Using Multi-class Classifiers to Predict Injuries and Identify Leading Indicators in Mine Industry Safety Report Corpus

Aadhithya Dinesh[1], Leonard D. Brown[1], Jefferey L. Burgess[1], and Hong Cui[2]

[1] Mel & Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ, USA
[2] School of Information, University of Arizona, Tucson, AZ, USA
ldbrown@arizona.edu

Abstract. Although progress has been made to improve health and safety (H&S) outcomes, many companies in high-risk industries struggle to identify health risks and avoid safety hazards. In this study, we use Machine Learning (ML) and Natural Language Processing (NLP) to predict potential injuries to ankle, back, eye, hand, knee, and shoulder in mine operators' safety reports. We assess the performance of four models, including SVM, Random Forest (RF), FastText, and BERT, on a transfer learning task. Each model is trained on the publicly available MSHA Accident Injuries dataset and then used to predict injuries in partners' internal safety interaction reports. These reports describe events where an injury might occur, but none was reported. Our tests showed that the models achieved 88-90% accuracy when predicting injuries within the MSHA dataset; although accuracy rates were lower in the transfer learning test (F1 scores of 61% best case for SVM and 48% worst case for BERT), three models (SVM, RF, and FastText) performed 81-97% as well as experts. Our results suggest that the models are comparable to human experts' intuition on this task. Association Rule Mining (ARM) was then used to discover relationships between injuries and keywords in operators' injury reports. ARM generated 149 rules with a confidence >80%; evaluation by four experts deemed 39 rules useful and intuitive. The outcomes of this project will help companies improve risk analysis and controls hierarchies, laying the foundation for a predictive H&S management system that will reduce injuries and fatalities throughout the industry.

Keywords: Health, Safety, Machine Learning, Natural Language Processing, Association Rule Mining.

1 Introduction

Health and safety (H&S) are key issues in many high-risk industries, such as mining, manufacturing, oil and gas, and construction. Although steady progress has been made to improve worker safety, companies struggle with methods to identify occupational health risks and avoid conditions giving rise to safety hazards, both of which are essential for reducing injuries, chronic health conditions, and deaths. The mining industry in particular is oriented toward reactive H&S management, both in the metrics used, such as reportable incidents, and the methods for achieving better performance – for example, addressing patterns of violation. Furthermore, the design of H&S software systems and technologies is equally oriented toward reaction, with dashboards and performance graphs prominently displaying "*lagging indicators*" such as the numbers of injuries, hours lost, and hazards on the worksite. There is a need to better utilize the wealth of data being collected on mine operations, environmental conditions, and incidents which may help managers and front-line workers address H&S issues before they become reportable.

Indeed, proactive H&S management systems are vital to address the human components of sustainable mining. This study uses computational intelligence strategies to learn the relationships between key *leading indicators* of H&S and their downstream lagging indicators. Leading indicators may include certain worker behaviors, work area states, equipment types, or other conditions to be determined. In this work, we focus on a subset of lagging indicators that are of high interest both to mine operators and companies in other high risk industries; specifically we focus on injuries to six commonly injured body parts: ankle, back, eye, hand (including fingers), knee, and shoulder.

Using valuable H&S data provided by industry partners, we employ machine learning (ML) and Natural Language Processing (NLP) to examine the underlying relationships between leading and lagging indicators. This work explores two research questions:

1) Can NLP and ML methods automatically predict injury outcomes from operators' internal reporting corpus?
2) Can useful and intuitive leading indicators be derived from reporting corpus to inform mitigation strategies for key injury types?

We evaluate four popular ML/NLP models, including Support Vector Machines (SVM), Random Forest (RF), FastText, and Bidirectional Encoder Representations from Transformers (BERT) to identify potential injuries. We employ a transfer learning approach; the models are trained on a large and readily available dataset and then used to predict injuries in a company's internal reporting corpus – a projected use case to enable future generations of safety management systems (SMS). Association Rule Mining (ARM) is then used to discover keyword-injury associations. The predictive technologies and leading indicators developed in this work will serve as a foundation to augment risk management programs and control hierarchies, and ultimately to develop a smarter SMS for mining and other high-risk industries.

2 Background

The Mine Safety & Health Administration (MSHA) serves as the regulatory body for the mining industry in the United States (US). As part of the Open Government initiative, MSHA offers a large, publicly available dataset summarizing all accident and injuries at US mine sites [1]. This dataset is frequently analyzed to determine injury types, health conditions, and clues on potential causes. The most commonly injured body parts have been identified as ankle, back, eye, finger, hand, knee, and shoulder [2][3][4].

Workforce H&S outcomes are well documented in the mining industry, although there remains an emphasis on negative metrics such as economic burden [5] and lost time injury rates [1] (Fig. 1). Although informative of industry trajectory, such metrics do not address the root causes of accidents and injuries, nor do they provide insights into how accidents may be prevented. Indeed, a focus on reportable injuries and regulatory violations – i.e. lagging indicators – have shown only limited correlation with improved H&S outcomes; additional study is needed to understand the complex causal pathways between leading and lagging indicators and to prevent negative H&S outcomes [6].

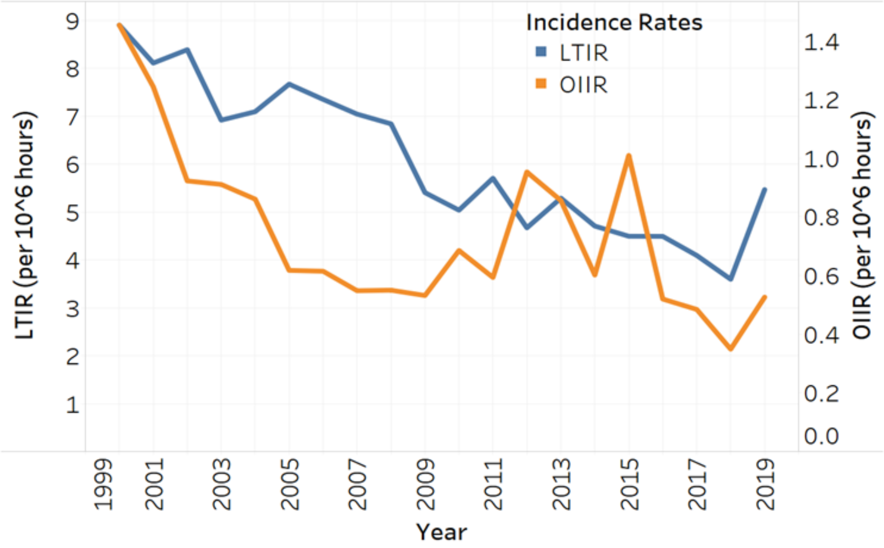

Fig. 1. Lost-time incidence rate (LTIR) and Occupational illness incidence rate (OIIR) for mines in western US from 2000-2018, excluding abandoned mines.

ML methods offer new ways of discovering leading indicators and developing predictive models that improve H&S. Such techniques are supported by recent work in a variety of industries, including as construction [7][8], energy [9], medicine [10], steel [11], and bioinformatics [12]. In particular, SVM, RF, and Artificial Neural Network (ANN) are among the most robust and high-performing methods. For example, Poh et al. employed several ML methods to develop 13 leading indicators of safety from a 7-year dataset of a construction company [7]. They also compared the performance of Decision Tree (DT), SVM, K-Nearest Neighbor, Logistic Regression, and RF in predicting construction site accidents on a 3-class classification task (no/minor/major accident). They found RF performed the best (accuracy=78%). Sarkar et al. found SVMs with features generated by DT (C5.0) outperformed ANN with a prediction accuracy of 90.7% on an occupational accident prediction task, after applying the same optimization techniques for both methods [11]. Recent popular NLP methods, such as FastText and BERT, have great potential to improve the utility of narrative data, such as incident descriptions and injury reports; these techniques have not yet been adequately explored in the context of H&S for high risk industries[13].

There is a need to better understand the relationship between injuries and their upstream indicators to facilitate risk management and improve training [14]. For example, Brown et al. used grounded theory to aggregate factors and develop a meta-model for behavioral risk factors, although the model as yet to be validated for mining [15], while Silva & Jacinto applied statistical methods to derive a model of the “typical accident” [16]. ML/NLP methods, including Topic Modeling [17], SVM, DT, and Naïve Bayes [8], as well as ARM [18], have also been applied to explore reporting corpus and identify causal factors of incidents in mining and allied industries; notably, Ganguli et al. used a transfer learning approach to train a set of RF models on an MSHA dataset, applying them to predict incident types based on mine operators’ incident report narratives [19]. Their system achieved a 96% overall accuracy on the operator’s incident reports. Note that these studies focused on realized incidents and their root causes. Further study is needed to investigate the circumstances and precursors of injuries and methods to detect them in the workplace.

3 Datasets

3.1 Training Data Source and Corpus Preprocessing

MSHA’s “Accident Injuries” dataset was used as a training data source for this study [1]. The MSHA dataset is a large, publicly available corpus describing over 245,000 injury-causing accidents at US mine sites since 2000. In this dataset, categorical values are expressed with standard codes and continuous values have standard units. In particular, the dataset includes reports that are already labeled by classes of injury which comprise 47 body parts. For this study, the number of classes was reduced to the six most frequently injured areas: ankle, back, eye, hand (inclusive of fingers), knee, and shoulder. The remaining body parts were classified as “other”. The MSHA accident narratives were used as features, while the six injury classes were used as labels to train each model.

For the ML algorithms (SVM and RF), the following preprocessing steps were taken:

- **Tokenization.** Each entry in the corpus was converted to lowercase and split into a set of words. All punctuation marks were removed.
- **Stop words removal.** Stop words are words that exist for semantics alone. Stop words were removed from our dataset using the NLTK (Natural Language Toolkit) open-source library [20].
- **Lemmatization.** Lemmatization is the process of grouping together inflected forms of a word such that they may be identified as a single entity. For example, ‘acts’, ‘acting’, and ‘acted’ are all derived from the root word ‘act’. The remaining words in the corpus were lemmatized.
- **Vectorization.** The words were then converted into numerical vectors using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization.

The two NLP models required additional consideration. For FastText, features were appended with injury class labels in the format *__label__[BODY PART].* As BERT relies on the context of an input stream, stop words removal and lemmatization were not performed.

3.2 Target Dataset Collection and Annotation

Safety Interaction Reports (SIR). The SIR dataset included proprietary H&S records from an anonymous mining industry partner. This partner was a mid-sized operator in the Metal/Non-Metal sector with four active mine sites across North America, including both surface and underground operations. Each SIR record contained a 1-2 sentence narrative describing an observation or interaction, but did not identify any specific safety incident or negative outcome (e.g. injury). For example, the text of a record might include, "*Observed a partially unguarded access to the sublevel, the manhole cover was ajar*" or "*Locked out SAG mill for down day, verified the mill would not start. Started removing SAG liners. Made a few changes in the way it was done.*" Five hundred records of the SIR dataset were considered in this study.

Three mine industry safety experts were recruited to evaluate the 500 safety interaction records based on their knowledge and experience in the domain. The first expert (EXP1) was a certified safety professional with a background in industrial hygiene and 10 years of field experience in occupational health; the second expert (EXP2) was a mine industry practitioner and certified trainer with 30 years of experience; and the third expert (EXP3) was a research scientist with over ten years of expertise in safety training and curriculum design. The three experts were asked to label each SIR record with one or more injury types (i.e., ankle, back, eye, hand, knee, shoulder or other) that could arise in that circumstance. In other words, the experts were asked to use their intuition to postulate injuries that might occur based on the narrative

The manually annotated SIR dataset provides an excellent basis to examine the quality of our predictive models in a transfer learning situation that is relevant to incident prediction and risk analysis. First, the purpose of the safety interaction dataset differs substantially from that of the MSHA training data; the MSHA data deals with incidents that have already occurred, while the interaction reports deal with situations where a safety incident might feasibly occur. Second, the phrasing and terminology of the interaction data are less formal and more grounded in the everyday language (vernacular) of the mining workplace. Finally, the interaction reports simulate difficult "what if" circumstances that the models could be reasonably expected to face in real-world deployment.

Incident Management Reports (IMR). IMRs describe incidents of varying severity, from "near misses" to fatalities. In particular, "near misses" indicate circumstances where either good luck or one critical control prevented an injury. Near misses and minor incidents are not reportable and therefore have no representation in the publicly available MSHA dataset. For this study, our industry partner provided access to 5,000 recent IMR records. As with the SIR records above, each IMR record consisted of a few sentences of text narrative describing the circumstances of an incident. Although the IMR records were given a 4-level risk rating (e.g. Low, Moderate, High, and Critical) by the industry partner, they were not categorized by injury type.

Studies suggest that, for every significant injury or fatality, there are numerous incidents with less serious injuries and even more with no injury at all [21]. The IMR reports are therefore a valuable tool for analyzing and preventing future injuries, if such conditions (i.e. leading indicators) can be linked to injury outcomes (lagging indicators); these narratives were used to discover association rules between incident keywords and injury types in Sec. 7.

4 Predictive Models

In this study, we evaluated the performance of four ML/NLP methods to classify injury outcomes based on the text of a mine operator's safety reports. Four popular classifier types are considered. SVM and RF are robust Machine Learning methods that perform well across a variety of classification tasks. Two NLP approaches, including FastText and BERT, were also selected due to their promising application in many domains and their large communities of active users; furthermore, both NLP models may be trained on a very large corpus (i.e. hundreds of millions of words). An overview of each model is given below.

4.1 Support Vector Machines (SVM)

SVM is a supervised learning algorithm which finds the hyperplane that best separates two classes based on the Structural Risk Minimization principle [22]. As SVM is inherently designed for binary classification, a One-vs-One (OvO) approach was used in this study since our dataset was roughly balanced. The OvO approach yields $n*(n-1)/2$ binary classification problems, where n is the number of classes. Each binary classification model predicts one class label; the predicted class is then the class receiving the most votes. A detailed discussion on SVM for text classification may be found in [23].

4.2 Random Forest (RF)

The RF classifier is a collection of multiple decision tree classifiers on various subsamples of the dataset. It uses bagging and feature randomness to create a forest of decision trees by ensuring low correlation among decision trees. The main advantage of this method is its ability to overcome overfitting. Furthermore, this factor becomes more important as we incorporate a transfer learning approach in our study. Details on the Random Forest classifier may be found in [24].

4.3 FastText

FastText is a library for learning word embeddings for text classification. It is essentially a shallow neural network with just one layer of a linear classifier. It represents text as a continuous bag of words (CBoW) and uses N-gram features coupled with hierarchical SoftMax, with negative sampling to reduce training time. A discussion of FastText may be found in [25].

4.4 Bidirectional Encoder Representations from Transformers (BERT)

BERT uses a deep learning model (ANN) specialized for natural language processing tasks. The transformers are based on an attention model architecture that can deal with long-range dependencies to solve sequential tasks. BERT leverages this transformers architecture for language modeling; it masks 15% of the words in each sequence and predicts the masked words by reading the sequence from both directions. In this study, we used the BERT base model (cased), as it has excellent support for everyday language and context, having been pre-trained on over 40 gigabytes of Wikipedia articles and five gigabytes of book corpus data. Details on the BERT model may be found in [26].

5 Methodology

5.1 The approach

A transfer learning approach was used in this study for two reasons: First, the amount of partner data that we could gather and manually annotate was not large enough to train models without overfitting. Second, each worksite and company uses its own variations of terminology and vernacular based on location, workforce demographics, and company operating procedures; conversely, the MSHA dataset provides an industry standard baseline for terminology and methods. This study will evaluate the performance of our models in a transfer learning task in the mining H&S domain. We believe that a well-trained model may yield good results across mine operators and datasets for injury prediction tasks.

The four ML/NLP models were trained on 50,000 rows of the preprocessed MSHA dataset (Sec. 3.1). Although each row of the MSHA data contains single labels for each incident narrative, the target dataset, as labeled by experts, had multiple labels. Thus the models had to be evaluated as multi-label models, yielding a two-fold challenge: Transferring to the model to the new type of dataset – from standard MSHA data to the operator's SIR dataset – and training multi-label models based on the single labels in the training dataset.

5.2 Experimental Setup and Hyperparameter Tuning

Three models, including SVM, RF, and FastText, were trained using a train-test split of 90-10, while BERT used a train-validate-test split of 80-10-10. Note that we used the `GridSearchCV` method to perform Ten-Fold Cross-Validation when training the SVM and RF models. To analyze the models' performance, they were first evaluated on 5,000 arbitrary rows of the MSHA dataset; the models were then evaluated against 500 rows of the SIR dataset (Sec. 3.2) during the transfer learning task.

Two computer configurations were used for training. To train the SVM, RF, and FastText models, we used an Apple computer running macOS 12.4 with an 8-core M1 CPU running at 3.2 GHz with 8GB of memory. To train BERT, a Nvidia Volta-based GPU with 32 GB of memory was used. The latter system provided 5,120 CUDA Cores and 640 Tensor Cores, offering a tensor performance of 112 TFLOPS and a memory bandwidth of 900GB/sec.

Each of the four models was configured and its hyperparameters tuned as follows:

SVM. Scikit-learn's Support Vector Classification (SVC) is an implementation based on libsvm and was used for this study. TF-IDF vectors were used as inputs for the model, as they performed better than TF alone. Tests were then run with two kernels: linear and radial basis function. The linear kernel provided the best fit for this study. Two additional hyperparameters where considered while configuring the model: `C value` and `Gamma`. The `C value` is a penalty parameter that determines the number of misclassifications allowed for the model, while `Gamma` controls the influence distance of each training point. To find the best parameters, Grid Search was performed with `C value` – {0.1, 1.0, 10, 100, 1000} and `Gamma` –{1.0, 0.1, 0.01, 0.001, 0.0001}. Among these values, setting both `C value` and `Gamma` at 1.0 gave the best model performance.

Random Forest. Scikit-learn's ensemble RF Classifier was used for this study. Similar to SVM, TF-IDF vectors were used as inputs. Five hyperparameters were considered: 1) `n_estimators` determines the number of decision trees in the forest, such that a greater number of trees increases accuracy but also leads to overfitting and longer training time; 2) `max_features` limits the maximum number of features considered while making a split in the decision tree; 3) `max_depth` limits the maximum depth of a tree to reduce overfitting; 4) `min_samples_split` determines the number of samples required to split a node in a tree; and 5) `min_samples_leaf` guarantees a minimum number of samples in the leaf node. Since experimenting with all combinations of parameters factorially increases training time, most hyperparameters were given default values. Grid Search was performed to find the best values for `n_estimators` – {100, 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000}, `max_depth` – { 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, none} and `max_features` – {`auto`, `sqrt`}. Best values were found to be `n_estimators` = 100, `max_depth` = none, and `max_features` = `sqrt`. We also considered the `criterion` function, which measures the quality of a split; `gini` was found to provide the best performance across all classification tasks.

FastText. The open-source FastText Application Programming Interface was used for this study. As discussed in Sec. 3, the model expects each line of text to start with a `__label__` prefix, followed by text of the narrative. The training and validation dataset were converted to lowercase before being passed as inputs to the model. FastText's autotune feature was used to automatically find the best hyperparameters. The parameter `autotune-duration` was used to determine the training duration. Different configurations of hyperparameters were tested, including `epoch` – {1 to 100}, `learning rate` – {0.01 to 5.0}, and `wordNgrams` – {1 to 5}. Note that `epochs` determines the number of training cycles to be completed, while the `learning rate` determines the step size at each iteration minimizing the loss function, and `wordNgrams` refers to a continuous sequence of *n* items from the given sample set. The best-performing hyperparameters were found as follows: `epoch` = 3, `learning rate` = 0.05, `wordNgrams` = 2, and `loss function` = SoftMax.

BERT. The BERT base model (cased) from the Hugging Face library was used for our multi-labeling problem. BERT accepts a maximum of 512 tokens as input. Since our narratives mostly exceeded the 512-token limit, they were truncated automatically. An alternate approach involved splitting each narrative into multiple subtexts and classifying the incident based on a voting mechanism. The truncation method was selected for this study due to computational constraints. While evaluating the model's performance on the MSHA dataset, the final output layer was passed through a SoftMax function to convert the prediction scores to probabilities. While evaluating the performance on the partner datasets, since multiple labels had to be determined, the final layer was also passed through a Sigmoid function. We used cross-entropy loss as the `criterion` function, as it generally yields superior performance to the square-loss function. Note that the `criterion` is used to signal the end of training for deep learning models; given an input and target, a gradient is computed according to the given loss function. We used the Adam optimizer [27] for this study and considered two hyperparameters – `learning rate` and `epochs`. The best performing values were found to be 1e-6 and 5, respectively.

6 Results and Discussion

To evaluate model performance, the top K (K=1 and K= 4) machine-predicted labels were checked against the actual labels – i.e., as provided by experts in the transfer learning task. K=1 essentially evaluated the models based on their multi-class classifying capability, while K=4 was used to check if a model had reasonable multi-label prediction capabilities. K=4 was chosen, as 4 was the maximum number of labels provided by an expert for any record. The evaluation metrics used included Precision, Recall, F1 Score, and Support. As this problem does not value Precision over Recall or vice versa, the F1 Score was used to evaluate the performance of each model. Furthermore, since our dataset was roughly balanced, we considered the micro average as the accuracy score of our models.

6.1 Results on the MSHA dataset

After the models were trained on the MSHA dataset, they were tested with an arbitrary 5,000 rows of unseen data. The performance results are given in Table 2. As there were single labels for each incident, these scores may be interpreted as multi-class classification scores.

Table 1. F1 Scores of four predictive models with MSHA dataset.

MODEL	PRECISION	RECALL	F1 SCORE
SVM	0.901	0.899	0.899
RF	0.902	0.897	0.898
FAST	0.857	0.863	0.854
BERT	0.886	0.884	0.884

Note that SVM yielded the highest F1 Score (89.9%), closely followed by RF and BERT. Note that Fast Text has only a ~4% lower F1 score than the best-performing model, SVM.

6.2 Results with SIR dataset

The F1 scores for the transfer learning task are reported for K=1 (Table 3). As before, this may be interpreted as multi-class classification scores.

Table 3. F1 Scores of four predictive models with SIR dataset, using three experts' labels as Ground Truth.

MODEL	EXP 1	EXP 2	EXP 3	AVG
	F1 SCORE	F1 SCORE	F1 SCORE	F1 SCORE
SVM	0.462	0.504	0.612	0.526
RF	0.484	0.512	0.646	0.547
FAST	0.382	0.426	0.478	0.429
BERT	0.39	0.456	0.516	0.454

Note that the performance of the models drops by 30%-40%. Unlike the MSHA dataset, the SIR data includes a collection of narratives where no injury was reported, yet an injury *could* occur. Indeed, predicting injuries in this context is a difficult task, even for domain experts who possess the power of intuition based on decades of training and experience; for this reason, we also allowed the experts to suggest more than one label for each report.

6.3 Comparison of Machine Results vs Expert Predictions

As a better indicator of model performance, it is to useful to examine the performance of each expert (Table 4). To interpret these scores, we may think of each expert as a (human) predictive model. Each expert's predictions may be compared against the ground truth of two other experts, and thus an average performance score calculated. Note that "Labels Predicted" is a measure of the total number of predicted labels for the 500 reports by each of the experts; furthermore, since more than one label can be present for a single report, all support values are greater than 500.

Table 4. Micro averages of Precision, Recall, and F1-Scores for each expert using two other experts as Ground Truth.

EXPERT	AVG	AVG	AVG	LABELS
	PRECISION	RECALL	F1 SCORE	PREDICTED
EXP 1	0.560	0.423	0.481	1014
EXP 2	0.548	0.534	0.537	842
EXP 3	0.503	0.654	0.568	704

Similarly, consider the (machine) model predictions at K=4, using the three experts' predictions as a ground truth with all labels considered (Table 5). Note that the Recall scores are considerably higher than the Precision scores. This outcome is expected, as the K value has increased from 1 to 4. At K=4, BERT performed considerably worse than the other models; possible reasons for this disparity may include the truncation of the narratives to 512 tokens or overfitting of the model to the MSHA data.

Table 5. Micro averages of Precision, Recall, and F1-Scores for all three experts at K = 4.

MODEL	EXP 1			EXP 2			EXP 3			Final Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SVM	0.367	0.729	0.49	0.346	0.816	0.483	0.289	0.794	0.413	0.332	0.787	0.471
RF	0.348	0.686	0.462	0.347	0.824	0.488	0.29	0.824	0.429	0.327	0.778	0.462
FAST	0.359	0.708	0.476	0.343	0.814	0.482	0.280	0.794	0.413	0.326	0.771	0.461
BERT	0.246	0.485	0.326	0.116	0.274	0.163	0.111	0.315	0.164	0.173	0.390	0.228

The Final Average (micro average) Precision, Recall, and F1 Scores (Table 5) may be compared to the Average F1 Score of the three experts (Table 4). In this context, three models perform favorably. Indeed, with the exception of BERT, the models have a final average F1 Score that is comparable to each experts' prediction scores when using the other experts as a Ground Truth; the machine models are 97% as good as the experts in the best case comparison (i.e. SVM vs. EXP1) versus 81% as good in the worst case (FastText vs. EXP3). Note that, in this analysis, the machine models' predictions are compared against three ground truths (all 3 experts), whereas the experts' predictions are compared against two ground truths (2 other experts).

7 Application in the Mining Industry

Leading indicators can provide H&S professionals with new avenues to prevent injuries and illnesses in the workplace by allowing them to improve risk management processes and controls hierarchies. To be effective, leading indicators must be concise, intuitive, and validated by empirical evidence. Our goal is to help mine operators answer several key questions relating to the lagging indicators (i.e. the injury classes) selected in this study:

- **Behaviors.** Are crews or workers fit for duty and adequately trained for work areas and jobs?
- **Environments**. Are work areas consistently safe, or are there issues suggestive of hazards?
- **Operations**. Are standard operating procedures adequate, and are personnel following them in practice?

As a step toward developing leading indicators, we used ARM to align key words in incident narratives with each injury class. Our industry partner's IMR dataset was used for this analysis. The IMR dataset considers incidents that did occur and thus may have caused negative outcomes ranging from negligible to fatal. The partner IMR records were used for this analysis, as opposed to the MSHA dataset, for three reasons: First, the IMR records include additional data and incidents which do not rise to the level of reportable and thus do not exist in the MSHA dataset; second, the reports are less formal and more grounded in vernacular; and third, the output rules should be more contextually relevant and thus of greater utility for domain users.

7.1 Association Rule Mining on FastText-Labeled IMR Dataset

ARM was used to develop simple association rules between keywords and injury types based on the empirical evidence in our partners' IMR dataset. As the IMR narratives were not specifically labelled by injury class, we selected a high-performing predictive model (Sec. 6) to automatically label 5,000 narratives each with an injury class. FastText was used for this task, as its performance compared favorably yet its computational performance exceeded the other models; a computational performance assessment of FastText may be found in [25].

After labeling each incident narrative with an injury class, we extracted keywords from the reports by combining two different methods: (1) Using Structural Topic Modeling (STM) with an 8-topic model; and (2) Using Topic Modeling with TF-IDF to generate a document-term matrix from which keywords were extracted. Note that eight classes of injuries were used in this analysis, which included the six classes identified previously, plus "Wrist" and "Other". Both the STM and TM with TF-IDF methods produced good results for approximately half of the injury classes and confusing results for the other half. We also noted considerable amounts of overlap between classes. For example, we found that, even after removing stop words, the narratives of any pair of injury classes had >30% keyword overlap with each other using the 8-topic STM. This finding may suggest that different injury classes occur under similar circumstances or contexts; it also informs the types of the goodness of fit measures that may be more suitable to ARM (discussed below) – in particular, Confidence and Lift may be better measures of goodness than the Kulczyinski Method or Imbalance Ratio.

Discovery of Association Rules. After labelling and Topic Modeling, the class labels and keywords of each record were merged to create "transactions" for the ARM method a priori. The ARM process discovers human-readable rules in the following format:

[left hand indicators] → [right-hand indicators] (support, confidence)

To understand this syntax, consider the rule "smoking -> lung cancer (0.05, 0.30)". This rule states that, within the dataset mined, 5% of the people smoke, and among those people, 30% of them get lung cancer. Details on the ARM technique may be found in [28]. The rules generation process yields an arbitrarily large number of candidate rules which may be ranked by a "measure of goodness" criterion. For reasons outlined above, Confidence was used as the measure of goodness for this study, with a Confidence threshold of 80% (0.80) as the inclusion criterion. In total, 149 rules satisfied this criterion across the seven classes of injury (i.e. excluding catch-all category "Other"). A visualization of the association rules and their supports may be found in Fig. 2.

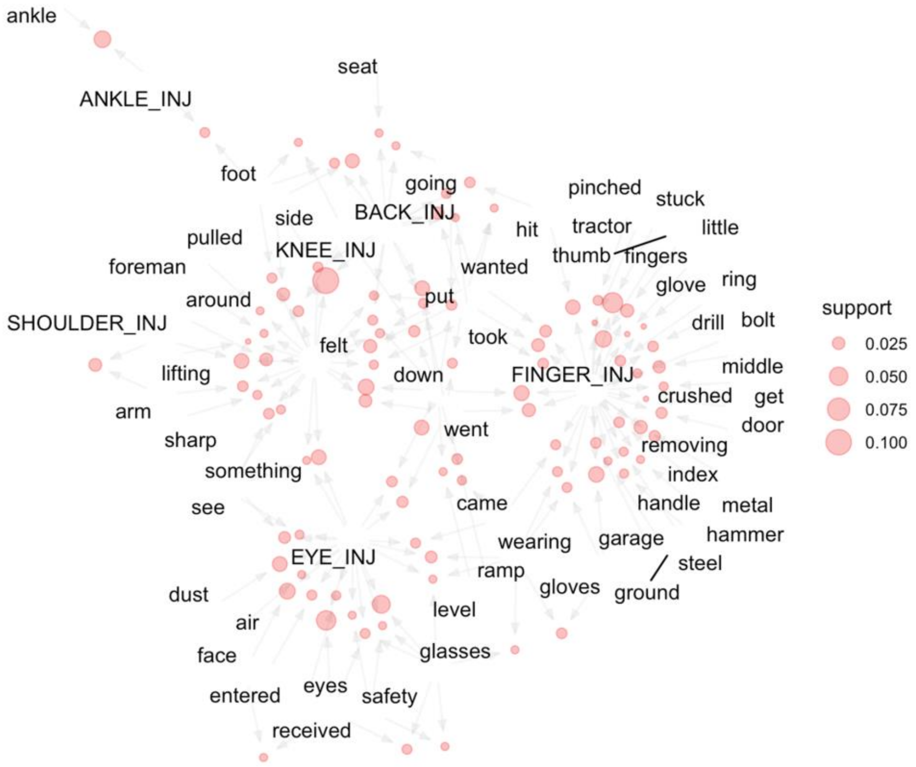

Fig. 2. A visualization of association rules and support values for six classes of injuries.

Evaluation of Association Rules. Although association rules may be discovered with high Confidence by the ARM technique, this does not mean that the rules are themselves meaningful for improving H&S. Domain experts were asked to evaluate the rules in terms of their intuitiveness – that is, the rules reveal interesting associations between classes of injury and potential correlating factors – and their applicability to safety management processes, such as for improving controls hierarchies or training.

For this evaluation, we invited four experts to survey the rules and evaluate them as being either “useful” (i.e., interesting and/or applicable) or “not useful” (i.e., too vague, or nonsensical). Three of these experts (EXP1-3) participated in the model performance assessment outlined previously (Sec. 6). A fourth expert (EXP4) had over 25 years of mining management experience and held a senior leadership position as head of a major operator’s H&S department. Each expert provided their ratings independently.

7.2 Correlation between Scores and the Usefulness of a Rule

A survey of the experts’ Usefulness ratings is given in Table 6. We used a binary classification to label each rule as either Useful=1 or Not Useful=0 for each of the four experts. The cumulative usefulness score indicates the number of rules satisfying each level of expert consensus. For instance, for the "Eye" class of injury, 5 rules (out of 61 candidates) were unanimously voted as useful by all four experts, while 17 rules were voted as useful by three or four experts. Notably, half of the association rules were rated as Useful by at least one expert (Cumulative Usefulness > 0) and 39 rules had a high level of consensus as being Useful (Cumulative Usefulness = 3 or 4) for safety management processes, such as for improving controls hierarchies or training.

Table 6. Number of association rules labeled as "Useful" by experts for each injury class.

Injury Class	# of Rules labeled as useful by N experts					Total Rules
	N = 0	N = 1	N = 2	N = 3	N = 4	
ANKLE	0	4	0	1	2	7
BACK	13	5	4	5	3	30
EYE	30	3	11	12	5	61
FINGER	24	5	6	7	4	46
HAND	1	0	0	0	0	1
SHOULDER	0	0	1	0	0	1
WRIST	1	0	2	0	0	3

Correlation analysis was then performed to identify the relationships between experts' Usefulness scores and the various measures of goodness. As shown in Table 7 below, Confidence, Lift, and Certainty all positively correlated with the experts' ratings, while an inverse relationship was noted for Support, Coverage, and Count.

Table 7. Number of association rules labeled as "Useful" by experts for each injury class.

MEASURE	CORRELATION
SUPPORT	-0.24
CONFIDENCE	0.28
COVERAGE	-0.25
LIFT	0.11
COUNT	-0.24
COSINE	-0.08
CERTAINTY	0.29
LEVERAGE	-0.22

7.3 H&S Outputs and Outcomes

Association Rules represent a significant contribution of this project, as they provide operators with insights from their own reporting processes. The 39 rules judged to be of high utility are given in Table 8. These rules are now being used by partners to revise their risk assessment and management plans. A sample of the high utility rules and outcomes included the following:

1) Back injuries were more often reported when “morning” was listed in the description, suggesting a need for stretching and warm-up before job activities;
2) The five most common terms associated with hand injuries all related to fingers, with “pinch” occurring 2nd most frequently, even though it may be less obvious as a mechanism of significant injury;
3) Eye injuries were frequently associated with the word “dust”, suggesting that many incidents were due to dust contamination and subsequent irritation; there may be a need for better training around eye PPE as well as additional areas where eyewear is warranted.

Mitigation strategies are being developed for each of the 39 rules, including new training tools for annual refreshers that target many of the contextual situations for injury that are suggested by these rules.

Table 8. Association rules judged to have high utility by four mining safety experts.

{keywords}	→	{injury-type}	{keywords}	→	{injury-type}
ankle, mud		ANKLE	dust, felt		EYE
ankle, rolled		ANKLE	air, face		EYE
ankle, hole		ANKLE	hose, received		EYE
felt, muck		BACK	glasses, wearing		EYE
felt, picked		BACK	glasses, safety, wearing		EYE
felt, heat		BACK	received, safety, wearing		EYE
bent, felt		BACK	glasses, received wearing		EYE
down, seat		BACK	glasses, received, safety, wearing		EYE
felt, seat		BACK	hit, pinch		FINGER
felt, morning		BACK	guardrail, stuck		FINGER
down, ice		BACK	scissor, stuck		FINGER
received, rinsed		EYE	frame, tractor		FINGER
face, sprayed		EYE	jack, stuck		FINGER
eyewash, received		EYE	gloves, index		FINGER
cement, received		EYE	index, stuck		FINGER
face, valve		EYE	gloves, pipe		FINGER
entered, hold		EYE	drill, stuck		FINGER
face, pressure		EYE	gloves, wearing		FINGER
dust, glasses		EYE	chisel, stuck		FINGER
dust, received		EYE			

8 Conclusions and Future Work

Occupational exposures and safety hazards remain key considerations in high risk industries. In this study, we have assessed the performance of four ML/NLP methods, including SVM, RF, FastText, and BERT, to predict injuries based on the safety narratives in a partner's H&S reporting corpus. Using a transfer learning approach to train the models with a large and publicly available MSHA dataset, we obtained performance rates in predicting injuries that are 81%-97% as good as domain experts, suggesting the power of ML/NLP approaches to match human intuition on this task. Furthermore, ARM was used to generate association rules linking keywords in incident reports with injury types. A total of 39 high confidence (>0.80) association rules were deemed useful and intuitive to domain experts; these rules represent a step toward linking positive H&S outcomes with actionable leading indicators that operators may use to improve their risk management protocols and controls hierarchies.

We are now evaluating the predictive models with other partners in the mining industry to verify the robustness of the transfer learning process. Future work will consider a new predictive dashboard that provides feedback to operators as new narratives are entered into their H&S management systems. Specific mitigations, including refresher training and controls, will be recommended by the dashboard to address any risks discovered in these narratives. Ultimately, we believe these NLP and ML-enabled technologies will lead to smarter H&S management systems in the workplace, with new predictive models that allow companies in high risk industries to proactively address hazardous conditions and reduce the risk of injury to their workforce.

Acknowledgements. This work was supported in part by NIOSH award 2U60 OH010014-11-00 and a University of Arizona School of Mining & Mineral Resources (SMMR) Seed Grant. The authors wish to thank the anonymous mining operators participating in these studies, without whom this work would not be possible.

References

1. Mine Safety & Health Administration. MSHA Open Government Dataset. Accessed Sept. 10, 2022.
2. Alessa, F. M., Nimbarte, A. D., & Sosa, E. M. Incidences and severity of wrist, hand, and finger injuries in the US mining industry. Safety science, 129, 104792 (2020).
3. Larsen, M., Whitson, A., Pollard, J., & Nasarwanji, M. Analysis of Shoulder Sprains and Strains in Mining. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 65, No. 1, pp. 1371-1375). Sage CA: Los Angeles, CA: SAGE Publications. (2021, September).
4. Roibal, J., Barreto, C., Chester, B., & Essomba, J. Analysis of Copper Mining Accidents in the United States, 2010-2019. (2021).
5. Nasarwanji, M. F., & Sun, K. Burden associated with nonfatal slip and fall injuries in the surface stone, sand, and gravel mining industry. Safety science, 120, 625-635 (2019).
6. Yorio, P. L., Willmer, D. R., & Haight, J. M. Interpreting MSHA citations through the lens of occupational health and safety management systems: Investigating their impact on mine injuries and illnesses 2003–2010. Risk analysis, 34(8), 1538-1553 (2014).

7. Poh, C.Q.X., Ubeynaranyana, C.U., & Goh, Y.M. Safety Leading Indicators for Construction Sites: A Machine Learning Approach. Automation in Construction, v. 93, Elsevier, pp. 375-386 (2018).
8. Zhang, F., Fleyeh, H., Wang, X., & Lu, M. Construction site accident analysis using text mining and natural language processing techniques. Automation in Construction, 99, 238-248 (2019).
9. Zendehboudi, A., Baseer, M. A., & Saidur, R. Application of Support Vector Machine Models for Forecasting Solar and Wind Energy Resources: A Review. Journal of Cleaner Production, v.199, pp. 272–285 (2018).
10. Claudino, J.G., et al. Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: A Systematic Review. Sports Medicine - Open, Springer, v.5 (28) (2019).
11. Sarkar, S., Vinay, S., Raj, R., Maiti, J. & Mitra, P. Application of Optimized Machine Learning Techniques for Prediction of Occupational Accidents. Computer Operations Research, v.106, pp. 210–224 (2019).
12. Couronné, R., Probst, P., & Boulesteix, A.L. Random Forest Versus Logistic Regression: A Large-scale Benchmark Experiment. BMC Bioinformatics, v.19(1), 270 (2018).
13. Ali, F., Ali, A., Imran, M., Naqvi, R. A., Siddiqi, M. H., & Kwak, K. S. Traffic accident detection and condition analysis based on social networking data. Accident Analysis & Prevention, 151, 105973 (2021).
14. Yorio, P. L., Haas, E. J., Bell, J. L., Moore, S. M., & Greenawald, L. A. Lagging or leading? Exploring the temporal relationship among lagging indicators in mining establishments 2006–2017. Journal of Safety Research, 74, 179-185 (2020).
15. Brown, L.D., Pham, N. & Burges, J.L., Toward a Systems Framework Coupling Safety Culture, Risk Perception, and Hazards Recognition for the Mining Industry. Advances in Human Factors in Simulation and Modeling, AHFE (2022).
16. Silva, J. F., & Jacinto, C. Finding occupational accident patterns in the extractive industry using a systematic data mining approach. Reliability Engineering & System Safety, 108, 108-122 (2012).
17. Passmore, D., Chae, C., Kustikova, Y., Baker, R., & Yim, J. H. An exploration of text mining of narrative reports of injury incidents to assess risk. In MATEC Web of Conferences (Vol. 251, p. 06020). EDP Sciences (2018).
18. Qiu, Z., Liu, Q., Li, X., Zhang, J., & Zhang, Y. Construction and analysis of a coal mine accident causation network based on text mining. Process Safety and Environmental Protection, 153, 320-328 (2021).
19. Ganguli, R., Miller, P., & Pothina, R. Effectiveness of Natural Language Processing Based Machine Learning in Analyzing Incident Narratives at a Mine. Minerals, 11(7), 776 (2021).
20. Natural Language Toolkit. NLTK Project. Accessed Sept 10, 2022.
21. Barnes, DF. Leading and Trailing Indicators: Occupational Health. ISSA/Chamber of Mines Conf. - Mines and Quarries: Prevention of Occupational Injury and Disease, South African Institute of Mining and Metallurgy, Johannesburg (2003).
22. Vapnik. V.N. The Nature of Statistical Learning Theory. Springer, New York (1995).
23. Joachims, T. Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science, vol 1398. Springer, Berlin, Heidelberg (1998).
24. Biau, G., Scornet, E. A random forest guided tour. *TEST* 25, 197–227 (2016).
25. Joulin, A., Grave, E., Bojanowski, P. and Mikolov, T. Bag of tricks for efficient text classification. arXiv preprint 1607.01759 (2016).

26. Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. BERT: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
27. Kingma, D.P. and Ba, J.L. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR) (2015).
28. Han, J., Kamber, M., & Pei, J. Data mining: concepts and techniques, Waltham, MA. Morgan Kaufman Publishers, 10, 978-1 (2012).