IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Nlp-Based Hazard Identification In Construction Reports

[1]Jarajapu Appalaraju, Student in Dept. Of Master of Computer Applications, at Miracle Educational Society Group of Institutions

[2]Mr. B. Mahendra Roy, Assistant Professor at Miracle Educational Society Group of Institutions

[3]Dr. Burada Venkata Rao, Associate Professor at Miracle Educational Society Group of Institutions

ABSTRACT

The construction sector is recognized as the most perilous profession around the globe and incurs huge human and economic costs. This research examines construction accident reports applying more text mining and natural language processing (NLP) approaches. It proposes a mixed hybrid classification approach that includes SVM, Naïve Bayes, and decision trees, all modified by the Sequential Quadratic Programming (SQP) for better model performance. In addition, a rule based chunker provides a list of potentially hazardous objects that may be present at accident sites. The study supports its conclusions empirically through the use of datasets provided by the Occupational Safety and Health Administration (OSHA) . This project intends to advance the construction industry towards greater chances of safety by providing techniques that deal with text data, and complement around traditional safety measures as well as AI enabled accurate decision making.

Keywords: NLP, SVM, SQP

INTRODUCTION:

Construction continues to rank among the most hazardous sectors across the globe, contributing significantly to the rate of fatalities in workplaces. Notably, the International Labor Organization has indicated that it is construction-related accidents that account for a considerable proportion of workplace deaths every year. These incidents result not only in dramatic human suffering, but also in huge economic waste. Tackling such issues requires learning from such accidents so that they do not happen again. While there is a plethora of reports on incidents, the effective use of such information is very low. Development in text mining and natural language processing suggests that useful interpretations can be derived from such non-structured texts. This project examines the incidence of construction site accidents using a machine learning model based on text mining to classify the causes of accidents and features of objects that are dangerous on the construction site. Using the OSHA dataset, this work addresses important issues related to the understanding of construction site activities, and consequently the creation of construction site protective measures based on the analysis of data.

GAP IDENTIFIED BASED ON LITERATURE SURVEY:

There is however a lot of work that has been done in the area of industrial safety, however, there are still gaps in the use of textual data in construction accident prevention strategies. The majority of the cited studies seem to disregard structural data, or manual analysis which is both time consuming and subjective. The ability to use older safety strategies as been reached, there is a need to review the older strategies in light of the advanced challenges faced while working in the construction industry today.

Most of the time, existing machine learning programs for safety do not adequately tackle the problem of unstructured narratives of accidents. Questions pertaining to causality, for instance, have hardly been asked in the context of construction, even when Bayesian networks and NLP systems have been used. There are a few of them that remain dormant, and these are as follows:

1. The Advanced NLP Techniques Are Not Fully Exploited: Most of the approaches are devoid of any advanced rule-based chunking techniques or sequence optimization techniques designed for language processing tasks.

2. Lack of A Comprehensive Ensemble Basic Models: The models include single classifiers without optimized weights for the purpose of combining these classifiers for better prediction accuracy.

3. Dataset Is Not Targeted Enough: Often researchers do not pay attention to some domains, like reports from OSHA and the like and come up with a catch-all conclusion.

PROBLEM STATEMENT:

Risk factors in the construction domain are heavily inter related as they often lead to safety challenges that are unable to be resolved because of the lack of assurances that existing methods hold. For instance, a single accident might include a multitude of factors that could include unsafe practices or dangerous objects in the area.

Key Challenges:

- The intricacies of high volumes of unstructured text accidents reports.

- The appropriate classification of accident causes using ML algorithms.

- Distinct contributors to accidents.

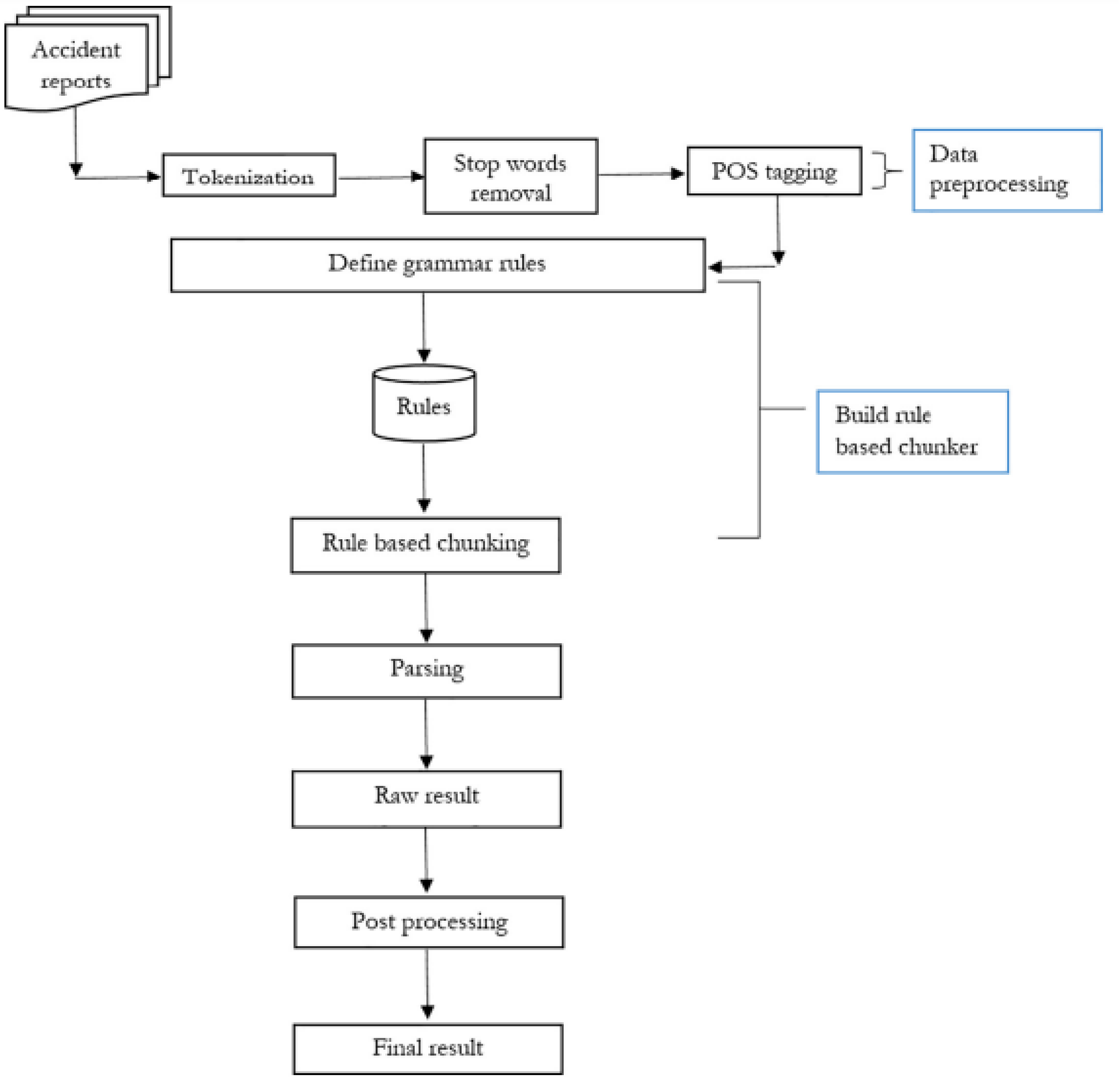
- Improving the accuracy of the model through ensemble modelling strategies.

- The domain and benchmarks specific datasets to facilitate the application of the methods in practice.

PROPOSED METHOD:

A systematic method of construction accident investigation through the application of machine learning and NLP techniques was developed in this study. The textual data was preprocessed by cleaning the reported OSHA accidents and clinical features were also extracted. Text information is transformed into a numerical form using the technique called TF-IDF. SVM, Naïve Bayes, and decision trees are used and there is an ensemble model of the combined classifiers with the weights that have been optimized by thorough SQP. In addition, a rule-based chunker extracts dangerous elements from the written text. The models are evaluated using the historical databases and measures including accuracy, precision, and recall are deployed to measure performance. This approach is expected to provide valuable information for improving risk management and safety measures on construction sites.

ARCHITECTURE:



DATASET:

The data set used in this paper consists of previously recorded incident reports from the OSHA databases. It gives unstructured free text which includes the narrative of the incidents, their causes and other associated information. Types of accident occupation, type of dangerous objects, and descriptions of the narration of event are among the primary features selected. Some of the word cleaning methods and data formatting techniques that are included are stop-word elimination, stemming, and TF-IDF transformations. This dataset represents a valuable opportunity for building and testing machine learning models, as it contains information on the occurrence of accidents and their analysis. It is however specialized enough implying that the findings made will be appropriate in the context of the construction industry and accurately portray construction safetyone.

METHODOLOGY:

Step 1: Data Preprocessing

Step 2: Collect the Textual Accidents and Injuries Report from OSHA Databases

Second, we performed text preprocessing in order to delete all punctuation marks, special characters and stop words from the text

Step 3: Applying stemming for word form reduction

Forth, the cleaned text was transformed into numeric form by employing the TF-IDF technique.

Step 2: Feature Engineering

Fifth, the key attributes such as the causes for the accidents and the hazards that were present during the occupation were identified

Seventh, part-of-speech tagging was used to flag actionable terms.

Step 3: Model Development

In order to generate sufficient data with which the classifiers will be trained, we trained multiple classifiers such as the SVM, Decision Tree, Naïve Bayes, Logistic Regression and KNN.

The eighth step involved building an ensemble model by bringing together the classifiers through majority voting.

Step 4: Optimization

1. The goal is to optimize weights using Sequential Quadratic Programming (SQP) algorithm in the ensemble model.

Step 5: Rule-Based Chunker Implementation

1. The intended purpose is to formulate rules in order to create a noun chunker that targets hazardous objects.
2. The purpose is to check whether the output of the chunker uses domain-specific words or not.

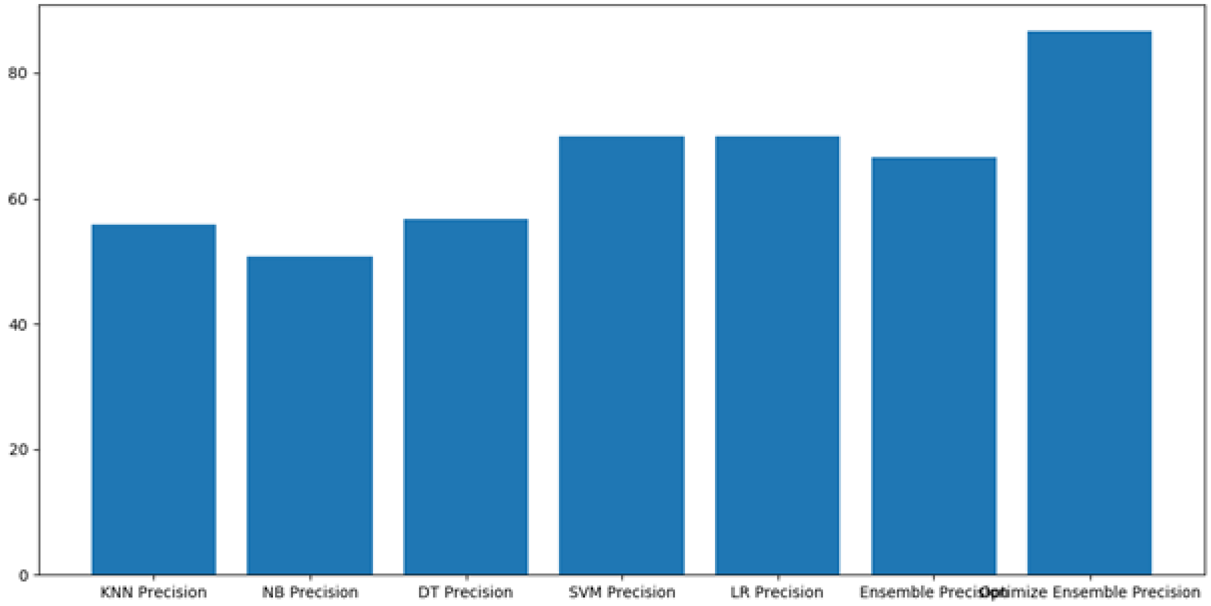
Step 6: Model Validation

1. Using 20% of the dataset as test data set while the remaining 80% is training data set.
2. Measure the performance with respect to accuracy, precision, recall and F1 score.

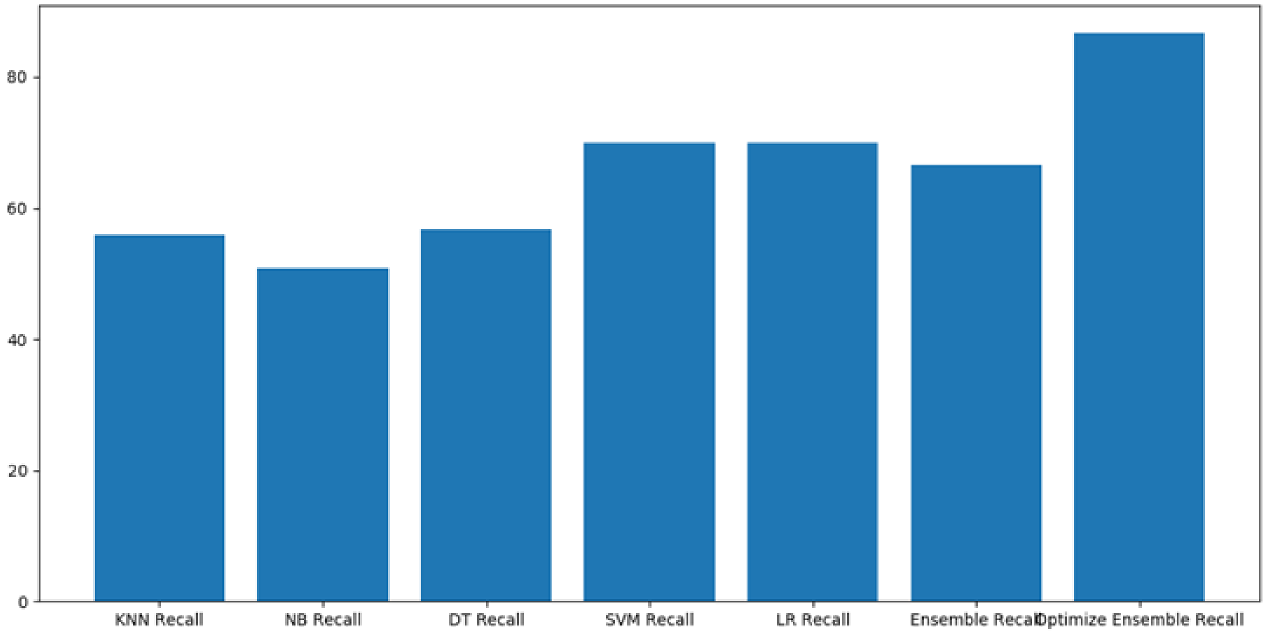
Step 7: Visualization and Reporting

1. The aim is to build safety dashboards for predicting the cause of accidents and also for the extraction of hazardous objects.
2. Further, provide a comprehensive analysis of the data that the beneficiaries will need in decision making.

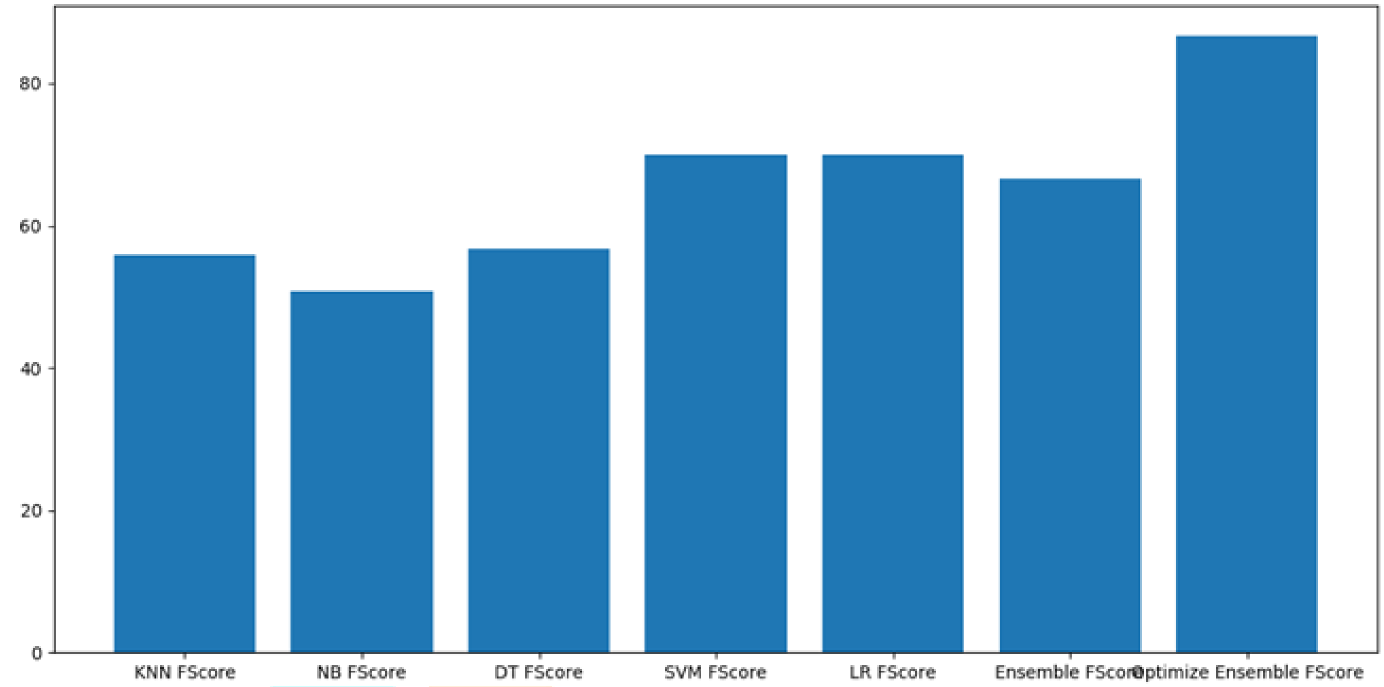
RESULTS:



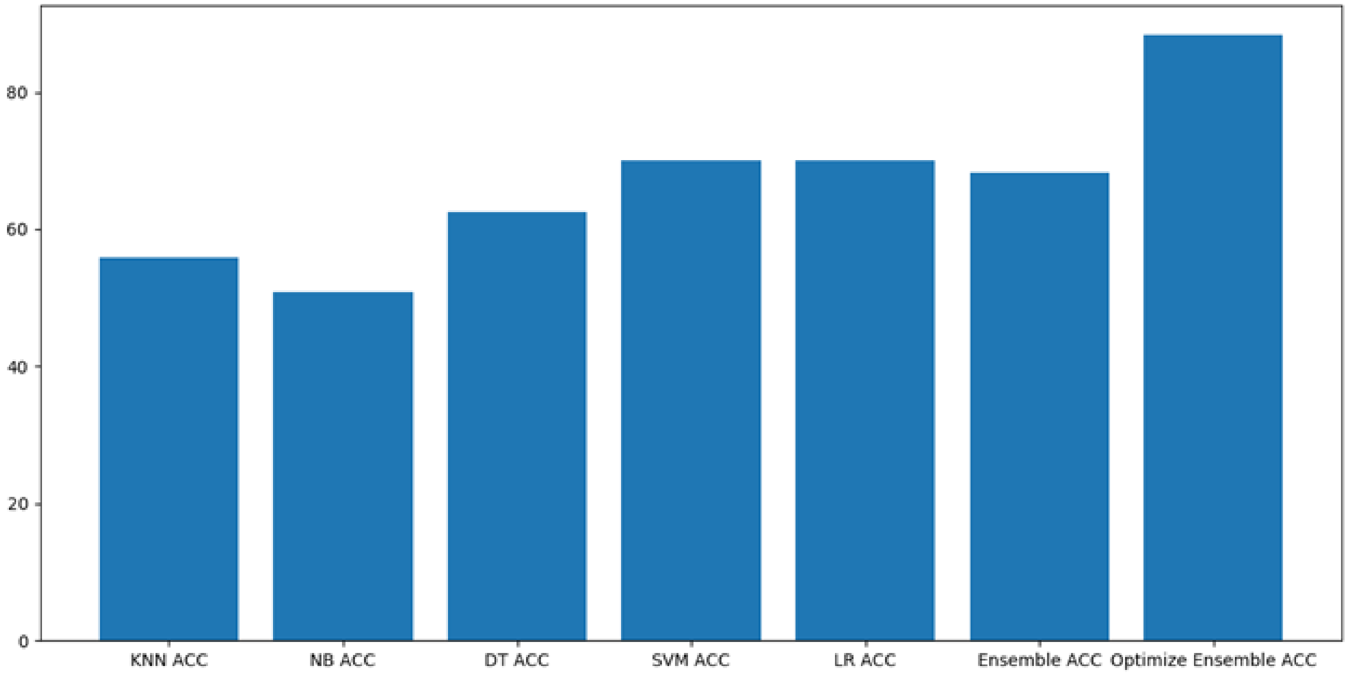
Graph x-axis represents algorithm names and y-axis represents precision of those algorithms. In above graph we can see Propose Optimize Ensemble (Voting Classifier) gave better performance..



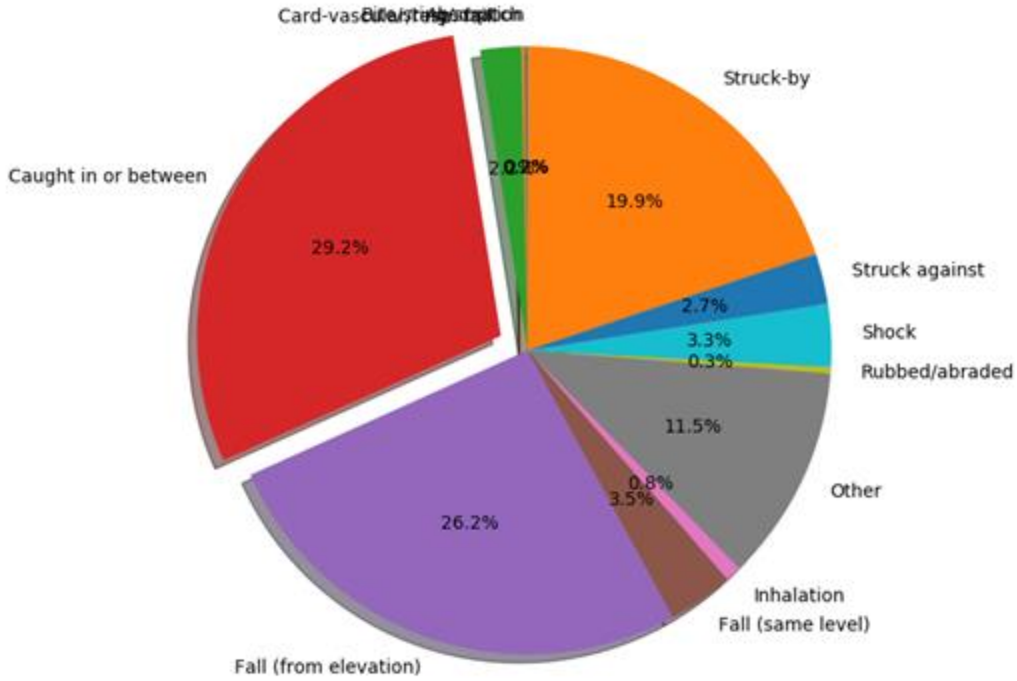
Recall Graph



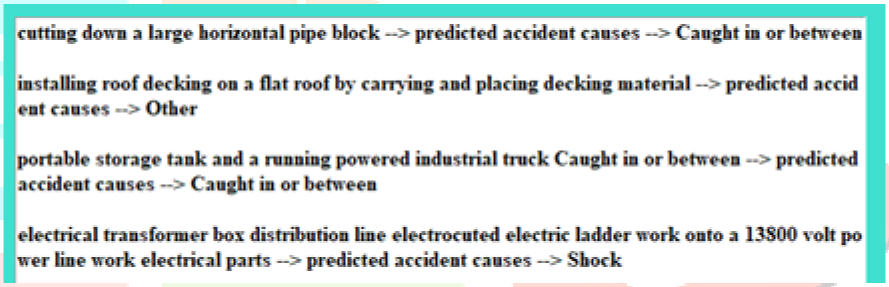
fscore graph



Accuracy graph x-axis represents algorithm name and y-axis represents accuracy of those algorithms.



Causes Accident Graphs



cutting down a large horizontal pipe block --> predicted accident causes --> Caught in or between

installing roof decking on a flat roof by carrying and placing decking material --> predicted accident causes --> Other

portable storage tank and a running powered industrial truck Caught in or between --> predicted accident causes --> Caught in or between

electrical transformer box distribution line electrocuted electric ladder work onto a 13800 volt power line work electrical parts --> predicted accident causes --> Shock

Predicted future accident that may occur while doing this work

CONCLUSION

The analysis of the construction accidents data, making use of machine learning with the integration of natural languages is seen in practice with this project. This optimally designed predictive in addition to the utilization of a rule-based chunker fills vital gaps as far as safety analysis is concerned. The robust nature of the approach in identifying the causes and the notions of hazardous objects has been demonstrated in validation on the OSHA datasets. These results can help the development of interventions to reduce risk factors and increase the security of work. Future work could investigate connectivity to live sources of information and extend the work of the model to other industries that have large risks and implications for the use of AI in occupational safety.

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-at-work/lang–en/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 fire 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/S0001-4575(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 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.

[13] Y.M. Goh, C.U. Ubeynarayana, Construction accident narrative classification: 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 Classification 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 (2017) 66–76, https://doi.org/10.1016/j.autcon.2017.04.003.

[18] G. Miner, J. Elder, T. Hill, D. Delen A Fast, Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications Academic Press, (2012) (ISBN: 9780123870117).

[19] 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.

[20] G.G. Chowdhury, Natural language processing, Annu. Rev. Inf. Sci. Technol. 37 (1) (2003) 51–89, https://doi.org/10.1002/aris.1440370103.