Automation in Construction 113 (2020) 103089

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Deep learning and network analysis: Classifying and visualizing accident narratives in construction

Botao Zhong[a,b], Xing Pan[a,b], Peter E.D. Love[c], Lieyun Ding[a,b,*], Weili Fang[a,b]

[a] Hubei Engineering Research Center for Virtual, Safe and Automated Construction, Huazhong University of Science & Technology, China
[b] School of Civil Engineering and Mechanics, Huazhong University of Science & Technology, China
[c] School of Civil and Mechanical Engineering, Curtin University, GPO Box U1987, Perth, WA 6845, Australia

ARTICLE INFO

Keywords:
Accident narratives
Deep learning
Text classification
Topic mining
Network analysis

ABSTRACT

If headway is to be made to improve safety performance in construction, then there is a need to learn from past accidents. Accident reports provide a useful source of information to make sense as to why and how events occurred. Analyzing such reports, however, can be a lengthy and challenging process as there is a tendency for data to be presented in an unstructured or semi-structured free-text format. Thus, being able to classify and analyze the narrative that surrounds accidents and to better understand their causal nature is a challenge. Text classification using shallow machine learning with sophisticated manual lexical, syntactic, and semantic features engineering has been typically used to mine accident data. However, this approach requires highly skilled experts with domain knowledge to undertake this task. A limited number of studies have employed deep learning models to examine the text of safety reports in construction. In consideration of this limitation, word embedding is used to model the semantic narratives of accidents. Then, a Convolution Neural Network (CNN) model is trained to automatically extract text features and classify accident narratives without manual feature processing. The Latent Dirichlet Allocation (LDA) model is used to examine the interdependency that exists between causal variables to visualize the accident narratives. The proposed automated classification model and LDA-based network analysis method provide a useful approach to enable machine-assisted interpretation of texts-based accident narratives. Moreover, the proposed approach can provide managers with much-needed information and knowledge to improve safety on-site.

1. Introduction

Safety analysis in construction can broadly be classified into two different categories, namely predictive and retrospective methods [1]. Retrospective methods rely on past experiences and accident records (e.g., lessons learned and safety checklists) to avoid them in the future. Retrospective management requires a mechanism to prevent occurrence of similar accidents and promote workplace safety [2]. The crucial function of such a mechanism is the ability to analyze accident narratives collected over a period of time and derive knowledge about what previously went wrong [3]. More often than not managers are not provided with timely and fact-based information about accident causation as it is typically in an unstructured or semi-structure format [4]. Having to manually analyze accident text is a time-consuming and inefficient task [5].

Text mining has been identified as a potential technique that can be used to analyze and classify data contained within safety reports [6]. Existing text classification approaches that have been used to examine safety reports have tended to combine lexical, syntactic, and semantic features manually [7–9]. Such approaches are referred as shallow machine learning and include Support Vector Machine (SVM), Naive Bayes (NB) and K-Nearest Neighbor (KNN) with Natural Language Processing (NLP). This manual feature extraction process is limited by a person's domain knowledge and only can learn using the human-specified shallow feature. Contrastingly, deep learning algorithms (e.g., Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN)) can automatically identify features and use multiple single functions to learn complex tasks with a nonlinear combination of parameters from training data [10].

While there have been a number of studies that have used deep learning methods in construction [11], there is a paucity of research that has focused its use with text classification to examine accident reports. Managers often glimpse over reports as they can be rich in content and in some cases lengthy. As consequence valuable

* Corresponding author at: Hubei Engineering Research Center for Virtual, Safe and Automated Construction, Huazhong University of Science & Technology, China.
E-mail address: dly@hust.edu.cn (L. Ding).

https://doi.org/10.1016/j.autcon.2020.103089
Received 9 August 2019; Received in revised form 7 December 2019; Accepted 14 January 2020
Available online 28 February 2020
0926-5805/

information that describes circumstances and conditions may be overlooked. Being able to analyze reports so that we can better understand the conditions and circumstances of an accident as well as their relationships others that have occurred can help managers make more informed-decisions about how manage safety. Thus, there is a need to automate the process of analyzing accident reports so that managers can learn and put in place processes to mitigate their future occurrence in a timely manner.

Against this contextual backdrop, we develop an Artificial Intelligence (AI) solution to automate the process of analyzing accident reports. We integrate therefore NLP with deep learning to automatically extract features and effectively classify accident narratives. The developed deep learning model incorporates topic mining and visual network to analyze and interpret texts-based accident narratives. The effectiveness of the deep learning model is verified using an experiment and compared with SVM, NB and KNN shallow machine learning methods. Each narrative category based on the CNN's classification is examined using the Latent Dirichlet Allocation (LDA) to garner further insights into the causes of the accident. The keywords derived from the LDA are analyzed using network analysis to identify and visualize the causal nature of an accident.

The aim of this paper is not to provide new insights into the causes of accidents per se, but demonstrate that deep learning can be used to extract unstructured safety data from accident text narratives automatically. As a result, managers will be better positioned to make timely and better-informed decisions about how to ensure the safety of their workforce on-site. The paper's contributions are twofold: (1) a novel deep learning approach is developed to analyze accident reports automatically; and (2) the narrative in form of text can be extracted and the causal variables of accidents visualized.

2. Related work

Text classification is the process of assigning tags or categories to text according to its content. Several studies have utilized text mining techniques to analyze injury texts. Its limited modeling and representational ability, however, makes it impossible to learn complex functions, such as those involved in text semantics [12]. Acknowledging this limitation, Ugray [13] used NLP and Bow-tie diagrams to form a semi-automated technique for classifying text-based ‘close call’ texts. Similarly, Bertke, et al. [14] used NB classifiers to classify workers' medical compensation claims into three ‘claim causation’ categories (i.e. musculoskeletal disorder, slips and falls).

In construction text mining has been widely used to address a number of problems such as cost overruns [15], disputes [16], document retrieval [17] and the classification inspection records [18]. Nevertheless, within the context of safety, there have been limited studies undertaken. Prevailing research has tended to focus on using shallow machine learning algorithms (e.g., SVM, NB or Hidden Markov Model) to automatically classify text [19–21]. For example, Tixier, et al. [22] used NLP to design a rule-based automated content analysis system to extract precursors and outcomes from unstructured injury texts. Relatedly, Chokor, et al. [23] applied a K-Means-based clustering approach to accident description texts to support safety inspections. While these shallow learning classifiers have been successfully applied in text classification problems, their ability to learn semantics from text has limitations [24]. Shallow learning models require complicated engineering to select effective features for better classification performance, which are extracted through manual text analysis.

An essential step in the text classification process is the engineering and extraction of features from free-text. Within construction, expert-driven feature extraction has been mainly used to examine accident data [19]. Keywords are manually extracted from the narrative as features for the classification process, which is profoundly reliant on the domain-specific knowledge and skills of the experts [17]. Such a reliance on expert decisions, often results in text being omitted and being fully utilized [25]. Additionally, this manual process is not easily extendable; for any new class or category, requiring experts to be engaged to extend the functionality of the existing model.

Fully automated feature extraction approaches, such as CNN's, are efficient and do not require any expert intervention to extract useful features from texts [26]. As a traditional feature selection method, term frequency-inverse document frequency (TF-IDF) is often used as the feature evaluation function [25]. The TF-IDF matrix has been widely used to train shallow learning models (e.g., SVM, KNN, and NB) [27]. A significant drawback, however, of the TF-IDF is that it ignores the semantics of words and cannot effectively extract their semantic features without manual features engineering [28].

Deep learning methods have been suggested to be a suitable approach for automatically extracting features for text classification [29]. When compared with handcrafted feature-based methods, deep learning can learn related features from given sentences without performing complicated feature engineering work. Furthermore, deep learning can learn complex functions, which cascade from multiple single functions, with a nonlinear combination of parameters using training data [30].

As deep learning can be used to extract features, in this study automatically, the skip-gram (word2vec) model is used to transform the sequence of words into a word feature vector matrix. The CNN model is then developed to extract the text feature and is trained as a classifier for classification. For each category of narratives that are classified by the CNN, LDA-based network analysis is developed to visualize the results. For the visual representation techniques, Mackinlay, et al. [31] suggested visual analysis can be used to support the interpretation of large amounts of text by graphical representation techniques that reduce the analyst's workload. In a similar vein, Robinson [32] used visual analysis to analyze texts from Aviation Safety Reporting System database.

Despite the potential benefits of visual representation, there has been limited research in construction that has used this approach to examine the nature of accident narratives. For the network analysis, it is essential to extract the keywords in text. Keywords can be identified manually or automatically and used as nodes of the network. A standard keyword extraction model is the term TF-IDF representation, but it has been generally used to determine word frequency [33]. A significant drawback of TF-IDF is that it ignores the semantics of words and cannot effectively extract keywords [28]. In comparison to TF-IDF, the LDA model considers the word frequency and semantic features of surrounding words to extract the keywords effectively [34]. Then, LAD-based network analysis method is developed for visual representation for accident narratives.

3. Research approach

The workflow for the research presented in this paper is presented in is Fig. 1. A CNN model is used to classify the data before preprocessing accident narratives. The accident narratives are then collated to form one document. The LDA technique with unsupervised learning approaches is used to mine the main topics and determine their respective (corresponding) keywords in the Co-occurrence Network. The LDA-based network analysis is employed to depict and visually to represent the causal relationships of accidents.

3.1. Data material and preprocessing

Akin to Yang and Ubeynarayana [27], the Occupational Safety and Health Administration (OSHA) website in the United States (US) is used as the primary data source for this research to demonstrate the efficiency and effectiveness of the developed CNN-based model. In this study, 4000 pieces of historical accident texts were randomly selected from the website of the OSHA. The original texts provide narratives of accidents that occurred on construction sites. Noteworthy, the text is

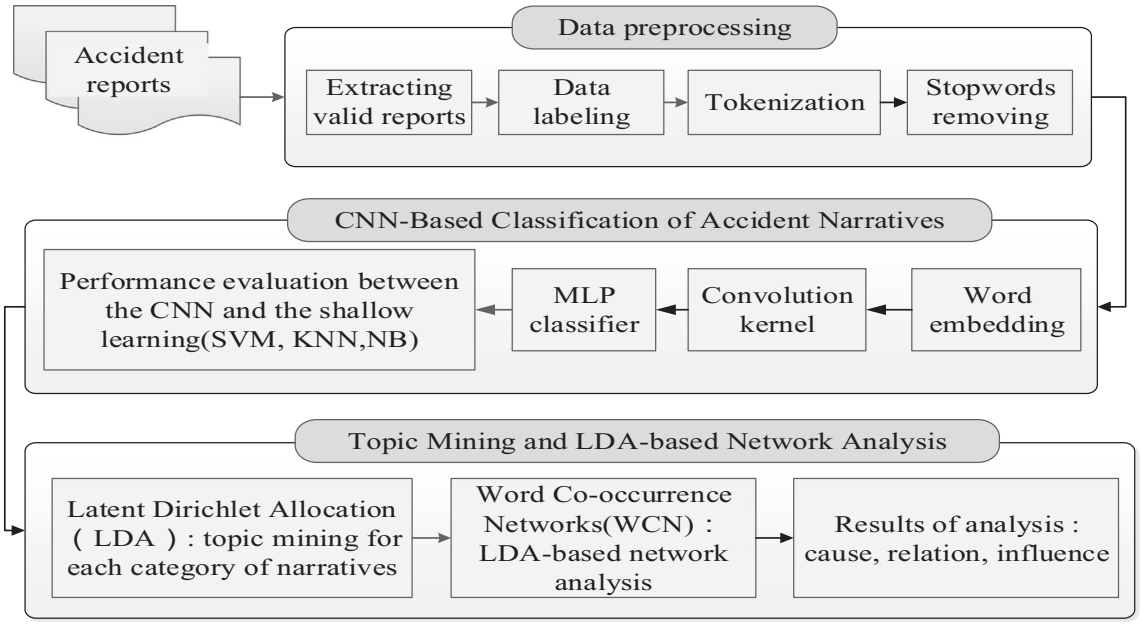


Fig. 1. Workflow of the research.

not labeled and therefore does not provide information about the causes, effects, and consequences of accidents.

A total of 2000 pieces of texts out of 4000 were again randomly selected and manually annotated following the Workplace Safety and Health Institute (2016) classification (Table 1). As more than one event can occur during an accident, the labeling of narratives followed the principle of identifying the primary and first occurrence of the uncontrolled or unintended action.

After labeling, the preprocessing for narratives was conducted using tokenization and stopwords removing. These labeled texts were divided into three data sets: (1) training data, which was used for optimizing the model; (2) validation data, which was used to select the optimized model's parameters; and (3) testing data, which was used to evaluate the performance of the established model. The data preprocessing process of accident narratives for the LDA model was used to connect multiple professional words with underscores to form word-group. For example, “safety helmet” was performed as “safety_helmet”, as it improved the model's ability to provide better outputs for word-groups rather than identifying the word for information identification and minimizing computational time [35,36].

The research presented in this paper used Python 3.6 in the Anaconda environment, and its main packages (such as tensorflow = 0.14, gensim, torch, numpy, and matplotlib) were used to design the CNN model design. Additionally, the main packages (such as genism, nltk, and openpyxl) were used for implementing the LDA model. Then, Gephi [37], was used to provide a visualization of the LDA-based network analysis.

4. CNN-based classification of accident narratives

As a deep learning model, a CNN can learn complex functions and related features from given texts without complicated feature engineering work [38]. By reducing manual interventions in pre-treatment and post-processing, a CNN model can automatically adjust parameters when a text classification task is performed. The CNN can automatically determine discriminative phrases in text using a max-pooling layer, instead of through manual feature engineering with domain knowledge [39]. A CNN classifier is based on multi-layer networks and a convolution architecture. The CNNs layers apply a convolution operation to an input matrix, which can compose different semantic fragments of sentences. As a consequence, interaction between composed fragments can be learnt, enabling the inter-modal semantic relations of accident narratives to be fully exploited. The CNN-based classification framework is shown in Fig. 2, which includes word

Table 1
Labels used for the data, their criteria and sample accident narrative.

ID	Label	Criteria	Narrative example
1	Caught in/between objects	According to the construction safety standards based on WHSI definitions, safety hazards in construction are categorized into 11 kinds with different label names.	Sample accident narrative: On April 9 2013 employee #1 was installing vinyl sidings on a single residence. The employee was standing an a-frame ladder that was set on a plank of a scaffold. The scaffold moved causing employee #1 to lose his balance. The employee fell from the ladder approximately 12-ft to the ground. Employee #1 was transported to an area hospital where he was treated for an abdominal fracture. The employee remained hospitalized. Label: falls
2	Collapse of object		
3	Electrocution		
4	Exposure to chemical substances		
5	Exposure to extreme temperatures		
6	Falls		
7	Fires and explosions		
8	Others		
9	Struck by falling object		
10	Struck by moving objects		
11	Traffic		

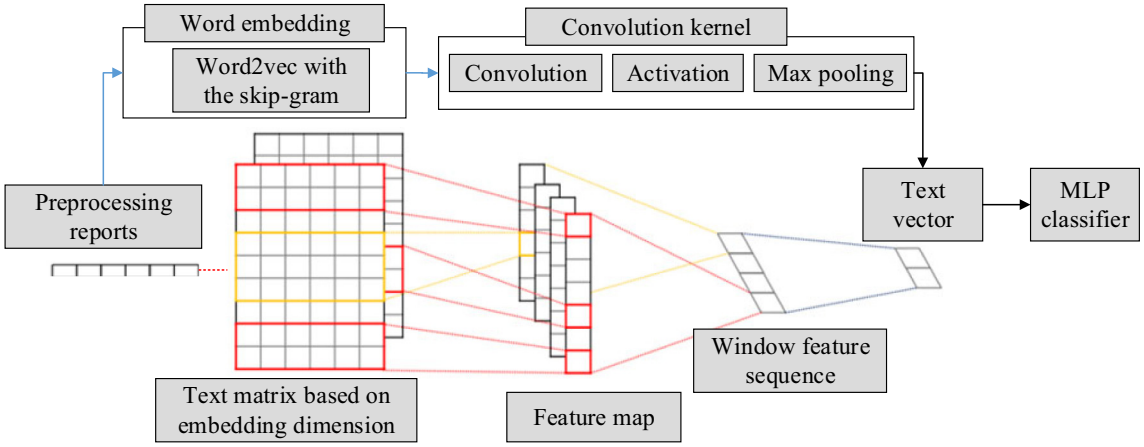


Fig. 2. The CNN-based classification framework.

embedding, convolution, activation, max pooling and a multi-layered perceptron (MLP) classifier. Word embedding is used to automatically transform the sequence of words into a vector matrix. The convolution kernel (i.e., convolution, activation, and max-pooling) then transforms the matrix to a vector to extract the text features. The MLP classifier is used to train the model to automatically classify text.

4.1. CNN-based deep learning model

4.1.1. Word embedding

Words themselves cannot be directly inputted into a CNN. So, word embedding is conducted to transform the sequence of words into a vector matrix that can be recognized and dealt with by computers [40]. The use of traditional word embedding, such as one-hot vectors, face two challenges: (1) loss of word order and (2) oversize of dimensionality. The theory of Maximum Likelihood Estimation is drawn upon to address these issues so that the distributed representations of word embedding can be better represented in semantic texts [30].

In comparison with contemporary representations of word embedding, distributed representations are considered to be more suitable and powerful [41]. In this research, each text inputted for feature extraction takes the form of a sequence of words, within a range of domain vocabularies. Zhu, et al. [42] proposed the use of word2vec techniques for efficient learning of high-quality distributed vector representations. Furthermore, the word2vec model can capture similarities with syntactic and semantic words. Word2vec exists in two models [48]: (1) skip-gram; and (2) continuous bag of words (CBOW). The CBOW model uses the context information to predict the current word. In stark contrast, the skip-gram model learns iteratively from existing words available in a sentence to predict the next word.

The context information contained with a narrative for an accident in construction is often missing, which results in sparse text features. In this instance, the skip-gram pattern is more suitable for accommodating sparse features. The word vector generated by the word2vec algorithm can contain semantic information and can be used to extract text features without complex manual analysis. The word2vec algorithm is described in detail in Zhu, et al. [42]. Notably, the skip-gram (word2vec) algorithm has two important parameters that need to be set: (1) the embedding dimension; and (2) filter size.

4.1.2. Convolution kernel and MLP classifier

A CNN is used to capture the semantic features of the domain vocabularies and to determine the relationships between them. After the process of word embedding and several times of convolution kernel calculations, different matrixes are computed. Thus, an MLP classifier with three layers (i.e. input layer, hidden layer and output layer) is used to classify accident texts. The calculation structure of the MLP classifier with a softmax layer is applied.

The filter size is set to five to avoid associating more semantically inappropriate words and shorten the training time. While the high-dimensional word vector can express the semantic characteristics of phrases, it will also increase the number of parameters of the CNN model and therefore increase the likelihood of over-fitting. For setting the optimal embedding dimension, different values (i.e. 100, 102, 104, 106 …, 200) are tested in this research using 10-fold cross-validation. The 128-embedding dimensional input-vector matrix is adopted for performing best in efficiency and accuracy. A dropout strategy was adopted to ensure neural nodes fail to avoid overfitting. A drop out probability of 0.5 has been identified by Kim, et al. [43] as being appropriate.

In sum the hyperparameters of the CNN model are set as follows: embedding dimension = 128; filter size = 5; number of filters = 128; drop out probability = 0.5; learning rate = 0.5; and regulation = 0, making the optimal CNN model. Learning text classification algorithms are tested to identify the best-performing models.

4.2. Model testing and evaluation

We compare it with the several shallow learning models (e.g., SVM, NB, and KNN) to demonstrate the effectiveness of the developed CNN model. For the same data resource, the parameter of these shallow learning models is described in detail Yang and Ubeynarayana [27]. Therefore, the parameters of SVM, NB and KNN are adopted in this study are same as those presented in Yang and Ubeynarayana [27]. Three key performance indicators are adopted to evaluate the model's performance, namely precision and recall rates and F_1 score metrics (Table 2):

1. *Precision and recall:* Precision-recall is used to measure the classification ability of a system. Performance is measured using precision rate and recall rate. The higher the score the better the system performance. Here, precision is the fraction of examples classified as positive that are truly positive. Recall, however, is the fraction of examples classified as positive that are correctly labeled. Thus, the value of precision rate and recall rate can be calculated by Eqs. (1) and (2).

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where, TP is the number of true positive, FN is the number of false negative, FP is the number of false positive, and TN is the number of

Table 2
Results of four text mining methods in the experimentations.

No.	Labels	F_1 Score				Precision				Recall			
		CNN	NB	SVM	KNN	CNN	NB	SVM	KNN	CNN	NB	SVM	KNN
1	Caught in/between objects	0.52	0.49	0.52	**0.65**	0.48	0.50	0.62	**0.63**	0.57	0.49	0.44	**0.67**
2	Collapse of object	0.46	0.22	**0.56**	0.40	**0.52**	0.36	0.50	0.36	0.40	0.16	**0.63**	0.44
3	Electrocution	**0.96**	0.80	0.92	0.84	**0.97**	0.90	0.94	0.78	**0.95**	0.72	0.89	0.91
4	Exposure to chemical substances	**0.84**	0.60	0.36	0.36	**0.90**	0.82	1.00	0.33	**0.78**	0.47	0.22	0.40
5	Exposure to extreme temperatures	**0.47**	0.14	0.33	0.29	0.47	**1.00**	0.33	0.5	**0.47**	0.08	0.33	0.20
6	Falls	**0.78**	0.68	0.69	0.68	0.72	0.53	**0.74**	0.69	0.86	**0.93**	0.65	0.67
7	Fires and explosion	**0.83**	0.70	0.65	0.40	0.84	**0.87**	0.53	0.50	**0.82**	0.59	0.83	0.33
8	Others	0.31	**0.48**	0.11	0.40	**0.34**	0.34	0.1	0.33	0.29	**0.80**	0.12	0.50
9	Struck by falling object	**0.35**	0.15	0.20	0.00	**0.33**	0.50	0.25	0.00	**0.36**	0.09	0.17	0.00
10	Struck by moving objects	0.38	0.00	**0.60**	0.29	0.47	0.00	**0.60**	0.31	0.31	0.00	**0.60**	0.28
11	Traffic	0.37	0.10	**0.57**	0.21	0.4	0.15	**0.57**	0.40	0.34	0.08	**0.57**	0.14
/	Average	**0.63**	0.47	0.59	0.51	**0.65**	0.51	0.62	0.52	**0.61**	0.53	0.59	0.52

(Note: All values in above table keep two decimal places behind. The numbers in bold are the highest F1 scores, precision and recall of corresponding labels.)

true negative.

2. *Average weight F_1*: To evaluate the model performance, F_1 score proposed by Price and Bouvier [44] has been widely adopted in literatures. However, support that denotes the number of true instances for each label is not considered in conventional F_1 score calculation. Therefore, the average weighted F_1 score is provided in Eq. (3) which is expressed as:

$$AvgF1_{weight} = \sum_{i=1}^{N} \left(\frac{S_i}{T} * F1 \right) \tag{3}$$

where, N denotes the total number of labels, S_i denotes the support of label i, T denotes the support of all labels and F_{1i} denotes the F_1 score of label i.

Based on Table 2, the testing results of different classifiers are summarized as follows:

- It can be seen from Table 2 that the CNN model outperforms all the other methods that are examined. With respect to the F_1 Score, the CNN model performs better for the label of No. 3, No. 4, No. 5, No. 6, No. 7 and No. 9 (i.e. “electrocution”, “exposure to chemical substances”, “exposure to extreme temperatures”, “falls”, “fires and explosion”, “struck by falling object”). Even though the SVM model performs better for the label of No.2, No.10 and No. 11 (i.e. “collapse of object”, “struck by moving objects”, “traffic”), and Bayes model performs better for the label of No. 8 (i.e. “others”), the gap between CNN and them is narrow.
- For the label No. 6 (“falls”), all models perform well. As for the label of No. 1 (i.e. “caught in/between objects”), despite the lower F_1 score of the CNN model than the KNN classifiers.

Even though the precision, recall, and F_1 scores demonstrate that the CNN out-performs the other algorithms, any information on how each category of narratives are misclassified is not provided. Thus, confusion matrices are introduced to identify the categories that have been misclassified. The diagonal figures presented in Fig. 3, represent the confusion matrices that represent correctly classified categories, where all other figures show the misclassified categories. Each row represents the actual categories, and columns represent that are predicted. As can be seen in Fig. 3, the top misclassified activities for the excavator are “collapse of object”. The “collapse of object” is confused 33 (in the sixth column of the second row) times with “falls”. During an accident, the “collapse of object” often leads to “falls”. Thus, the confusion among “collapse of object” and “falls” can be explained by the co-occurrence tendency. The confusion matrix is also widely used to analyze each precision rate and recall rate, such as: the recall of “electrocution” is 0.95, the precision of “electrocution” is 0.97.

In comparison with the shallow learning models identified above, the performance of the proposed CNN method is comparable. It should be noted that no single text classification model can apply to all domains, as text features vary with their application [10]. The results presented in Table 2 show that the classifiers could produce better performance for different labels.

There is, however, a fundamental difference between the use of the CNN and shallowing learning methods. As an end-to-end learning method, the proposed deep learning model performs well, even without manual rules or features. Thus, the developed CNN model can be used for similar text processing tasks (e.g. safety document classification) with specific domain features.

The overall category distribution is applied to compare the differences between manual and machine classification (Fig. 4). It can be seen in Fig. 4 “falls” are clearly the most commonly identified cause of accidents. In stark contrast “exposure to extreme temperatures” and “exposure to chemical substances” are considerably less frequent. There is minimal difference between manual and machine classification.

5. Topic mining and LDA-based network analysis

The LDA model is a useful method for mining topics and their respective (responding) keywords [45]. LDA models are particularly useful for minimizing the time required to examine data that does not possess a label [34]. The accident keywords and the categories obtained from the CNN model are combined to form a single document, as shown in Fig. 5. Then, the LDA model is used to mine the main topics and identify their respective (responding) keywords. The keywords under a topic are treated as the features hidden in the topic. The keywords then reflect a correlation with a topic.

The LDA has two main steps: (1) topic distribution per document; and (2) word distribution per topic. Using the Dirichlet distribution with a parameter of α, the probability of the topic in the document is derived. Then, the Dirichlet distribution with a parameter of β, is used to obtain the probability of the word in the topic. Another important step is to obtain the topic's number K. In accordance with Kim, et al. [43], the empirical value of K is $\alpha = K/50$ and $\beta = 0.01$.

The perplexity (P) approach [46] is used to determine the topic number K. An appropriate probability distribution has a relatively low perplexity. The perplexity (P) can be calculated using Eq. (4):

$$P = \exp\left(-\frac{\sum_d \sum_i^{N_d} \ln p(w_{d,i})}{\sum_d N_d} \right) \tag{4}$$

where, N_d means word frequency in the d document. $w_{d,i}$ means the nth word in the d document, α is the parameter of the Dirichlet prior on the per-document topic distributions, and β is the parameter of the Dirichlet prior on the per-topic word distribution.

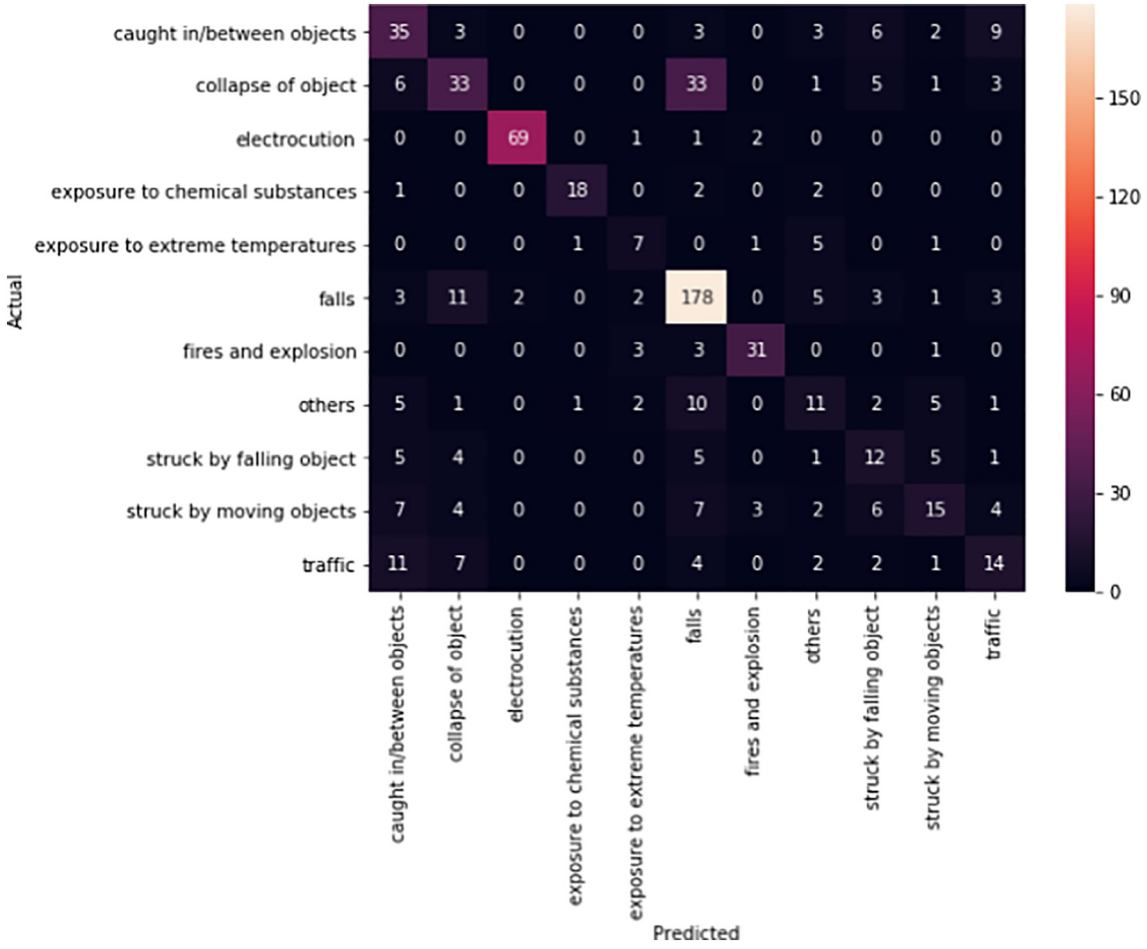


Fig. 3. The confusion matrix of the CNN model.

The appropriate number of topics of all preprocessing narratives without labels are as shown in Fig. 6. The appropriate numbers of topics of documents are 14 respectively by calculating the perplexity. To analyze the perplexity of all topics, the result of LDA has five keyword distributions about “falls”, which happens most among all accidents. Finally, calculated through LDA model, the 10 probability distributions for different per-topic keywords are listed in Table 3.

For the topic “falls”, for example, the keyword “falls” (with the maximum probability distribution of 0.22) can determine the title of topic as “falls” to the greatest extent. With the keyword distributions

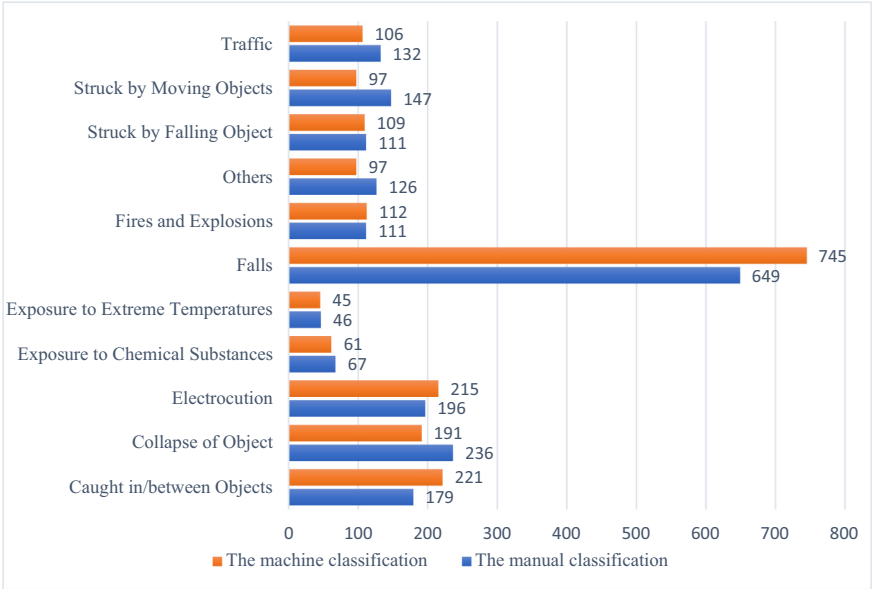


Fig. 4. Distribution of categories for all causes of accidents.

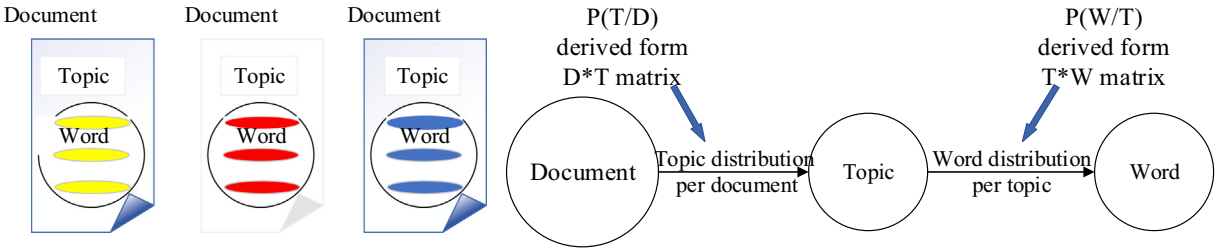
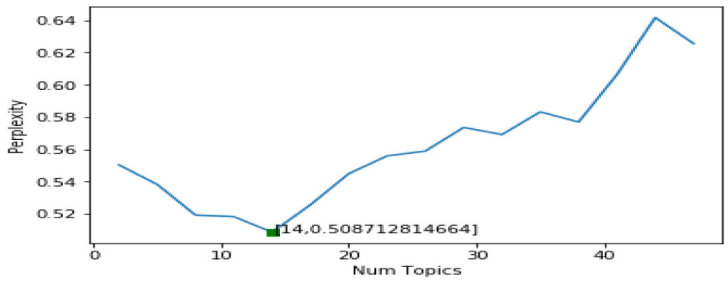
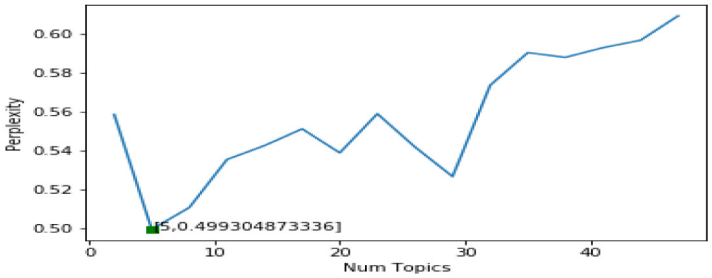


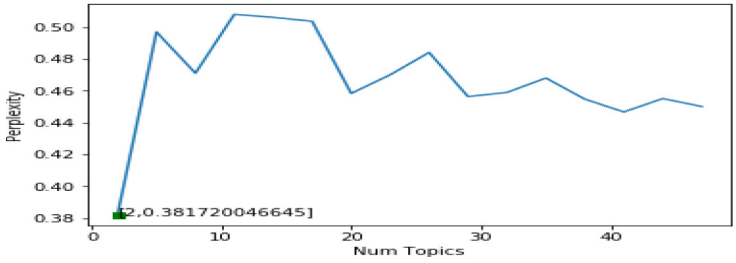
Fig. 5. Concept of LDA.
(Kim et al. 2018:p.58–64 [43]).



(a) Perplexities of all topics



(b) Perplexities of "falls"



(c) Perplexities of "collapse of object"

Fig. 6. Number of topics of all preprocessing narratives without labels.

with the probabilities in each topic, the titles of the topics are derived: "caught in/between objects", "collapse of object", "electrocution", "exposure to chemical substances", "exposure to extreme temperatures", "falls", "fires and explosions", "struck by moving objects", "struck by falling objects", "traffic", and "others". Noteworthy the topics extracted from the above align with the linkage results for the causes of accidents presented in Section 4.2. However, in this instance the per-topic keyword distribution provides a more in-depth analysis of the relationship between topics. For example, topic "falls" may represent a closely co-occurrence relationship in these keywords ("falls", "traffic", "collapse of object"). Topic "caught in/between object" may reflect a closely influential relation that some body parts ("leg", "hand" and "finger") get injured easily in the accident of "caught in/between object".

As "falls" and "collapse of object" commonly result in a serious accident or death, the probability distributions for them are respectively listed in Tables 4 and 5.

According to the Fig. 6(b) and (c), five topics and two ones are identified respectively. After analyzing the corresponding results generated by the LDA model, three out of the five topics and two of two topics are selected. In Table 4 for example, the topic (object about "falls") includes seven most common objects that can result in the occurrence of an accident, which are "ladder", "tower", "scaffold", "elevator", "bridge", "floor" and "tower".

Table 3
Probability distributions for different per-topic keywords.

Topic	Keywords and probabilities										
Falls	Falls	Collapse of object	Install	Traffic	Ladder	Strike	Held	Scaffold	Construction	Killed	
	0.22	0.10	0.07	0.07	0.07	0.07	0.07	0.07	0.04	0.04	
Caught in/between object	Caught in/between object	Leg	Falls	Hand	Finger	Laceration	See	Run	Machine	Toe	
	0.14	0.11	0.10	0.04	0.04	0.03	0.03	0.02	0.01	0.01	
Electrocution	Electrocution	Power line worker	Burn	Utility work	Overhead power line	Induced current	Electrocuted	Lighting	Metal bar	Protective grounding	
	0.08	0.06	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.04	
Collapse by object	Collapse by object	Falls	Kill	Construction	Falls	Crane	Fracture	Head	Steel	Unsecured	
	0.08	0.08	0.04	0.04	0.03	0.03	0.03	0.02	0.02	0.01	
Struck by moving objects	Struck by moving objects	Machine	Work	Scaffold	Falls	Protection	Equipment	Moving	Fail	Fracture	
	0.08	0.05	0.05	0.05	0.04	0.03	0.03	0.03	0.03	0.02	
Fires and explosion	Fires and explosion	Heater	Caught	Explosion	Natural gas	Fire extinguisher	Fuel oil	Trench	Collapse	Excavation	
	0.06	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02	0.02	
Struck by falling object	Struck by falling object	Struck	Go	Collapse	Heat	Move	Object	Fall	Struck	Fail	
	0.18	0.14	0.14	0.12	0.12	0.07	0.07	0.06	0.00	0.00	
Traffic	Traffic	Leg	Truck	Rule	Construction	Fracture	Crush	Vehicle	Burn	Highway	
	0.05	0.04	0.04	0.04	0.03	0.02	0.02	0.02	0.02	0.02	
Exposure to extreme temperatures	Exposure to extreme temperatures	Heat	Hot tar	Skin	Overheated	High temperature	Burn	Go	Work	Construct	
	0.05	0.05	0.05	0.04	0.04	0.03	0.03	0.03	0.03	0.03	
Exposure to chemical substances	Exposure to chemical substances	Chemical vapor	Asphyxiate	Poison	Carbon monoxide	Methylene chloride	Chemical	Inhalation	Unconscious	Ventilation	
	0.05	0.04	0.03	0.02	0.02	0.02	0.02	0.01	0.01	0.01	

5.1. LDA-based network analysis

Visual presentation should be infused for text analysis to identify information [49]. The Word Co-occurrence Network (WCN) uses graphs as a means to represent words as nodes and identify their relationships with one another. Considering the advantages of the LDA model, which make full use of the word frequency and semantic features of surrounding words to effectively extract the keywords [34], the keywords of the WCN are obtained from the model. Then, the LDA-based network analysis is developed to measure the interweaving relations of the keywords in the accident narrative.

During the model's processing, a set of keywords, as noted in Table 3 are collated and used as a document, d∈D, where $D = \{d_1, d_2, \ldots, d_n\}$ is the collection of N documents. Consider the set of words w∈d, where $w = \{w_1, w_2, \ldots, w_t\}$ is collection of words in a document d∈D. The representation of word co-occurrence network for the accident narratives is shown in Fig. 7.

The LDA-based network analysis is used to identify the relationships between keywords visually. These relations can potentially unearth essential patterns that can be used to understand the nature of accidents better. For example, in the case of “falls”, a relationship between objects and actions is depicted (Fig. 7). The degree of centrality and eigenvector centrality are shown in Table 6. The degree of centrality is used to measure the number of directed or undirected relationships that a node has with others in a network. This indicator can be used to calculate a node's number of directly connected neighbors in a network. A node with a higher degree of centrality indicates that it has more influence and importance than others in a network. The Eigenvector centrality [47] is an appropriate measure when the status of a node is assumed to be a positive function of the status of the others to which it is connected.

As shown in Fig. 7, the word co-occurrence network is visualized. In general, the keywords that have a co-occurrence relationship are closely linked. The degree of centrality and eigenvector centrality of the top 20 keywords are presented in Table 6. The nodes with a higher degree of centrality in Table. 6, such as “fall”, “collapse of object”, indicate that they have more influence and importance than other nodes. Take the “scaffold” as an example, while its degree centrality is 0.05, it has a lower influence and importance than other nodes in the network, but its eigenvector centrality is ranked highly. Thus, accidents on scaffolds are infrequent. When such accidents occur, they are severe and are likely to result in someone being killed.

With the assistance of data visualization techniques, managers can intuitively understand the potential relationship between the causal variables that result in an accident occurring. As a result, this provides them with the ability to put in place safety mechanisms to mitigate accidents on their sites.

6. Discussion

With the increasing emergence of AI and digital technologies, the construction industry is beginning to embrace their use to improve productivity and performance of operations on-site. The AI techniques of machine and deep learning, for example have had a significant influence, but use cases in construction are still relatively nascent. To demonstrate how AI can be used to improve safety, this paper develops a deep learning CNN approach to analyze unstructured accident text automatically. In doing so, it can potentially provide managers with the knowledge to better understand the characteristics of accidents.

The developed CNN model performance is comparable with shallow models, though without manual feature engineering. The combination with the LDA-based network analysis, however, enables invaluable insights into the nuances contained within the text narratives of accidents. The developed approach is sophisticated due to the CNNs architecture and the hyperparameter tuning that was undertaken. From this perspective, it is essential to consider the nuanced tradeoffs

Table 4
Probability distributions for “falls”-topic keywords.

Topic	Keywords and probabilities									
Objects about “falls”	Falls	Traffic	Ladder	Tower	Scaffold	Elevator	Bridge	Floor	Roof	Rules
	0.06	0.04	0.04	0.03	0.03	0.02	0.02	0.02	0.02	0.02
Actions about “falls”	Falls	Collapse	Install	Struck	Hold	Run	Work	Exposure	Open	Install
	0.09	0.06	0.05	0.05	0.04	0.03	0.02	0.02	0.02	0.01
Workers' feature about “falls”	Falls	Death	Kill	Unsecured	Head	Run	Roof	Unstable	Lost	Install
	0.08	0.04	0.04	0.04	0.03	0.03	0.03	0.02	0.02	0.02

Table 5
Probability distributions for “collapse of object”-topic keywords.

Topic	Keywords and probabilities									
Actions about “collapse of object”	Collapse of object	Falls	Work	Protect	Strike	Caught in/between objects	Trike	Install	Run	Crush
	0.05	0.05	0.04	0.04	0.03	0.03	0.03	0.02	0.02	0.02
Object about “collapse of object”	Collapse of object	Subway	Construction	Fracture	Equipment	Scaffold	Crane	Collapse	Excavation	Ladder
	0.07	0.05	0.05	0.04	0.03	0.03	0.03	0.03	0.03	0.02

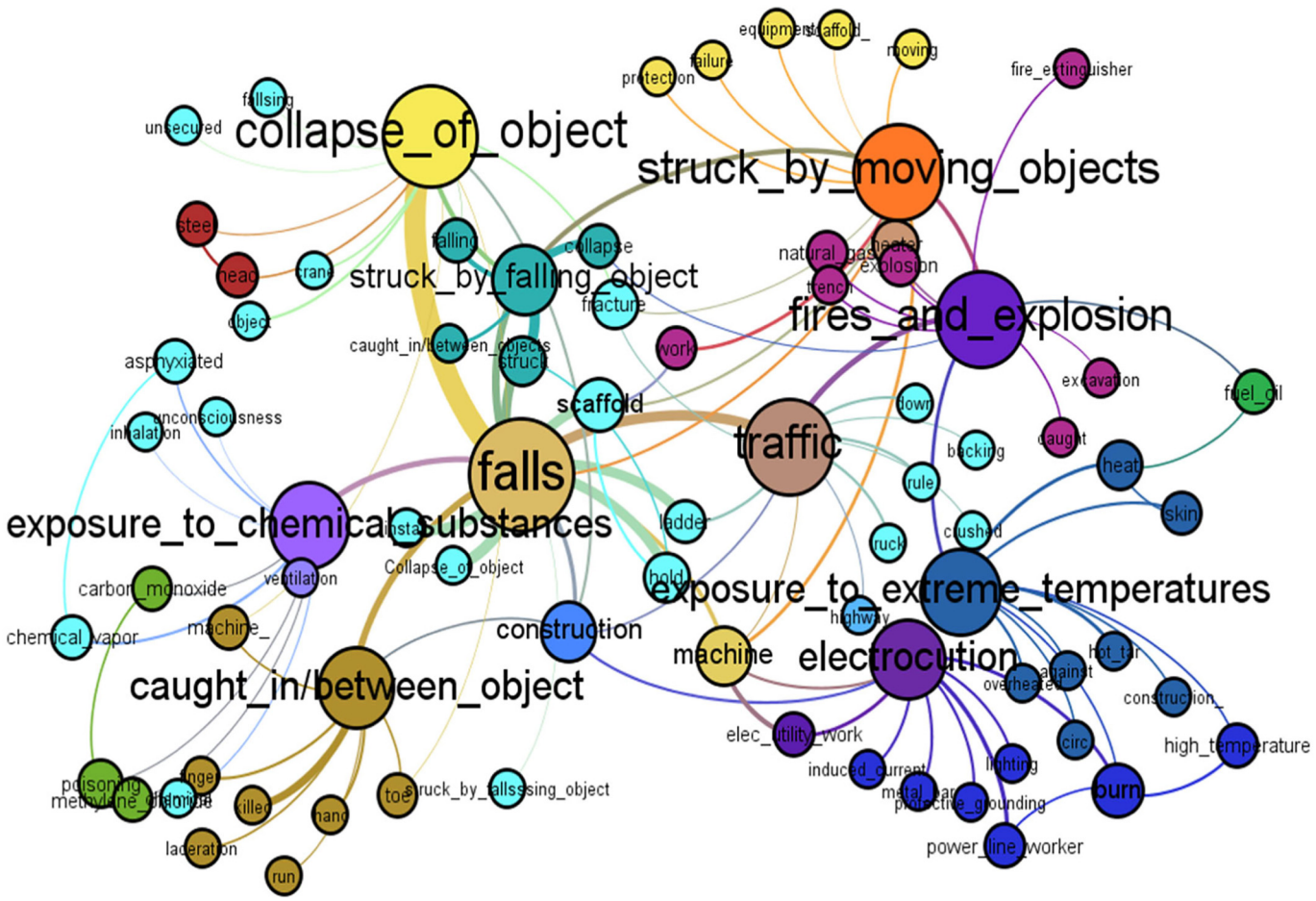


Fig. 7. Representation of word co-occurrence network.

between architecture engineering and feature engineering.

Within this mind, it suggested that there is a need to determine if the integration of advanced semantic and syntactic features and domain-specific knowledge into CNN models would improve the process of classification and provide even more accurate results that compare to human experts. For example, within a text narrative, words with similar meaning can be compressed into the same term or use a domain-specific ontology to map them. Also, the ontology can be used to model the complex semantics of in the narratives and simultaneously evaluate its feasibility to be combined with deep learning and NLP.

7. Limitations

The research presented in this paper, however, is not without its limitations. As the CNN model was constructed on top of generic and basic features (i.e. words), it was expected the model would perform well for classifying similar accident texts. However, in this study, only the text of accidents occurring on construction sites from the OSHA

Table 6
Degree centrality, and eigenvector centrality of the top 20 keywords.

Label	Degree of centrality	Label	Eigenvector centrality
Falls	0.19	Falls	1.00
Collapse of object	0.18	Collapse of object	0.79
Traffic	0.15	Struck by moving objects	0.67
Struck by moving objects	0.15	Traffic	0.64
Fires and explosion	0.15	Caught in/between object	0.56
Caught in/between object	0.14	Struck by falling object	0.52
Exposure to extreme temperatures	0.13	Construction	0.51
Exposure to chemical substances	0.13	Machine	0.47
Electrocution	0.11	Fires and explosion	0.47
Struck by falling object	0.09	Scaffold	0.36
Machine	0.06	Exposure to chemical substances	0.31
Construction	0.06	Strike	0.30
Scaffold	0.05	Electrocution	0.29
Burn	0.05	Hold	0.28
Strike	0.04	Ladder	0.28
Hold	0.04	Exposure to extreme temperatures	0.23
Ladder	0.03	Install	0.17
Elec utility work	0.03	Burn	0.14
Power line worker	0.03	Elec utility work	0.13
Install	0.01	Power line worker	0.08

website was selected to train and test the effectiveness of the proposed method. Thus, future work is needed to test the algorithms on a much larger sample of accident narratives from other sources.

Besides the sample size, another limitation that comes to the fore relates to labeling. Sometimes the classification may be a multi-label text classification task. However, to avoid having an accident with multiple categories, a label was assigned according to the principle of identifying the primary and first occurrence of the uncontrolled or unintended action, if more than one event can occur during an accident. In future, therefore, there is a need to develop a multi-label classifier to process accident texts with multiple labels.

8. Conclusion

Unstructured and semi-structured free-texts are widely produced and used in construction. Such text provides practitioners with essential sources of information that can be used to retrospectively inform decision-making and improve the management of safety in projects. Typically, however, the process used to decipher and garner an understanding of accident texts is a manual and time-consuming process. Consequently, managers may overlook some important and recurring issues that are embedded in the narratives that have been formulated.

In addressing this issue, the paper has proposed a method by integrating NLP with a CNN deep learning and then using visual network analysis. The combining of NLP and CNN's enables accident narratives to be automatically classified and visualized and therefore help improve the effectiveness of decision-making. The developed CNN is trained and evaluated by comparing its performance with traditionally shallow machine learning methods that have been typically used to analyze accident narratives. The proposed model outperforms the other models, while only using the raw narrative text as inputs, without requiring manual feature engineering.

As an end-to-end learning method, the proposed word embedding and CNN-based deep learning model can automatically learn high-level representative features through the network, instead of being manually designed, and have more powerful feature representation and learning ability. The LDA-based network analysis method is proposed to analyze and provide visual representations of the data. The results demonstrate that the method can provide invaluable insights from the unstructured text data by enabling a machine-assisted interpretation of large volumes of accident narratives.

The research effort and results represent an initial step towards developing a new accident management capability by quickly analyzing a large number of accident texts over a more considerable period, with exploiting the potential power of deep learning and network analysis and visual representation technique.

Declaration of competing interest

The authors declared that they have no conflicts of interest to this work.

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted. We declare that the work presented is original research that has not been published previously, and is not under consideration for publication elsewhere, in whole or in part.

Acknowledgments

This research is partly supported by the National Natural Science Foundation of China (Grant No. 51878311, No. 71732001, No. 51978302).

References

[1] M. Fiedler, P. Renner, J. Schubert, F. Weber, A. Hartmann, H. Iro, V. Vielsmeier, C. Bohr, M. Gerken, T.E. Reichert, Predictive value of FHIT, p27, and pERK1/ERK2 in salivary gland carcinomas: a retrospective study, Clin. Oral Investig. (2019) 1–9, https://doi.org/10.1007/s00784-019-02809-z.
[2] P. Thepaksorn, S. Thongjerm, S. Incharoen, W. Siriwong, K. Harada, A. Koizumi, Job safety analysis and hazard identification for work accident prevention in para rubber wood sawmills in southern Thailand, J. Occup. Health 59 (6) (2017) 542–551, https://doi.org/10.1539/joh.16-0204-CS.
[3] K. Mckenzie, M.A. Campbell, D.A. Scott, T.R. Discoll, J.E. Harrison, R.J. Mcclure, Identifying work related injuries: comparison of methods for interrogating text fields, BMC Medical Informatics and Decision Making 10 (2010) 19, https://doi.org/10.1186/1472-6947-10-19.
[4] A.M. Aitken, Managing unstructured and semi-structured information in organisations, IEEE/ACIS International Conference on Computer and Information Science (2007) 712–717, https://doi.org/10.1109/ICIS.2007.129.
[5] M. Behm, A. Schneller, Application of the Loughborough construction accident causation model: a framework for organizational learning, Constr. Manag. Econ. 31 (6) (2013) 580–595, https://doi.org/10.1080/01446193.2012.690884.
[6] L. Tanguy, N. Tulechki, A. Urieli, E. Hermann, C. Raynal, Natural language processing for aviation safety reports: from classification to interactive analysis, Comput. Ind. 78 (2016) 80–95, https://doi.org/10.1016/j.compind.2015.09.005.
[7] M. Goudjil, M. Koudil, M. Bedda, N. Ghoggali, A novel active learning method using SVM for text classification, Int. J. Autom. Comput. (2018) 1–9, https://doi.org/10.1007/s11633-015-0912-z.
[8] W. Fang, B. Zhong, N. Zhao, P.E.D. Love, H. Luo, J. Xue, S. Xu, A deep learning-based approach for mitigating falls from height with computer vision: convolutional neural network, Adv. Eng. Inform. 39 (2019) 170–177, https://doi.org/10.1016/j.aei.2018.12.005.

[9] S. Ahmad, R. Varma, Information extraction from text messages using data mining techniques, Malaya Journal of Matematik (2018) 26–29, https://doi.org/10.26637/MJM0S01/05.

[10] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.

[11] Z. Jiang, L. Li, D. Huang, L. Jin, Training word embeddings for deep learning in biomedical text mining tasks, 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2015, pp. 625–628, , https://doi.org/10.1109/BIBM.2015.7359756.

[12] V. Cherkassky, Y. Ma, Practical selection of SVM parameters and noise estimation for SVM regression, Neural Netw. 17 (1) (2004) 113–126, https://doi.org/10.1016/s0893-6080(03)00169-2.

[13] Z. Ugray, L. Lasdon, J. Plummer, F. Glover, J. Kelly, R. Martí, Scatter search and local nlp solvers: a multistart framework for global optimization, INFORMS J. Comput. (2007) 328–340, https://doi.org/10.1287/ijoc.1060.0175.

[14] S.J. Bertke, A.R. Meyers, S.J. Wurzelbacher, J. Bell, M.L. Lampl, D.J.J.R. Robins, Development and evaluation of a Naïve Bayesian model for coding causation of workers' compensation claims, J. Saf. Res. 43 (5–6) (2012) 327–332, https://doi.org/10.1016/j.jsr.2012.10.012.

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

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

[17] W.D. Yu, J.Y. Hsu, Content-based text mining technique for retrieval of cad documents, Autom. Constr. 31 (5) (2013) 65–74, https://doi.org/10.1016/j.autcon.2012.11.037.

[18] N.W. Chi, K.Y. Lin, N. El-Gohary, S.H. Hsieh, Evaluating the strength of text classification categories for supporting construction field inspection, Autom. Constr. 64 (2016) 78–88, https://doi.org/10.1016/j.autcon.2016.01.001.

[19] S. Bahassine, A. Madani, M. Kissi, Technology, Arabic text classification using new stemmer for feature selection and decision trees, Journal of Engineering Science 12 (6) (2017) 1475–1487, https://doi.org/10.1016/j.jksuci.2018.05.010.

[20] F.S. Al-Anzi, D. AbuZeina, Toward an enhanced Arabic text classification using cosine similarity and latent semantic indexing, Journal of King Saud University-Computer and Information Sciences 29 (2) (2017) 189–195, https://doi.org/10.1016/j.jksuci.2016.04.001.

[21] E. Jadon, R. Sharma, Data mining: document classification using naive Bayes classifier, International Journal of Computer Applications 167 (6) (2017) 13–16, https://doi.org/10.5120/ijca2017913925.

[22] 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, Autom. Constr. 62 (2016) 45–56, https://doi.org/10.1016/j.autcon.2015.11.001.

[23] A. Chokor, H. Naganathan, W.K. Chong, M.E. Asmar, Analyzing Arizona OSHA injury reports using unsupervised machine learning, Procedia Engineering 145 (2016) 1588–1593, https://doi.org/10.1016/j.proeng.2016.04.200.

[24] J. Suto, S. Oniga, Efficiency investigation from shallow to deep neural network techniques in human activity recognition, Cogn. Syst. Res. 54 (2019) 37–49, https://doi.org/10.1016/j.cogsys.2018.11.009.

[25] P. Tao, L. Lu, W. Zuo, PU text classification enhanced by term frequency–inverse document frequency-improved weighting, Concurrency and computation: practice and experience 26 (3) (2014) 728–741, https://doi.org/10.1002/cpe.3040.

[26] J. Piper, E. Granum, On fully automatic feature measurement for banded chromosome classification, Cytometry: The Journal of the International Society for Analytical Cytology 10 (3) (2010) 242–255, https://doi.org/10.1002/cyto.990100303.

[27] M.G. Yang, C.U. Ubeynarayana, Construction accident narrative classification: an evaluation of text mining techniques, Accid. Anal. Prev. 108 (2017) 122, https://doi.org/10.1016/j.aap.2017.08.026.

[28] G. Forman, BNS feature scaling: An improved representation over tf-idf for svm text classification, Acm Conference on Information and Knowledge Management, ACM, 2008, pp. 263–270, , https://doi.org/10.1145/1458082.1458119.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, G. Louppe, Scikit-learn: machine learning in python, Machine Learning Research 12 (10) (2013) 2825–2830, https://doi.org/10.1524/auto.2011.0951.

[30] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828, https://doi.org/10.1109/TPAMI.2013.50.

[31] J. Mackinlay, P. Hanrahan, C. Stolte, Show me: automatic presentation for visual analysis, IEEE Trans. Vis. Comput. Graph. 13 (6) (2007) 1137–1144, https://doi.org/10.1109/TVCG.2007.70594.

[32] S.D. Robinson, Visual representation of safety narratives, Saf. Sci. 88 (2016) 123–128, https://doi.org/10.1016/j.ssci.2016.05.005.

[33] Y. Zhu, Z. Wen, P. Wang, Z. Peng, A method of building Chinese basic semantic lexicon based on word similarity, 2009 Chinese Conference on Pattern Recognition, IEEE, 2009, pp. 1–4, , https://doi.org/10.1109/CCPR.2009.5344041.

[34] H. Jelodar, Y. Wang, Y. Chi, F. Xia, X. Jiang, Y. Li, L. Zhao, Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey, Multimed. Tools Appl. (2017) 1–43, https://doi.org/10.1007/s11042-018-6894-4.

[35] A. Gledson, J. Keane, Using web-search results to measure word-group similarity, Proceedings of the 22nd International Conference on Computational Linguistics, Volume 1 Association for Computational Linguistics, 2008, pp. 281–288, , https://doi.org/10.3115/1599081.1599117.

[36] L.M. Steacy, D.L. Compton, Examining the role of imageability and regularity in word reading accuracy and learning efficiency among first and second graders at risk for reading disabilities, J. Exp. Child Psychol. 178 (2019) 226–250, https://doi.org/10.1016/j.jecp.2018.09.007.

[37] S. Heymann, B. Le Grand, Visual analysis of complex networks for business intelligence with gephi, 2013 17th International Conference on Information Visualisation, IEEE, 2013, pp. 307–312, , https://doi.org/10.1109/IV.2013.39.

[38] B. Benjamin, Shifting the focus of strategic occupational injury prevention: mining free-text, workers compensation claims data, Saf. Sci. 46 (1) (2008) 1–21, https://doi.org/10.1016/j.ssci.2006.09.006.

[39] A. Bernabe, E. Martina, J. Alvarez-Ramirez, C. Ibarra-Valdez, A multi-model approach for describing crude oil price dynamics, Physica A: Statistical Mechanics and Its Applications 338 (3–4) (2014) 567–584, https://doi.org/10.1016/j.physa.2004.03.007.

[40] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, H. Hao, Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification, Neurocomputing 174 (2016) 806–814, https://doi.org/10.1016/j.neucom.2015.09.096.

[41] J. Liu, W.-C. Chang, Y. Wu, Y. Yang, Deep learning for extreme multi-label text classification, Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2017, pp. 115–124, , https://doi.org/10.1145/3077136.3080834.

[42] L. Zhu, G. Wang, X. Zou, A study of Chinese document representation and classification with Word2vec, 2016 9th International Symposium on Computational Intelligence and Design (ISCID), 1 IEEE, 2016, pp. 298–302.

[43] B.S. Kim, S. Chang, Y. Suh, Text analytics for classifying types of accident occurrence using accident report documents, Journal of the Korean Society of Safety 33 (3) (2018) 58–64, https://doi.org/10.14346/JKOSOS.2018.33.3.58.

[44] T.D. Price, M.M. Bouvier, The evolution of F1 postzygotic incompatibilities in bieds, Evolution 56 (10) (2002) 2083–2089, https://doi.org/10.1111/j.0014-3820.2002.tb00133.x.

[45] T. Iwata, T. Yamada, N. Ueda, Modeling noisy annotated data with application to social annotation, IEEE Trans. Knowl. Data Eng. 25 (7) (2013) 1601–1613, https://doi.org/10.1109/TKDE.2012.96.

[46] H. Fan, L. Chufan, G. Hanqi, S. Enya, Y. Xiaoru, L. Sikun, FLDA: latent dirichlet allocation based unsteady flow analysis, IEEE Trans. Vis. Comput. Graph. 20 (12) (2014) 25–45, https://doi.org/10.1109/TVCG.2014.2346416.

[47] B. Ruhnau, Eigenvector-centrality — a node-centrality? Soc. Networks 22 (4) (2000) 357–365, https://doi.org/10.1016/S0378-8733(00)00031-9.

[48] D. Jatnika, M.A. Bijaksana, A.A. Suryani, **Word2Vec model analysis for semantic similarities in english words**, Procedia Computer Science 157 (2019) 160–167, https://doi.org/10.1016/j.procs.2019.08.153.

[49] J.J. Thomas, K.A. Cook, A visual analytics agenda, IEEE Computer Graphics Applications 26 (2006) 10–13, https://doi.org/10.1109/MCG.2006.5.