see examples in Table 2). The lexicon was further enriched using online resources, books and common terms related to construction accidents.

For the first incident report given as example in Table 2, the source of hazard was identified as natural factors due to the key phrase in the sentence describing the accident was the word phrase 'gust of wind displaced a tree'. In the second example, accident occurred as the worker 'jumped off' the trailer. This action is performed by the worker and thus, it is categorised under worker factor. In the third example, accident was occurred as the motor vehicle driven in the closed road got stuck on the worker. Thus, the root cause of the accident is a surrounding activity which is beyond the control of the worker. The root cause of the accident in the fourth example is poor safety management of the site. Accident was occurred as the fall protection system was not used. Hence, it can be categorised under organisational factor.

In the final example, accident occurred as the jib attachment slipped off. It is a fault of the equipment used and can thus be categorised under technological factor. In the fifth example, however, final narrative description does not include any description that implies the root cause of the accident. Only the accident itself was mentioned. Such cases were named as 'Null'. For example, there were records such as 'Finger cut accident', 'Employee broke a leg at a construction yard', Employee's right thumb was crushed', 'Employee caught hand in system', 'Employee's lower left arm was amputated' etc. Therefore, such data reports were neglected while building the lexicon.

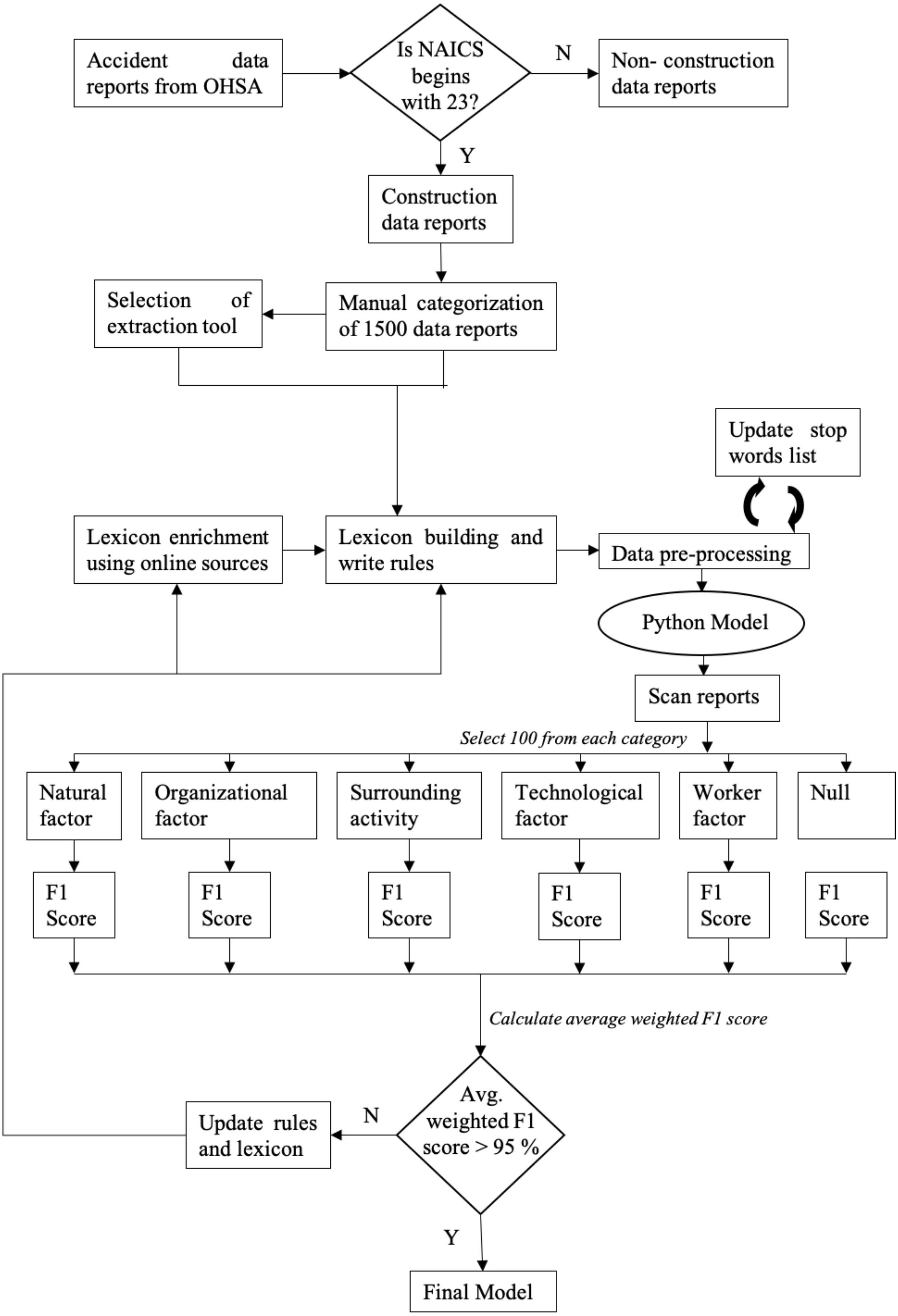
Accident data reports from OHSA
Is NAICS begins with 23?
N
Non- construction data reports
Y
Construction data reports
Manual categorization of 1500 data reports
Selection of extraction tool
Update stop words list
Lexicon enrichment using online sources
Lexicon building and write rules
Data pre-processing
Python Model
Scan reports
Select 100 from each category
Natural factor
Organizational factor
Surrounding activity
Technological factor
Worker factor
Null
F1 Score
F1 Score
F1 Score
F1 Score
F1 Score
F1 Score
Calculate average weighted F1 score
Update rules and lexicon
N
Avg. weighted F1 score > 95 %
Y
Final Model

Figure 1. Development of extraction tool and validation

Table 2. Example of Extracted Key Phrases for Lexicon Building

Final Narrative	Phrase for Lexicon Building	Source of Hazard
Example 1: A worker was installing a retaining wall next to Highway 285 when a gust of wind displaced a tree next to the highway. The tree struck the worker, resulting in back, pelvis, and ankle injuries.	gust of wind displaced a tree	Natural factor
Example 2: Employee was loading plastic pipes onto a utility trailer. He jumped off the trailer and broke his ankle.	jumped off	Worker factor
Example 3: At approximately 4:30 PM on January 28, 2015, an employee removing cones in the 'closed lane' was struck by a motor vehicle driven by a member of the general public. The employee's left leg sustained a fractured femur and fibula.	vehicle driven by a member of the general public	Surrounding activity
Example 4: Employee fell approximately 15 feet. No fall protection was being used at the time.	No fall protection	Organisational factor
Example 5: Employee partially amputated index finger with sledge hammer.	Null	Null
Example 6: Employee was transferring a mini dumpster into a larger dumpster when the jib attachment slipped off, hit the ground, bounced, and hit the employee.	jib attachment slipped off	Technological factor

Identification of N-grams

N-gram is a series of n number words taken from a given document or speech. These N-grams can be a letter, word, syllables or base pairs according the administration of the term [66].

In this study, N-grams were identified as unigrams, bigrams and trigrams with the help of identified key phrases as presented in Table 3. These lists of N-grams were obtained by manually analysing the reports which were extracted using the approach presented in Table 2. These lists consist of only contributing words to extract the cause of accident. Further, N-grams were enriched by identifying probable common words related to each category for future anticipated cases. For instance, for the source of hazard natural factors, unigrams such as 'thunderstorm', 'landslide', 'tornado' and 'tsunami' was added even though they were not found in first observed reports.

Table 3. Examples for N-Grams

Source of Hazard	Unigram	Bigram	Trigram
Natural factor	gust, cyclone, earthquake, flood, tsunami	Wind blew, ground collapse, heat wave, strong wind, wet weather	Gust wind caused, ice/slick roads, strong gust wind, wind displaced tree, violent thunderstorm caused
Organisational factor	Bitten, stung, unguarded, unhooked, wasp	Floor opening, heat exhaustion, not worn, bacterial infection, management issues	guardrail not placed, no fall protection, not wearing seatbelts, high work pressure, fell through hole
Surrounding activity factor	Dog, motorcycle, robbed, robbed, struck, automobile	dog chased, general public, privately driven, remote control, motor vehicle	struck by car, backhoe ran over, fell from upper, unknown object fell, foreign body injected
Technological factor	Failed, malfunction, loose, collapsed, dislodged	blew apart, blade broke, lost power, cord snapped, became unstable	ladder slid away, broke from choke, carbon monoxide poisoning, chain came off, dolly tripped over
Worker factor	Attempted, forgot, slipped, fainted, dizziness	began cramping, accidentally stepped, jumped off, employee jumped, experience cramping	lost his balance, unseen by employees, fist fight another, he became overheated, he let off

Composing Extraction Rules

Table 4 demonstrates the comprehensive list of extraction rules according to the order of implementation and every rule was written as a combination of statements using the ‘if else if’ function in Python.

According to the manually extracted data set, attempting to perform work without proper knowledge is a popular cause of accidents and this always led to worker factor. Therefore, before searching for any other unigram, the word ‘attempt’ was initially searched in final narrative unigram set, and that word was categorised under worker factor. However, the word ‘attempt’ can be either written exactly as ‘attempt’ or can be written as ‘attempting’, ‘attempted’ etc. These types of suffixes and prefixes were removed while creating N-grams lists. The removal of these terms along with the conversion of unstructured data into structured data is discussed below in data pre-processing.

After categorising all the records containing the word ‘attempt’ under worker factors, bigrams of natural factors were checked. This was done to remove the conflicts which are likely to arouse by having N-grams that can fall into two or more categories. As an example, final narrative ‘As the employee was holding the window unit, a gust of wind blew and the employee lost his grip on it’ had bigrams of [‘wind’, ‘blew’] which fall into natural factor and [‘lost’, ‘his’, ‘grip’] which fall into worker factor. However, it is clear that the worker lost his grip due to the gust of wind and root cause for the accident is the gust of wind. By checking the bigrams of the natural factors as the second step, root cause of the accident was clearly identified and categorised under the most accurate sources of hazard.

Then, in the third and fourth step [‘he’, ‘slip’] and [‘employee’, ‘slip’] was checked with the bigrams of final narrative. This can be explained using the following example. ‘The injured employee slipped and fell in front of the other employee who was operating a Lull at the time. The Lull ran over the injured employee's foot’. Here, the accident was occurred as the Lull ran over the employee. This may fall into surrounding activity if trigrams were checked before checking ‘employee’, ‘slip’ in bigram. The accident occurred as the employee slipped on. Therefore, this way of ordering rules helps the program to categorise these final narratives under worker factor.

After that, trigrams, bigrams and unigrams were checked respectively for each category. However, in each category, worker factors were checked at last to minimise the co-occurrences as discussed above. In the final step, all the final narratives which do not fall into any other category were named as null. This category includes the final narratives which do not state the cause of the accident directly.

Table 4. Comprehensive List of Extraction Rules

Statement Returns True if... (elif)	**Description**
('attempt') in unigram	Checks whether the word ‘attempt’ is in the unigram list of final narrative data (unigram)
any(check in bigram for check in bigramnf)	Checks whether any of the words in final narrative bigrams (bigram) are presented in bigrams of natural factors (bigramnf)
('he', 'slip') in bigram	Checks whether the words ‘he’, ’slip’ are in the bigram list of final narrative data
('employe', 'slip') in bigram	Checks whether the words ‘employe’, ’slip’ are in the bigram list of final narrative data
any(check in trigram for check in trigramnf)	Checks whether any of the words in final narrative trigrams (trigram) are presented in trigrams of natural factors (trigramnf)
any(check in trigram for check in trigramof)	Checks whether any of the words in final narrative trigrams (trigram) are presented in trigrams of organisational factors (trigramof)
any(check in trigram for check in trigramsa)	Checks whether any of the words in final narrative trigrams (trigram) are presented in trigrams of surrounding activity (trigramsa)
any(check in trigram for check in trigramtf)	Checks whether any of the words in final narrative trigrams (trigram) are presented in trigrams of technological factors (trigramtf)
any(check in trigram for check in trigramwf)	Checks whether any of the words in final narrative trigrams (trigram) are presented in trigrams of worker factors (trigramwf)

any(check in bigram for check in bigramof)	Checks whether any of the words in final narrative bigrams (bigram) are presented in bigrams of organisational factors (bigramof)
any(check in bigram for check in bigramsa)	Checks whether any of the words in final narrative bigrams (bigram) are presented in bigrams of surrounding activity (bigramsa)
any(check in bigram for check in bigramtf)	Checks whether any of the words in final narrative bigrams (bigram) are presented in bigrams of technological factors (bigramtf)
any(check in bigram for check in bigramwf)	Checks whether any of the words in final narrative bigrams (bigram) are presented in bigrams of worker factors (bigramtf)
any(check in unigram for check in unigramtf)	Checks whether any of the words in final narrative unigrams (unigram) are presented in unigrams of technological factors (unigramtf)
any(check in unigram for check in unigramof)	Checks whether any of the words in final narrative unigrams (unigram) are presented in unigrams of organisational factors (unigramof)
any(check in unigram for check in unigramsa)	Checks whether any of the words in final narrative unigrams (unigram) are presented in unigrams of surrounding activity (unigramsa)
any(check in unigram for check in unigramwf)	Checks whether any of the words in final narrative unigrams (unigram) are presented in unigrams of worker factors (unigramwf)
any(check in unigram for check in unigramnf)	Checks whether any of the words in final narrative unigrams (unigram) are presented in unigrams of natural factors (unigramnf)
else	Final narratives which do not fall into any of above

Data Pre-processing

Data pre-processing consists of five main steps to improve the quality of the unstructured raw data before performing any TM task and prepare raw data for further processing.

Punctuation Removal

This step includes removing all the characters except alphabetical characters. Hence, word complexity due to ‘Employee1’ and ‘Employee2’ was eliminated and treated as one word ‘Employee’. However, black spaces among words are kept as it helps the computer to identify one word from another.

Uppercase to Lowercase

All the uppercase letters are then converted into lower case letters. After the transformation, as an example ‘Machine’ and ‘machine’ is treated as one word ‘machine’.

Tokenization

The document is broken down into words to create a token for each word. For example, after the tokenization, the sentence ‘employee was fallen down’ will turn in to array of words [‘employee’, ‘was’, ‘fallen’, ‘down’].

Stop Word Removal

These are the most common words that exist in a sentence which adds low value to the meaning of the sentence in TM [67]. Generally, stop words are determined by the frequency of words appearing in the document and then by filtering the most frequent terms such as ‘a’, ‘an’, ‘is’, ‘the’, ‘not’, ‘didn’t’, ‘and’, ‘be’ and so on. However, to maintain the sense of the N-grams, some of these stop words were removed from the stop word list and some additional words were added to the stop word list. For instance, an accident occurred due to ‘not having fall protection system’ will be considered as an organisational factor. This will not be accurately identified if the stop word list consists of ‘not’. Therefore, word ‘not’ was removed. Likewise, words such as ‘not’, ‘didn’t’, ‘wasn’t’, ‘it’, ‘through’, ‘lower’, ‘volt’, ‘degree’, ‘himself’, ‘herself’, and ‘themselves’ were removed from the original stop word list. Moreover, to reduce the complexity of the meaning of the sentence, words such as ‘ft’, ‘feet’, ‘inches’ etc. were added to the stop words list.

Stemming and Lemmatization

In this process, a word family such as (‘collect’, ‘collectively’, ‘collection’) based on ‘collect’ can be expressed as the single word ‘collect’. Also, word complexity due to their plurality and singularity was eliminated.

Append

In the final data pre-processing step, usually all the documents are appended in to a corpus. However, in this study, each sentence was appended to lists of N-grams and separate lists of unigrams, bigrams and trigrams were created.

Model Validation

Validation process was carried out each and every time after a classification of sources of hazards by using the output written to a Microsoft Excel file. A random number was then assigned to each injury report and separated into different Excel sheets according to the source of hazard. This iterative process is shown in the latter part of Figure 1. However, to eliminate the display of manually trained data in the beginning, these outputs were then sorted according to the random number. Finally, the first 100 data reports were examined carefully in each category and accuracy is calculated through F1 score in Equation 1 which is a measure of test accuracy of a classification. These were recorded at each repetition as shown in Table 5. This validation process played a crucial role in establishing a decidedly accurate system. After careful examination over seven repetitions, model performance was evaluated through counting all attributes for average weighted F1 score as shown in Equation 2. However, it should be noted that imperfections are possible as the manual text analysis was done by the author and the model can be further investigated and fixed by refining the N-gram files accordingly.

$$\text{F1 score} = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})} \tag{1}$$

Precision = TP / (TP + FP),

Recall = TP / (TP + FN)

Where TP = True positive, FP = False positives, FN= False negatives

$$\text{Average weighted F1 score} = \sum_{i=1}^{N} \left(\frac{S_i}{T} \times F1_i \right) \tag{2}$$

Where N is the total number of labels, S_i is the number of true instances in i^{th} label, T is the true predictions of all labels, $F1_i$ is the F1 score of i^{th} label. N in this study was six (6), which is the number of categories. True positives (TP) in this study are the ones which were extracted correctly from the injury data report through the rule-based TM model. False positives (FP) refer to the cases in which an identified category does not correctly represent the true category. False negatives (FN) refer to those cases in which a particular category that is not identified turn out to be the true category.

Results and Discussion

Model Performance

For the validation of the model, the 0.95 threshold value was selected. The summary of the model performance is presented in Table 5. Seven iterations were required to achieve the threshold of 0.95. However, the fourth iteration shows that it has achieved the threshold. Nevertheless, average weighted F1 score testing for the fifth iteration was reduced as the F1 score of surrounding activity of the fifth iteration went below the threshold of 0.95. Therefore, the sixth and seventh iterations were performed.

Table 5. Summary of Model Performance at Each Iteration

Iteration	F1 Score						
	Avg. Weighted	NF	OF	SA	TF	WF	Null
1	0.85	0.59	0.92	0.88	0.89	0.85	0.84
2	0.93	0.75	0.97	0.96	0.95	0.94	0.91
3	0.94	0.86	0.98	0.95	0.96	0.96	0.92
4	0.97	0.97	0.99	0.96	0.99	0.96	0.94
5	0.94	0.97	0.98	0.89	0.97	0.93	0.92
6	0.98	0.97	0.99	0.96	0.99	0.99	0.98
7	0.98	1	0.98	0.95	1	0.99	0.99

Note: NF = Natural factor, OF = Organisational factor, SA = Surrounding activity, TF = Technological factor, WF = Worker factor

The F1 score obtained through rule-based model was compared with the existing statistical classifiers by using the final rule-based model. SVM obtained the highest F1 score of 0.81 while RF, kNN, kernel SVM and NB obtained 0.71, 0.53, 0.47 and 0.28 respectively. These scores comparably better than the most scores achieved by statistical classifiers found in existing literature. For instance, optimised ensembled model achieved only 0.68 F1 score to extract causes of accidents while SVM, DT and NB achieved 0.58, 0.52 and 0.44 respectively [32]. This shows that the rule-based classifier outperforms the other classifiers for the extraction of the sources of hazards from the final narratives taken from the 8,940 datasets.

Implication of Extracted Factors on Construction Site Accidents

The outcome obtained through analysing the data of 8,940 injury reports over four years is presented in Figure 2. It can be seen that the worker factor has the highest contribution to construction site accidents (35%) followed by technological factor (20%), surrounding activities (11%), organisational factor (3%) and natural factor (1%). It should be noted that the 'Null' category is excluded from the interpretation. It is worth reiterating here that this category refers to the cases in which the causes of accidents cannot be determined. Such cases only contain the description of the nature of the accident (e.g., 'Finger cut accident', 'Employee broke a leg at a construction yard', Employee's right thumb was crushed', and 'Employee caught hand in system') without enough information on what may cause such accident.

The significance of the worker factor (35%) and technological factor (20%) identified in this study as the top two contributors of construction accidents can be corroborated by previous research. The study on causes of construction accidents in the UK by Haslam et al., [70] revealed that worker-related hazards and equipment issues are ranked second and third, respectively (after 'lack of risk management'). A study by Feyer et al., [68] also found that equipment failures (i.e. technological factor) was the third most contributing factors.

Being ranked third, accidents occurred due to surrounding activities (11%) are caused by an unexpected struck by something falling from outside the workspace, misconduct or mistake of other employee, blasting activities, surrounding animal attacks, general public activities and remote-control works. Mostly, accidents due to surrounding animal attacks and general public activities are connected with road construction, highway maintenance and traffic controlling activities. It is interesting to note that this factor has not been highlighted much in past research, despite it showing significant contribution in this study.

The organisational factor was found to account for 3% of the construction accidents (ranked fourth). This factor mainly includes unsafe work space (unguarded deep openings), poor usage of PPE, poor site planning and allowing untrained workers to perform work which requires certain training. The above-cited study by Haslam et al., [70] also found workplace issues as the fourth contributing factor to construction accidents. Although not found to contribute highly to accidents in relation to others factors in this study, some activities under the organisational factor have been found as significant causes of construction accidents. For example, Boamah [69] found that poor planning at site, poor usage of PPE and unsafe work conditions represented the top three accident-causing factors.

Natural factor contributed to 1% of the accidents in which all of the accidents occurred as a result of strong wind. Despite the fact that construction safety guidelines usually caution against working in windy conditions, it can be questioned that these accidents may be associated with the misconduct of the workers which eventually leads to worker factor. Moreover, it can be suggested that, allowing workers to engage in construction work in windy situation is poor safety management and high work pressure which can be considered organisational factors. However, there is no enough evidence in the injury reports to make such a conclusion.

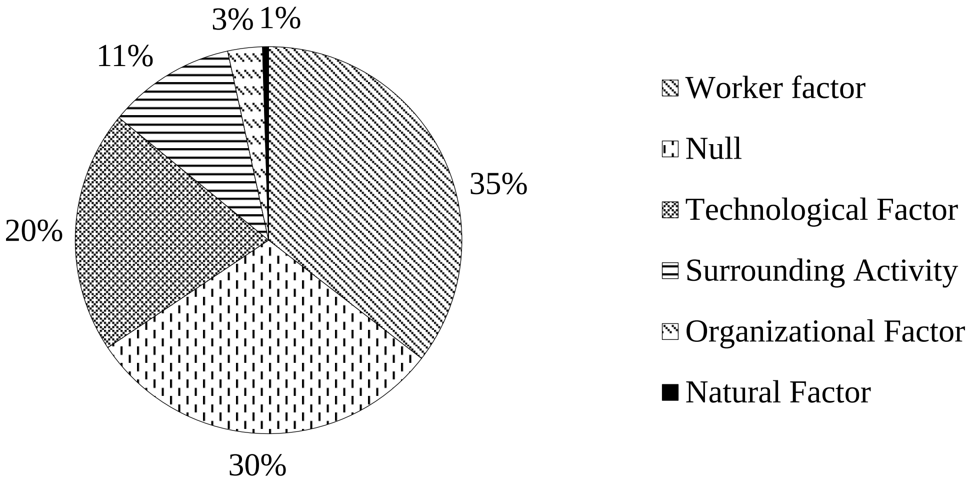

Figure 2. Distribution of extracted sources of hazard

Concluding Remarks

According to the existing literature, construction health and safety has been a widely discussed topic. Thus, recording of construction accidents and analysing the health and safety behaviour is vastly discussed. There has been limited literature focusing on extracting sources of hazard in the construction industry, despite the fact that it could help reveal the root cause of the accident, especially by learning from the accidents that have occurred and reported. This study has developed a rule-based extraction tool which allows user to input textual data from accident reports to obtain the sources of hazards found in the construction industry. In the validation process, the tool achieved 95% accuracy as indicated by the average weighted F1 score, which outperformed other existing statistical classifiers. However, the tool requires basic literacy on Python as the process is not yet fully automated.

Factors which contribute to the accident are the most essential during an accident investigation. Many researchers have addressed various techniques for mitigation of accidents, but analysis of sources of hazards uncovers root causes for accidents which allows better safety management. According to the findings from this study, it shows that 35% of the construction accidents are due to the worker factors, including mistakes, errors, misbehaviour or the behavioural issue related to the workers. Also, 20% of the accidents occurred due to technological factors which include the equipment, methodology and design failures. A proportion of 11% of the construction accidents are due to the surrounding activities which are beyond the control of the victims themselves. Another 3% of the accidents occurred due to the organisational factors such as management decisions, planning and controlling issues and health and safety management issues. The remaining 1% is due to the natural factors. Thus, this tool helps to uncover the basic causes of past accidents, which are beneficial for construction companies to understand systematic and underlying issues that needs addressing for accident prevention. In summary, the contribution made by this research can be summarised as follows:

- Through NLP and the open data provided by OSHA, this research identified the 'high-level' factors representing main sources of construction accident hazards, rather than simply identifying the type of accidents (fall, struck by objects, caught in between, exposure to environmental heat, traffic etc.), which most of the existing literature has already identified utilising the same data source and similar technique. These high-level factors could help companies to understand, from the strategic level, the critical factors that requires attention when it comes to construction site safety improvement.
- The identified hazard factors were derived from a large volume of open data of accident records maintained by OSHA. This is considered more comprehensive compared to previous similar studies attempting to identify causes of construction accidents using case studies or questionnaire survey.
- The methodology presented in this study can be further modified and utilised to extract any other reports in various domains by adjusting the N-gram files accordingly, provided that the N-grams be enriched with relevant words and phrases. Also, to accomplish accuracy threshold, N-gram files should be refined with more vocabulary related to each factor.

References

[1] M. Yılmaz, and R. Kanıt, "A practical tool for estimating compulsory OHS costs of residential building construction projects in Turkey," *Safety Science,* Vol. 101, pp. 326-331, 2018.

[2] R. Sacks, O. Rozenfeld, and Y. Rosenfeld, "Spatial and temporal exposure to safety hazards in construction," *Journal of Construction Engineering and Management,* Vol. 135, No. 8, pp. 726-736, 2009.

[3] Q. Chen, and R. Jin, "Safety4Site commitment to enhance jobsite safety management and performance," *Journal of Construction Engineering and Management,* Vol. 138, No. 4, pp. 509-519, 2012.

[4] R.M. Choudhry, D. Fang, and H. Lingard, "Measuring safety climate of a construction company," *Journal of Construction Engineering and Management,* Vol. 135, No. 9, pp. 890-899, 2009.

[5] Q. Li, C. Ji, J. Yuan, and R. Han, "Developing dimensions and key indicators for the safety climate within China's construction teams: A questionnaire survey on construction sites in Nanjing," *Safety Science,* Vol. 93, pp. 266-276, 2017.

[6] D. Fang, Z. Jiang, M. Zhang, and H. Wang, "An experimental method to study the effect of fatigue on construction workers' safety performance," *Safety Science,* Vol. 73, pp. 80-91, 2015.

[7] N. Xia, P.X. Zou, X. Liu, X. Wang, and R. Zhu, "A hybrid BN-HFACS model for predicting safety performance in construction projects," *Safety Science,* Vol. 101, pp. 332-343, 2018.

[8] International Labor Organization (ILO), *World Statistic,* ILO, Available: https://www.ilo.org/moscow/areas-of-work/occupational-safety-and-health/WCMS_249278/lang--en/index.htm. [Accessed: June 2020]

[9] International Labor Organization (ILO), *Health and Safety at Work,* ILO, Available: https://www.ilo.org/global/topics/safety-and-health-at-work/lang--en/index.htm [Accessed: October 2019]

[10] B. Hoła, "Methodology of hazards identification in construction work course," *Journal of Civil Engineering and Management,* Vol. 16, No. 4, pp. 577-585, 2010.

[11] T.K. Fredericks, O. Abudayyeh, S.D. Choi, M. Wiersma, and M. Charles, "Occupational injuries and fatalities in the roofing contracting industry," *Journal of Construction Engineering and Management,* Vol. 131, No. 11, pp. 1233-1240, 2005.

[12] A.A. Hassanein, and R.S. Hanna, "Safety performance in the Egyptian construction industry," *Journal of Construction Engineering and Management,* Vol. 134, No. 6, pp. 451-455, 2008.

[13] R. Liaudanskiene, N. Varnas, and L. Ustinovichius, "Modelling the application of workplace safety and health act in Lithuanian construction sector," *Technological and Economic Development of Economy,* Vol. 16, No. 2, pp. 233-253, 2010.

[14] S. Debios, Advantages and disadvantages of questionnaires [Blog post], Available: https://surveyanyplace.com/questionnaire-pros-and-cons [Accessed: June 2019]

[15] N.K.A.H. Rupasinghe, and K. Panuwatwanich, "Extraction and analysis of construction hazard factors from open data," In: *IOP Conference Series*, Malayasia, Vol. 849, 2019. doi: 10.1088/1757-899X/849/1/012008

[16] C. Balaguer, and M. Abderrahim, *Robotics and Automation in Construction*, BoD–Books on Demand, 2008.

[17] V. Faghihi, A. Nejat, K.F. Reinschmidt, and J.H. Kang, "Automation in construction scheduling: a review of the literature," *The International Journal of Advanced Manufacturing Technology,* Vol. 81, No. 9-12, pp. 1845-1856, 2015.

[18] S.S. Kamaruddin, M.F. Mohammad, and R. Mahbub, "Barriers and impact of mechanisation and automation in construction to achieve better quality products," *Procedia-Social and Behavioral Sciences,* Vol. 222, pp. 111-120, 2016.

[19] P.O. Alumbugu, W.W. Shakantu, T.A. John, and A.W. Ola-Awo, "Automation in construction materials handling: The case study North Central Nigeria," In: *Proceedings of WABER 2019 Conference,* West Africa Built Environment Research (WABER) Conference, Accra, Ghana, pp. 200-213, 2019.

[20] Z. Liu, Y. Cao, Y. Wang, and W. Wang, "Computer vision-based concrete crack detection using U-net fully convolutional networks," *Automation in Construction,* Vol. 104, pp. 129-139, 2019.

[21] X. Li, G.Q. Shen, P. Wu, and T. Yue, "Integrating building information modeling and prefabrication housing production," *Automation in Construction,* Vol. 100, pp. 46-60, 2019.

[22] S.T. Matarneh, M. Danso-Amoako, S. Al-Bizri, M. Gaterell, and R. Matarneh, "Building information modeling for facilities management: A literature review and future research directions," *Journal of Building Engineering,* Vol. 24, pp. 100-755, 2019.

[23] G.B. Ozturk, "Interoperability in building information modeling for AECO/FM industry," *Automation in Construction,* Vol. 113, pp. 103-122, 2020.

[24] S. Tang, D.R. Shelden, C.M. Eastman, P. Pishdad-Bozorgi, and X. Gao, "A review of building information modeling (BIM) and the internet of things (IoT) devices integration: Present status and future trends," *Automation in Construction,* Vol. 101, pp. 127-139, 2019.

[25] Z. Jiao, P. Yao, J. Zhang, L. Wan, and X. Wang, "Capability construction of C4ISR based on AI planning," *IEEE Access,* Vol. 7, pp. 31997-32008, 2019.

[26] H. Baker, M.R. Hallowell, and A.J.-P. Tixier, "AI predicts independent construction safety outcomes from universal attributes," *Automation in Construction,* Vol. 118, No. 2, p. 103146, 2020.

[27] D. Nozaki, K. Okamoto, T. Mochida, and X. Qi, “AI management system to prevent accidents in construction zones using 4K cameras based on 5G network,” In: 21*st International Symposium on Wireless Personal Multimedia Communications (WPMC),* IEEE, pp. 462-466, 2018.

[28] C.-H. Ko, and M.-Y. Cheng, “Hybrid use of AI techniques in developing construction management tools,” *Automation in Construction,* Vol. 12, No. 3, pp. 271-281, 2003.

[29] A.J.-P. Tixier, M.R. Hallowell, B. Rajagopalan, and 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,* Vol. 62, pp. 45-56, 2016.

[30] M. Al Qady, and A. Kandil, “Automatic clustering of construction project documents based on textual similarity,” *Automation in Construction,* Vol. 42, pp. 36-49, 2014.

[31] J.-Y. Hsu, “Content-based text mining technique for retrieval of CAD documents,” *Automation in Construction,* Vol. 31, pp. 65-74, 2013.

[32] A. Rai, What is Text Mining: Techniques and Applications? [Blog post], Available: https://www.upgrad.com/blog/what-is-text-mining-techniques-and-applications/ [Accessed: September 2020]

[33] F. Zhang, H. Fleyeh, X. Wang, and M. Lu, “Construction site accident analysis using text mining and natural language processing techniques,” *Automation in Construction,* Vol. 99, pp. 238-248, 2019.

[34] Y.M. Goh, and C. Ubeynarayana, “Construction accident narrative classification: An evaluation of text mining techniques,” *Accident Analysis & Prevention,* Vol. 108, pp. 122-130, 2017.

[35] K. Kang, and H. Ryu, “Predicting types of occupational accidents at construction sites in Korea using random forest model,” *Safety Science,* Vol. 120, pp. 226-236, 2019.

[36] S. Bertke, A. Meyers, S. Wurzelbacher, J. Bell, M. Lampl, and D. Robins, “Development and evaluation of a Naïve Bayesian model for coding causation of workers’ compensation claims,” *Journal of Safety Research,* Vol. 43, no. 5-6, pp. 327-332, 2012.

[37] R. Akhavian, and A.H. Behzadan, “Smartphone-based construction workers' activity recognition and classification,” *Automation in Construction,* Vol. 71, pp. 198-209, 2016.

[38] J.-H. Chen, “KNN based knowledge-sharing model for severe change order disputes in construction,” *Automation in Construction,* Vol. 17, No. 6, pp. 773-779, 2008.

[39] H. Fan, and H. Li, “Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques,” *Automation in Construction,* Vol. 34, pp. 85-91, 2013.

[40] L. Chen, K. Vallmuur, and R. Nayak, “Injury narrative text classification using factorization model,” *BMC Medical Informatics and Decision Making,* Vol. 15, No. S1, p. S5, 2015.

[41] H. Luo, C. Xiong, W. Fang, P.E. Love, B. Zhang, and X. Ouyang, “Convolutional neural networks: Computer vision-based workforce activity assessment in construction,” *Automation in Construction,* Vol. 94, pp. 282-289, 2018.

[42] P. Kulkarni, S. Londhe, and M. Deo, “Artificial neural networks for construction management: a review,” *Journal of Soft Computing in Civil Engineering,* Vol. 1, No. 2, pp. 70-88, 2017.

[43] M.-Y. Cheng, D. Kusoemo, and R.A. Gosno, "Text mining-based construction site accident classification using hybrid supervised machine learning," *Automation in Construction*, Vol. 118, pp. 103-265, 2020.

[44] S. Hide, S. Atkinson, T.C. Pavitt, R. Haslam, A.G. Gibb, and D.E. Gyi, *Causal Factors in Construction Accidents*, Research Report 156, Loughborough University, 2003.

[45] V.J. Davies, and K. Tomasin, *Construction Safety Handbook*, Thomas Telford, 1996.

[46] R.W. King, and R. Hudson, *Construction Hazard and Safety Handbook*, Butterworth-Heinemann, 1985.

[47] A. Suraji, A.R. Duff, and S.J. Peckitt, "Development of causal model of construction accident causation," *Journal of Construction Engineering and Management*, Vol. 127, No. 4, pp. 337-344, 2001.

[48] P. Mitropoulos, T.S. Abdelhamid, and G.A. Howell, "Systems model of construction accident causation," *Journal of Construction Engineering and Management*, Vol. 131, No. 7, pp. 816-825, 2005.

[49] J. Reason, "The contribution of latent human failures to the breakdown of complex systems," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, Vol. 327, No. 1241, pp. 475-484, 1990.

[50] P. Rezakhani, "Classifying key risk factors in construction projects," *Buletinul Institutului Politehnic din lasi. Sectia Constructii, Arhitectura,* Vol. 58, No. 2, p. 27, 2012.

[51] J.C. Gill, and B.D. Malamud, "Hazard interactions and interaction networks (cascades) within multi-hazard methodologies," *Earth System Dynamics*, Vol. 7, No. 3, pp. 659, 2016.

[52] R. Kerr, M. McHugh, and M. McCrory, "HSE management standards and stress-related work outcomes," *Occupational Medicine*, Vol. 59, No. 8, pp. 574-579, 2009.

[53] D. Woods, and S. Dekker, "Anticipating the effects of technological change: A new era of dynamics for human factors," *Theoretical Issues in Ergonomics Science*, Vol. 1, No. 3, pp. 272-282, 2000.

[54] T.P. Williams, and J. Gong, "Predicting construction cost overruns using text mining, numerical data and ensemble classifiers," *Automation in Construction*, Vol. 43, pp. 23-29, 2014.

[55] H. Baker, M.R. Hallowell, and A.J.-P. Tixier, "Automatically learning construction injury precursors from text," *Automation in Construction*, Vol. 118, pp. 103-145, 2020.

[56] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," In: *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational linguistics: Human Language Technologies*, pp. 1480-1489, 2016.

[57] R. Prabowo, and M. Thelwall, "Sentiment analysis: A combined approach," *Journal of Informetrics*, Vol. 3, No. 2, pp. 143-157, 2009.

[58] H. Muir, and L. Thomas, "Passenger safety and very large transportation aircraft," *Measurement and Control*, Vol. 37, No. 2, pp. 53-58, 2004.

[59] J. Garrett, and J. Teizer, "Human factors analysis classification system relating to human error awareness taxonomy in construction safety," *Journal of Construction Engineering and Management*, Vol. 135, No. 8, pp. 754-763, 2009.

[60] T.S. Abdelhamid, and J.G. Everett, "Identifying root causes of construction accidents," *Journal of Construction Engineering and Management*, Vol. 126, No. 1, pp. 52-60, 2000.

[61] S.W.A. Gunn, “The language of disasters,” *Prehospital and Disaster Medicine,* Vol. 5, No. 4, pp. 373-376, 1990.

[62] L.A. Owen, U. Kamp, G.A. Khattak, E.L. Harp, D.K. Keefer, and M.A. Bauer, “Landslides triggered by the 8 October 2005 Kashmir earthquake,” *Geomorphology,* Vol. 94, No. 1-2, pp. 1-9, 2008.

[63] J. Reason, *Managing the Risks of Organizational Accidents,* Routledge, 2016.

[64] OSHA, *Severe Injury Reports | Occupational Safety and Health Administration* [Online], Available: www.osha.gov/severeinjury/ [Accessed: September 2019]

[65] M. Desvignes, *Requisite Empirical Risk Data for Integration of Safety with Advanced Technologies and Intelligent Systems,* Thesis (Master's), University of Colorado, 2014.

[66] A.Z. Broder, S.C. Glassman, M.S. Manasse, and G. Zweig, “Syntactic clustering of the web,” *Computer Networks and ISDN Systems,* Vol. 29, pp. 1157-1166, 1997.

[67] M. Sanderson, D.M. Christopher, P. Raghavan, and S. Hinrich, *Introduction to Information Retrieval,* Cambridge University Press, 2008.

[68] A.M. Feyer, A.M. Williamson, and D.R. Cairns, “The involvement of human behaviour in occupational accidents: Errors in context,” *Safety Science,* Vol. 25, No. 1-3, pp. 55-65, 1997.

[69] F.A. Boamah, “Measures and strategies for managing safety on construction sites,” *International Journal of Advanced Research,* Vol. 7, No. 8, pp. 96-102, 2019.

View publication stats