www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

# MACHINE LEARNING-BASED AUTOMATED CLASSIFICATION OF WORKER-REPORTED SAFETY REPORTS IN CONSTRUCTION

SUBMITTED: December 2021
REVISED: September 2022
PUBLISHED: November 2022
EDITOR: Esther Obonyo
DOI: 10.36680/j.itcon.2022.045

***Nikhil Bugalia, Assistant Professor,***
***Department of Civil Engineering, Indian Institute of Technology Madras, India***
***E-mail: nikhilbugalia@gmail.com***

***Vurukuti Tarani, Graduate Student***
***Department of Civil Engineering, Indian Institute of Technology Madras, India***
***E-mail: ce19m012@smail.iitm.ac.in***

***Jai Kedia, Undergraduate Student***
***Department of Civil Engineering, Indian Institute of Technology Madras, India***
***E-mail: ce17b037@smail.iitm.ac.in***

***Hrishikesh Gadekar, Graduate Student***
***Department of Civil Engineering, Indian Institute of Technology Madras, India***
***E-mail: ce17b116@smail.iitm.ac.in***

***SUMMARY:*** *Limited academic attention has been paid to the applicability of Machine Learning (ML) approaches for analyzing worker-reported near-miss safety reports, as opposed to injury reports, at construction sites. Although resource-efficient analysis through ML of large volumes of such data at construction sites can help guide practitioners in decision-making to prevent injuries. The current study addresses this research gap by evaluating the relevance of ML approaches through quantitative and qualitative methods for scaling efficient near-miss reporting programs at construction sites. The study uses an extensive experimentation strategy consisting of input data processing, n-gram modeling, and sensitivity analysis. It first tests the proposition that, despite the data-quality challenges, the high performance of different ML algorithms can be achieved in automatically classifying the textual near-miss observations. The study relies on worker-reported near-miss data collected from a real construction site in Kuwait. The classification performance of various ML approaches is evaluated using F1 scores for three academically novel but commonly used category labels at the sites - "Unsafe Act (UA)," "Unsafe Condition (UC)," and "Good Observation (GO)." In addition, the practitioner's input was utilized to assess the practical applicability of ML classifiers for construction sites. The conventional Logistic Regression (LR) classifiers have a comparatively high F1 score of 0.79. However, ML classifiers faced challenges in distinguishing between UA and UC. Further, the analysis reveals that optimal ML classifiers may lose on being acceptable to human decision-makers. Overall, despite the promising performance of ML tools for the near-miss data, the sites with low maturity of reporting systems may find themselves unable to leverage ML to scale their reporting systems. A simplified experimentation strategy like the current study could help practitioners identify the data-specific optimal ML approaches in future applications.*

***KEYWORDS:*** *Automated construction; Machine Learning; Construction; Safety; Near-Miss Reporting; Neural Networks; n-grams.*

***REFERENCE:*** *Nikhil Bugalia, Vurukuti Tarani, Jai Kedia, Hrishikesh Gadekar (2022). Machine learning-based automated classification of worker-reported safety reports in construction. Journal of Information Technology in Construction (ITcon), Vol. 27, pg. 926-950, DOI: 10.36680/j.itcon.2022.045*

# 1. INTRODUCTION

Globally, the construction sector is infamous for high Occupational Health and Safety (OHS) issues (Manu et al., 2019). In this regard, a proactive safety management approach relying on learning from near-miss events instead of accidents and injuries is deemed necessary (Bugalia et al., 2021; Xu et al., 2021). Occupational Safety and Health Administration (OSHA) (https://www.osha.gov/incident-investigation) in the United States defines near-misses as incidents in which a worker might have been hurt if the circumstances had been slightly different. Subsequently, researchers and academicians have recommended developing a near-miss reporting program at construction sites to prevent accidents, injuries, and fatalities (Oswald et al., 2018). An efficient near-miss reporting program is where many workers proactively report near-misses at sites. These reports are analyzed to identify accident or injury precursors trends and implement corrective actions at the site to prevent injuries or fatalities (Bugalia et al., 2021).

However, construction organizations across the globe continue to face issues in scaling and sustaining an efficient near-miss reporting program, mainly due to its resource intensiveness (Bugalia et al., 2021; Oswald et al., 2018). With an increase in the number of reports, more efforts are also needed by the construction organizations to analyze the reports and provide timely feedback to the reporters. Without such activities by the management, employee motivation to report near misses can decrease significantly, putting the whole program in jeopardy (Bugalia et al., 2021). Hence, efforts are also needed to develop analytical tools to help expedite the analysis process for the large volume of near-miss reports generated on construction sites (Goh and Ubeynarayana, 2017).

Near-miss reports at construction sites, also known as Safety Observations (SOs), are generally unstructured textual data. Such reports often require substantial manual classification efforts before reliable trends about the various near-misses can be developed (Goh and Ubeynarayana, 2017). Even though the classification of SOs using computer-based automated techniques may not generate new knowledge per se, the high performance of the automatic classification approaches is deemed necessary for enhancing the scale of the near-miss reporting program (Goh and Ubeynarayana, 2017). Furthermore, such classification is also essential for further analyzing the data, such as for risk-factor identification and prediction (Baker et al., 2020; Tixier et al., 2016a). To solve similar practical issues faced by the decision-makers, efficient processes for analyzing text-based information (such as the SOs) using Machine-Learning (ML) approaches have been suggested in the literature (Goh and Ubeynarayana, 2017; Sarkar and Maiti, 2020; Tixier et al., 2017, 2016b, 2016a).

However, in the literature focusing on the construction sector, only limited attention has been given to applying ML approaches for data containing near-miss SOs, as opposed to accident and injury reports (Baek et al., 2021; Fang et al., 2020; Sarkar and Maiti, 2020). To the best of the author's knowledge, only one previous study, i.e., (Fang et al., 2020), focused on analyzing near-miss reporting data obtained from a database managed by safety professionals across multiple metro construction sites in China (Baek et al., 2021). Even in that study, the language of the original dataset, i.e., Chinese, is deemed a limitation to represent the large quantity of data generated in the English language in construction sites worldwide (Baek et al., 2021). Furthermore, there is a concern that such standardized and well-maintained injury databases often utilized in previous studies rarely represent the large quantities of data generated on real-construction sites (Baek et al., 2021; Yan et al., 2020). In previous studies, commonly used data sources are often well-processed, well harmonized, industry-wide data sources managed by industry-wide bodies (Baek et al., 2021). For example, Goh and Ubeynarayana (2017) utilize the accident reports maintained by safety professionals certified by OSHA. However, the literature also supports workers' dominant role in reporting SOs instead of certified safety professionals (Zhou et al., 2019). Under such circumstances, language constraints typical of many large-scale sites in the world (Ne'Matullah et al., 2021; Trajkovski and Loosemore, 2006) and a lack of safety awareness among construction workers could affect the quality of textual description and the classification labels in the original SO data (Bugalia et al., 2021; Fang et al., 2020). Such quality issues in the original data may challenge the ML algorithms from performing well in classification tasks (Fang et al., 2020). Hence, the applicability of ML approaches for near-miss reporting data with numerous quality issues that could represent the reality of the construction sites must be tested (Baek et al., 2021; Sarkar and Maiti, 2020). However, such a topic has not received much attention in the previous literature (Fang et al., 2020; Goh and Ubeynarayana, 2017; Yan et al., 2020) and is an essential academic gap that the current study aims to address.

Therefore, to understand the relevance of ML approaches for scaling and sustaining near-miss reporting programs for construction sites, the current research aims to comprehensively evaluate ML approaches using a mix of quantitative and qualitative methods. The study's primary objective is to test whether the high performance of different ML approaches in automatically classifying the text in near-miss SOs, obtained from real construction sites, can be achieved. The study relies on rigorous experimentation-based on technical features of ML approaches and quantitative comparison of classification performance to reveal the optimal ML classifiers. Beyond technical features of classification performance, the study qualitatively assesses the relevance of ML approaches for scaling and sustaining near-miss reporting programs by considering organizational decision-making factors and practitioners' perspectives pertinent to ML implementation. The main contributions of the study are listed below.

- Towards technical contribution, the study provides essential validation for ML approaches' capabilities in achieving high classification performance for usually poor-quality near-miss data obtained directly from construction sites. Previous studies have rarely focused on large quantities of near-miss data representative of site realities (Baek et al., 2021). Further, due to its analytical focus on a novel type of data, i.e., near-misses instead of injury narratives, the study also presents insights for classifications tasks related to previously unexplored labels, i.e., Unsafe Acts (UAs), Unsafe Conditions (UCs), and Good Observations (GOs) (Baek et al., 2021). These labels are typical to the near-miss reporting systems globally but are rarely included in the previous literature.
- The study also provides insights on several contemporary topics in ML applications of construction safety-specific data, such as performance comparison between conventional and deep-learning methods (Baker et al., 2020) and clarifying the relationship between dataset size and classification performance.
- For organizational and management aspects, the study comprehensively assesses the relevance of ML approaches for near-miss reporting systems and reveals strategies that could be used to enhance the implementation of ML approaches in practice.

The study is structured to demonstrate all the contributions mentioned above. Section 2 summarizes the essential experimentation strategies for improving the classification performance of various ML approaches while highlighting the essential literature gaps supporting the study's contribution. Section 3 describes the data utilized, its relevance for construction sites across the globe, and the adopted methodology. Section 4 presents the results from the analysis of the textual data. Section 5 offers the main discussions centered around the study contributions. Conclusions have been summarized in section 6.

# 2. LITERATURE REVIEW

The current study relies on previous studies focusing on ML applications for safety-related textual data to identify technical details necessary for improving the classification performance of ML classifiers. Such technical details are then utilized for the current study's design. Considering the necessity to reduce the resource intensiveness of analysis methods, only fully automated ML approaches that do not require any intermediate manual analysis inputs are deemed suitable for the scope of this study. Section 2 begins with a brief overview of the critical concepts involved in improving the performance of ML classifiers.

## 2.1. Overview of technical details of various ML approaches

The two main techniques for solving text-mining tasks are – Bag-of-Words (BoW) representation and word embeddings representation (Baker et al., 2020). A brief description of these techniques is as follows.

### 2.1.1. BoW representation-based techniques

A BoW representation is often utilized to transform the unstructured free-text data into a structured representation. As per BoW, a given text row, known as a document, is represented as a vector of terms. This vector's size is equal to the size of a dictionary consisting of all the unique terms (or tokens) in the pre-processed data used in the study. In this vector, all values are zero except for the dimensions corresponding to the terms in the document (Goh and Ubeynarayana, 2017). Term-Frequency-Inverse-Document-Frequency (tf-idf) vectorizer (Peng et al., 2014) is used to convert this meaningful text into an array that can be used as a feature to develop the model. To date, the BoW-based methods continue to be one of the most prominent approaches for text mining in various sectors, including construction (Sarkar and Maiti, 2020). Commonly used supervised learning algorithms utilizing tf-idf representations are – Support Vector Machine (SVM), Logistic Regression (LR), Bernoulli Naïve Bayes (BNB), Multinomial Naïve Bayes (MNB), Random Forest (RF), and Decision Tree (DT). The K-means algorithm is a popular unsupervised learning algorithm that utilizes tf-idf representation (see Table 1 and (Baek et al., 2021)).

*TABLE 1: Summary of text classification literature for construction safety (source: authors)*

| ***S.No*** | ***Reference*** | ***Data Source*** | ***Data Size (Spilt ratio)*** | ***Pre-processing*** | ***Algorithms used*** | ***Labels*** | ***Performance*** | ***Recommendation for performance improvement*** |
|---|---|---|---|---|---|---|---|---|
| 1 | (Chokor et al., 2016) | OSHA, Arizona, accident | 513 (NA) | Stop-word removal, lemmatization | K-Means clustering (unsupervised) | 4 (falls, struck by objects, among others) | Four distinct clusters could be identified representing different accident class | Increase sample size |
| 2 | (Tixier et al., 2016a) | Injury data from 470 contractors | 4398 (95-5) | Attribute identification (Tixier et al., 2016a) | RF, SGTB | 7 injury type labels | RPSS. RF (0.068), SGTB (0.236) Typical good range for RPSS [0.05,0.2] | The attribute-based framework is viable in producing structured accident data from unstructured reports |
| 3 | (Tixier et al., 2016b) | The manually analyzed incident report from the literature | 2201 (94 – 6) | The rule-based automated content analysis system | - | > 80 attributes | F1 score. 95% | Manual rule-based content analysis |
| 4 | (Goh and Ubeynarayana, 2017) | OSHA, accident | 1,000 (80-20) | Manual labeling, stop-word removal, word stemming, tf-idf | SVM, LR, RF, KNN, DT, NB | 11(caught in between falls and fire, among others.) | F1 score, LSVM. Average – 0.67. Max – 0.92. Min – 0.52 | N-gram modeling; Optimization; increase the sample size |
| 5 | (Marucci-Wellman et al., 2017) | Insurer, incident reports | 30000 (50-50) | A small list of stop-words, pre-processing | SVM, LR, NB (Unigram, Bi gram) | Explosion, falls and many more BLS, OIICS 2012 | Recall – LR (70%) – Algorithm alone; 93% - With about 30 – 40% manual coding | Human-machine paiNLP-based rules |
| 6 | (Zhang et al., 2019) | OSHA, accident | 1,000 (80-20) | Stop-word removal, stemming and lemmatization, part of speech tagging, tf-idf | SVM, KNN, DT, LR, NVB | 11 (caught in between falls, fire) | F1 score, SVM, Average – 0.68. Max – 0.87 | Rule-based chuker; increase the sample size |
| 7 | (Baker et al., 2020) | Global industrial partner in the oil and gas sector, incident reports | 90,000 (90-10) | Translation of 25% of accounts, pre-training | Deep-Learning (HAN, RNN);<br>Tf-idf -SVM | Incident type (7), Injury type (4), Body part (6), Severity (2) | F1 scores. Incident type (71.55, tf-idf SVM), Injury type (82.26, tf-idf SVM), Body part (86.34, tf-idf SVM), Severity (82.88, tf-idf SVM) | None |
| 8 | (Fang et al., 2020) | Wuhan metro group, near-miss reports | 3280 (80-20) | Translation to English, | Deep-Learning tf-idf(BERT) | 170 categories | F1 score. 86.91% | Hyperparameter tuning |
| 9 | (Zhang, 2022) | OSHA, accident | 1280 (80-20) | Stop-word removal, part-of-speech tagging, lemmatization, Word2Vec encoding | Deep-Learning (ANN)m tf-idf – SVM, DT, LR, KNN | 11(caught in between falls and fire, among others.) | F1 score, Word2Vec ANN. Average – 0.69. | Construction-specific word embeddings generation, increase the sample size, hyperparameter tuning |
| Spilt-ratio: x – y denotes x% of the total data used for training the algorithm, whereas the remaining y% (or 100 – x%) is used for testing purposes. Abbreviations – OSHA – Occupational Safety and Health Administration, USA. SVM – Support Vector Machine, LR – Logistic Regression, RF – Random Forest, DT – Decision Tree, NM – Naïve Bayes, KNN – K-nearest neighbors algorithm, HAN – Hierarchical Attention Network, RNN – Recurrent Neural Network, BERT - Bidirectional Transformers for Language Understanding, SGTB – Stochastic Gradient Tree Boosting, RPSS – Rank Probability Skill Score | | | | | | | | |

### 2.1.2. Word embedding-based techniques

The enormous size (also known as dimensionality) of vector space generated from processing a large quantity of unstructured textual data in BoW representation is considered an important limitation for improving classification performance (Baker et al., 2020). Alternatively, in word embeddings-based representation of textual data, each word in the vocabulary is denoted using a small, dense vector (compared to BoW vector) in the space of shared concepts. Therefore, a document can be represented using the corresponding word embeddings (Baker et al., 2020). Literature supports using the word embeddings-based representation to be used with deep-learning-techniques, such as Convolutional Neural Network (CNN) (LeCun et al., 1998), to achieve high performance in text-mining tasks (Baker et al., 2020).

## 2.2. Experimentation with automated approaches for improving classifier performance

Typically, the classification performance of a classifier is measured using an F1 score. F1 score, an overall measure of the prediction performance of a given classifier, is the harmonic mean of both Precision and Recall. Precision measures how accurate the actual predictions are, whereas Recall measures the proportion of true positives identified by the classifier. The F1 score could range between 0 and 1, and a high F1 score corresponds to better classification performance (Goh and Ubeynarayana, 2017). It is essential to note that despite significant work in the field, there are still no specific ML approaches that perform consistently across all types of textual data. Hence, an experimentation approach is recommended to identify the most suitable ML algorithm specific to a data set (Baek et al., 2021). The details of various experiments reported in the literature to achieve a higher classification performance for ML classifiers are discussed below.

### 2.2.1. Performance for BoW representation-based classifiers

For conventional BoW-based automated algorithms using supervised learning, average F1 scores range from 0.67 to 0.8 (see Table 1). Furthermore, the performance is also better for classification categories constituting a significant proportion of the total data (Baker et al., 2020; Goh and Ubeynarayana, 2017a).

### 2.2.2. Modifications of BoW – topic modeling and n-gram modeling

BoW representation poses several significant challenges limiting its performance for various applications. First, the vector space thus generated for a large quantity of unstructured text is enormous, limiting the prediction capabilities of different ML classifiers (Baker et al., 2020). Therefore, unsupervised learning techniques have been proposed to cope with the large dimensionality of BoW representation and improve performance (Chokor et al., 2016; Sarkar and Maiti, 2020).

Second, BoW ignores word order, limiting the semantic meaning derived from such representations (Baker et al., 2020). Therefore, a combination of tokens instead of single tokens can be taken to capture the words locally. A combination of tokens is typically referred to as n-grams, with "n" representing the number of tokens taken together. Previous studies report a positive contribution of n-gram modeling for higher classification performance (Goh and Ubeynarayana, 2017; Zhang et al., 2019).

### 2.2.3. Moving beyond BoW towards word embeddings

Word embeddings-based representation has also been proposed as an innovative alternative to BoW representation. Notably, Baker et al. (2020) implemented word embeddings-based representation using Deep-learning techniques for construction safety data. Although in their study, the tf-idf-based SVM classifier performed better than the deep-learning techniques. Zhang (2022) developed construction-specific word embeddings and reported marginal improvement in the performance of the deep-learning techniques compared to the conventional methods. At present, only a limited number of studies in the construction sector have utilized deep-learning algorithms for accident or incident classification, and more applications should be explored (Baek et al., 2021; Baker et al., 2020; Fang et al., 2020).

### 2.2.4. Variation of data size and its effect on performance

Several previous studies (see Table 1) have also suggested that an increase in dataset size is expected to improve the classification performance (Géron, 2019; Ng, 2019). A considerable variation is observed in the size of the datasets used in the studies within the literature. However, none of the previous studies have tested the variation of the F1 score as the sample size changes (see Table 1). Moreover, even when data sources as big as 90,000 observations were utilized, the F1 scores thus obtained were only marginally better (see results from (Baker et al., 2020) and (Goh and Ubeynarayana, 2017) in Table 1). Therefore, the relationship between the prediction performance and sample size needs to be further examined, especially for their relevance to improving performance for near-miss data.

## 2.3. Organizational decision -making factors for ML implementation in near-miss reporting

In addition to high performance, several organizational decision-making factors pertinent to ML implementation should also be considered to ensure the relevance of ML approaches towards scaling and sustaining a near-miss reporting program for the construction sector (Demirkesen and Tezel, 2022). Some factors potentially affecting organizational decision-making for the adoption of digital tools in construction include – concerns for the tool's application for data representative of real-site conditions, factors causing variability in the performance of techniques, resource constraints of techniques, and acceptability of the approaches by existing decision-makers, among others (Wang et al., 2020). Moreover, ethical considerations such as privacy and biased decision-making against individuals in an organization are among other relevant factors governing the adoption of digital tools in construction (Wang et al., 2020).

However, evaluating the various ML classifiers on a full range of organizational decision-making factors is still challenging as the literature lacks a comprehensive framework of such factors specific to ML applications (Demirkesen and Tezel, 2022; Wang et al., 2020). Developing such a comprehensive framework is beyond the current study's primary scope. However, the analytical results obtained from the experimentation with various ML classifiers can still be leveraged to make an objective assessment for establishing the relevance of ML approaches in the construction sector (Baker et al., 2020; Goh and Ubeynarayana, 2017). For example, a performance comparison between conventional (such as LR and LSVM) and computationally complex (CNN) ML approaches can make the relative resource benefits of the two approaches evident to practitioners. Furthermore, information on a factor potentially causing variability in the performance of ML approaches, i.e., dataset size, can be obtained by clarifying the relationship between prediction performance and the dataset size. Such information can help establish the ML approach's relevance for practical applications, especially in small-scale construction sites.

Similarly, the lack of visibility and interpretability of intermediate processing steps of various ML algorithms tasks has been received with doubts by organizational decision-makers, affecting ML adoption (Demirkesen and Tezel, 2022). However, a recent study proposed a word saliency-based approach that can identify the regions of a given report that significantly influence the predicted category label for a trained ML classifier (Baker et al., 2020). Such an approach can provide a sneak peek into the ML algorithm's functioning. Depending on the intuitive nature of these salient words to the human analysts, the acceptability of the ML models to human decision-makers could be assessed (Baker et al., 2020).

Furthermore, the literature also calls for adopting a participatory approach with industry stakeholders to seek input on various organizational decision-making factors pertinent to ML implementation (Demirkesen and Tezel, 2022; Poh et al., 2018; Yan et al., 2020). However, previous academic studies on applications of ML approaches in construction have focused on improving performance and rarely on understanding the organizational decision-making factors around these approaches (Poh et al., 2018; Sarkar and Maiti, 2020).

Finally, for the current scope of the study, the authors do not expect to observe differences in ethical considerations across different ML algorithms. None of the ML algorithms used in the current research rely on any personal information of the workers or organizations involved to predict classification labels. Furthermore, the performance of all the algorithms is benchmarked for the collective dataset rather than the reports obtained from individuals or groups, precluding any biases in decision-making if the proposed ML approach were to be implemented in the construction organization. An in-depth understanding of such ethical considerations could have been obtained through practitioners' knowledge of implementing ML algorithms in their respective organizations. While such attempts to engage with practitioners have been made in the current study, such ethical issues were not raised during the open-ended discussions in the current study and hence cannot be evaluated further.

## 2.4. Main research gaps

Synergistic to the essential contribution of the study highlighted previously, the review presented so far helps identify two prominent research gaps related to technical and organizational aspects.

For the technical aspects, a lack of literature for utilizing the near-miss data for ML applications has been identified (Baek et al., 2021). Therefore, by assessing the effectiveness of the various ML approaches in achieving high prediction performance for the near-miss reports collected from real construction sites, the current study clarifies the potential contribution of such approaches in assuring a scalable and sustainable near-miss reporting system. In implementing so, the current study can also reveal the specifics of classification-related tasks for a new type of data and classification labels representative of the construction sites' realities. Instead of developing new ML approaches, the primary methodology adopted in this study is guided by the various experiments conducted in previous studies (as summarized in section 2.2). Implementing such experimentation strategies will also help the study contribute to contemporary technical topics for which the academic debate is not yet settled. For example, the relationship between the dataset size and classifier prediction performance is a topic requiring further attention. Similarly, the relative performance of computationally advanced deep learning-based methods over conventional methods also needs further examination (see Section 2.2.3 and 2.2.4).

Further, the previous literature notes the usefulness of the analytical information obtained from the classification process in identifying organizational decision-making factors. For example, the word saliency-based approach can help assess the acceptability of the developed ML models to human decision-makers. In conjunction with the participatory approach with industry stakeholders, such analysis can provide a comprehensive and objective assessment of the relevance of ML approaches for scaling and sustaining near-miss reporting programs for construction sites. However, a paucity of previous studies systematically examining organizational decision-making factors relevant to ML implementation in construction has been identified. It is one of the essential gaps that the current study addresses.

## 3. METHODOLOGY

To achieve the technical objectives of the study, the automated analytical approaches that could lead to high classification performance were used for experimentation, as discussed in section 2.2. Various results obtained through multiple classifiers of varying sophistication and investigations related to input data, such as dataset size and mislabel corrections, further help illustrate the inherent challenges of applying ML tools in near-miss data. These results were then shared with industry stakeholders to gather rich information on the organizational decision-making factors pertinent to using ML for near-miss reporting in the construction sector. The analysis results combined with inputs from industry stakeholders obtained through in-depth discussion sessions help clarify the relevance of ML approaches for near-miss reporting systems. The details of the various steps have been described as follows.

### 3.1. Data

The data used in this study is collected from a large-scale construction site managed by a consortium of international contractors on a natural gas plant in Kuwait. The near-miss data availability and sharing-related constraints prevalent in safety-critical industries influenced the site selection for data collection. The front-end immigrant workers from several non-English speaking developing countries such as India, Bangladesh, Sri Lanka, Pakistan, and Venezuela represented most workers at the site. As per the prevalent reporting practice at this site, the focus is to promote reporting from the front-end workers as much as possible, rather than obtaining SOs only from safety supervisors (Kedia et al., 2021). The workers provide SOs by writing them in a Safety Observation Card (SOC), then handing them over to the safety staff to convert them to digital textual format. SOC provides the workers the opportunity to write a brief description of the SO and categorize the SO into the categories such as UA, UC, or GO. UAs refer to individual acts that could negatively affect safety, such as not wearing protective equipment. UCs refer to dangerous conditions (site-related, management, environment-related) that could lead to accidents, such as an on-site dig conducted without barricading. GOs refer to the good behaviors or conditions observed at the construction site that help promote and improve safety. The current study assesses ML classifiers' performance only for UA, UC, and GO categories. The focus on these categories is also academically novel, as none of the previous studies have focused on such categories (see Table 1). From an academic perspective, such a classification of SOs (i.e., UA, UC) may be deemed oversimplified. However, from a construction practice perspective, such a classification scheme is commonly found for near-miss reporting systems globally, partially due to the continued focus on behavior-based safety approaches in construction (Bugalia et al., 2021; Oswald et al., 2018; Zhou et al., 2019).

Authors could access the data of about 12,500 SOs made in three months at the given site (see Fig. 1). About 50% of the SOs were categorized as UCs, about 30% as UAs, and 20% as GOs. Such a large volume of data represents a large-scale construction site with a relatively mature reporting culture (Bugalia et al., 2021). For example, about 21,000 workers worked at the site during the three months of data collection, reporting about 4 million manhours per month. Correspondingly, 4000 SOs a month represents a large-scale construction site, even from the global standards (Oswald et al., 2018).

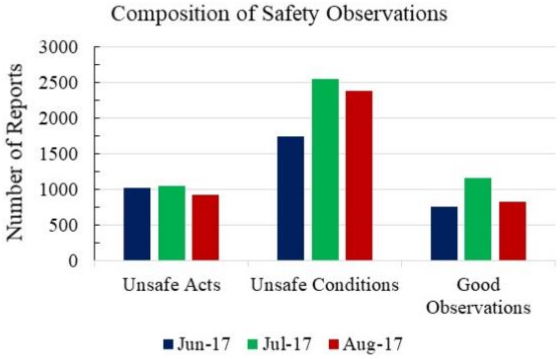


*FIG. 1: SOs obtained from the site by observation type*

The data obtained in the current study is also marred by several data-quality-related issues such as short descriptions, grammatical errors in sentences, and misspelled words. For example, as discussed in a preliminary version of the current study, about 7.26% of the data words were misspelled (Kedia et al., 2021). The proportion of misspelled words in the current dataset is very high compared to 1.87% of words in the OSHA data from the USA (Goh and Ubeynarayana, 2017). The current data's average word count for each SOs stood at 8.8 compared to 50+ for each SO in the OSHA database (Kedia et al., 2021). Such short descriptions may hamper the ML algorithms' feature extraction abilities from the input data. Some other issues in data quality have also been listed in Table 2.

As per the authors' discussion with the practitioners, the characteristics of the textual data obtained in the current study could be attributed to immigrant workers' poor English language ability. The following excerpts from the interview with practitioners show how the presence of non-English speaking workers affects the quality of the safety-related data collected at the site –

*"…at any international construction site, you have people coming from different countries, and not all people are very conversant or very versed with the language or in articulating their thoughts. They are experts in their field, but that need not be their expertise in, you know, expressing or English language…because they have to write something in English, and they may not be so comfortable writing long sentences or picking the right word and correct grammar, etc. So, that's definitely a challenge [for data quality]."*

While the data was obtained from a large-scale construction site in a developed country, this study's data and analysis can reveal lessons applicable to construction sites worldwide, as the constraints related to the usage of the English language on construction sites are somewhat universal. For example, many of these sites in developed countries rely on immigrant workers, making the issue of language barriers and the subsequent impact on data quality somewhat ubiquitous (Ne'Matullah et al., 2021; Trajkovski and Loosemore, 2006). Similar challenges exist for construction sites in the developing world, where workers and safety professionals predominantly may not be fluent in English. However, despite their lack of English proficiency, workers and safety supervisors in many developing countries may still use English to report near-misses, impacting the quality of the safety reports. As also shared by practitioners in our study, English has become a primary operative language for many developing countries, such as India and countries in Africa, because of the vast diversity in regional languages coupled with intra-regional labor migration (Emuze and James, 2013; Samanta and Gochhayat, 2021). Such a situation could also be commonly found in large-scale construction sites in developing countries, involving joint ventures with foreign construction companies, often bringing their safety systems, including near-miss reporting systems (Auffray and Fu, 2015). Although, it is possible that in a few developing countries, reporting responsibilities are executed by safety professionals with marginally better English language abilities. However, a global review of the literature supports an increasing focus on the idea that workers should be essential in reporting for a scalable near-miss reporting system instead of relying only on resource-intensive safety professionals (Marks et al., 2014; Zhou et al., 2019). Hence, the study's findings for such a dataset may have implications for the worldwide construction industry. However, the authors also acknowledge that such a large quantity of data may not represent the data obtained from medium to small-scale sites prevalent in developing countries. Nevertheless, the current study can generate recommendations for construction sites worldwide by analyzing the current data and the relationship between data size and prediction performance.

*TABLE 2: Example of characteristics of the input dataset*

| ***Example of Raw Data*** | ***Potential concerns related to Input data*** |
|---|---|
| Brown Field New Units Area, Unit-137, New Valve Pit<br>Valve Pit concrete. | For the research team, such sentences with little or no contextual information are difficult to classify as UA or UC. In the first example, no information about the valve pit concrete is available. In the second sentence, the research team cannot ascertain the presence of a trailer without a beacon light at the construction site is UC or the fact that a driver is moving such a vehicle is UA. Further, contextual information could also influence the classification. For example, any information on the factors influencing the trailer driver's decision to operate the machine could potentially hint toward the event being classified as UC. Wherever such cases were present, they were not relabelled to avoid inducing unnecessary errors in the classification. |
| Trialer driver moving the trialer without becon light. | |
| While debling after hydro test vaccum tanker shoulder be available for sucking oily water while deblinding called vaccum tanker. | |
| was observed worker whilst helping the welder during weld the truses support he is not wearing any face protection | Unstructured English, Grammar patterns, spelling mistakes, lot of context-specific abbreviations (AE, PTW, KNPC) |
| Observe AE ptw holder maintain and following knpc work permit system fond no devoation. | |
| As heat is raising there is a need of supply of ors to the workers | The observation was initially labeled as a good observation, but it seems to be mislabelled. |
| Improper House keeping inside the working area | |
| Protruded rebars used for barricadding the RCC trench under curing,this is unsafe and may injur the workers | Examples of mislabelled observations were initially labeled as UA but have been corrected to be UC. |
| was observed road closure barricade & signages was collapse due to strong wind. | |

## 3.2. Experimentation

The steps involved in a typical ML analysis for a given data are shown in Fig. 2(a). First, the input data is pre-processed, essentially to simplify the textual information so that the high performance of the ML algorithms can be achieved. Various ML algorithms are then applied to the pre-processed data. Finally, the results obtained from the ML algorithm are evaluated for their effectiveness using the F1 score. Naturally, the characteristics and quality of the input data (see Table 2) and the aspects of the ML approaches can affect the F1 score. Hence, various inputs and parameters in the overall process can be experimented with until a stable and desired performance on the F1 score is obtained. The details of these steps and the corresponding experiments implemented in the current study are described later. These experiments can be associated with the input data and the automated approaches for classifier performance enhancement.

### 3.2.1. Scenarios related to input data, and pre-processing steps

The data containing a description of the near-miss event is used as a basis for classification. The entire entry in each row was stripped to individual words (known as Tokens), and all these individual words were lowercased (Bird et al., 2009). Next, lemmatization, a linguistic process that returns the base or the dictionary form of a word, was employed for the tokens thus generated (Bird et al., 2009) to ensure uniformity in the text. These initial pre-processing steps are common to all scenarios; however, the differences in various scenarios conceptualized in the current study are discussed below.

#### 3.2.1.1 Base-Case scenario

The lemmatized tokens still had punctuations, numbers, and many other words with no significant meaning to the overall textual observation. Therefore, for the *Base-Case* scenario of the experiments, the words with little lexical significance were removed using a commonly used list of stop-words (Bird et al., 2009) that are excluded from being passed in pre-processed data. Even at this stage, many unknown words remained in the dataset. Levenshtein distance was utilized to find similarities between the unknown words and the known set of words and then replace them with the best alternative. In addition to the inbuilt English dictionary, a manual external dictionary consisting of construction-specific terms (Goh and Ubeynarayana, 2017) was created to correct a significant proportion of the unknown words.

As part of the iterative improvement of F1 scores for a given set of classifiers (Fig. 2(a)), errors, i.e., observations with wrong classifications, are identified. Patterns from these errors provide valuable inputs for the authors to identify the aspects where the ML process could be improved (Géron, 2019; Ng, 2019). For example, two authors independently performed error analysis for some of the early classification models used in the study (Kedia et al., 2021). Upon an in-depth examination of these erroneous observations, the authors conceptualized two more scenarios that could help improve the classifier performances.

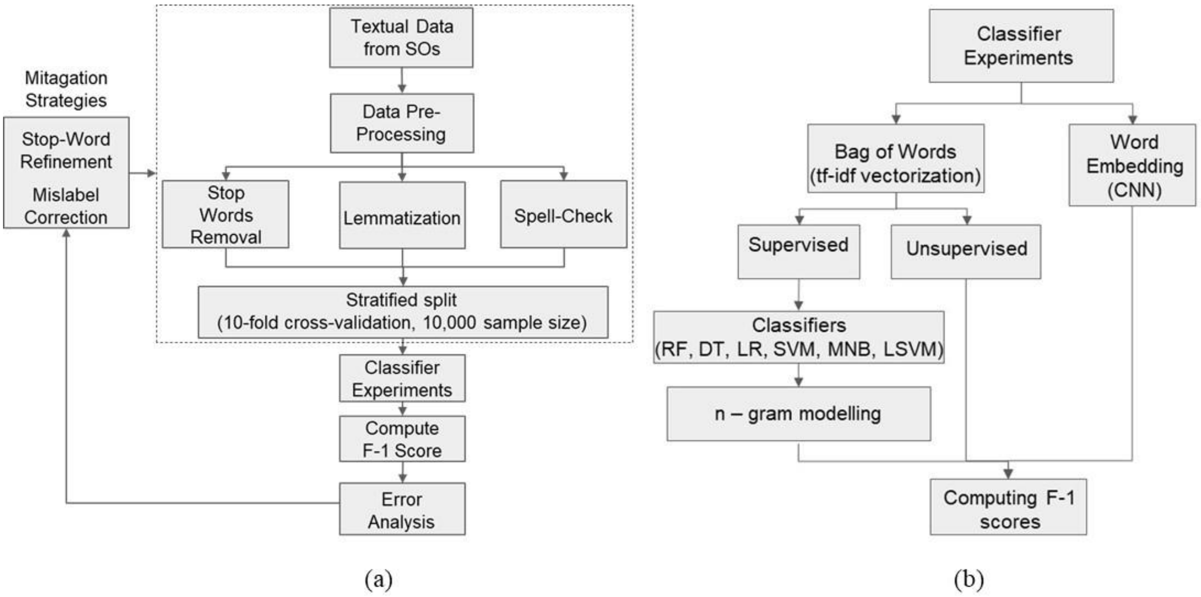


*FIG. 2: Analysis and experimentation strategy used in the current study*

### 3.2.1.2 Stop-word refinement scenario

The issues relating to excluding certain important words as part of the stop-word removal process were identified during the error analysis. Such exclusion of certain stop-words could affect the meaning of specific observations; for example, upon removing the stop-word "No," the issue of having no signboard changed to the presence of a signboard at the site. Such an observation could have been wrongly classified as the GO by the classifier instead of the original UC (Kedia et al., 2021). Through mutual discussions, three authors then identified instances of words removed during the pre-processing in the *Base-Case*, which could have potentially contributed to the erroneous classification by the algorithm. All such words were removed from the original list of stop-words, and the results from this scenario were termed *Stop-Word Refinement.* The supplementary material also shows a complete list of stop-words removed from the original list (Appendix A).

### 3.2.1.3 Mislabel correction scenario

Another prominent issue related to mislabeled classification in the original data was also identified during the error analysis. The authors corrected labels for all the GOs after examining the extent of mislabeling in the original data. To ascertain the reliability of the label corrections, first, an inter-rater reliability metric (Cohen's kappa) was estimated for the three authors independently rating the same 600 observations selected randomly (McHugh, 2012). A kappa score of 0.66 and a percentage agreement of 89.5 were obtained. Such a kappa score signifies moderate agreement and is deemed acceptable even for highly sensitive decisions such as patient safety (McHugh, 2012). Furthermore, the assumptions made in kappa calculations could lower the estimate excessively despite a high percentage agreement. Therefore, for a slightly less sensitive classification such as the one in the current study, such a kappa score was deemed acceptable, assuring the consistency of the classification across different authors (McHugh, 2012). Then, all the GOs were divided among the three authors for label corrections. The total number of mislabeling corrections from each category through this exercise has been summarized in Table 3. Overall, the process adopted was excessively time-consuming. Further, the authors also faced difficulty distinguishing UAs and UCs, as the description provided was insufficient to classify the observations (see Table 2 for examples). Because of these reasons, the authors did not correct the mislabels for the whole dataset, as it may introduce unwanted errors due to the authors' lack of familiarity with the ground reality of the site. The results obtained from this improvement for various classifiers are then termed *Mislabel Correction*.

*TABLE 3: Number of observations relabelled in each category*

| ***Original Labels*** | ***Number of observations after relabeling*** | | | ***Total in the original dataset*** |
|---|---|---|---|---|
| | ***UA*** | ***UC*** | ***GO*** | |
| ***UA*** | 2843 | 213 | 7 | 3063 |
| ***UC*** | 52 | 6339 | 11 | 6402 |
| ***GO*** | 107 | 240 | 2597 | 2944 |
| ***Total in the relabeled dataset*** | 3002 | 6792 | 2615 | 12409 |

### 3.2.2. Experiments with automated classification approaches

### 3.2.2.1 BoW-based classifiers and their variations

Consistent with the recommendations from the literature, experiments were conducted to identify the best-performing classifier for the data used in the current study. Commonly adopted BoW-based ML classifiers (as listed in Table 1 and described in the literature (Baek et al., 2021; Goh and Ubeynarayana, 2017) have been used for experimentation in the current study (as shown in Table 4).

*TABLE 4: Information on various classifiers used in the current study*

| ***Classifier Name*** | ***Hyperparameters Values*** |
|---|---|
| RF | Max_depth = 90; n_estimators = 100; min_samples_split = 0.01; min_samples_leaf= 2; max_features = auto; remaining parameters = default. |
| DT | min_samples_leaf = 2; min_samples_split = 0.01;max_depth = 100; max_features = auto; remaining parameters = default. |
| BNB | 'alpha': 1.0; 'binarize': 0.0; 'class_prior': None; 'fit_prior': True. |
| LR* | Solver = liblinear; c = 10; 'class_weight': None; 'dual': False; 'fit_intercept': True; 'intercept_scaling': 1;'l1_ratio': None; 'max_iter': 100; 'multi_class': 'auto'; 'n_jobs': None; 'penalty': 'l2'; 'random_state': None. |
| Lagrangian SVM (LSVM)* | C= 1 |
| MNB | alpha=0.2; class_prior=None; fit_prior=True. |
| CNN | Number of filters – 100; Number of convolution layers – 3 with filter sizes 3,4 and 5, respectively; Activation function – Rectified Linear of ReLU; Pooling strategy – max-pooling; Dropout Rate – 0.5; remaining parameters – default. |
| For LR and LSVM, the Hyperparameters are also verified using the grid-search optimization technique (Goh and Ubeynarayana, 2017), further enhancing the validity of the sensitivity tests. | |

Experiments were also conducted utilizing the unsupervised learning approaches based on BoW representation, taking cues from the literature (Chokor et al., 2016). However, in the current study, these approaches did not improve performances compared to the usual BoW implementation (Kedia et al., 2021); hence, more results are not discussed here. Additionally, the n-gram modeling was also used to assess whether the combination of two or more words taken together could improve classification performance (Goh and Ubeynarayana, 2017) (see Fig. 2(b)). Results from these experiments have been discussed later.

For any ML algorithm, the parameters known as hyperparameters control its learning process. The prediction performance for a given classifier can show significant sensitivity to the value of hyperparameters chosen. Therefore, in the current study, the value of these hyperparameters was selected by referring to the previous ML applications on textual data and confirmed through a sensitivity analysis. Information on hyperparameters for each of the classifiers is also summarized in Table 4. Detailed information on the functioning of various algorithms and their hyperparameters has been summarized in the supplementary material.

#### 3.2.2.2 Word embeddings-based classifiers

Since the unstructured data obtained in the current study contains only a few words in each document, the opportunities to learn from the text semantics are limited. Therefore, the current study relies on CNN-based classification using a word embeddings-based representation (Baker et al., 2020). The details of the CNN architecture are consistent with a well-cited study describing the CNN architecture (Zhang and Wallace, 2015).

### 3.2.3. Experiments with sample size

For each experiment described in section 3.2.1 and section 3.2.2, a stratified sample of 10,000 out of 12,500 observations was utilized to calculate performance. In stratified sampling, the proportion of observations belonging to each category, i.e., UA, UC, and GO, is kept the same as the original data. To obtain the relationship between the prediction performance and the sample size, experimentation was conducted by varying the sample size in increments of 1000, ranging between 1000 and 12,000.

## 3.3. Computing F1 score

Consistent with the literature, the current study uses a 10-fold cross-validation strategy for computing the average F1 score (Bouckaert and Frank, 2004). All results were calculated using Python programming language executed with Google Colaboratory in the web browser. Scores and a weighted total are also calculated for each category. The literature also guides the utilization of the F1 score computed for different parts of the same data to identify the reliability and generalizability of the classifier performance for the whole data (Géron, 2019; Ng, 2019). Training data refers to the part of data used for training a given classifier. Testing data corresponds to the data part used for testing the already trained classifier. For a given iteration of 10-fold cross-validation, 10% of the data is used for testing the classifier trained on the remaining 90% of the data. The F1 scores for the training data indicate the classifier's bias, i.e., the extent to which the classifier can model the true relationship between the inputs and the outputs (Géron, 2019; Ng, 2019). A high F1 score for training data is an indicator of the low bias of the classifier. Further, the difference between the F1 scores obtained for the test data (the commonly reported F1 score) and the training data is an indicator of the variance of the classifier, which defines the extent of dependence of the classifier on the training data. A high variance suggests significant changes in F1 scores with modifications to the training dataset, marking lower generalizability of the classifier for different data, potentially from another source (Géron, 2019; Ng, 2019). Suppose a classifier shows low bias but high variance; it is overfitting, showing low generalizability of the classifier training. In addition, for cross-validation strategies, the F1 score on the test data is not a true representative of the generalizability of the classifier. Since the classifier, in most cases, has seen a part of the test data before. Therefore, independent test data can also be prepared to contain data rows that have never been included in cross-validation iterations. For a given classifier, a high level of generalizability can be shown if the differences between the independent test data scores and test data scores are insignificant (Géron, 2019; Ng, 2019). Various interpretations related to F1 scores are then utilized later to comment on the generalizability of the findings.

## 3.4. Discussions with industry stakeholders

The results obtained from experimentation were then shared with the practitioners through open-discussion sessions to establish the relevance of the ML classification approaches in sustaining and scaling a near-miss reporting system at construction sites. In total, three different sessions were held with five experienced professionals in the construction industry. Each of the sessions lasted between 1.5 and 2.5 hours. The sessions were divided into two parts. In the first part, all the participants were asked questions about their experience in the construction industry, safety, and experiences implementing digital solutions for construction at their respective sites. Details of the questions are included in the supplementary material. All five professionals had construction-related experience ranging between 10 and 30 years for sites located across developed and developing parts of the world. In addition, one of the five professionals had about 30 years of experience working as a safety professional. Two others had worked in numerous project management and project consultant positions, where safety-related work was a part of their direct responsibility. The two remaining professionals had expertise in implementing various digital and data-driven applications at construction sites. Such diversity in professional profiles allowed the authors to develop a richer understanding of the relevance of the study's results in the construction industry across the world.

All study participants were then shown a presentation containing a summary of the results and detailed examples of the various challenges that authors had faced while attempting to increase the performance of various automated ML classifiers. The practitioners were asked to share their thoughts on the relevance of the data and the classifier performance obtained in the study for their applicability at construction sites in both developed and developing countries. Questions also explicitly focused on identifying the challenges for ML implementation on the sites, potential opportunities, and potential ideas for mitigating some of the challenges faced based on the results obtained in the current study. The supplementary material (Appendix A) details a complete set of questions and the presentation containing the preliminary results used for the discussion sessions.

The information thus obtained from these discussion sessions was then utilized to confirm the relevance of the ML classifier and the data toward near-miss reporting systems at construction sites. As the authors observed a good convergence in practitioners' responses across different sessions, an in-depth analysis of the information obtained, such as transcribing and codification, was not deemed necessary.

# 4. RESULTS

## 4.1. Category-wise classification performance

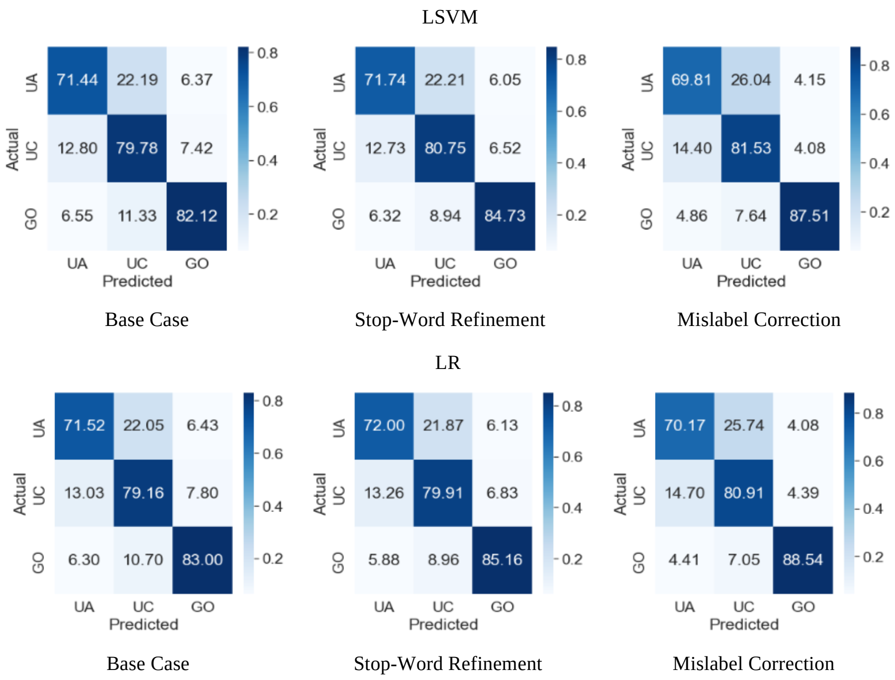
Table 5 shows the F1 scores for the *Base-Case* scenario for all classifiers for all data types (training, test, and independent test).

TABLE 5: F1 scores for the *Base-Case* scenario

| | *Training Data score* | | | | *Testing Data scores* | | | | *Independent Test Data score* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *UA* | *UC* | *GO* | *Total* | *UA* | *UC* | *GO* | *Total* | *UA* | *UC* | *GO* | *Total* |
| RF | 0.68 | 0.84 | 0.76 | 0.79 | 0.57 | **0.79** | 0.68 | 0.73 | 0.55 | **0.78** | 0.71 | 0.72 |
| DT | 0.75 | 0.86 | 0.80 | 0.82 | 0.55 | 0.74 | 0.66 | 0.68 | 0.55 | 0.74 | 0.68 | 0.68 |
| LR | 0.77 | 0.88 | 0.87 | 0.85 | 0.60 | **0.79** | 0.74 | **0.73** | 0.58 | **0.78** | 0.77 | **0.73** |
| LSVM | 0.77 | 0.88 | 0.87 | 0.84 | 0.60 | **0.79** | 0.74 | **0.73** | 0.58 | **0.78** | 0.77 | **0.73** |
| MNB | 0.69 | 0.84 | 0.77 | 0.79 | 0.59 | 0.78 | 0.66 | 0.71 | 0.56 | 0.76 | 0.69 | 0.70 |
| BNB | 0.67 | 0.81 | 0.77 | 0.77 | 0.61 | 0.77 | 0.69 | 0.71 | 0.59 | 0.75 | 0.72 | 0.70 |
| CNN | **0.89** | **0.94** | **0.94** | **0.93** | **0.64** | **0.79** | **0.77** | **0.74** | **0.62** | 0.77 | **0.79** | **0.74** |

The numbers highlighted in **Bold** represent the maximum F1 score across different classifiers for a given category within scores for a given data type

Results demonstrate that the highest performance (F1 score) on test data was achieved for category UC (range 0.74 – 0.79), followed by GO (range 0.66 – 0.77) and UA (range 0.55 – 0.64). Further, distinguishing between UA and UC was challenging. Such difficulty in distinguishing between UAs and UCs was consistent across different classifiers, as shown in the confusion matrices for LR and LSVM classifiers on testing data for all scenarios (see Fig. 3).



*The entries in each cell for a given confusion matrix represent a percentage, such that row totals are 100%.*
Example reading : For Base Case, LSVM – 22.19% of the observations that were originally labeled as UA are predicted as UC.

*FIG. 3: Confusion Matrices for LSVM and LR (Testing Data, All scenarios)*

## 4.2. Results from scenario analysis

Fig. 4 and Table 6 present the total F1 score (weighted average for UAs, UCs, and GOs) for the three scenarios related to input changes and pre-processing steps, as discussed in section 3.2.1. Comparing the *Stop-Word Refinement* scenario with the *Base-Case* across datasets and classifiers reveals only a modest improvement (1 – 2 percentage points) in the total F1 score. Although, the gain is significant for the category GO (3 – 12 percentage points) (compare Table 6 and Table 5). Like the previous result, comparing the *Mislabel Correction* scenario with the other cases reveals only a modest improvement in the total F1 score (1 – 4 percentage points). However, in all the cases, there is a significant improvement in GO performance.

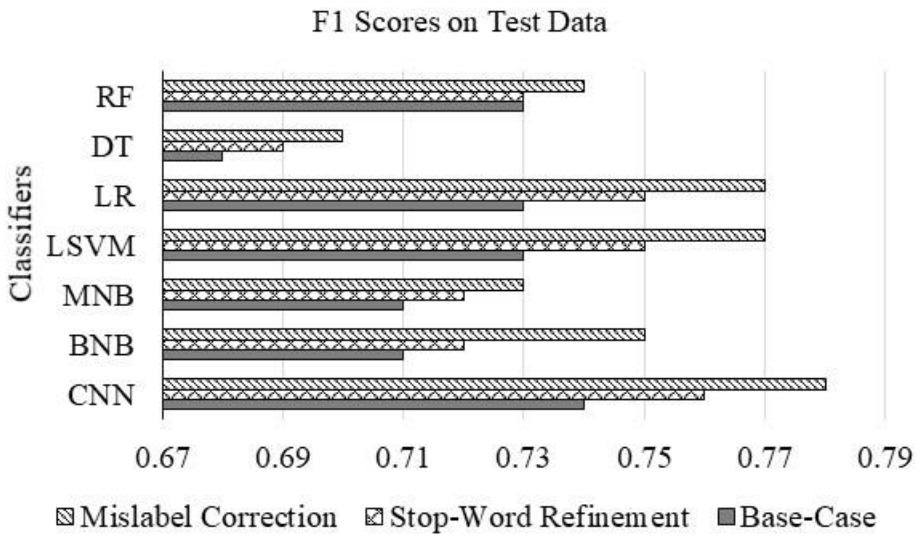


*FIG. 4: Results from the Scenario Analysis*

*TABLE 6: F1 scores across scenarios for classifier experimentation*

| | *Stop-Word Refinement* | | | | *Mislabel Correction** | | | |
|---|---|---|---|---|---|---|---|---|
| | *UA* | *UC* | *GO* | *Total* | *UA* | *UC* | *GO* | *Total* |
| *Training Data scores* | | | | | | | | |
| RF | 0.69 | 0.79 | 0.84 | 0.80 | 0.65 | 0.85 | 0.82 | 0.80 |
| DT | 0.75 | 0.81 | 0.87 | 0.82 | 0.73 | 0.87 | 0.83 | 0.83 |
| LR | 0.77 | 0.88 | 0.89 | 0.86 | 0.76 | 0.90 | 0.93 | 0.87 |
| LSVM | 0.77 | 0.88 | 0.89 | 0.86 | 0.76 | 0.90 | 0.93 | 0.87 |
| MNB | 0.69 | 0.78 | 0.84 | 0.79 | 0.69 | 0.85 | 0.81 | 0.80 |
| BNB | 0.67 | 0.78 | 0.82 | 0.77 | 0.67 | 0.83 | 0.83 | 0.79 |
| CNN | **0.90** | **0.95** | **0.95** | **0.94** | **0.88** | **0.94** | **0.98** | **0.93** |
| *Testing Data scores* | | | | | | | | |
| RF | 0.58 | 0.71 | **0.80** | 0.73 | 0.52 | 0.81 | 0.74 | 0.74 |
| DT | 0.55 | 0.69 | 0.75 | 0.69 | 0.52 | 0.76 | 0.73 | 0.70 |
| LR | 0.61 | **0.80** | 0.77 | **0.75** | 0.59 | **0.82** | **0.84** | **0.77** |
| LSVM | 0.61 | **0.80** | 0.77 | **0.75** | 0.59 | **0.82** | **0.84** | **0.77** |
| MNB | 0.60 | 0.68 | 0.79 | 0.72 | 0.58 | 0.80 | 0.70 | 0.73 |
| BNB | 0.61 | 0.71 | 0.78 | 0.72 | 0.61 | 0.80 | 0.77 | 0.75 |
| CNN | **0.64** | **0.80** | **0.80** | **0.76** | **0.63** | 0.81 | **0.85** | **0.78** |
| *Independent Test Data scores* | | | | | | | | |
| RF | 0.55 | 0.73 | 0.79 | 0.73 | 0.50 | 0.80 | 0.75 | 0.73 |
| DT | 0.55 | 0.71 | 0.75 | 0.69 | 0.52 | 0.76 | 0.74 | 0.70 |
| LR | 0.58 | **0.79** | 0.79 | **0.74** | 0.56 | **0.81** | **0.84** | **0.76** |
| LSVM | 0.58 | **0.79** | 0.80 | **0.74** | 0.56 | **0.81** | **0.84** | **0.76** |
| MNB | 0.57 | 0.70 | 0.77 | 0.71 | 0.55 | 0.78 | 0.72 | 0.72 |
| BNB | 0.60 | 0.73 | 0.76 | 0.71 | 0.59 | 0.78 | 0.77 | 0.73 |
| CNN | **0.63** | 0.78 | **0.81** | **0.75** | **0.60** | 0.79 | **0.85** | **0.76** |
| The numbers highlighted in **Bold** represent the maximum F1 score across different classifiers for a given category within scores for a given data type<br>**Mislabel Correction* scenario constitutes a combined effect of mislabeling correction activity as well as *Stop-Word Refinement* activities implemented by authors | | | | | | | | |

## 4.3. Results from experiments with classifier selection

### 4.3.1. BoW-based classifiers

Table 5 and Table 6 show that among different BoW-based classifiers, LSVM and LR performed consistently better across different scenarios and datasets. F1 scores for LSVM and LR for test data across scenarios are in the range (0.73 – 0.77), at least four percentage points higher than other BoW classifiers.

The results from n-gram modeling are shown in Table 7. The results indicate that an association of multiple high-order sequences of tokens taken together (such as (1,3) and (1,4)) led to better classification performance when compared to single n-gram sequences (such as (1,1) and (2,2)).

*TABLE 7: Results of n-gram modeling for LR*

| Total F1 Score<br>(*Mislabel Correction*, testing data,<br>10-fold cross-validation) | | *The upper boundary of n-gram: "**b**"* | | | |
|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** |
| *The lower boundary of n-gram: "**a**"* | **1** | 0.77 | 0.786 | 0.789 | 0.792 |
| | **2** | - | 0.74 | 0.74 | 0.75 |
| | **3** | - | - | 0.64 | 0.64 |
| | **4** | - | - | - | 0.54 |
| The n-gram range is denoted as (a,b) where "a" and "b" refers to the lower and upper boundary of the range respectively of n-values that need to be extracted for different n-grams. For example, (1,3) consists of all unigrams, bigrams, and trigrams. | | | | | |

The n-gram (1,4) resulted in the highest F1 score of 0.79. Results also show that removing uni-grams from the analysis led to a significant decrease in classifier performance. Further, results identifying the top 10 most significant n-grams (obtained using a saliency-based approach (Baker et al., 2020)) are summarized in Table 8 (more in Appendix B in the supplementary material). In the later sections, such salient n-grams have been evaluated for their intuitiveness and human acceptability.

*TABLE 8: Top-10 n-grams for LR (1,4) and (2,4)*

| **LR Top 10 Uni/Bi/Tri/Quadgrams (1,4)** | | | | | |
|---|---|---|---|---|---|
| **UA** | | **UC** | | **GO** | |
| **Weight** | **Feature** | **Weight** | **Feature** | **Weight** | **Feature** |
| 7.711 | not wearing | 10.456 | not | 15.543 | good |
| 7.411 | without | 7.078 | no | 11.695 | proper |
| 6.15 | not | 6.118 | not properly | 10.771 | properly |
| 5.881 | shortcut | 5.944 | without proper | 9.361 | safe |
| 5.339 | walking | 5.463 | no proper | 8.694 | well |
| 4.813 | poor ppe | 5.33 | need | 6.618 | tie |
| 4.72 | worker | 5.202 | many | 6.51 | ppe |
| 4.386 | horseplay | 5.065 | ppe work | 6.488 | provided |
| 4.22 | observed during excavation | 4.912 | scattered | 6.413 | complete |
| 4.18 | faceshield | 4.634 | poor | 5.997 | slowly |
| **LR Top 10 Uni/Bi/Tri/Quadgrams (2,4)** | | | | | |
| **UA** | | **UC** | | **GO** | |
| **Weight** | **Feature** | **Weight** | **Feature** | **Weight** | **Feature** |
| 10.574 | not wearing | 6.029 | no proper | 7.037 | proper ppe |
| 8.275 | hand glove | 5.63 | not properly | 6.996 | complete ppe |
| 6.345 | not using | 5.409 | not provided | 6.543 | good barricade |
| 5.714 | without helmet | 4.828 | no barricade | 6.346 | good housekeeping |
| 5.002 | face shield | 4.113 | during erection | 5.967 | area barricaded |
| 4.929 | chin strap | 4.083 | without proper | 5.5 | good observation |
| 4.926 | not used | 4.051 | no barrication | 5.431 | good barrication |
| 4.864 | worker horseplay | 4.002 | wooden plank | 5.335 | using proper |
| 4.732 | not wear | 3.954 | brown field | 5.319 | nh worker have |
| 4.691 | nbc worker | 3.935 | excavation without | 5.125 | proper segregation |

### 4.3.2. Word embeddings-based classifier

Results from Table 5 and Table 6 also highlight that in most cases, the results from CNN were at par with LSVM and LR, if not surpassed. On the other hand, the higher performance for the training test scores for CNN suggests overfitting compared to LR and LSVM.

## 4.4. Results from experiments with sample size

A summary of the results for experimentation with the sample size for the two best classifiers, i.e., LSVM and LR, is shown in Fig. 5. The average F1 score for the training dataset is shown in Fig. 5(a). The distribution for F1 scores for the testing dataset obtained for each of the ten iterations of 10-fold cross-validation is shown in Fig. 5(b). As the sample size increases from 1000 to 6000 observations, there is a significant improvement in the performance of the two classifiers. The median-test F1 score for LSVM increased by 12 percentage points, whereas for LR, it increased by eight percentage points. There is a steep decline in training set F1 scores in the same range. Such trends suggest overfitting the two classifiers for the low sample size, which appears to be balanced as the sample size reaches 6000 observations. Such trends then indicate that the increase in sample size plays a significant role in assuring the robustness and generalizability of the two classifiers for a sample range between 1000 and 6000 observations. However, beyond 6000 observations, an increase in sample size (up to 12000 observations) results in only a marginal improvement for the F1 scores on test data. At the same time, there is a steady decline in training data performance. Such trends suggest that there appears to be an underfitting for both the classifiers beyond 6000 observations.

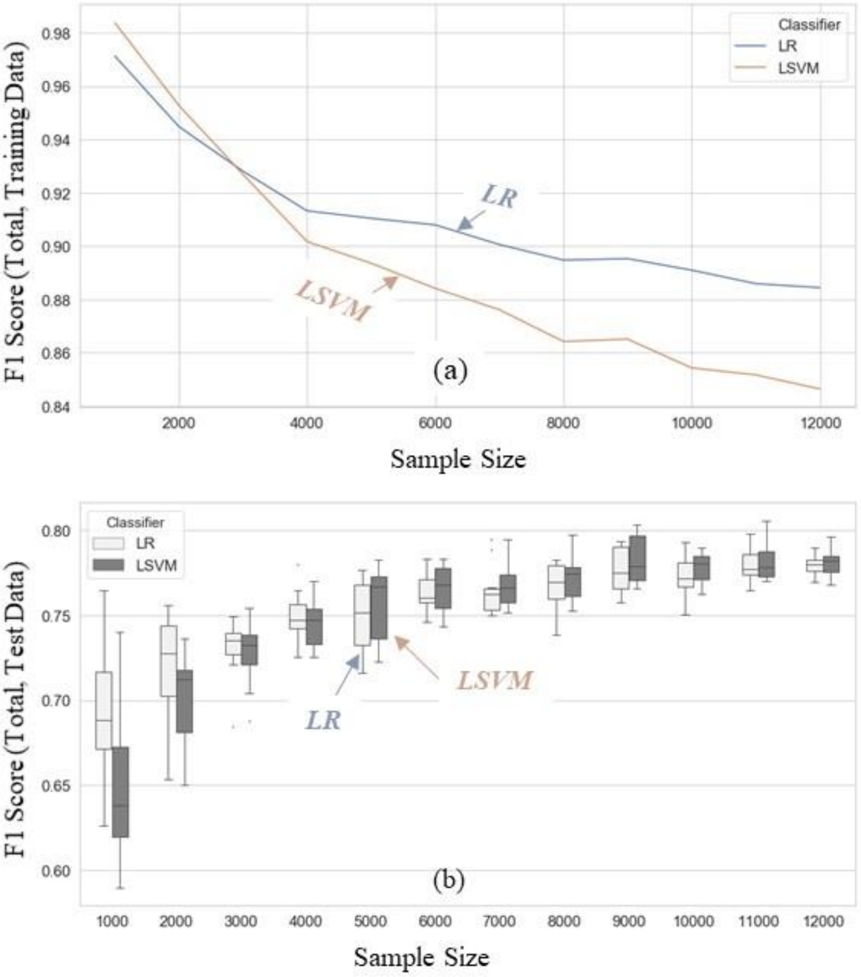


*FIG. 5: Variation in F1 scores with data size for LR and LSVM*

# 5. DISCUSSIONS

The discussions summarize the study's main findings from experimentation results and novel findings compared to the existing literature. The discussions are also extended by utilizing the data obtained through in-depth discussion sessions with practitioners to reveal the critical implications for designing near-miss reporting systems for construction sites.

## 5.1. Summary of the results and main contributions

### 5.1.1. Specific technical aspects and capability of automated ML approaches

The results obtained in this study provide a novel and important validation for the ML tools' real-world applicability and demonstrate the ML approaches' capabilities for usually poor-quality near-miss data obtained directly from construction sites. The study also presents novel insights for classifications related to labels rarely included in the previous literature, i.e., UA, UC, and GO. For the SO data reported by the workers at the construction site, the averaged total F1 score across different classifiers and scenarios ranged between 0.68 and 0.79. Such a performance range is consistent with the previous literature utilizing automated classification techniques for the textual data relevant to the construction sector using BoW-based approaches (Goh and Ubeynarayana, 2017) and more recent word embeddings-based approaches (Baker et al., 2020; Zhang, 2022). However, applications of recently developed ontology-based text-mining approaches such as BERT have shown better classification performance (F1 score of 0.87) (see Table 1) (Fang et al., 2020). Hence, a comprehensive evaluation of the BERT approach on classification performance and organizational decision-making aspects for worker-reported construction safety data could be an exciting research area for future exploration.

On the other hand, the results from n-gram modeling also show interesting effects of data association on classifier performance. In the current study, the inclusion of higher-order n-grams in addition to uni-grams positively affected classifier performance. Such a finding is consistent with previous literature (Goh and Ubeynarayana, 2017), highlighting the importance of the relative arrangement of tokens in a document, particularly for near-miss data.

Furthermore, the study confirms that the quality-related issues inherent in the data representative of real-construction sites can affect classification performance. In the current study, to avoid overinflating the F1 score in actual conditions, the authors relied on pre-processing steps that could be easily implemented automatically by programming languages. However, even after pre-processing, word-spelling and grammatical errors could also be commonplace in the data. For example, variation in the word "barrication" observed in the post-processed data is shown in the supplementary material (see Appendix C). With such variations in tokens, the overall dimensionality of the tf-idf vectorizer cannot be reduced significantly, potentially contributing to the inability of the classifiers of even higher performance (Baker et al., 2020). Such results also highlight the continued need for manual intervention in ML applications in near-miss data from real construction sites. For example, manual steps are necessary to eliminate the potential variations of a single word, which cannot be captured through standardized, automated processes.

On the other hand, a considerable variation in classification performance across different categories was also observed. The most proportionate category (about 50%) UC consistently and expectedly showed high performance (F1 scores ranging between 0.71 and 0.83) across different scenarios. The prediction for GOs was also increased to a great extent (F1 score reaching as high as 0.85) across different scenarios. Such results are expected as several generic stop-words, such as "no," "did not," signified the presence or absence of a particularly safe or unsafe system state. Therefore, their consideration in the analysis could have improved the classification performance of the category GO. The findings suggest that a conventionally popular standardized list of stop-words can no longer satisfy the requirement for suitably classifying the near-miss report data. A more context-specific list of stop-words should be developed for processing near-miss data.

Further, all classifiers showed consistently low performance for UAs (F1 score ranging between 0.55 and 0.64), despite their relatively high proportion in the data. Such a trend is consistent with the authors' observations regarding challenges faced during the mislabel correction exercise related to distinguishing between UAs and UCs (see examples in Table 2). This interaction between UAs and UCs is further examined through qualitative analysis of the errors (mismatch in prediction compared to the label in the original dataset), taking one of the testing datasets as an example. The results are summarized in Table 9. One of the experienced authors examined all the errors of UAs and UCs that were mispredicted by the optimal ML algorithm as UCs and UAs, respectively. These errors were marked based on the potential cause of misprediction. Errors were marked as Mislabeled in cases where the inaccuracy in category-label in the original data could have contributed to misprediction. All other errors were marked as Misclassified (see Table 9). Overall, the error-analysis results obtained from the example dataset could be deemed generalizable for a dataset size of 10,000, as the F1 score does not show significant variation in cross-validations (see Fig. 5(b)).

Of the errors marked as Misclassified, 60 belonged to the category UA and 15 to UC (see Table 9). A closer examination of the high-saliency tokens for errors occurring in each of the two categories (UAs and UCs) (Baker et al., 2020) reveals that some of the high-saliency words constituting UAs were also persistent for UCs (see examples in Table 9). Such commonality could create issues in achieving high performance for UAs, as just the presence of these tokens in the document can skew their prediction towards being UC, a category constituting 50% of the total data. The analysis also reveals that high-saliency words also affect the UC category. However, the issue is more prominent for the category UA (see Table 9). Such results highlight the importance of efforts that must be made to include keywords that could enable unique identification for each category as much as possible at the data origin.

On the other hand, 49 and 50 cases of *Mislabeled* errors were identified for the category UA and UC, respectively (see Table 9). Overall, *Mislabeled* cases constitute a significant proportion (99) of total errors (174). Theoretically, the extensive label correction exercise could help improve the classification performance. However, the results from the input and pre-processing related scenarios also helped reveal the challenges in improving classification performance despite the analysts' resource-intensive interventions for mislabel correction. In this study, even the human analysts faced difficulties differentiating between UAs and UCs (see examples in Table 2), potentially due to missing details of the contextual information required to differentiate between them. Under such circumstances, there is always a possibility that analysts are inducing more errors in the data while removing some. Such errors, coupled with the issues related to the commonality of high-saliency tokens across different categories, make it highly challenging to improve the classification performance on realistic datasets if good controls for data quality are not assured at its origin.

*TABLE 9: Example SOs not classified correctly for LR (1,4) and potential causes of misclassification*

| | ***Type of Error*** | | | |
|---|---|---|---|---|
| | ***Mislabeled*** | | ***Misclassified*** | |
| | ***No.*** | ***Example and Remarks*** | ***No*** | ***Example and Remarks*** |
| For SOs initially labeled as UAs but predicted as UC | 49 | "Excavation trench need hard barrication it's un safe condition"<br>*(Lack of hard barricading is indicative of a UC at the site. However, the same was labeled as UA in the original dataset)*<br>"Found wooden materials with protruding nails are scattered on the area causing tripping hazard."<br>*(Material disarrangement potentially causing a hazard is an example of UC that was labeled as UA in the original dataset)* | 60 | "Using **plank** for the **access**"<br>*("Plank" and "access" are words that have a higher weightage for the UC category )*<br>"worker working dark area **poor** visibility"<br>"scaffolder **crew** working **without light** mpr rd floor u"<br>*(The specific keywords such as "poor," "no," "crew," and "without light" weigh more towards UCs and not towards UAs)* |
| For SOs initially labeled as UCs but predicted as UA | 50 | "Civil worker doing manual excavation without safety glass."<br>"Sitting at unsafe location"<br>*(The above two examples tend to indicate that these could have been potentially mislabelled as being UC instead of UA)* | 15 | "Unwanted **material** on **access area**"<br>*(The term "material access area," as it would appear in the processed original observation, has a high positive weightage for the UA category )* |
| Text in quotes "is the SO as obtained in the original data. Text written in *Italics* are remarks related to the SO. The words highlighted in **Bold** represent some of the high-saliency tokens indicated in the list of top 100 n-grams in appendix B of the supplementary material. | | | | |

### 5.1.2. Comparison between conventional and deep-learning ML approaches

The current study also demonstrates the continued relevance of the conventional LSVM and LR even for the near-miss data compared to more sophisticated deep-learning techniques. The relatively high performance of CNN for training data indicates that it could be prone to overfitting. Such findings are meaningful in a context where the potential for deep-learning techniques in construction-specific task classification remains debated (Baek et al., 2021; Baker et al., 2020).

### 5.1.3. Relationship between data size and classifier performance

As opposed to a general expectation of improvement in classifier performance with dataset size (see section 2.2.4), the current study reveals a decreasing marginal improvement in classifier performance with respect to an increase in dataset size. Such results are novel as none of the previous studies had attempted to observe the results from experimentation with sample size. Implications of such a relationship for performance improvement and relevance toward near-miss reporting systems at various construction sites have been discussed later.

### 5.1.4. Potentials for improving classifier performance

Experience and results from various experiments also help identify essential strategies that could be explored to improve the F1 scores for multiple classifiers further. Based on the experimentation with the dataset size, the current study has obtained classifiers with increasing bias and constant variance as the size increases. The literature suggests that for classifiers with increasing bias and constant variance, the performance cannot be increased further by adding more data for the training set (Géron, 2019; Ng, 2019). Hence, contrary to the general recommendation in the literature (see Table 1), further efforts are needed to improve the feature learning abilities of the classifier from the existing training data instead of adding more data points. For example, the context-specific information could also be coded by various rules, such as distinguishing between UAs and UCs, for future studies focusing on classification performance. Natural language process-based rules could further help improve classification performance. However, such manual approaches can become intensive on the analyst’s part and reduce the ML approach’s generalizability (Goh and Ubeynarayana, 2017). Removing the text fragments contributing most towards classification in the first run and retraining the model could also be a potential approach for developing ML classifiers that learn additional features and potentially provide higher performance (Baker et al., 2020). On the other hand, dimensionality reduction could be an essential direction for such an extensive dataset to improve the performance of BoW-based conventional classifiers (Baker et al., 2020). In this regard, more comprehensive automatic approaches such as the Principal Component Analysis could be explored in future studies (Wold et al., 1987).

### 5.1.5. Generalizability of the findings

Overall, the good synergy between the F1 score for testing data and independent testing data for all classifiers (see Table 5 and Table 6) at a 10,000-sample size indicates that the results obtained are generalizable for all parts of the data from the same source for the relatively larger dataset. However, such generalizability is also compromised for the smaller datasets (see Fig. 5).

On the other hand, the generalizability of the study results beyond the data source used in the study is also difficult to conclude. On the one hand, the current study's data is deemed to represent the large quantities of SO data at construction sites across the globe (see section 3.1). Hence, results obtained from the current study potentially apply to a broader set of construction sites. On the other hand, the dominance of context-specific uni-grams towards the optimal classification performance (see Table 7) may also limit the generalizability of the classifier to other data sources where such specific tokens are not used. The classification performance of a classifier trained for data from one source and tested for data from another source containing similar categories remains to be analyzed through formal approaches. To address the question of classifier generalizability, a summary of a non-standardized attempt by the authors has been provided in the supplementary material (see Appendix D). Based on the best ML classifier, i.e., LR (1,4) and LR (2,4), a web-based tool that can give a probability score of a textual description belonging to each of the three categories was developed. The scores are estimated such that sum of probabilities for each of the three categories adds to 1. The authors then extracted a set of 30 different sentences from the OSHA injury database reported in (Goh and Ubeynarayana, 2017), stripped them for outcome-related information, and manually labeled them as being UA, UC, and GO. As previously described, the OSHA dataset is distinct compared to the dataset used in the current study. Additionally, four textual descriptions were generated by the authors. The classifier LR (1,4) could correctly classify 47% of the 34 observations in vastly different data. Interestingly, LR (2,4) provided better results in correctly classifying about 56% of the observations, hinting at potentially high generalizability of classifiers when the effect of the relative positioning of tokens is considered. Such reasonable results on a completely different dataset are also promising for assuring the generalizability of the study results to a wide variety of data. Readers of the study can also explore the web-based tool to get a sense of the generalizability of the findings at https://mlcsafety.herokuapp.com/.

## 5.2. Organizational decision-making aspects related to ML implementation

The analytical information readily available as part of the classification task for various ML classifiers can still be leveraged to assess ML approaches' relevance in the construction sector. The results comparing the conventional and deep learning-based ML approaches suggest that adopting a computationally sophisticated (and likely resource-intensive) technique may not always result in optimal classification performance. Such results are positive, as simplified approaches could still be practical.

On the other hand, a sneak peek into the most prominent features contributing to classifications (see Table 8) reveals that certain high-frequency uni-grams, with low significance towards safety, tend to dominate the classifier performance. For example, the presence of tokens such as "slowly" or "tie" in a document can lead to a high chance that such an observation will be classified as GO by the ML classifier. However, these words, in isolation, cannot represent the presence of GO to human analysts, making it difficult to assure the human validity and acceptance of the results. On the other hand, n-gram (2,4) provides good classifying performance, even though not optimal. However, its tokens, such as "proper ppe," "complete ppe," and "good barricade," being classified as GO, are also intuitive from a human analyst's perspective. Similar patterns for the impact of token association on classification performance are also seen for LSVM. The presence of uni-grams in the classifier led to specific tokens having a dominant effect, potentially affecting the generalizability of the study's results to other similar data sources that may exist across different construction sites. Such results raise concerns about the validity of the optimal classifiers to human analysts and organizational decision-making (Wang et al., 2020).

Experiments on sample size highlight an essential aspect for ML applications, specifically in construction sites with varying maturity of reporting systems. The prediction performance across different classifiers drops considerably as the data size reaches about 3000 observations. Such data size then corresponds to about 1000 observations per month for a given construction site, which was deemed considerably high by practitioners for many small and medium-size projects or with projects with low maturity of the reporting systems.

Overall, these results provide necessary guidance for practitioners to not instinctively follow the recommendations from the literature and identify the best possible set of algorithms for the given data, size, and organizational decision-making factors. A simplified experimental strategy like the one developed in this study could be helpful for practitioners in identifying the most suitable ML algorithms specific to their application. Such results also hint toward only a moderately positive outlook for ML implementation in the field at construction sites. Although, these implications are further complemented by the information obtained through engagement with industry stakeholders.

### 5.2.1. Implications for scaling and sustaining near-miss reporting system using ML for construction sites

The interviews with practitioners in our study confirm that text mining approaches for the near-miss reports provide immense opportunities for post-processed empirical studies and project statistics at construction sites. The trends across various categories and the prominent features observed within data can guide organizational decision-making to improve safety. In this regard, the F1 scores of around 0.78 with less-resource intensive ML classifiers were deemed more than sufficient by all the practitioners to start implementing such tools at the construction site for the quick and effective analysis of SOs at sites with mature reporting systems, although with caution. The practitioner's understanding also reaffirms the relevance of human interpretability of the results from ML classifiers and, therefore, the need for a relatively comprehensive assessment of ML approaches rather than an acute focus on classification performance.

For example, one of the experts shared, "*The outcome of the NLP program is going to be made available to the management to decide further action, right! Because there is so much of the data, so much information, processing of the data manually is ruled out, and that's where you want to use NLP….But you cannot rely entirely on NLP either….I think you have challenges. The results from NLP should be polished and manually re-checked. Because the results are so much affected by some wrong word being featured by spell-check or the wrong word being deleted [by stop-words], the entire meaning change. So you cannot go blindly by the conclusion of the program….But by and large, I think the accuracy[F1 score] of 70-80% that you have achieved is a very good accuracy[F1 score].*"

The discussions with practitioners also helped identify numerous potential applications of ML approaches beyond the immediate project statistics-related analysis. One applicability relates to developing a safety training tool for the workers. The ML tool could provide the approximate classification to the worker based on the sample inputs provided by the worker. The practitioner noted that in large construction sites, where the specialized workforce keeps changing depending upon the ongoing activity, such a tool could help accelerate the training process for the workers in report classification. An in-depth examination of the potential effect of such a tool on workers' ability to recognize UA, UC, and GO should also be carefully examined for future work.

However, the results from the study also highlight significant barriers to utilizing and implementing the ML approach for small and medium-sized construction firms with relatively immature reporting systems. The classification performance levels for the small dataset size in these firms can be expected to remain low. Only when a critical maturity of the near-miss reporting system for an organization is achieved and many observations are available can ML tools readily be used to accelerate the reporting program further. However, small construction firms without such a mature reporting system will likely be locked in for the existing resource-intensive manual classification practices. More sophisticated ML applications for improving performance for smaller datasets may require more resources and produce results specific to the given dataset. Such efforts may not be scalable across reporting systems at different sites, even for a single organization. Due to such effects, the application of the beneficial ML approaches is unlikely to see rapid adoption, especially in small or medium-scaled projects. In such a scenario, policy intervention is likely to be the one way to promote the accelerated implementation of ML-based near-miss reporting systems worldwide. Government facilitation to create a country-wide database for storing near-misses from several construction sites, such as the one developed in China (Fang et al., 2020), could be one solution. Such a database could allow all construction sites facing similar contextual factors to leverage the ML models trained on the large quantities of data points in the shared database.

## 5.3. Limitations and future work

The current study has utilized a comprehensive experiment strategy to conduct a comparative analysis to reveal the most suitable ML approaches for classifying textual near-miss reports. However, the authors acknowledge that such comparisons could rely on more rigorous quantitative methods in the study. For example, no comparisons for different F1 scores have been made using rigorous statistical tests suitable for 10-fold cross-validations schemes (Bouckaert and Frank, 2004). Similarly, hyperparameters are also identified using sensitivity analysis and not more strict approaches such as bayesian optimizations (Goh and Ubeynarayana, 2017). Hence, rigorous criteria could also be accommodated in the overall experimentation methodology adopted in this study for future studies. However, the authors do not expect significant differences in the study's main findings relying on qualitative criteria, as several of the results obtained from the study are already generally consistent with the previous literature. Future work could also focus on a comprehensive evaluation of the newly emerging classification approaches, e.g., BERT, for worker-reported construction safety data. Future studies should also systematically examine the lack of adoption of ML tools in construction sites.

# 6. CONCLUSIONS

The current study presents the first of its kind validation for the applicability of ML tools for the classification of near-miss observations obtained directly from construction sites, representative of actual site conditions. Despite the several data-quality related challenges prevalent for such data, relatively high performance for several ML classifiers could be obtained. LR, LSVM, and CNN approach achieved an average F1 score of 0.79 for the whole data (10,000 observations). For categories such as UC and GO, the average F1 score as high as 0.85 could be obtained. Such results are at par with previous studies, which relied on data obtained from standardized sources and have relatively lower quality-related issues. The study also confirmed the generalizability of the classification performance over the entire dataset.

Based on the experimentation presented here, the study concludes that conventional BoW classifiers such as LR and LSVM remain relevant compared to computationally intensive CNN approaches. The study also clarified the relationship between the overall sample size and the F1 score, revealing a decreasing marginal improvement in the F1 score as the data size improves. Therefore, for future studies, further improvement in the performance can only be achieved through enhanced feature learning capability, potentially requiring tremendous manual inputs for developing context-specific rules.

Through a sneak peek into prominent features contributing to ML classifiers, the study also confirms that the most optimal classifiers may not always be acceptable to human and organizational decision-makers. These results also guide practitioners not to blindly follow the recommendations from the literature and identify the best possible set of algorithms for the given data, size, and organizational decision-making factors. A simplified experimentation strategy like the current study could also identify data-specific conclusions in future works.

Finally, engagements with industry stakeholders in this study also highlight the potential training tools that could be created using such classification algorithms. On the other hand, the study also provides insight into the potentially low adoption of such tools at sites where the availability of large-scale databases, such as the one used here, could be challenging. Construction industry practitioners may find themselves locked in for not leveraging tools like ML to scale their reporting systems' analysis capabilities rapidly. Therefore, the current study recommends a government facilitation program for data sharing across different companies to enable each partner organization to quickly achieve a mature reporting system for improved safety.

# ACKNOWLEDGEMENTS

The authors would like to acknowledge JGC corporation for sharing the data. We also acknowledge support from Mr. Marcos Morales of JGC corporation, Mr. Suhas Bhagwat of Avroglide Consultants, Mr. Ajithkumar Vadakkoot, Mr. Arun Sivasankara Pillai, and Mr. Amit Saraswat of Bexel India Consulting for in-depth candid discussions. The authors would also like to acknowledge Ms. Vishakha Shankar's contributions for her help in independently proofreading the article. The authors are also grateful to the two anonymous reviewers whose comments helped shape the article in its final form.

## REFERENCES

Auffray C. and Fu X. (2015). Chinese MNEs and managerial knowledge transfer in Africa: the case of the construction sector in Ghana. *Journal of Chinese Economic and Business Studies.* Vol. *13,* No. 4, 285–310. https://doi.org/10.1080/14765284.2015.1092415

Baek S., Jung W. and Han S.H. (2021). A critical review of text based research in construction: Data source, analysis method, and implications. *Automation in Construction*. Vol. 132, 103915. https://doi.org/10.1016/j.autcon.2021.103915

Baker H., Hallowell M.R. and Tixier A.J.-P. (2020). Automatically learning construction injury precursors from text. *Automation in Construction.* Vol. 118, 103145. https://doi.org/10.1016/j.autcon.2020.103145

Bird S., Klein E. and Loper E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. *O'Reilly Media, Inc.* Sebastopol, CA, USA.

Bouckaert R.R. and Frank E. (2004). Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In: Dai H., Srikant R. and Zhang, C. (eds) *Advances in Knowledge Discovery and Data Mining*. PAKDD 2004. *Lecture Notes in Computer Science*, Vol. 3056. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24775-3_3

Bugalia N., Maemura Y. and Ozawa K. (2021). A system dynamics model for near-miss reporting in complex systems. *Safety Science*. Vol. 142, 105368. https://doi.org/10.1016/j.ssci.2021.105368

Chokor A., Naganathan H., Chong W.K. and el Asmar M. (2016). Analyzing Arizona OSHA injury reports using unsupervised machine learning. *Procedia Engineering,* Vol. 145, 1588–1593. https://doi.org/10.1016/j.proeng.2016.04.200

Demirkesen S. and Tezel A. (2022). Investigating major challenges for industry 4.0 adoption among construction companies. *Engineering, Construction and Architectural Management*, Vol. 29, No. 3, 1470-1503. https://doi.org/10.1108/ECAM-12-2020-1059

Emuze F. and James M. (2013). Exploring communication challenges due to language and cultural diversity on South African construction sites. *Acta Structilia: Journal for the Physical and Development Sciences.* Vol. 20, No. 1, 44–65. https://hdl.handle.net/10520/EJC141555

Fang W., Luo H., Xu S., Love P.E.D., Lu Z. and Ye C. (2020). Automated text classification of near-misses from safety reports: An improved deep learning approach. *Advanced Engineering Informatics*. Vol. 44, 101060. https://doi.org/10.1016/j.aei.2020.101060

Géron A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. *O'Reilly Media, Inc.* Sebastopol, CA, USA.

Goh Y.M. and Ubeynarayana C.U. (2017). Construction accident narrative classification: An evaluation of text mining techniques. *Accident Analysis and Prevention*. Vol. 108, 122–130. https://doi.org/10.1016/j.aap.2017.08.026

Kedia J., Vurukuti T., Bugalia N. and Mahalingam A. (2021). Classification of safety observation reports from a construction site: An evaluation of text mining approaches, *in: PMI Research & Academic Virtual Conference 2021. Indian Institute of Technology Bombay, Mumbai*, 50–66.

LeCun Y., Bottou L., Bengio Y. and Haffner P. (1998). Gradient based learning applied to document recognition. *Proceedings of the IEEE*. Vol. 86, No. 11, 2278–2324. doi: 10.1109/5.726791

Manu P., Emuze F., Saurin T.A. and Hadikusumo B.H.W., 2019. Construction Health and Safety in Developing Countries. *Routledge*. New York, USA. https://doi.org/10.1201/9780429455377

Marks E., Teizer J. and Hinze J. (2014). Near-Miss Reporting Program to Enhance Construction Worker Safety Performance. *Construction Research Congress 2014*: *Construction in a Global Network.* 2315-2324. https://doi.org/doi:10.1061/9780784413517.235

Marucci-Wellman H.R., Corns H.L. and Lehto M.R. (2017). Classifying injury narratives of large administrative databases for surveillance—a practical approach combining machine learning ensembles and human review. *Accident Analysis & Prevention.* Vol. 98, 359–371. https://doi.org/10.1016/j.aap.2016.10.014

McHugh M.L. (2012). Interrater reliability: the kappa statistic. *Biochemica Medica*, Vol. 22, No. 3, 276–282. https://hrcak.srce.hr/89395

Ne'Matullah K.F., Pek L.S. and Roslan S.A. (2021). Investigating Communicative Barriers on Construction Industry Productivity in Malaysia: An Overview. *International Journal of Evaluation and Research in Education*. 10, No. 2, 476–482. DOI: 10.11591/ijere.v10i2.21163

Ng A. (2019). Machine learning yearning: Technical strategy for ai engineers in the era of deep learning, *Deepleanring.ai*. https://itbook.store/books/1001590486081

Oswald D., Sherratt F. and Smith S. (2018). Problems with safety observation reporting: A construction industry case study. *Safety Science*, Vol. 107, 35–45. https://doi.org/10.1016/j.ssci.2018.04.004

Peng T., Liu L. and Zuo W. (2014). PU text classification enhanced by term frequency–inverse document frequency-improved weighting. *Concurrency and computation: practice and experience*. Vol. 26, No. 3, 728–741. https://doi.org/10.1002/cpe.3040

Poh C.Q.X., Ubeynarayana C.U. and Goh Y.M. (2018). Safety leading indicators for construction sites: A machine learning approach. *Automation in Construction*, Vol. 93, 375–386. https://doi.org/10.1016/j.autcon.2018.03.022

Samanta S. and Gochhayat J. (2021). Critique on occupational safety and health in construction sector: An Indian perspective. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.05.707

Sarkar S. and Maiti J. (2020). Machine learning in occupational accident analysis: a review using science mapping approach with citation network analysis. *Safety Science*, Vol. 131, 104900. https://doi.org/10.1016/j.ssci.2020.104900

Tixier A.J.-P., Hallowell M.R., Rajagopalan B. and Bowman D. (2017). Construction safety clash detection: identifying safety incompatibilities among fundamental attributes using data mining. *Automation in Construction*. Vol. 74, 39–54. https://doi.org/10.1016/j.autcon.2016.11.001

Tixier A.J.-P., Hallowell M.R., Rajagopalan B. and Bowman, D. (2016a). Application of machine learning to construction injury prediction. *Automation in Construction*, Vol. 69, 102–114. https://doi.org/10.1016/j.autcon.2016.05.016

Tixier A.J.-P., Hallowell M.R., Rajagopalan B. and Bowman, D. (2016b). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Automation in Construction.* Vol. 62, 45–56. https://doi.org/10.1016/j.autcon.2015.11.001.

Trajkovski S. and Loosemore M. (2006). Safety implications of low-English proficiency among migrant construction site operatives. *International Journal of Project Management.* Vol. 24, No. 5, 446–452. https://doi.org/10.1016/j.ijproman.2005.11.004

Wang M., Wang C.C., Sepasgozar S. and Zlatanova S. (2020). A Systematic Review of Digital Technology Adoption in Off-Site Construction: Current Status and Future Direction towards Industry 4.0. *Buildings*. Vol. 10, No. 11, 204. https://doi.org/10.3390/buildings10110204

Wold S., Esbensen K. and Geladi P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*. Vol. 2, No. 1-3, 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Xu J., Cheung C., Manu P. and Ejohwomu O. (2021). Safety leading indicators in construction: A systematic review. *Safety Science*. Vol. 139, 105250. https://doi.org/10.1016/j.ssci.2021.105250

Yan H., Yang N., Peng Y. and Ren Y. (2020). Data mining in the construction industry: Present status, opportunities, and future trends. *Automation in Construction.* Vol. 119, 103331. https://doi.org/10.1016/j.autcon.2020.103331

Zhang F. (2022). A hybrid structured deep neural network with Word2Vec for construction accident causes classification. *International Journal of Construction Management.* Vol. 22, No. 6, 1120–1140. https://doi.org/10.1080/15623599.2019.1683692

Zhang F., Fleyeh H., Wang X. and Lu M. (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction.* Vol. 99, 238–248. https://doi.org/10.1016/j.autcon.2018.12.016

Zhang Y. and Wallace B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. https://doi.org/10.48550/arXiv.1510.03820.

Zhou Z., Li C., Mi C. and Qian L. (2019). Exploring the Potential Use of Near-Miss Information to Improve Construction Safety Performance. *Sustainability*. Vol. 11, No. 5, 1264. https://doi.org/10.3390/su11051264