e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

PREDICTION OF CONSTRUCTION SITE ACCIDENTS USING TEXT MINING AND NLP

B. Saritha[*1], T. Meghana[*2], P. Yaswanth Sai[*3], N. Sai Deekshith[*4]

[*1]Assistant Professor, Department Of Computer Science & Engineering Malla Reddy College Of Engineering & Technology Hyderabad, India.

[*2,3,4]Final Year Student, Department Of Computer Science & Engineering Malla Reddy College Of Engineering & Technology Hyderabad, India.

DOI : https://www.doi.org/10.56726/IRJMETS50057

ABSTRACT

Workplace safety is a major concern in many countries. The construction sector is regarded as the most dangerous workplace among various industries. Construction site accidents result in significant financial loss in addition to suffering for human beings. Analysis of accidents is crucial for developing scientific risk control strategies and preventing the recurrence of similar mishaps in the future. Summaries of investigation reports detailing fatalities and catastrophic incidents from previous accidents within the construction industry are available. Given that hazardous objects represent a primary cause of construction accidents, detecting them and scrutinising past accident reports can offer invaluable insights for preventing future mishaps, spanning across diverse industries. Taking everything into account, this research contributes to our understanding of location strategy and its impact on organisational expansion.

I. INTRODUCTION

The construction industry remains one of the most perilous workplaces globally. As per projections by the International Labour Organisation (ILO), approximately 60,000 fatal accidents occur on construction sites worldwide annually, equating to one fatal incident every 10 minutes. Accidents occurring at construction sites not only lead to significant health issues but also substantial financial ramifications. It's important to analyze past accidents to avoid repeating them and to improve workplace safety. Safety experts can take the proper steps to lessen the likelihood of such occurrences happening in the future. Accidents on building sites are frequently caused by the presence of dangerous objects. Such accidents might be avoided by identifying these tools and using them carefully.

A casualty investigation report that provides a thorough account of the tragedy is created following a tragic accident in the construction industry; this text data can be used for further analysis.

Past accident data contains specifics about accidents, and by developing a machine learning algorithm model, we may analyse data to discover accidents' causes and prevent them in the future by providing test data for new work that can forecast accidents' causes and help avoid them. These machine learning (AI) algorithms can aid in the extraction of hazardous materials, such as improperly used tools, surrounding sharp objects, broken equipment, etc. Safety experts can take the appropriate actions to eliminate or decrease the identified causes based on the findings of the cause analysis. The chance of such incidents can be decreased by increasing awareness and demanding routine checks before utilising a machine that malfunctioned and caused an accident earlier.

This study involves the analysis of data from the Occupational Safety and Health Administration (OSHA) using natural language processing (NLP) techniques to evaluate construction site incidents. Firstly, the data is pre-processed and then required text is extracted and vectorisation is performed. Multiple Machine Learning(ML) models are built to analyze the performance and accuracy. Rule (DSCCR), Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), and Two-Dimensional Uncertain Language Variables (2DULVs) was used to.

II. LITERATURE REVIEW

To examine workplace accidents, numerous researchers have employed techniques from natural language processing (NLP) and machine learning models.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

1. In the context of aviation accident prediction, the NLP technique of sentiment analysis is employed to ascertain the sentiment of data, whether it is positive, negative, or neutral. Additionally, various machine learning algorithms such as Naïve Bayes, Random Forest, Support Vector Machine (SVM), Logistic Regression, and Ada Boosting are utilised to enhance the accuracy of predictions regarding aircraft accidents. Despite the imbalanced nature of values in the "fatality" column, the classification algorithms demonstrated notably high f1-scores and accuracy values.
2. Proposed that based on Cosine Similarity and Term Frequency and Inverse Document Frequency (TF IDF), sentences of input are grouped together. Industry specific dictionary is derived, and an optimiser module is used to combine multiple ML algorithms to achieve best performance in categorising and finding the reason for an accident. This strategy demands a large vocabulary of derived keywords and industry-specific terms.
3. To address sequential challenges while considering the textual attributes of construction accident narratives, this paper proposes a hybrid model known as Symbiotic Gated Recurrent Unit (SGRU), which integrates Symbiotic Organisms Search (SOS) with Gated Recurrent Unit (GRU). It represents the inaugural application of a Recurrent Neural Network variant to accurately classify construction site accidents.
4. Introduced a unique model for enhancing mine safety, wherein the data underwent characterisation through descriptive analysis, including measures such as mean and standard deviation. The chi-square test was employed to explore the association between demographic characteristics and mine hazards, while the independent samples t-test was utilised to compare the perspectives of surface miners with those of underground miners.
5. Utilising Feature Selection Algorithms to extract the most critical features, the study employs Decision Tree, KNN, Naive Bayes, and Ada Boost Machine Learning Algorithms to construct models for classifying causes into four categories. Two experiments were conducted to identify the top-performing model, with AdaBoost demonstrating superior performance in both instances.
6. This paper proposes Case-Based Reasoning (CBR) which involves study of previous cases and prevents further incidents. This research suggests a method for risk case retrieval system that combines NLP and vector space modelling. Python is used to construct a prototype system to analyze an input case and return top similar cases.
7. Accident reports sourced from the US OSHA website are classified utilising six machine learning models. The linear SVM exhibited precision ranging from 0.5 to 1, recall from 0.36 to 0.9, and F1 scores spanning from 0.45 to 0.92 across the 11 labels for accident causes or categories. To improve classification performance, the study recommends employing an ensemble technique.
8. It was proposed to use an unsupervised, data-driven K-Means-Based Clustering Approach to divide the collected reports, examining the various discovered groupings and highlighting data patterns Only four categories, each reporting a specific accident type, make up the accident reports.
9. The initial approach involves the standard automatic voting of class labels by all six classifiers, while the second approach relies on manually devised voting rules derived from observations of the predictions generated by the six classifiers. Both methods aim to evaluate the effectiveness of an ensemble approach employing well-established machine learning (ML) algorithms, including Decision Tree, Logistic Regression, Support Vector Machine, Naive Bayes, K Nearest Neighbours, and neural networks. To facilitate a more efficient comparison of the accuracy of the six classifiers, 3-fold cross-validation is conducted. The utilisation of the ensemble classifier leads to more effective categorisation of accident narratives and enables the identification of safety trends and solutions with greater accuracy.
10. R is recommended as the programming language for developing a Natural Language Processing system that utilises manually crafted rules and keyword dictionaries to extract insights from unstructured injury reports stored in databases, aiming to enhance safety management.

III. METHODOLOGY

DATASET

In the proposed approach dataset present in OSHA website is used. The dataset contains abstracts of the accidents and injuries of construction workers during 2015-2017. The dataset consists of the title of accident

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

report, abstract of the accident, type of accident. This dataset contains 4847 records. The dataset is labelled with 11 causes of accidents.

DATA PREPROCESSING

Data preparation, the process of refining raw data to make it suitable for machine learning models, is considered the crucial initial stage in model development. While it's not uncommon to encounter unclean or unprepared data during machine learning projects, it's essential to clean and format the data before proceeding. Hence, data preprocessing is employed to address this requirement whenever working with data.

It involves below steps:

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Dividing the dataset into training and testing subsets.
- Feature scaling

NLP

Natural language processing (NLP), a machine learning technique, empowers computers to comprehend, manipulate, and analyze human language. NLP software processes text data, analysing both sentiment and meaning of messages, and offers real-time responses during human conversations. To ensure comprehensive and efficient text analysis, various approaches in natural language processing are employed, including K-Means Clustering, 2DULVs, TOPSIS, and DSCCR.

LOWER CASING

The initial and commonly adopted text processing technique is converting the data to lowercase. This entails transforming variations such as 'DATA', 'Data', 'DaTa', and 'DATa' into 'data', maintaining consistency with the casing style of the input text. Tokenization refers to breaking running text into phrases and words, essentially segmenting text into token-sized units while eliminating specific characters like punctuation. Removing punctuation from text facilitates equitable treatment of each text segment.

STOPWORD REMOVAL

Stop words removal includes eliminating words like "and" "the," and "to" that are commonly used in English-speaking identifiers and prepositions. Some very common words that seem to offer little to no value to the NLP goal are filtered and removed from the text to be processed during this process, thereby removing widespread and frequent terms that are unhelpful of the corresponding text. By performing a lookup in a predefined list of keywords, stop words can be safely disregarded, freeing up database space and speeding up processing. No complete collection of stop words exists. These may be chosen beforehand or created from the beginning.

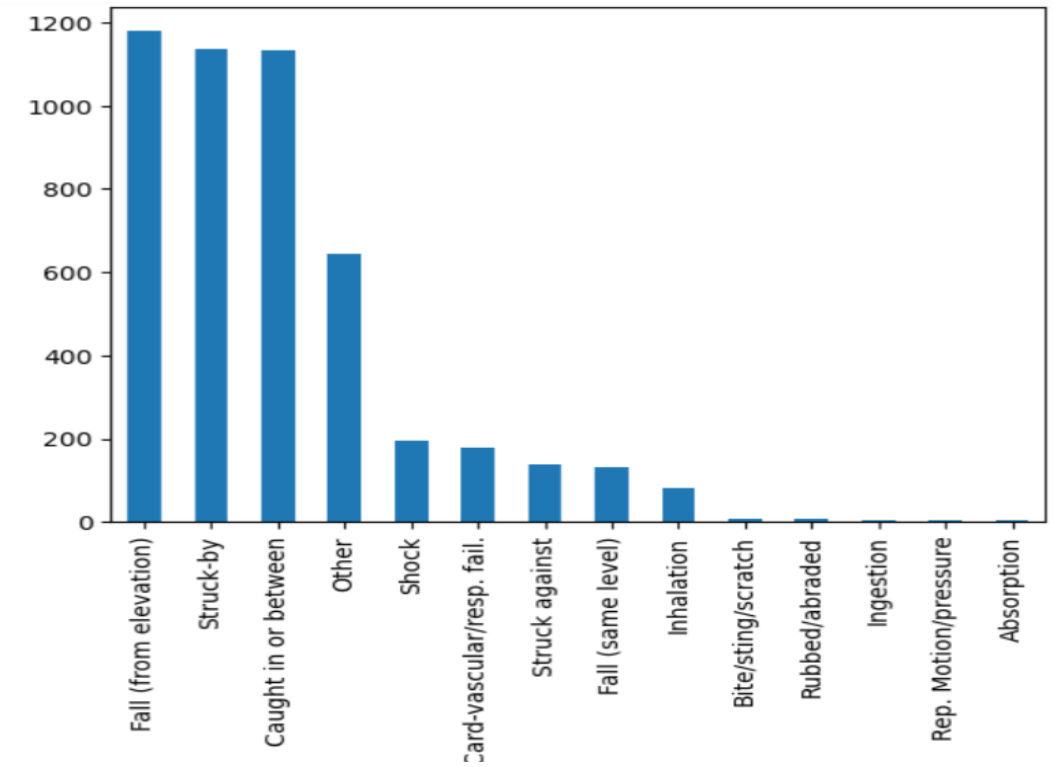
VECTORIZATION

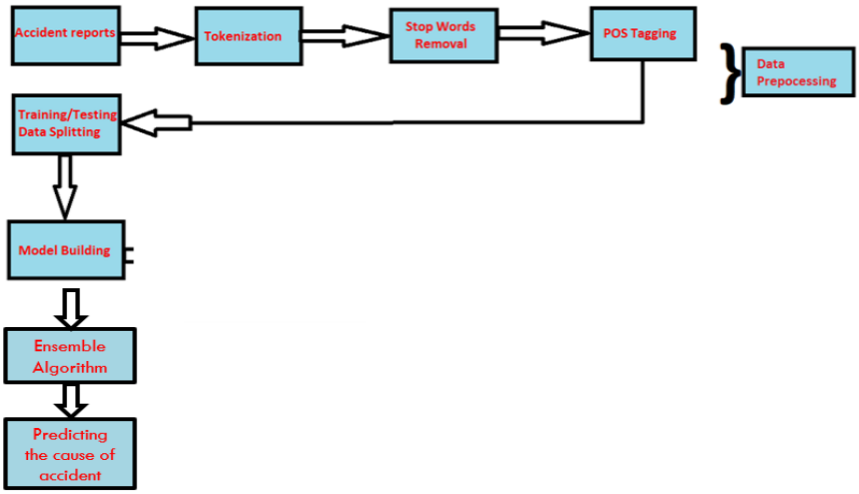
To advance with subsequent processing, the text undergoes transformation into numerical vectors via a method known as vectorisation. These vectors are subsequently employed as inputs for various machine learning algorithms.

The Python library scikit-learn provides a valuable tool named Count Vectoriser for this purpose. Count Vectoriser converts the given text into a vector by counting the occurrences of each word across the entire text. Each text sample within the document corresponds to a row in the matrix generated by Count Vectoriser, with each distinct word represented by a column in the matrix.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

DATA VISUALIZATION





IV. RESULT AND ANALYSIS

This section discusses and compares the outputs of each machine learning algorithm, analyzing their performance on the available dataset. The assessment metrics used for comparison include precision, recall, accuracy, and F1-Score.

Classification accuracy, defined as the percentage of properly classified samples over all samples obtained, was employed for evaluation. Among the algorithms, SVM and Logistic Regression achieved the highest accuracy rate of 70%. Logistic Regression demonstrated 46% precision, 47% recall, and a 46% F1 score, while SVM exhibited 45% precision, 45% recall, and a 44% F1 score.

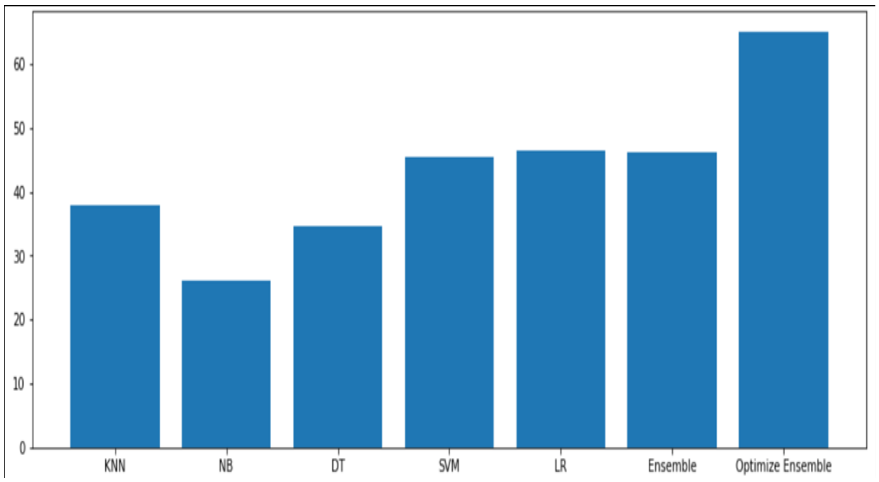
The Random Forest algorithm yielded an accuracy of 69%, with precision, recall, and F1 values of 46%, 47%, and 46%, respectively. Decision Tree achieved an accuracy of 60%, with precision, recall, and F1 scores of 34%, 32%, and 32%, respectively.

KNN demonstrated an accuracy of 55%, with precision, recall, and F1 scores of 38%, 31%, and 31%, respectively. The Naive Bayes model achieved an accuracy of 51%, with precision, recall, and F1 scores of 26%, 27%, and 25%, respectively.

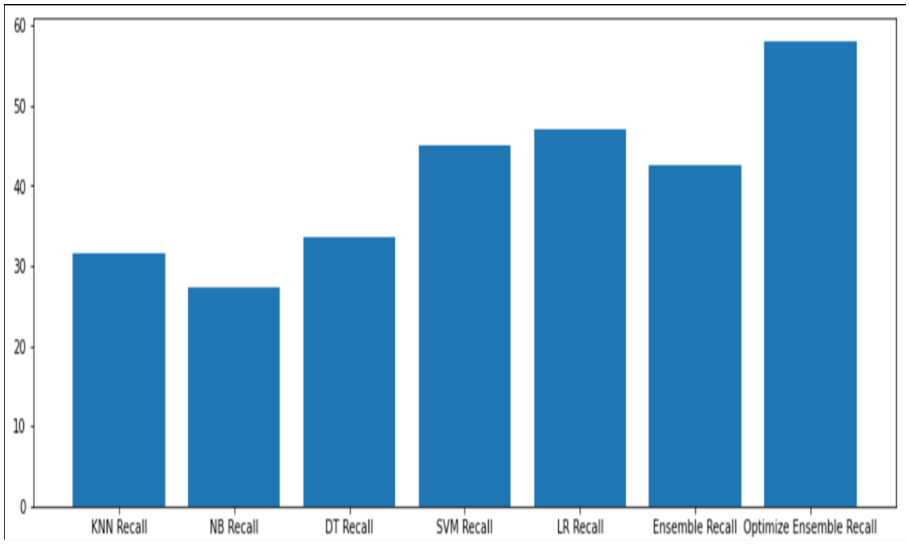
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

Comparatively, the Voting Classifier, when compared to individual classifiers, exhibited slightly lower accuracy at 68%, which was closest to the highest accuracy achieved. It demonstrated a precision rate of 46%, a recall rate of 40%, and an F1 score of 38%.

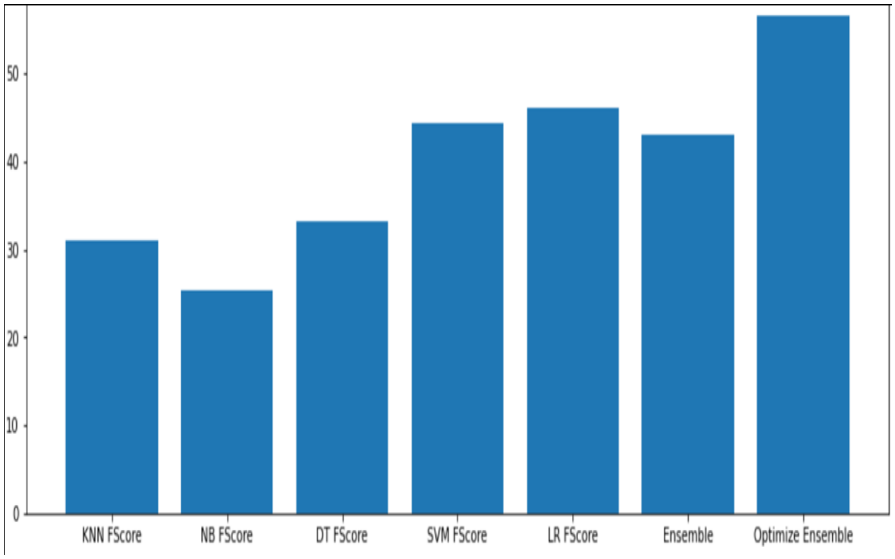
Precision graph



Recall graph

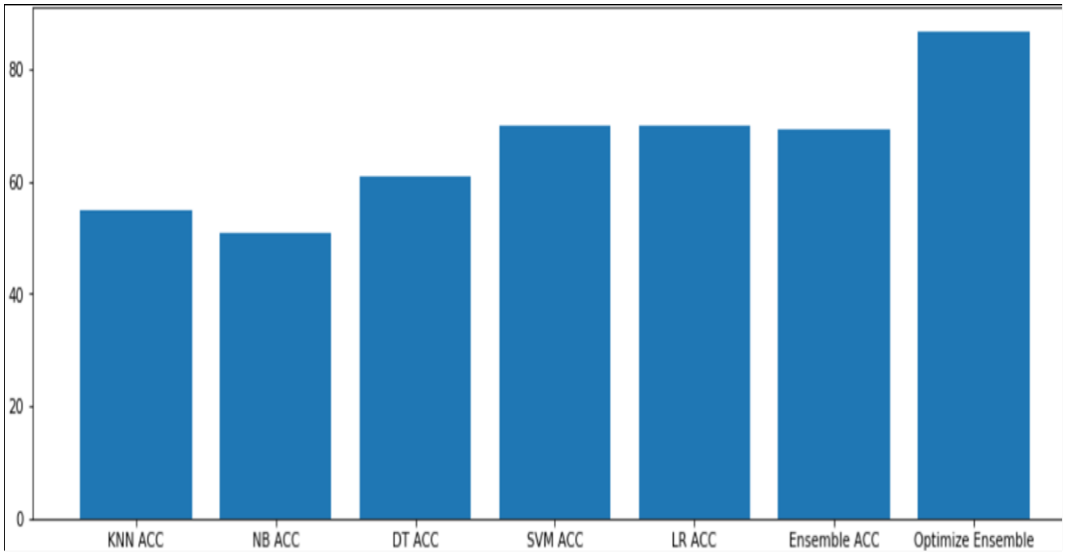


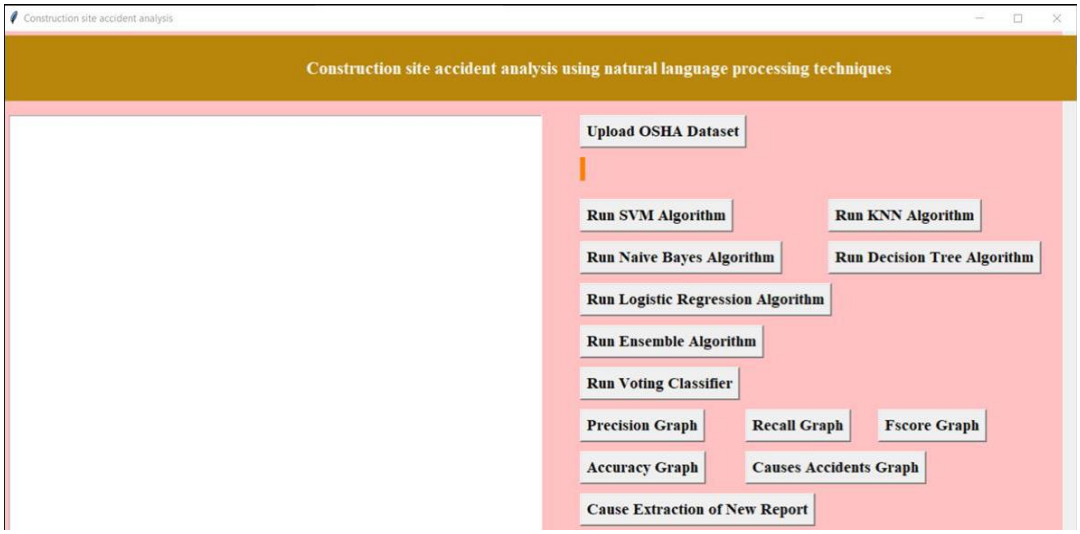
Accuracy Graph



e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

Score graph





V. CONCLUSION

Analysing construction disaster reports is crucial for preventing future mishaps by gaining valuable insights into past occurrences. Categorising accident causes is essential, considering the variety of factors involved in developing prevention techniques. Identifying hazardous elements is equally important for enhancing workplace safety, as it enables proactive measures to mitigate risks associated with these objects. However, manual categorisation of accident reports and examination of hazardous materials at construction sites require significant time and effort.

VI. REFERENCES

[1] Zhang, F., Fleyeh, H., Wang, X., & Lu, M. (2019). Utilising text mining and natural language processing techniques for analysing construction site accidents. Automation in Construction, 99, 238-248.

[2] Labib, M. F., Rifat, A. S., Hossain, M. M., Das, A. K., & Nawrine, F. (2019, June). Road accident severity analysis and prediction using machine learning in Bangladesh. In 2019 7th International Conference on Smart Computing & Communications (ICSCC) (pp. 1-5). IEEE.

[3] Sankarasubramanian, P., & Ganesh, E. N. (2020). Analysis of industrial accident reports employing natural language processing techniques. International Journal of Scientific & Technology Research, 9(6), 470-475.

[4] Cheng, M. Y., Kusoemo, D., & Gosno, R. A. (2020). Construction site accident classification based on text mining and hybrid supervised machine learning. Automation in Construction, 118, 103265.

[5] Jiskani, I. M., Cai, Q., Zhou, W., ChangZ., Chalgri, S. R., Manda, E., & Lu, X. (2020). Developing a distinctive model for sustainable mining safety in Pakistan. Mining, Metallurgy & Exploration, 37(4), 1023-1037.

[6] Chokor, A., Naganathan, H., Chong, W. K., & El Asmar, M. (2016). Analysis of Arizona OSHA injury reports using unsupervised machine learning. Procedia Engineering, 145, 1588-1593.

[7] Rupasinghe, N. K. A. H., & Panuwatwanich, K. (2021). Understanding construction site safety hazards through open data: A text mining approach. ASEAN Engineering Journal, 11(4), 160-178.

[8] Pirge, G. (2021). Analysis of F-16 accidents using natural language processing. January.

[9] Zou, Y., Kiviniemi, A., & Jones, S. W. (2017). Retrieving similar cases for construction project risk management using natural language processing techniques. Automation in Construction, 80, 66-76.

[10] Goh, Y. M., & Ubeynarayana, C. U. (2017). Classification of construction accident narratives.