CONSTRUCTION SITE ACCIDENT ANALYSIS USING TEXTMINING AND NATURAL LANGUAGE PROCESSING TECHNIQUES

A.Vijay Kumar[1], G.Swapna[2], Ganji Nikitha[3], Chinthala Sheshanthika[4], Gaddala Sreeja[5], Guntaka Akanshha[6]

[1]Assistant Professor, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

[2]Assistant professor, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

[3,4,5,6] IVth Btech Student, Department of CSE, Sri Indu Institute of Engineering & Technology, Hyderabad

ABSTRACT

Workplace safety is a major concern in many countries. Among various industries, construction sector is identified as the most hazardous work place. Construction accidents not only cause human sufferings but also result in huge financial loss. To prevent reoccurrence of similar accidents in the future and make scientific risk control plans, analysis of accidents is essential. In construction industry, fatality and catastrophe investigation summary reports are available for the past accidents. In this study, text mining and natural language process (NLP) techniques are applied to analyse the construction accident reports. To be more specific, five baseline models, support vector machine (SVM), linear regression (LR), K-nearest neighbour (KNN), decision tree (DT), Naive Bayes (NB) and an ensemble model are proposed to classify the causes of the accidents. Besides, Sequential Quadratic Programming (SQP) algorithm is utilized to optimize weight of each classifier involved in the ensemble model. Experiment results show that the optimized ensemble model outperforms rest models considered in this study in terms of average weighted F1 score. The result also shows that the proposed approach is more robust to cases of low support. Moreover, an unsupervised chunking approach is proposed to extract common objects which cause the accidents based on grammar rules identified in the reports. As harmful objects are one of the major factors leading to construction accidents, identifying such objects is extremely helpful to mitigate potential risks. Certain limitations of the proposed methods are discussed and suggestions and future improvements are provided.

1. Introduction

Construction industry remains globally the most dangerous work place. There are > 2.78 million deaths every year caused by occupational accidents according to the International Labor Organization (ILO) . Among which approximately one of six fatal accidents occur in the construction sector. Construction accidents not only cause severe health issues but also lead to huge financial loss. To prevent occurrence of similar accidents and promote workplace safety, analysis of past accidents is crucial. Based on the results of cause analysis, proper actions can be taken by safety professionals to remove or reduce the identified causes. It is also noted that one major factor contributing to the risk of an accident is the presence of harmful objects such as misused tools, sharp objects nearby, damaged equipment. Mitigating strategies can be made accordingly after identification of such objects. For example, raising awareness, performing mandatory regular checks before operation of the machine which went wrong and caused the accident earlier. In construction industry, a catastrophe investigation report is generated after a fatal accident which provides a complete description of the accident, such text data can be utilized for further analysis. Studies of text mining, NLP and ensemble techniques for the analysis of construction accidents report are rare. Motivation of this paper

is to fill this research gap. In this study, text mining and NLP techniques are applied to analyse the construction site accidents using the data from Occupational Safety and Health Administration (OSHA). Aan ensemble model is proposed to classify the causes of accidents. While in conventional majority voting mechanism, equal weights are assigned to each base classifier involved in the ensemble model. In this study, the weight of each base classifier is optimized by Sequential Quadratic Programming (SQP) algorithm. Moreover, a rule based chuker is developed to identify common objects which cause the accidents. Neither SQP optimization nor chuker algorithm is found to be applied in this field in any existing literatures.

Major contributions of this work are:

- Various texting mining and NLP techniques are explored with respect to construction site accident analysis.
- Ensemble algorithm which has not been well studied in this field is proposed to classify the causes of accidents and SQP algorithm is utilized to search for optimal weighs of the ensemble model.
- A rule based chuker is developed for dangerous objects extraction. Neither SQP optimization algorithm nor rule based chuker with regard to this field is found in the state of the art.
- Case studies are designed using OSHA dataset and effectiveness of the proposed approaches is verified by the experiment result

2. Literature review

There are several studies which utilize text mining or natural language process (NLP) approaches for occupational accidents analysis. et al. developed a Naïve Bayesian model to classify the compensation claims causation due to work related injuries. The proposed model achieved an overall accuracy of approximately 90%, however the accuracy of claims belongs to minor injury categories dropped. Taylor et al. applied Naïve Bayesian and Fuzzy models to categorize the injury outcome and mechanism of injury for fire service incident reports extracted form from the National Fire fighter Near-Miss Reporting System. Results showed that Fuzzy model achieved a sensitivity of 0.74 while sensitivity of Naïve Bayesian model is 0.678. Wellman et al. proposed a Fuzzy Bayesian model to classify injury narratives into external-cause-of-injury and poisoning (E-code) categories. Data used in this study is the injury reports from US National Health Interview Survey (NHIS) during 1997 and 1998. The proposed model achieved an accuracy of 87.2%. Ab data et al. applied Bayesian network to extract recurrent serious Occupational Accident with Movement Disturbance (OAMD) scenarios from narrative texts. It is noted that data pre-processing of this approach is time consuming and expert knowledge is required. Wellman et al. proposed an approach which combined manual coded rules with machine learning algorithms for injury narratives classification. Results showed that using Logistic Regression (LR) and filtering out the bottom 30% of its predictions reviewed manually resulted an overall sensitivity of 0.89. et al. compared the methods of Naïve Bayesian and Regularized Logistic Regression for auto coding the causation of injury narratives. Dataset used was from Occupational Injury and Ill-ness Classification System (OIICS), and results showed that the logistic model achieved an overall accuracy of 71% for 2-digit OIICS event/ exposure classification system and 87% for first digit respectively. In terms of the analysis of construction related accidents, et al. applied Random Forest (RF) and Stochastic Gradient Tree Boosting (SGTB) algorithms to predict type of energy involved in the accident, injury type, body part affected, and injury severity using construction injury reports. Rank Probability Skill Score (PRSS) of the proposed methods ranked from 0.236 to 0.436. et al. proposed a NLP approach based on hand crafted rules and keywords dictionary to extract outcomes and precursors from unstructured injury reports and achieved a recall of 0.97 and precision of 0.95, however the proposed approach was not robust to unanticipated situations. Goh et al. applied support vector machine (SVM), linear regression (LR), random forest (RF), K-nearest neighbour (KNN), decision tree (DT) and Naive Bayes (NB) algorithms for construction accident narrative classification. Among which, SVM achieved a F1 score ranged from 0.45 and 0.92 and outperformed the other classifiers. The author further presented an ensemble approach for construction accident narrative classification . Choker et al. applied a K-means based approach to classify injury reports. Four clusters were identified and each cluster represented a type of accident. Identified accident types were ‘falls', ‘struck by objects’, ‘electrocutions’ and ‘trenches collapse’ respectively. Fan et al.

3. Methodology

3.1. Text mining and natural language processing

Text mining, also referred to as text data mining, is defined as the process of deriving information from text data which is not previously known and not easy to be revealed . It involves transforming text into numeric data which can be used in data mining algorithms then . Natural language processing (NLP) involves the techniques of multiple areas in artificial intelligence, computational linguistics, mathematics and information science, it the approach to make computer understand natural language and perform certain tasks . NLP can be utilized to analyse semantic and grammatical sutures of text while such analysis cannot be performed by text mining. In this work, five single classifiers are evaluated along with the proposed ensemble model for accident causes classification and a rule based chunking approach is proposed to identify common objects which cause the accident. Before applying the aforementioned classifiers to text data, certain pre-processing and feature extraction steps are needed. Common steps to process text are: Lower case and punctuation removal: This step transforms the text into lower case which reduces variation of same word, e.g., after transformation ‘Employee’ and ‘employee’ are treated as the same word. Punctuations increase the size of training data and usually do not contribute much to text analysis, thus are removed. Stop words removal: Stop words are extremely common words which are of little value in helping select documents and such words are excluded. Some published stop words lists are available for example in Snowball stop word list published with the Snowball Stemmer and Terrier stop word list published with the terrier package. However, stop words of different domains are different. For medical domain, words like ‘pill’, ‘patient’ occur in most documents and such words are considered stop words while for computer product domain, potential stop words list consists words such as ‘CPU’, ‘memory’, etc. Generally, common stop words list does not cover such terms, a domain specific stop words list can be complied base +on acquired domain knowledge. Tokenization: Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, certain characters, such as punctuation is filtered out during the process . Stemming and lemmatization: In a document, same word can be expressed in different forms, e.g. ‘kill’, ‘kills’, ‘killing’. Moreover, words can be represented in different syntactic categories that have the same root form and are semantically related, e.g. ‘irony’, ‘ironic’. The two aforementioned scenarios are common due to grammatical reasons. Stemming and lemmatization are used to reduce in flectional and derivationally related form of a word and converting it to a base form . E.g. ‘am’, ‘is’, ‘are’ are converted to ‘be’, ‘dog’, ‘dogs’, ‘dog's’ are converted to ‘dog’. Part of speech tagging: (POS tagging) is the process of assigning parts of speech tag to each token, such as noun, verb, adjective, etc. A comprehensive list of part of speech tags can be found in Penn Treebank .

3.2. KNN

K-Nearest Neighbour (KNN) algorithm is widely used for pattern classification based on feature similarity. For a given unclassified sample point, it is classified by a majority vote of its neighbours, with the point being assigned to the class most common among its K nearest neighbours. KNN is a lazy algorithm. Unlike most statistical methods which elaborate a model from the information available in the historic data, KNN considers the training set as the model itself. Thus there is no explicit training phase for KNN algorithm and during the testing phase, all training data is needed due to the lack of generalization. A KNN algorithm is characterized by issues such as number of neighbours, adopted distance, etc. More details of the KNN fundamental theory can be found in .

3.3. Decision tree

Decision tree is a hierarchical tree-based classifier. It is represented by a set of nodes, a directional graph that starts at the base with a single node and extends to many leaf nodes that represent the categories that the tree can classify. It classifies a given sample data by applying a series of rules to features of the sample data. Each rule is represented by a node and each internal node points to one child node for each possible outcome corresponding to the applied rule.

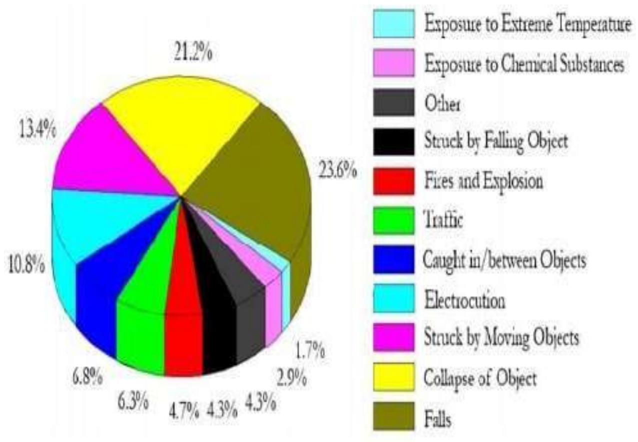

Fig. 3. Distribution of the different causes of accidents.

4. Experiments and results

4.1. Experiment tools and data description

Two experiments are designed in this study. In the first experiment, an ensemble classifier is developed to classify the cause of construction accident while the second experiment is designed to identify common objects which cause the accident. Developing tool used is Python 2.7, main packages used for algorithms design are learn v0.19.1, pandas v0.22.0, v3.2.5 and matplotlib v2.1.2 package for visualization. The original dataset from the Occupational Safety and Health Administration (OSHA) website is free to download. It contains 16,323 records of construction site accidents (happened between 1983 and 2016) without labelling the cause of accidents. The report provides a detailed description of the incident, including causal factors and events which lead to the incident. In this study, case summary are used for classification while for analysing objects caused the accidents, only case title information of the dataset is used. Data pre-processing process for accident cause classification and object identification are different. The major difference is that supervised learning approach used for classification task requires label data while for object identification task an unsupervised rule based chunking approach is adopted, and hence the dataset is not necessarily to be annotated. Thus, details of the data pre-processing steps are discussed in the corresponding experiment sections separately.

4.2. Accident classification

Since classification requires label data and label the whole OSHA dataset is a tedious work due to resource construction, another dataset is invoked. In an early study of Goh , a processed dataset which consists of 1000 label records was published . Therefore, this dataset is utilized instead of using the original dataset from OSHA website. A sample case is depicted in . The cases are annotated according to labels used in Workplace Safety and Health Institute (2016). Furthermore, to avoid having a case with multiple categories the label is assigned according to the first incident if multiple incidents leading to one accident. For example, in the case summary in , the first incident is “second story collapsed” followed by the second incident “Bricks struck Employee #1's head and neck”, thus this case is label as “Collapse of object” accordingly. Meanwhile to reduce the number of labels representing similar causes, the cases are annotated in a more general and standard fashion. For example, the cause of the case in is label as “Collapse of object” instead of “Collapse of building story”

4.2.1. Results and discussions

To evaluate the model performance, F1 score proposed by Buckland et al. has been widely adopted in literatures. However, support which denotes the number of true instances for each label is not considered in conventional F1 score calculation. Therefore, the average weighted F1 score given by Eq. (8) is adopted for performance measure.

(8) where N denotes the total number of labels, Si denotes the support of label I , T denotes the support of all labels and F1i denotes the F1 score of label I . Other performance measures such as precision, recall of each model, and support for each label is also measured. Detailed results are presented in Table 2. The F1 score for each cause of the accident for each classifier is depicted in Fig. 4, while Fig. 5 depicts the overall F1 score for each classifier. In Table 2, the weighted average F1 score of each model and the highest F1 score for each label are highlighted in bold. It is noted that the highest weighted average F1 score for labels is 0.68 achieved by the proposed ensemble model with optimized weights. While the second best model is SVM, the

overall performance of Naïve Bayesian model is the worst. The result also showsthat ensemble model using simple majority voting mechanism without optimization doesn't effectively improve the overall performance. Besides, it can be seen from the result that the highest F1 score is achieved by the proposed ensemble model for almost all labels. Except label ‘electrocution’, ‘fires and explosions’ and ‘exposure to extreme temperatures’. To be more specific, for label ‘electrocution’ and ‘fires and explosions’ classification, SVM out per forms the rest models. For ‘exposure to extreme temperatures’, the highest F1 score is achieved by decision tree model. It is worth noticing that, although support of this label is extremely low, decision tree achieves the most satisfying classification result followed by the proposed model. The results also show that the performance of both Naïve Bayesian model and Logistic Regression model are poor when classifying cases with a low support while the proposed ensemble model is more robust to the value of support. It is worth noting that the proposed model achieves high F1 score for most labels.

4.3. Identification of common objects causing accidents

In this experiment, rule based chunking approach is adopted to extract common objects which cause accidents from ‘title’ data. As it is an unsupervised learning approach, unb original dataset is invoked. Data preprocessing involves three steps, tokenization, stop words removal and POS tagging which are the same as used for cause classification experiment. After the POS tagging step, it is noted that a few words which can be critical to identifying objects in the context. Are annotated with wrong POS tags, e.g. ‘injures’, ‘breaks’, ‘crush’, ‘swings’ are tagged with ‘NN’ by the POS tagger. Such errors are manually corrected.

Table 4 shows the sample title data after POS tagging. After the 'title' data, certain syntactic structure is observed. From sample POS tagged title data 1,2,3,4 shown in Table 4, the target object is a noun or noun phase appears after a past tense verb followed by a proposition and from title 5,6, the target object is a noun or noun phase appears after a verb followed by a proposition. A chuker is built using regular expressions according to the identified rules. Then the text data is parsed into a tree consists of a set of connected label nodes. A sample parsed tree using title data is shown in Fig. 6. The original text before parsing is 'Employee is splashed with hot water and is burned'. The root node 'S' represent sentence, leaves of 'NP' node which is under the 'TARGET CLAUS' node compose the target object, i.e. 'hot water' is the object which causes the accident in this context. extracting the target objects using the proposed , it is found some extracted noun phases are actually not legitimate objects, e.g. 'height', 'exposure', 'fall'. Thus, a post process is performed to filter out such words from the result.

4.3.1. Results and discussions

The 10 most common objects, which are 'ladder', 'root', 'truck', 'machine', 'forklift', 'scaffold', 'vehicle', 'fire', 'press', 'tree', are shown in . The corresponding word cloud is depicted in . It is noted that the proposed approach involves certain manual inspections and corrections to improve the results.

Due to the dynamic characteristics of the natural language, sentences of same meaning can be expressed differently in terms of the F. Zhang et al. Automation in Construction 99 (2019) 238–248 246structure or wording. Thus, developing exhaustive rules to all cover variations is not feasible. As a consequence, certain objects in the documents are missed out and some extracted objects are actually not legitimate. Moreover, vagueness of natural language is common and results in various interpretations from different people. In fact, it is challenging even for a human to identify the object which cause the accidents in some cases. For example, for sentence ‘Employee dies of brain aneurism’, the cause of accident is ‘brain aneurism’, however, ‘brain aneurism’ is not an object. For sentence ‘Employee faints in trench’, it is difficult to tell if ‘trench’ is the actual object that causes the accident without giving more context. Apart from exhaustive rules, need to be hand crafted when dealing with dynamic structured cases. Another challenge like other unsupervised learning approaches is that the correct result is not available. In other words, the result needs to be manually checked.

5. Conclusions and future work

Analysing the construction accident reports leads to valuable knowledge of what went wrong in the past in order to prevent future accidents. To be more specific, accident causes classification is essential as prevention strategies should be developed based on different causes accordingly. Besides, identification of dangerous objects plays a crucial role in improving the safety of the working environment as well, as preventive actions can be implemented to eliminate or mitigate the potential risks of identified objects. However, manual classification of accident reports and investigation of dangerous objects involved in accidents are time consuming and labour intensive. In this work, an ensemble model with optimized weights is proposed for construction accident causes classification. The results show that the proposed model outperforms other single model in terms of the average weighted F1 score. Further, the proposed model is proved to be more robust to the cases of low support. Moreover, a rule based approach is explored to identify the common objects which cause the accidents. Therefore, the aforementioned labour intensive tasks are effectively automated by the proposed approaches. Besides, the proposed approaches support the informed culture and play an important role in improving the safety information system proposed by Reason which enhance the construction site safety in the long run. Several possible future improvements can be considered, for example, data balancing techniques such as under sampling, over sampling or a combination of both can be applied. Compiling a stop words list specific to construction accident domain which reduces stop words more accurately is also an approach can be considered to improve the data quality. Besides, missing corresponding context in formation between tokens can also cause the mis classification problem. In this study, only unigrams is used when building the classifiers, while bigrams and trigrams can preserve more context information and probably lead to a better performance of the classifier. Optimization algorithms such as GA, PSO, DE can be utilized to better select the weight and model parameters of each single classifier when forming the ensemble model. Besides, instead of ensemble of weak learners, more advanced recurrent neural network model such as long short term memory (LSTM) neural network can be explored in a future study. It is also noted some POS tags are not annotated properly by the published POS tagger, as POS tags are most critical information for chunking, utilizing a domain specific POS tagger is also benefit to the performance of built eventually. To chunk an unable large dataset, supervised learning approach requires large amount of annotated data while rule based approach requires manual checks of the results. One potential technique to explore is semi supervised learning approach. Last but not the least, more NLP frameworks such as Natural Node/natural , and Stanford NLP can be explored in the future research.

6. Acknowledgement

The dataset used in this study is published and processed by Yang Goh and . Download link is: https:// github.com/safety hub/OSHA _Acc. git. The original data can be downloaded from below link: https://www.osha.gov/pls/imis/ accidentsearch.html (Occupational Safety and Health Administration,

2016) .

7.References

[1] H.M. Al-Humaidil, F.H. Tan, Construction safety in Kuwait, J. Perform. Constr. Facil. 24 (1) (2010) 70–77, https://doi.org/10.1061/(ASCE)CF.1943-5509.0000055.

[2] R. Navon, R. Sacks, Assessing research issues in automated project performance control (APPC), Autom. Constr. 16 (4) (2007) 474–484, https://doi.org/10.1016/j.autcon.2006.08.001.

[3] International Labor Organization (ILO), Safety and Health at Work, http://www.ilo.org/global/topics/safety-and-health-atwork/langen/index.html (Accessed: Oct. 2nd, 2018).

[4] R.A. Haslam, et al., Contributing factors in construction accidents, Appl. Ergon. 36 (4) (2005) 401–415, https://doi.org/10.1016/j.apergo.2004.12.002.

[5] S.J. Bertke, A.R. Meyers, S.J. Wurzelbacher, J. Bell, M.L. Lampl, D. 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.

[6] J.A. Taylor, A.V. Lacovara, G.S. Smith, R. Pandian, M. Lehto, Near-miss narratives from the fifire service: a Bayesian analysis, Accid. Anal. Prev. 62 (2014) 119–129, https://doi.org/10.1016/j.aap.2013.09.012.

[7] H.M. Wellman, M.R. Lehto, G.S. Sorock, G.S. Smith, Computerized coding of injury narrative data from the National Health Interview Survey, Accid. Anal. Prev. 36 (2) (2004) 165–171, https://doi.org/10.1016/S00014575(02)00146-X.

[8] F. Abdat, S. Leclercq, X. Cuny, C. Tissot, Extracting recurrent scenarios from narrative texts using a Bayesian network: application to serious occupational accidents with movement disturbance, Accid. Anal. Prev. 70 (2014) 155–166, https://doi.org/10.1016/j.aap.2014.04.004.

[9] H.R. Marucci - wellman, H.L. Corns, M.R. Lehto, Classifying injury narratives of large administrative databases for surveillance - a practical approach combining machine learning ensembles and human review, Accid. Anal. Prev. 98 (2017) 359–371, https://doi.org/10.1016/j.aap.2016.10.014.

[10] S.J. Bertke, A.R. Meyers, S.J. Wurzelbacher, A. Measure, M.P. Lampl, D. Robins, Comparison of methods for auto-coding causation of injury narratives, Accid. Anal. Prev. 88 (2016) 117–123, https://doi.org/10.1016/j.aap.2015.12.006.

[11] A.J. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Application of machine learning to construction injury prediction, Autom. Constr. 69 (2016) 102–114, https://doi.org/10.1016/j.autcon.2016.05.016.

[12] A.J. Tixier, M.R. Hallowell, B. Rajagopalan, D. Bowman, Automation in construction automated content analysis for construction safety: a natural languageprocessing 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.

[13] Y.M. Goh, C.U. Ubeynarayana, Construction accident narrative classifification: an evaluation of text mining techniques, Accid. Anal. Prev. 108 (2017) 122–130,

https://doi.org/10.1016/j.aap.2017.08.026.

[14] C.U. Ubeynarayana, Y.M. Goh, An Ensemble Approach for Classifification of Accident Narratives ASCE International Workshop on Computing in Civil Engineering 2017, (2017), pp. 409–416, https://doi.org/10.1061/9780784480847.051.

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

[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] Y. Zou, A. Kiviniemi, S.W. Jones, Retrieving similar cases for construction project risk management using Natural Language Processing techniques, Autom. Constr. 80 F. Zhang et al. Automation in Construction 99 (2019) 238–248 247