Accident Analysis and Prevention 108 (2017) 122–130

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Construction accident narrative classification: An evaluation of text mining techniques

Yang Miang Goh*, C.U. Ubeynarayana

Safety and Resilience Research Unit (SaRRU), Dept. of Building, School of Design and Environment, National Univ. of Singapore, 4 Architecture Dr., 117566, Singapore

ARTICLE INFO

Keywords:
Accident classification
Construction safety
Data mining
Support vector machine
Text mining

ABSTRACT

Learning from past accidents is fundamental to accident prevention. Thus, accident and near miss reporting are encouraged by organizations and regulators. However, for organizations managing large safety databases, the time taken to accurately classify accident and near miss narratives will be very significant. This study aims to evaluate the utility of various text mining classification techniques in classifying 1000 publicly available construction accident narratives obtained from the US OSHA website. The study evaluated six machine learning algorithms, including support vector machine (SVM), linear regression (LR), random forest (RF), k-nearest neighbor (KNN), decision tree (DT) and Naive Bayes (NB), and found that SVM produced the best performance in classifying the test set of 251 cases. Further experimentation with tokenization of the processed text and non-linear SVM were also conducted. In addition, a grid search was conducted on the hyperparameters of the SVM models. It was found that the best performing classifiers were linear SVM with unigram tokenization and radial basis function (RBF) SVM with uni-gram tokenization. In view of its relative simplicity, the linear SVM is recommended. Across the 11 labels of accident causes or types, the precision of the linear SVM ranged from 0.5 to 1, recall ranged from 0.36 to 0.9 and F1 score was between 0.45 and 0.92. The reasons for misclassification were discussed and suggestions on ways to improve the performance were provided.

1. Introduction

Workplace safety and health is a major concern in the construction industry in many countries (Zhou et al., 2015). To improve the industry's safety and health performance, the industry needs to learn from past accidents effectively (Chua and Goh 2004). However, accident reports are typically unstructured or semi-structured free-text data that require significant manual classification before statistical analyses can be conducted to facilitate interventions. These classification tasks are typically conducted at organizational and national levels. Due to the resource-intensiveness of the classification process, significant amount of resources need to be spent on classifying accident narratives, but the consistency of the classification is hard to be ascertained. On the other hand, organizations that choose not to classify accident narratives, will suffer loss of precious data for learning and accident prevention.

There had been an increased interest in automatic classification or auto-coding of accident narratives through the application of text mining techniques. These studies typically aim to improve the consistency, productivity and efficiency of accident narrative classification (e.g. Chen et al., 2015b; Marucci-Wellman et al., 2011; McKenzie et al., 2010b; Taylor et al., 2014; Tixier et al., 2016; Vallmuur 2015; Vallmuur et al., 2016). The results appear to be promising, but there are concerns that the success of automatic classification of accident narratives is very sensitive to the dataset and the effectiveness of classification algorithms may not be consistent across different datasets. There is also a wide range of text mining techniques and the usefulness of different techniques in the context of accident narrative classification need to be evaluated. Even though automatic classification of accident narratives does not generate new knowledge per se, it may be argued that with higher efficiency, more incident data can be collected and more detailed analytics can be conducted to produce useful insights that would not be available when fewer incidents were classified by human coders.

This study aims to evaluate the utility of various text mining classification techniques in classifying publicly available accident narratives obtained from the US OSHA website (Occupational Safety and Health Administration, 2016). This study also contributes to future studies on accident narrative classification by making available a dataset of 4470 construction accident narratives to other researchers (see Appendix A). The dataset includes 1000 narratives labelled in this study and 3470 narratives that were not labelled. The subsequent sections provide an overview of current text mining research on accident narratives, the text data that were used in this study, an overview of the

* Corresponding author.
E-mail address: bdggym@nus.edu.sg (Y.M. Goh).

http://dx.doi.org/10.1016/j.aap.2017.08.026
Received 15 May 2017; Received in revised form 24 July 2017; Accepted 26 August 2017
Available online 01 September 2017
0001-4575/

text mining techniques implemented in this study, the results of the evaluation, and discussion and recommendations for future research on text mining of accident narratives.

2. Literature review

2.1. Text mining techniques

Text mining is a well-researched field. One of the common tasks in text mining is classification of text data (Sebastiani 2002). Text classification is the task of assigning one or more class labels to a document using a predefined set of classes or labels. The supervised machine learning approach to text classification relies on an initial set of corpus (or collection of documents) with known class labels. This corpus is split into training and testing datasets in order to train and then ascertain the performance of the classifier. The classifier is trained by observing the characteristics of the training dataset through different machine learning algorithms.

Data for text classification is typically represented using vector space model. In this model, each document is represented as a vector of terms. Another way to look at these terms is that they are essentially a bag of words (Bird et al., 2009). Terms are features that represent a document, which could be a single word, phrase, or string. To distinguish between documents in a corpus, each feature for each document is given numeric values to show the importance of that term to the document (Keikha et al., 2008).A commonly used vector space model is the term frequency–inverse document frequency (tf-idf) representation (Peng et al., 2014). In the tf-idf representation, values, or x_{ik} weights, reflecting the importance of each given feature of a document is given by

$$x_{ik} = f_{ik} \times \log\left(\frac{N}{n_i}\right) \tag{1}$$

where f_{ik} is the frequency of feature i in document k, N is the number of documents in the corpus, and n_i is the number of documents where feature i occurs. Once the document is represented using a suitable vector space representation model, the data can be trained and classified using typical data mining techniques such as decision tree, neural network, support vector machine and Bayesian network (Raschka 2015; Witten 2011).

2.2. Performance metrics

This study adopts the use of recall, precision and F1 score (or F-measure) (Buckland and Gey 1994) to evaluate the performance of the machine learning algorithms experimented. Table 1 and Equations (2) to (4) define these metrics. Essentially, precision is a measure of how accurate the positive predictions are and recall is a measure of how many of the actual positives the model can identify (Williams 2011). F1 score combines precision and recall to provide an overall assessment of performance of the classifier. As these metrics are widely used and discussed in the literature, readers can refer to text mining or machine learning textbooks (e.g. Bird et al., 2009; Witten 2011) for their detailed description.

$$Precision = \frac{TP}{(TP + FP)} \tag{2}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{3}$$

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Table 1
True and false positives and negatives (adapted from Bird et al. (2009)).

	Relevant	Irrelevant
Retrieved	True Positives (TP)	False Positives (FP) (Type I error)
Not retrieved	False Negatives (FN) (Type II error)	True Negatives (TN)

2.3. Past studies on accident narrative classification

There were several other studies that applied text mining techniques in the analysis of injury narratives. In Chen et al. (2015a), the study aimed to automatically classify narratives in emergency room medical reports into common injury cause codes. The authors argued that injury narratives have unique characteristics that make them different from general documents and a detailed experiment is needed to evaluate the usefulness of different text mining techniques for their dataset. It was found that the use of matrix factorization coupled with support vector machine (SVM), gave the best classification performance. The authors reported recall ranging from 0.48 to 0.94 and precision ranging from 0.18 to 0.95 for different classification labels. McKenzie et al., (2010a) also attempted to classify emergency department injury narratives for the purpose of injury surveillance to support an evidence-based public health response. The study compared keyword search, index search, and text mining. Text mining was conducted using a content text mining software, Leximancer (Leximancer Pty Ltd, 2016), and it was found that text mining approach provided the best performance. Bertke et al. (2012) made use of Naïve Bayesian classifiers to classify workers' medical compensation claims into three "claim causation" categories, i.e. musculoskeletal disorder (MSD), slip trip fall (STF), or others (OTH). The study found that the Naïve Bayesian classifier was able to achieve "approximately 90% accuracy" for MSD, STF and OTH classifications. However, it was observed that when OTH was being broken up into lower level classifications, the performance of the classifier dropped significantly.

Tanguy et al. (2015) evaluated an aviation safety report corpus, which contains 136,861 documents, using support vector machine. As part of the pre-processing and experimental design, numerous forms of text units were created. Some of the text units explored include word, word stems (e.g. "falling" is converted to its word stem, "fall"), and N-grams of words and stems. An N-gram refers to a set of N adjacent words (Bird et al., 2009). The study found that use of bi-gram and tri-gram of stemmed narratives produced the best results in their preliminary classifications. They constructed a binary classifier for each target label (e.g. air traffic management, bird strike, runway excursion and glider towing related event) and that means 37 classifiers were trained. However, the authors only reported the results for seven of the classifiers, the precision ranged from 0.6 to 0.96, recall was 0.36–0.93, and F1 score was 0.45–0.95. The authors highlighted that the performance for each classifier is dependent on issues such as "rarity, difficulty and inconsistency" of the text data.

Taylor et al. (2014) trained Fuzzy and Naïve Bayesian models to assign mechanism of injury and injury outcome for a set of fire-related near miss narratives obtained from the National Firefighter Near-Miss Reporting System. Their algorithms achieved sensitivity (same as recall) of between 0.602 and 0.74. Taylor et al. (2014) also made a comparison with five other studies and claimed that their findings are "are comparable with the growing body of seminal studies on narrative autocoding".

For the construction industry, there were several studies that utilized text mining approaches in areas such as dispute resolution (Fan and Li 2013), cost overrun (Williams and Gong 2014) document retrieval (Yu and Hsu 2013) and classification of field inspection records (Chi et al., 2016). Specifically in the domain of construction accident narrative classification, Tixier et al. (2016) made use of a customized term lexicon (keyword dictionary) as well as a set of rules to automatically classify construction incident narratives. The study was conducted on a dataset with 2201 accident narratives. The lexicons were used to reduce the wide range of terms into a smaller set of keywords and rules were then manually crafted to improve the classification performance of the approach. The authors reported a F1 score of 0.95; which is significantly better than several text mining papers that they reviewed. Even though hand crafting of rules has the advantage of being more accurate; there are concerns that the rule-creation process is very tedious. The approach may not be sustainable when the rule set grows in proportion to the size of the dataset. In addition; the generalizability of the rules will need to be evaluated against other datasets.

It can be observed that despite the successes of machine learning algorithms in classifying emergency department narratives, there is a lack of study on the use of machine learning algorithms in classifying narratives in construction accident reports. Thus, this study aims to experiment with a variety of text mining machine learning algorithms and evaluate their potential in automatically classifying accident narratives.

3. Dataset

3.1. Data source

The dataset used in this study is based on accident narratives collected from the US OSHA website (Occupational Safety and Health Administration, 2016). A dataset of 16,323 accident records (occurring between 1983 and 2013) was downloaded and a dataset consisting of 4471 construction industry cases were compiled based on the Industry Code (SIC code) (Goh 2016).For the purpose of this study, only the title and accident narrative information were used to classify the cases.

Fig. 1 shows a sample record, which includes the title, narrative and label. It should be noted that the original data was not labelled and the labels were created during data pre-processing.

3.2. Data pre-processing

One thousand cases were manually labelled based on the labelling rules described in Table 2. These labels were based on the classifications used in Workplace Safety and Health Institute (2016). Since more than one event can occur during an accident, the labelling of cases followed the principle of identifying the first occurrence of uncontrolled or unintended action. The number of possible labels was also reduced by combining labels with similar meanings. For example, in the case of Fig. 1, the narrative could have been labelled as “Struck by flying object”, but was classified as “Struck by moving object”, which is more general. Another example is the combination of “falling from height” and “fall on same level” into “falls”.

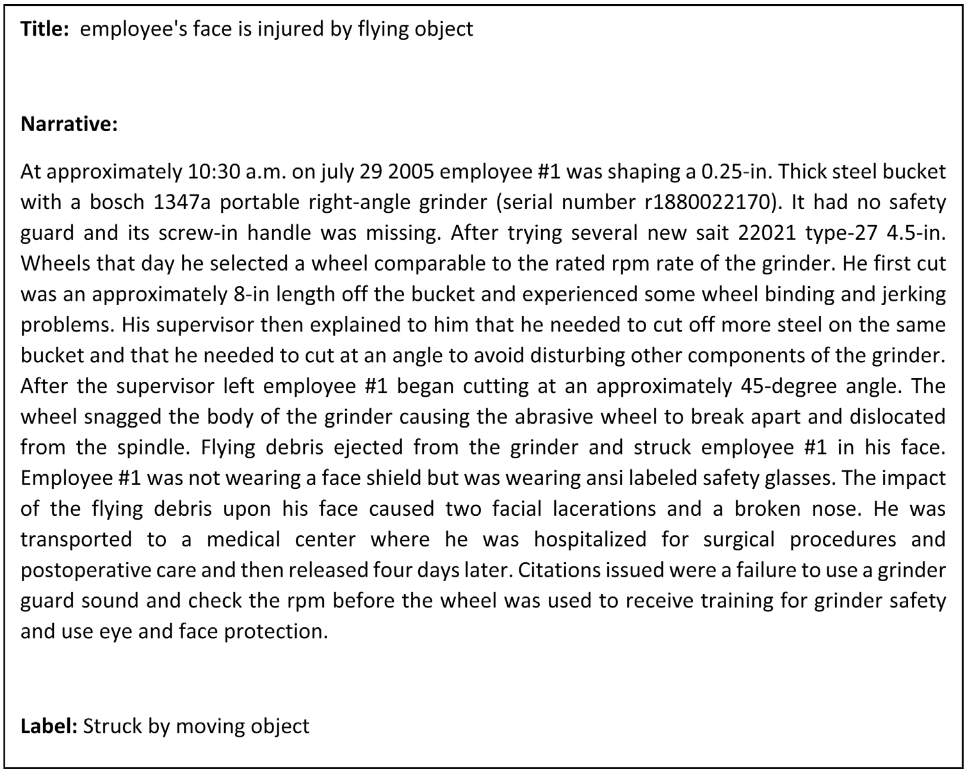
The narrative and title were then combined to form a single paragraph of text. Stopwords, which are high-frequency words with little lexical content (Bird et al., 2009), were removed. Words could exist in different forms, but are semantically similar, e.g. plural, and past tense forms. The use of several different forms of the same word during training and can lead to poorer performance. Thus, word stemming was carried out using the Snowball Stemmer (Shibukawa 2013), to obtain word stems in which the different forms of the words is aggregated to a single type. As an example, Fig. 2 is the processed text of the narrative in Fig. 1. Fig. 3 shows the word count for the original text and processed text for each label. It can be observed that the number of words in the original text had been compressed after processing. Fig. 3 also shows that the word count for the text narratives is varied and there are many outliers for each label which may cause difficulties for the classifiers. The outliers, in this context, refer to a narrative with word count that lies outside the expected range of word count. The expected word count range for each label is defined as

$$[(Q1 - 1.5 \times IQR), (Q3 + 1.5 \times IQR)]$$

where Q1 is the 25th percentile, Q3 is the 75th percentile and IQR is the inter-quartile range.

4. Experiments conducted

Prior to experimentation, each stemmed narrative was broken into strings of N-consecutive words, N-grams, in a process called



Title: employee's face is injured by flying object

Narrative:

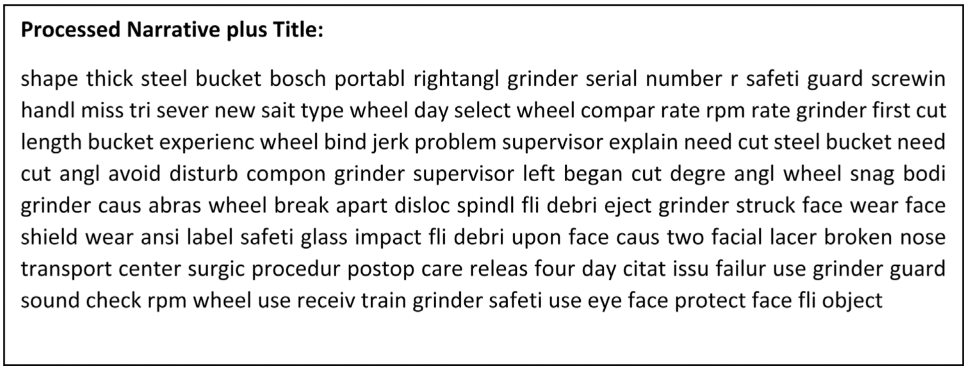
At approximately 10:30 a.m. on july 29 2005 employee #1 was shaping a 0.25-in. Thick steel bucket with a bosch 1347a portable right-angle grinder (serial number r1880022170). It had no safety guard and its screw-in handle was missing. After trying several new sait 22021 type-27 4.5-in. Wheels that day he selected a wheel comparable to the rated rpm rate of the grinder. He first cut was an approximately 8-in length off the bucket and experienced some wheel binding and jerking problems. His supervisor then explained to him that he needed to cut off more steel on the same bucket and that he needed to cut at an angle to avoid disturbing other components of the grinder. After the supervisor left employee #1 began cutting at an approximately 45-degree angle. The wheel snagged the body of the grinder causing the abrasive wheel to break apart and dislocated from the spindle. Flying debris ejected from the grinder and struck employee #1 in his face. Employee #1 was not wearing a face shield but was wearing ansi labeled safety glasses. The impact of the flying debris upon his face caused two facial lacerations and a broken nose. He was transported to a medical center where he was hospitalized for surgical procedures and postoperative care and then released four days later. Citations issued were a failure to use a grinder guard sound and check the rpm before the wheel was used to receive training for grinder safety and use eye and face protection.

Label: Struck by moving object

Fig. 1. Sample accident narrative and title.

Table 2
Labels used for labelling the dataset, their criteria and frequency.

	Case Labels	Description	Count	%
1	Caught In/Between Objects	Fracture, cuts, lacerations, amputations caused by being caught in between objects, generally referring to hand tools	68	6.8%
2	Collapse of Object	Involving cases that started with structural failure	212	21.2%
3	Electrocution	Direct electric shock or any burns caused by electrical faults	108	10.8%
4	Exposure to Chemical Substances	Contact with any toxic/corrosive chemical substances	29	2.9%
5	Exposure to Extreme Temperatures	Extreme temperatures caused by frost, hot liquid or gases, including hypothermia	17	1.7%
6	Falls	Involving slip, trip cases and where the victim is falling from elevation but not due to structural failure	236	23.6%
7	Fires and Explosion	Injuries caused by direct fires and explosion but not electrical burns	47	4.7%
8	Struck by Falling Object	victim is hit by falling object from height but object is not due to structural failure	43	4.3%
9	Struck by Moving Objects	The victim is hit by a moving object (that is not in free fall)	134	13.4%
10	Traffic	Injury happen due to worker's driving a vehicle or a moving vehicle hits worker.	63	6.3%
11	Other	To accommodate cases that do not fall in the above categories. Some of the less occurring categories were merged into others as the number of occurrences was too low (drowning, suffocation)	43	4.3%
		Total	1000	100.0%



Processed Narrative plus Title:

shape thick steel bucket bosch portabl rightangl grinder serial number r safeti guard screwin handl miss tri sever new sait type wheel day select wheel compar rate rpm rate grinder first cut length bucket experienc wheel bind jerk problem supervisor explain need cut steel bucket need cut angl avoid disturb compon grinder supervisor left began cut degre angl wheel snag bodi grinder caus abras wheel break apart disloc spindl fli debri eject grinder struck face wear face shield wear ansi label safeti glass impact fli debri upon face caus two facial lacer broken nose transport center surgic procedur postop care releas four day citat issu failur use grinder guard sound check rpm wheel use receiv train grinder safeti use eye face protect face fli object

Fig. 2. Sample processed text.

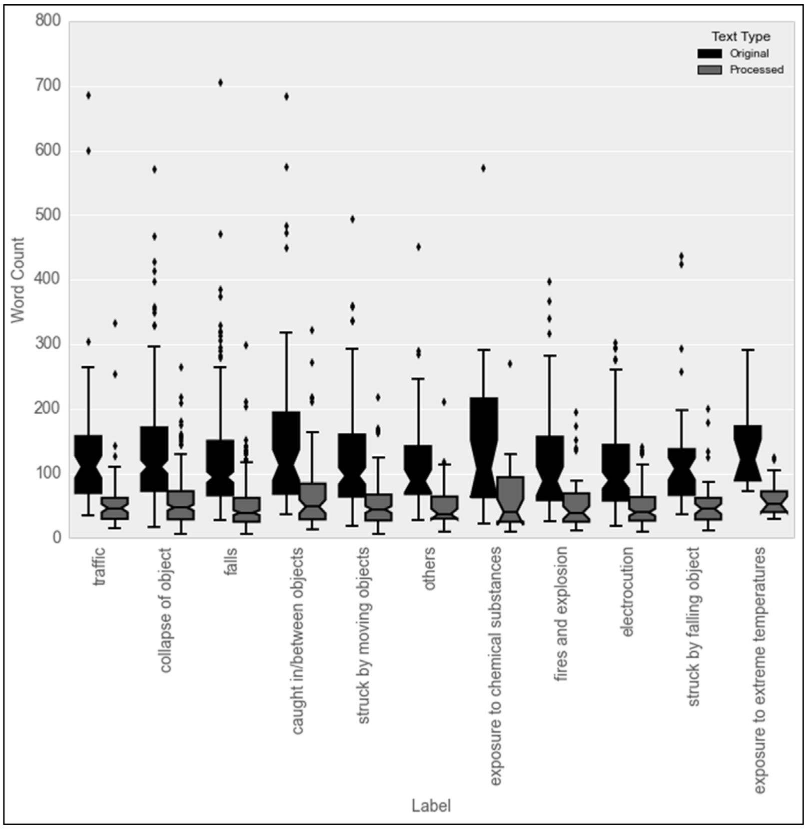


Fig. 3. Both box plots for word counts on 11 class labels.

tokenization (Bird et al., 2009). Tokenization is the process of breaking up a sentence of strings into pieces such as words, phrases which are then referred to as ‘tokens’. Tokens are used as the input data in text mining. Uni-gram tokenization breaks a longer string into one word tokens, bi-gram is a token consisting of two words and tri-gram is a token with three words. A corpus of documents can thus be represented by a matrix with one row per document and one column per token (e.g. one word). The general process of turning a collection of text documents into a set of numerical feature vectors is called vectorization. Tokenization is also called the “n-grams representation” (scikit-learn Community, 2016). Uni-grams and bi-grams, which consist of one and two, word length tokens, were used in this experiment. These were then converted into the tf-idf document term matrix representation (see Equation (1)). This tf-idf matrix, coupled with the case labels, was then used to train the classifiers.

Using stratified sampling, about 25% of the 1000 labelled cases (251) were set aside as an independent test set to compare the performance of the different approaches evaluated. The open-source algorithms used in this study were primarily derived from Python 2.7 (Python Software Foundation, 2016), scikit-learn library version 0.17 (scikit-learn Community, 2016) and Natural Language Toolkit (nltk) library version 3.1 (Bird et al., 2009). A preliminary experiment was conducted on the dataset to identify the best performing learning algorithm to focus on. The preliminary multinomial classification experiment included six machine learning classifiers: support vector machine (SVM), linear regression (LR), random forest (RF), k-nearest neighbor (KNN), decision tree (DT) and Naive Bayes (NB). SVM and NB were used and recommended in previous studies and the remaining algorithms were arbitrarily selected. Each of these algorithms are well established and are carefully described in machine learning textbooks (e.g. Bishop 2006). Thus, the remainder of this paragraph only provides brief descriptions of those algorithms. SVM algorithm creates hyperplanes based on training data and then optimizes hyperplanes, which are the basis for classifying data points among classes. Logistic regression algorithm predicts the probability of occurrence of an event by fitting data to a logit function. Naïve Bayes algorithm uses Bayes rule with strong independent assumptions between features. It simplifies the calculation of probabilities by assuming that the probabilities of each feature belonging to a given class label are independent of all other features. K nearest neighbor classifier is a simple algorithm that stores all available class labels and classifies new class labels based on a distance vector when there is no prior knowledge about the underlying distribution of the data. A decision tree starts with a root node, where all the training data is located. An initial split is made using a test on a feature thus, separating the data into child nodes. Further splitting can be made to the child nodes depending on the outcome of multiple tests. The branches signify the result of the test on a feature and the leaf nodes represent the class labels. The nodes where the final class labels are fixed is known as the terminal node. A random forest classifier constitutes a set of decision trees. Each tree in the forest predicts their final class label. The collection of trees then voted for the most popular class as the final class label.

Throughout all the experiments, the same set of training cases was used to train each machine learning classifier independently to label each narrative to the 11 labels in Table 2. The classifiers were then tested on the test set and the best performing classifier was used in subsequent experiments aiming to improve the performance of the classifier.

Consistent with other studies like Tanguy et al. (2015) and Chen et al. (2015b), the SVM classifier (Burges 1998) was the best performer during the preliminary experiment. The initial SVM was set up as a One-vs-Rest (OVR) linear classifier with parameters C = 10 and “balanced” class weight (scikit-learn Community, 2016). A OVR technique extends a binary classifier for a multi-class problem by training one classifier for each label or class (Raschka 2015). The “balanced” mode automatically adjusts weights inversely proportional to the frequencies of the labels and this helps to account for the imbalance frequency of different labels (see Table 2). Subsequently, a grid search function (scikit-learn Community, 2016) was implemented to identify the suitable tuning parameters or hyperparameters by implementing a “brute-force exhaustive search” (solving a problem in which by analyzing nearly all possible solutions) (Raschka 2015)through many models in sequence and returning the best model as measured by a pre-determined performance metric.

In the linear SVM model, the C value is the key hyperparameter that needs to be tuned. The C parameter is essentially a regularization parameter, which determines how much the classifier adapts to the training samples (scikit-learn Community, 2016). C is the parameter, which helps to strengthen the misclassification of training data. Lower values of parameter C will smoothen the decision surface with longer running times. Higher values of parameter C will correctly classify the training data with lesser the running times. Seven C values on the log scale ranging from 10^{-3} to 10^{3}were used and the selection was based on performance measured using F1 score. In addition, ten-fold cross-validation was utilized during the search. The same test set was then used to measure the performance of the SVM model. During the linear SVM experimentation, different combinations of uni-gram, bi-gram and tri-gram were experimented with. Finally, the radial basis function (RBF), a non-linear function, was experimented with search function based on variations in the C and gamma parameters. The gamma parameter has similar function as the C parameter and it can be seen as the “inverse of the radius of influence of samples selected by the model as support vectors” (scikit-learn Community, 2016). Nonlinear classifications can be handled with the help of the parameter gamma. The parameter gamma belongs to the Gaussian kernel family, a small gamma indicates a Gaussian kernel with large variance, where as a large gamma reveals a Gaussian kernel with low variance. More importantly the kernel function in SVM classifier contains all the information about the relative positions of the input data in the feature space. The actual learning algorithm of SVM based only on the kernel function and thus the algorithm can be carried out without explicit use of the feature space.

5. Results and findings

5.1. Preliminary classification

As highlighted earlier, an initial round of classification was conducted using six machine learning classification algorithms, so as to select the best performer for further experimentation. The classification results, as represented by F1 score, are shown in Table 3.

Table 3
F1 scores for preliminary experiment.

	F1 Score					
Labels	SVM	LR	RF	KNN	DT	NB
caught in/between objects	0.60	0.36	0.53	0.34	0.46	***0.89***
collapse of object	***0.66***	0.48	0.54	0.52	0.56	0.25
electrocution	***0.95***	0.91	***0.95***	0.69	0.92	0.67
exposure to chemical substances	***0.62***	0.00	0.00	0.40	0.40	0.40
exposure to extreme temperatures	***0.67***	0.00	0.00	0.40	0.25	0.00
falls	***0.78***	0.63	0.74	0.66	0.76	0.17
fires and explosion	***0.74***	0.56	0.71	0.50	0.64	0.63
others	0.43	0.00	0.29	0.20	***0.48***	0.00
struck by falling object	*0.14*	0.00	0.00	0.11	0.27	***0.61***
struck by moving objects	***0.58***	0.48	0.45	0.44	0.55	0.48
traffic	***0.67***	0.48	0.40	0.54	0.54	0.36
Average	***0.62***	0.35	0.42	0.44	0.53	0.41

SVM – Support Vector Machine, LR – Linear Regression, RF – Random Forest, KNN – K-Nearest Neighbour, DT – Decision Tree, NB-Naïve Bayes; numbers in bold and italics are the highest value for the label.

Table 4
F1 Score for experiments using different tokenisation and optimisation.

	F1 Score					
	SVM–Linear					SVM–RBF
Label	(1,1)	(1,2)	(1,3)	(2,2)	(2,3)	(1,1)
caught in/between objects	***0.61***	0.58	0.58	0.36	0.20	0.56
collapse of object	0.67	0.62	0.60	0.40	0.39	***0.71***
electrocution	0.92	***0.94***	0.93	0.75	0.72	0.90
exposure to chemical substances	0.55	***0.60***	0.25	0.25	0.25	***0.60***
exposure to extreme temperatures	***0.67***	***0.67***	***0.67***	0.00	0.00	***0.67***
falls	***0.83***	0.76	0.78	0.70	0.66	***0.83***
fires and explosion	0.80	0.80	0.75	0.47	0.27	***0.83***
others	***0.45***	0.29	0.17	0.15	0.17	***0.45***
struck by falling object	***0.52***	0.29	0.15	0.17	0.17	0.50
struck by moving objects	0.57	0.60	***0.62***	0.44	0.37	0.57
traffic	***0.76***	0.62	0.59	0.42	0.36	***0.76***
Average	***0.67***	0.61	0.55	0.37	0.32	***0.67***

SVM − Support Vector Machine; RBF − radial basis function; (1,1) − uni-gram; (1,2) − uni-gram and bi-gram; (1,3) − uni-gram and tri-gram; (2,2) − bi-gram; (2,3) − bi-gram and tri-gram; numbers in bold and italics are the highest value for the label.

A best macro average F1 score, or the average F1 score for all labels, of 0.62 was achieved using the SVM classifier. Even though the second best classifier (in accordance to average F1 score) was DT, NB classifier performed very well for two of the labels, “caught in/between objects” and “struck by falling object”. However, NB performed very poorly for most of the other labels.

5.2. Tokenization and optimization

With reference to Table 4, another six experiments were conducted to improve the performance of the SVM classifier. The linear SVM with optimization was conducted by varying different tokenization of the processed text. The variations include use of uni-gram, bi-gram and combinations of uni-gram, bi-gram and tri-gram. As can be observed from Table 4, uni-gram gave the highest average F1 score.

In addition, a RBF SVM with uni-gram tokenization was trained and tested. As can be seen from Table 4, among the six classifiers in Table 4, the RBF SVM classifier had the highest F1 score for seven of the 11 labels, while the linear SVM had the highest F1 score for six labels. Thus, the overall performance of the non-linear SVM is similar to the linear SVM with uni-gram. In accordance to the Occam's Razor, the linear SVM with unigram tokenization is recommended due to its relative simplicity. Table 5 shows the precision, recall and F1 score of the best performing linear SVM model.

Table 5
Precision, recall and F1 score of best performing SVM model.

Label	Precision	Recall	F1 Score
electrocution	***1.00***	***0.85***	***0.92***
exposure to chemical substances	***0.75***	0.43	0.55
fire and explosion	***0.77***	***0.83***	***0.80***
struck by moving objects	0.53	0.62	0.57
exposure to extreme temperatures	***1.00***	0.50	0.67
others	0.57	0.36	0.45
falls	***0.77***	***0.90***	***0.83***
struck by falling object	0.50	0.55	0.52
traffic	***0.85***	***0.69***	***0.76***
collapse of object	0.67	***0.66***	0.67
caught in/between objects	0.63	0.59	0.61
Average	0.73	0.63	0.67

Values above average were highlighted in bold and italics.

5.3. Misclassification

For the linear SVM model, misclassification is less likely in narratives labelled as ‘electrocution’, ‘falls’, ‘fire and explosion’, and ‘traffic’, but more likely in cases with labels, ‘others’, ‘struck by falling object’, ‘exposure to chemical substances’ and ‘struck by moving objects’. The commonly mislabeled cases were evaluated by creating a confusion matrix, and qualitative evaluation of the mislabeled cases were recorded in Table 6.

6. Discussion

MisclassificationOne of the key reasons for the misclassification was probably due to the tf-idf vector space model not being able to capture the context in which the words are used in the incident. For example, the tf-idf vector space does not retain any information about the order of terms after it is converted into the tf-idf matrix. Even though the tokenization of terms into bi-gram or tri-gram may help to retain some of these context, the experimentation conducted shows that uni-gram provided the best results. Thus, the classifier may give too much attention to terms that do not provide clear linkage to the direct cause of the accident. Another key challenge for classification of accident narratives is that the narrative might be overly focused on aspects not directly related to the cause of the accident. For instance, for the narrative in Fig. 4, the number of words describing the work environment was excessive in relation to the number of words describing the actual direct cause of the accident, ‘fall into the trench’.

Some labels were fairly tricky to classify even for human readers, in particularly the labels, “collapse of object”, “struck by falling object”, “struck by moving object”. This is because natural language is not precise and there are different ways to word a sentence, giving rise to different interpretation. For instance, in Fig. 5, it is not very clear whether it is a case of being struck by a moving object, or was the victim caught in between objects. In this particular case, the label was given as “struck by moving object”. The inherent difficulty in classifying accident narrative was also discussed in Tanguy et al. (2015), who found that significant portion of their corpus were misclassified by the human classifiers.

Despite the credible performance of the linear SVM for some of the categories, it is not possible to automatically label all the narratives without human intervention. If a machine learning approach is to be taken for labelling of accident narratives, it is suggested that the autocoding of narratives be done only for labels with good F1 score, e.g. ‘electrocution’ and ‘falls’. Labels with moderate level of F1 scores, e.g. ‘exposure to extreme temperatures’ and ‘collapse of object’, can be autocoded, but additional checks by human classifiers will be needed.

6.1. Industry implications

The proposed text mining approach has the ability to support an informed culture (Reason 1997) in construction organizations. As argued by Reason (1997), an informed culture, “one in which those who manage and operate the system have current knowledge about the human, technical, organizational and environmental factors that determine the safety of the system as a whole”, is synonymous to safety culture. An informed culture is dependent on an effective “safety information system” that is able to effectively and efficiently capture, analyze and disseminate information from incidents, near misses and proactive checks on the safety management system. To materialize the safety information system and informed culture proposed by Reason (1997), the ability to efficiently and consistently classify incidents and near misses is critical. This is especially challenging in the construction industry, which consists of temporary and resource-scarce projects that may not be prepared to develop and maintain a safety information system.

It is proposed that the construction industry should create a web-

Table 6
Qualitative evaluation of commonly mislabeled labels.

	Actual Case Label	Most Commonly Mislabelled Label	Qualitative Evaluation of Samples
1	electrocute	fire and explosion	The narrative focused too much on the effects of electrocution, e.g. burn, flash burn, and fires, which overlaps with fire and explosion
2	exposure to chemical substances	others	In the mislabeled cases, there were a lot of cardio-related terms in the narrative from the forensics diagnosis, which contributed to an'others' category label, and also, the chemical described in the narrative was unique.
3	fire and explosion	electrocution	The narrative in the fire and explosion case described electrical equipment, but the cause of accident was not related to the equipment.
4	struck by moving objects	collapse of object	Object involved in the incident are trees and beams which are not typically associated under the "struck by moving" objects. Also, some objects broke/sheared which were confused with structural collapse.
5	exposure to extreme temperatures	fire and explosion	The mislabeled cases were mostly about scalds and burns, while the correctly labelled extreme temperature cases were about heat strokes.
6	collapse of object	falls	Ladder "collapse" leading to a fall is common problem.
7	falls	collapse of object	The victim was in an environment that was frequently associated with structural collapse e.g. trench. However, the fall incident does not have any relation with the environment.
8	struck by falling object	collapse of object	The classifier was unable to distinguish between objects falling as a result of structural failure or improper securing
9	traffic	collapse of object	Some seldom-seen vehicles e.g. train and bus, were not linked to 'traffic' cases. In some cases, vehicle overturning was associated with 'collapse of object' than 'traffic'
10	others	falls	'others' is not a well-defined category and it has no strong defining features
11	caught in/between objects	struck by moving objects	Classifier was unable to pick up the cause; possibly due too much noise from other labels.

based safety information system to allow construction companies and their employees to submit descriptions of incidents and near misses anonymously. It is believed that if the data is collected industry wide, a large pool of data can then be collected to understand construction hazards more effectively; this will encourage reporting, learning and informed culture, which will lead to better safety performance in the long run. To manage the large pool of text data, the proposed text mining approach becomes essential. Concurrently, on-going calibration, training and testing of the algorithms will improve the accuracy of the automatic classification systems. The safety information system will then allow other analytics and machine learning techniques to be applied.

6.2. Future work and limitations

The study showed that different classifiers were able to produce better performance for different labels. None of the classifiers were unanimously the best across all labels. Thus, an ensemble approach (Witten 2011), where several classifiers are trained independently and then used collectively to select the best label for each narrative, may be able to produce better classification performance. Another technique that should be explored is the co-training approach (Blum and Mitchell 1998). It is a semi-supervised learning technique that seeks to make use of an additional large number of unlabeled cases, as in the case of this study, and use them to improve the classification performance. To be able to exploit the unlabeled cases would likely to be an advantage in any classification system because labelling cases require additional effort.

A common problem in classification of text, which was not examined in this study, is the excessive number of terms (features) in each set of narrative. To reduce the number of features and possibly improve performance, a set of ontology and lexicon (e.g. Tixier et al., 2016) can be applied to the corpus during pre-processing. Words with similar meaning can be compressed into the same term and allow the learning algorithm to spot patterns across the corpus. Future works in this area could involve the creation of a more domain specific construction-related dictionary to better identify specific terms. A more intelligent pre-processing of the narrative, such as using rule-based methods, could help strip away elaborate narratives that are not related to the incident, and improve classification accuracy. In addition, dimension reduction techniques such as the Latent Semantic Analysis (LSA) (Turney and Littman 2003)can be applied to the corpus. LSA considers the relationship between words in the data and it makes use of the Singular Vector Decomposition (SVD) mathematical technique to produce a new

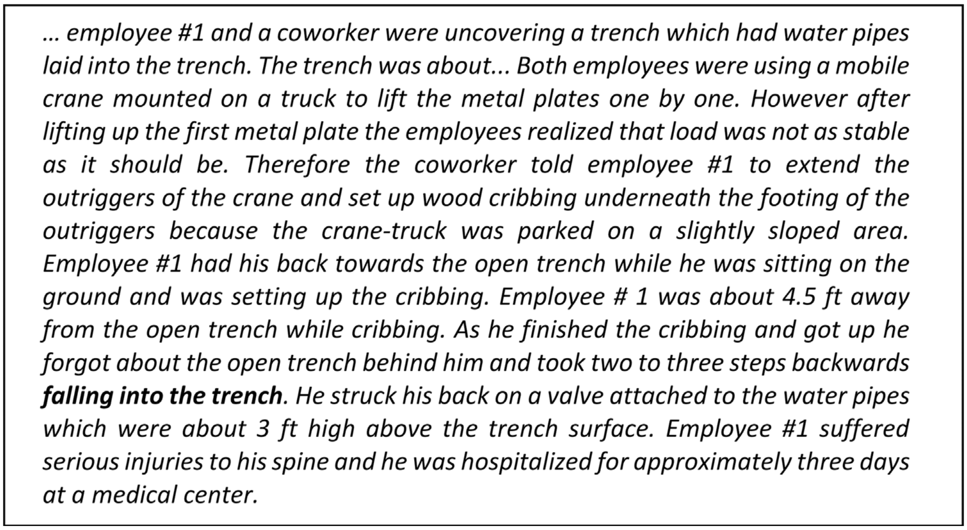
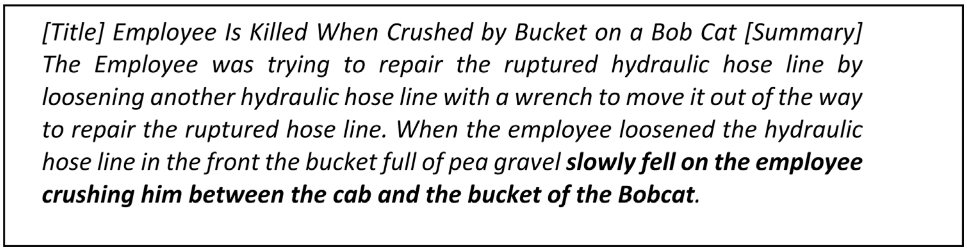

… employee #1 and a coworker were uncovering a trench which had water pipes laid into the trench. The trench was about… Both employees were using a mobile crane mounted on a truck to lift the metal plates one by one. However after lifting up the first metal plate the employees realized that load was not as stable as it should be. Therefore the coworker told employee #1 to extend the outriggers of the crane and set up wood cribbing underneath the footing of the outriggers because the crane-truck was parked on a slightly sloped area. Employee #1 had his back towards the open trench while he was sitting on the ground and was setting up the cribbing. Employee # 1 was about 4.5 ft away from the open trench while cribbing. As he finished the cribbing and got up he forgot about the open trench behind him and took two to three steps backwards ***falling into the trench****. He struck his back on a valve attached to the water pipes which were about 3 ft high above the trench surface. Employee #1 suffered serious injuries to his spine and he was hospitalized for approximately three days at a medical center.*

Fig. 4. Sample text to illustrate how narrative can over-focus on content not related to the label.



[Title] Employee Is Killed When Crushed by Bucket on a Bob Cat [Summary] The Employee was trying to repair the ruptured hydraulic hose line by loosening another hydraulic hose line with a wrench to move it out of the way to repair the ruptured hose line. When the employee loosened the hydraulic hose line in the front the bucket full of pea gravel ***slowly fell on the employee crushing him between the cab and the bucket of the Bobcat****.*

Fig. 5. Sample text to illustrate how some narratives can be difficult to classify.

set of concepts that is drawn from the words present in the data.

In addition, this study had not explored the potential inconsistency in classification when the machine learning models are applied across different datasets. Each dataset may have its own characteristics and the machine learning models developed using different dataset may produce different classification when tested with a fresh set of data. It is recommended that further studies be conducted to reduce the inconsistency that may arise when different datasets are used.

7. Conclusions

This study produced comparable results as previous text mining studies that were focused on corpus from other industries (e.g. Tanguy et al., 2015; Taylor et al., 2014). Based on the overall F1 score, this study recommends the use of linear SVM with uni-gram tokenization to automatically code accident narratives, in particular the U.S. OSHA accident database. A suitable alternative was non-linear SVM using radial basis function, but the improvement in F1 score was not significant. Text mining could alleviate the difficulty in labelling large volume of text by selectively labelling a portion of the cases with high confidence and leaving the ones that are harder to label to human classifiers, who would have their workload reduced. Periodic updating of the training set will be required to check and update the classifiers.

It is believed that different corpus has different characteristics and it will be useful for researchers to share the dataset that they have evaluated, so that other researchers can assess the performance of different machine learning techniques on the same dataset. Since application of machine learning approaches in classifying accident narratives provides the foundation for improved ability to learn from past accidents, it is an important area within construction management research. Thus, the dataset collected in this study was made publicly available to motivate other researchers to search for better approaches to automatically label accident narratives (see Appendix A).

Acknowledgement

This study is funded by Singapore Ministry of Education (MOE) Tier 1 grant (Grant No.: R-296-000-161-112). The author acknowledges Mr Xianglong Liu's assistance in pre-processing the data and developing portions of the scripts for this study. Ms Zhenzhen Fan's advice and inputs are also appreciated.

Appendix A

The dataset used in this study can be downloaded from https://github.com/safetyhub/OSHA_Acc.git. Please acknowledge this article, if the dataset is used. The original data were obtained from https://www.osha.gov/pls/imis/accidentsearch.html (Occupational Safety and Health Administration, 2016).

References

Bertke, S.J., Meyers, A.R., Wurzelbacher, S.J., Bell, J., Lampl, M.L., Robins, D., 2012. Development and evaluation of a naive bayesian model for coding causation of workers compensation claims. J. Safety Res. 43.

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python. O'Reilly Media Inc.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York.

Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, ACM. Madison, Wisconsin, USA. pp. 92–100.

Buckland, M., Gey, F., 1994. The relationship between recall and precision. J. Am. Soc. Inf. Sci. (1986–1998) 45 (1), 12.

Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discovery 2 (2), 121–167.

Chen, L., Vallmuur, K., Nayak, R., 2015a. Injury narrative text classification using factorization model. BMC Med. Inform. Decis. Mak. 15 (1), 1–12.

Chen, L., Vallmuur, K., Nayak, R., 2015b. Injury narrative text classification using factorization model. BMC Med. Inf. Decis. Making 15 (1), S5.

Chi, N.-W., Lin, K.-Y., El-Gohary, N., Hsieh, S.-H., 2016. Evaluating the strength of text classification categories for supporting construction field inspection. Autom. Constr. 64, 78–88.

Chua, D.K.H., Goh, Y.M., 2004. Incident causation model for improving feedback of safety knowledge. J. Constr. Eng. Manage. – Am. Soc. Civ. Eng. 130 (4), 542–551.

Fan, H., Li, H., 2013. Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques. Autom. Constr. 34, 85–91.

Goh, Y.M., 2016. Accident Narratives Dataset Obtained from Occupational Safety and Health Administration (OSHA) Fatality and Catastrophe Investigation Summaries. OSHA, U.S.

Keikha, M., Razavian, N.S., Oroumchian, F., Razi, H.S., 2008. Document representation and quality of text: an analysis. In: Berry, M.W., Castellanos, M. (Eds.), Survey of Text Mining II: Clustering, Classification, and Retrieval. Springer, London, pp. 219–232.

Leximancer Pty Ltd, 2016. Leximancer. http://info.leximancer.com/ (April 11, 2016).

Marucci-Wellman, H., Lehto, M., Corns, H., 2011. A combined Fuzzy and Naïve Bayesian strategy can be used to assign event codes to injury narratives. Inj. Prev. 17 (6), 407–414.

McKenzie, K., Campbell, M.A., Scott, D.A., Discoll, T.R., Harrison, J.E., McClure, R.J., 2010a. Identifying work related injuries: comparison of methods for interrogating text fields. BMC Med. Inform. Decis. Mak. 10 (1), 1–10.

McKenzie, K., Scott, D.A., Campbell, M.A., McClure, R.J., 2010b. The use of narrative text for injury surveillance research: a systematic review. Accid. Anal. Prev. 42 (2), 354–363.

Occupational Safety and Health Administration, 2016. Fatality and Catastrophe Investigation Summaries. https://www.osha.gov/pls/imis/accidentsearch.html (Mar 30, 2016).

Peng, T., Liu, L., Zuo, W., 2014. PU text classification enhanced by term frequency–inverse document frequency-improved weighting. Concurrency Comput. Pract Experience 26 (3), 728–741.

Python Software Foundation, 2016. Python Language Reference, Version 2.7. http://www.python.org (Apr 13, 2016).

Raschka, S., 2015. Python Machine Learning. Packt Publishing, Birmingham.

Reason, J., 1997. Managing the Risks of Organizational Accidents. Ashgate, Aldershot.

Sebastiani, F., 2002. Machine learning in automated text categorization. ACM Comput. Surv. 34 (1), 1–47.

Shibukawa, Y., 2013. Snowball Stemming Library Collection for Python. https://gi-thub.com/shibukawa/snowball_py (Apr 12, 2016).

Tanguy, L., Tulechki, N., Urieli, A., Hermann, E., Raynal, C., 2015. Natural language processing for aviation safety reports: from classification to interactive analysis. Comput. Ind. 78, 80–95.

Taylor, J.A., Lacovara, A.V., Smith, G.S., Pandian, R., Lehto, M., 2014. Near-miss narratives from the fire service: a Bayesian analysis. Accid. Anal. Prev. 62, 119–129.

Tixier, A.J.P., Hallowell, M.R., Rajagopalan, B., Bowman, D., 2016. Automated content analysis for construction safety: a natural language processing system to extract precursors and outcomes from unstructured injury reports. Autom. Constr. 62, 45–56.

Turney, P.D., Littman, M.L., 2003. Measuring praise and criticism: inference of semantic orientation from association. ACM Trans. Inf. Syst. 21 (4), 315–346.

Vallmuur, K., Marucci-Wellman, H.R., Taylor, J.A., Lehto, M., Corns, H.L., Smith, G.S., 2016. Harnessing information from injury narratives in the 'big data' era: understanding and applying machine learning for injury surveillance. Inj. Prev. 22 (Suppl 1), i34–i42.

Vallmuur, K., 2015. Machine learning approaches to analysing textual injury surveillance data: a systematic review. Accid. Anal. Prev. 79, 41–49.

Williams, T.P., Gong, J., 2014. Predicting construction cost overruns using text mining: numerical data and ensemble classifiers. Autom. Constr. 43, 23–29.

Williams, G., 2011. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R!). Springer, Canberra, ACT, Australia.

Witten, I.H., 2011. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington, MA.

Workplace Safety and Health Institute, 2016. Workplace Safety and Health Report 2015. . https://www.wsh-institute.sg/ (Mar 30, 2016).

Yu, W.D., Hsu, J.Y., 2013. Content-based text mining technique for retrieval of CAD documents. Autom. Constr. 31, 65–74.

Zhou, Z., Goh, Y.M., Li, Q., 2015. Overview and analysis of safety management studies in the construction industry. Saf. Sci. 72 (February), 337–350.

scikit-learn Community, 2016. Scikit-learn – Machine Learning in Python. http://scikit-learn.org/ (Apr 13, 2016).