**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 analyze the construction accident reports. To be more specific, five baseline models, support vector machine (SVM), linear regression (LR), K-nearest neighbor (KNN), decision tree (DT), Naïve 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.

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| S.No | Figure description | Page no. |
| 1 | Data set summary statistics | 3 |
| 2 | Event date statistics | 3 |
| 3 | Abstract text statistics | 4 |
| 4 | Event description statistics | 4 |
| 5 | Event keyword statistics | 4 |
| 6 | Nature of injury statistics | 5 |
| 7 | Part of the body statistics | 5 |
| 8 | Event type statistics | 5 |
| 9 | Human factor statistics | 6 |
| 10 | Task statistics | 6 |
| 11 | Architecture diagram | 14 |
| 12 | Class diagram | 19 |
| 13 | Activity diagram | 20 |
| 14 | Sequence diagram | 21 |
| 15 | Use case diagram | 22 |
| 16 | SVM algorithm | 24 |
| 17 | KNN algorithm | 25 |
| 18 | Decision tree algorithm | 26 |
| 19 | Analysis on task assigned | 31 |
| 20 | Analysis on degree of injury | 31 |
| 21 | Analysis on nature of injury | 31 |
| 22 | Analysis on environment factor | 32 |
| 23 | Analysis on human factor | 32 |
| 24 | Analysis on part of the body | 32 |
| 25 | Main window | 42 |
| 26 | Loading data set | 42 |
| 27 | Result for SVM algorithm | 43 |
| 28 | Result for KNN algorithm | 43 |
| 29 | Result for naïve bayes algorithm | 44 |
| 30 | Result for decision tree algorithm | 44 |
| 31 | Result for logistic regression algorithm | 45 |
| 32 | Result for ensemble algorithm | 45 |
| 33 | Result for optimised ensemble algorithm | 46 |
| 34 | Accident prediction | 46 |
| 35 | Dangerous object extraction | 47 |
| 36 | Cause of accidents | 47 |
| 37 | Accuracy graph | 48 |
| 38 | F-score graph | 48 |
| 39 | Recall graph | 49 |