*Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference*
*Edited by* Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P6833-cd

# A Semi-Automated Framework for Coding Fatal Accident Data in Mines Using Natural Language Processing

Amit Sharma
*Department of Mining Engineering, Indian Institute of Technology Kharagpur, India, E-mail: amitfdh@gmail.com*

Ashish Kumar
*Department of Mining Engineering, Indian Institute of Technology Kharagpur, India, E-mail: aksaini4@kgpian.iitkgp.ac.in*

Shubham There
*Talaipalli Coal Mining Project, NTPC Ltd, India, Email: SHUBHAMTHERE@ntpc.co.in*

Bibhuti Bhusan Mandal
*Department of Mining Engineering, Indian Institute of Technology Kharagpur, India, E-mail: bbmandal@gmail.com*

Mining is among the most hazardous industries, with frequent fatalities resulting from various occupational hazards. Traditionally, identifying the causes of such fatalities has relied on manual coding of accident reports, which is time-consuming, inconsistent, and prone to human error. With the increasing volume of accident reports, particularly in data-intensive environments, automation is crucial for timely safety interventions. Advances in Natural Language Processing (NLP) and Machine Learning (ML) provide promising solutions for semi-automated coding, reducing manual effort while improving accuracy. This study utilizes NLP and ML models to predict the causes of fatalities in Indian mines using accident data from the Directorate General of Mines Safety (DGMS) reports from 2016 to 2022. The dataset consists of 401 fatal accident descriptions spanning seven years. Accident descriptions were pre-processed and vectorized using the Bag of Words approach. Five machine learning models were compared: Naïve Bayes, Logistic Regression with and without adjusted weights, Support Vector Machines, and Random Forest. Each model was trained to predict accident causes based on textual descriptions. The models were assessed based on their accuracy in classification, using an 80/20 train-test split for validation. The study utilized a semi-automatic classification approach. Instances with a high-confidence classification (above a predefined probability threshold) are assigned automatically, while lower-confidence cases are flagged for manual review. Conversely, if the maximum probability is below the threshold, the instance is filtered for manual review. Among the models evaluated, Logistic Regression with Adjusted Weights outperformed the standard Logistic Regression model with a precision of 80%, a recall of 83%, and an F1-score of 80%, demonstrating its robustness in handling imbalanced data and effectively identifying positive cases. This approach significantly reduces manual coding workload, accelerates data processing, and strengthens safety oversight in mining operations.
*Keywords*: Occupational safety, Predictive modelling, Mining hazard mitigation

## 1. Introduction

Mining is one of the most hazardous industries globally, where workers are often exposed to life-threatening risks due to the dangerous working conditions both underground and at surface mines (Sharma and Mandal, 2021). Mining operations have seen a significant number of fatal accidents, making it essential to investigate and prevent the underlying causes to improve worker safety. Fatal accident descriptions, as recorded in the Directorate General of Mines Safety (DGMS) reports, provide detailed narratives of incidents that have occurred across various mining sites in the country. These descriptions, while rich in content, present a major challenge when it comes to analyzing and classifying the causes of accidents due to the unstructured nature of the text. Accurate identification and classification of these causes are critical to implement preventive

measures and to develop safety protocols that could potentially save lives.

Manually coding accident narratives is an arduous and time-consuming task, traditionally requiring skilled personnel to read through each report, comprehend the nature of the incident, and classify it under specific event codes. However, with the sheer volume of accident data generated every year, manual classification is inefficient, error-prone, and often inconsistent. Furthermore, human interpretation can vary, leading to discrepancies in how similar incidents are classified. Thus, there is a growing need for a semi-automated system that can quickly and accurately assign cause codes to these narratives, while reducing human error and streamlining the analysis process. The use of machine learning (ML) and natural language processing (NLP) techniques offers a promising solution to this challenge, enabling the efficient processing and classification of large datasets such as accident and fatality reports (Gupta et al., 2022).

Several studies have explored the use of accident narratives stored in administrative databases, such as national surveys, accident reports, and worker's compensation claims, to extract critical information for analyzing workplace injuries and fatalities (Abdat et al., 2014; Das et al., 2024). These studies have shown that leveraging narrative data provides deep insights into accident causes, revealing patterns that could otherwise be missed in structured data. For example, narratives in Occupational Safety and Health Administration (OSHA) logs in the U.S. have been used to identify the nature of workplace hazards and propose preventive safety measures (Liu and Yang, 2022). Similarly, the narratives in Indian mining accident reports present an opportunity to analyze fatal incidents to improve safety conditions in mines. However, this rich data source has not been fully exploited due to the manual labor required for its analysis.

Manually coding fatal accident descriptions in the Indian mining sector might greatly benefit from automation due to the increasing volume of data recorded each year. Automating the coding process has several advantages. Firstly, machine learning models, once trained on a sufficient dataset, can process large volumes of data at a fraction of the time required by human coders. This allows for the analysis of years of accident reports within a short span of time, making it possible to detect patterns in accident causes and trends more efficiently (Lombardi et al., 2009). Secondly, machine learning algorithms offer consistency in coding that is often difficult to achieve manually. Since human coders may interpret the same narrative differently, this can lead to variation in coding outcomes, which affects the reliability of any subsequent data analysis (Verma et al., 2014). In contrast, machine learning models can be trained to follow a specific set of rules or classifications, ensuring consistent outcomes across similar reports. Finally, automation reduces the labor cost associated with manual coding, freeing up expert human resources to focus on more complex tasks that require expert judgment, such as handling ambiguous or difficult-to-classify narratives.

In this study, we aim to apply a range of machine learning models to predict the causes of fatal accidents in Indian mines, based on accident descriptions provided in DGMS reports from 2016 to 2022. Specifically, we explore the performance of several machine learning models, including Naive Bayes, Logistic Regression, Random Forest, and, Support Vector Machine for the task of classifying accident narratives into predefined cause categories.

## 2. Materials and Methods

### 2.1. *Data collection*

This study utilized fatal accident data from Directorate General of Mines Safety (DGMS) reports spanning 2016 to 2022, with a primary focus on coal mines in India. These reports provide detailed textual descriptions of mining accidents, making them a valuable resource for identifying patterns in fatality causes. The study aimed to forecast accident causes from narrative descriptions, and the data extraction process carefully targeted accident-related text fields while also collecting essential information. The principal data collected comprised the mine type (classified as Opencast (OC) or Underground

(UG)), the accident timing (morning, afternoon, or night shift), the employment of the deceased, and the worker's age to investigate potential demographic risk factors. Most importantly, the immediate cause of fatality, as described in the narrative accident reports, was systematically coded for classification. Since the DGMS reports are unstructured, preprocessing was required to clean and standardize the textual descriptions before applying Natural Language Processing (NLP) and Machine Learning (ML) techniques. This structured approach ensured that the dataset was optimized for cause classification, forming the foundation for developing an automated predictive framework to categorize fatality causes in Indian coal mines.

### 2.2. *Data preprocessing*

Since the DGMS accident reports contain unstructured textual descriptions, preprocessing was essential to transform the raw text into a format suitable for Natural Language Processing (NLP) and Machine Learning (ML) models. The preprocessing pipeline included multiple steps to clean, standardize, and extract meaningful features from the accident narratives while preserving critical contextual information.

The first step involved text cleaning, where punctuation, special characters, and extra whitespace were removed. This was followed by tokenisation, which split the text into individual words to facilitate further processing. Stopword removal was applied to eliminate commonly occurring words (e.g., “the”, “is”, “in”) that do not contribute to cause classification. Additionally, we performed lemmatization to transform words into their root forms, such as “operating” or “operate”, to maintain consistency among similar words.

To prepare the text for classification, different text representation techniques were applied to convert narratives into numerical features. The Bag of Words (BoW) model was implemented with unigram and bigram tokenization to capture both individual word frequency and short sequences of words relevant to accident causes. These preprocessing steps ensured that the textual descriptions were clean, structured, and converted into numerical representations, making them suitable for ML-based cause prediction. The processed data was then used as input for training and evaluating various classification models to identify fatality causes with high accuracy.

### *2.3. Machine learning models*

To predict accident causes based on DGMS accident narratives, multiple supervised machine learning models were implemented. Each model follows a distinct mathematical foundation, ensuring robustness in text classification. Below, we provide a detailed explanation of each model along with its mathematical formulation.

#### *2.3.1 Naïve Bayes Model*

The Naïve Bayes classifier is a probabilistic model based on Bayes’ theorem, assuming that the features (words) used for classification are conditionally independent given the target class (Murty and Devi 2011). Considering an accident narrative consisting of n words, represented as:

$$W = \{ w_1, w_2, w_3 \ldots., w_n\}$$

and a set of m possible accident causes (E):

$$E = \{E_1, E_2, E_3, \ldots., E_m\}$$

the probability of assigning a specific cause $E_m$ to a given text is computed as:

$$P(W) = \frac{P(E_m)P(E_m)}{P(W)}$$

**where:**

- $P(W)$ is the posterior probability of assigning class Em for the accident narrative W,
- $P(E_m)$ is the likelihood of observing words W given class Em,
- $P(E_m)$ is the prior probability of class Em.,
- P(W) is the evidence or marginal probability of words appearing in all documents.

Since accident narratives consist of multiple words, assuming word independence, the likelihood is calculated as:

$$P(E_m) = \prod_{i=1}^{n} P(w_i|E_m)$$

where $P(E_m)$represents the probability of word $w_i$ appearing in narratives belonging to class $E_m$. This is computed as:

$$P(E_m) = \frac{count(E_m) + \alpha.count(w_i)}{count(E_m) + \alpha.N}$$

where:

- $count(E_m)$ is the number of times word $w_i$ appears in class $E_m$ ,
- count ($w_i$) is the total occurrences of word $w_i$ in the corpus,
- count ($E_m$)is the number of narratives assigned to class $E_m$,
- N is the total number of narratives,
- α is the smoothing constant (Laplace smoothing) to prevent zero probabilities.

In this study, **α is set to 0.05** to ensure a low level of smoothing.

### *2.3.2 Logistic Regression Model*

Logistic regression is a discriminative model that predicts the probability of a fatality cause belonging to a particular class Em using the sigmoid function (Hosmer et al., 2013):

$$P(W) = \frac{1}{1 + e^{-(w^T X + b)}}$$

where:

- w is the weight vector,
- X is the feature vector representing the accident narrative,
- b is the bias term,
- e is Euler's number.

For multi-class classification, the softmax function generalizes logistic regression as follows:

$$P(X) = \frac{e^{w_m^t X}}{\sum_{j=1}^{M} e^{e^{w_j^t X}}}$$

where M is the total number of accident cause classes. The model is optimized using cross-entropy loss.

2.3.3 Random Forest Model

Random Forest is an ensemble-based learning algorithm that constructs multiple decision trees and aggregates their predictions (Breiman, 2001). Considering an input feature vector X, each decision tree predicts an accident cause:

$$h_m(X)$$

where m represents a specific tree in the forest. The final predicted class "y" is determined by majority voting:

$$y = arg\ arg\ max \sum_{m=1}^{M} I(h_m(X) = E_k)$$

where:

- I(·) is an indicator function that counts the number of times class Ck is predicted,
- M is the total number of trees.

This approach reduces overfitting compared to single decision trees and improves classification accuracy

### *2.3.3 Support Vector Machines*

SVM is a margin-based classifier that finds the optimal hyperplane to separate different accident cause categories. The decision boundary is defined as:

$$w^T X + b = 0$$

where w and b are the weight vector and bias term, respectively. The goal is to maximize the margin between the closest support vectors:

$$\frac{2}{||w||}$$

For non-linearly separable data, a kernel function K (Xi, Xj) transforms data into a higher-dimensional space:

$$K(X_i, X_j) = e^{-\gamma||X_i - X_j||^2}$$

where γ is a hyperparameter.

Each of these models contributes unique advantages for text classification in mining accident reports. Naïve Bayes and Logistic Regression provide fast, interpretable results, Random Forest enhances generalization, SVM is effective for high-dimensional data. By comparing these models, the study aims to identify the most accurate and efficient method for predicting accident causes from narrative descriptions.

### *2.4 Model training and evaluation*

The dataset was partitioned (80% training, 20% testing) using stratified sampling to maintain class distribution consistency. Models were trained using Bag of Words (BoW) with unigram and bigram representations. Hyperparameter tuning was performed via grid search, except for Naïve Bayes, which follows a probabilistic approach. Logistic Regression models were optimized by varying regularization strength (0.01–10), Random Forest was tuned for tree count (100–500) and depth (10–50), and SVM was tested with different kernels (linear, RBF, polynomial) adjusting hyperparameters (0.1–100).

We evaluated the models using Accuracy, Precision, Recall, and F1-Score, with a special focus on Recall to account for class imbalance. Relying on Accuracy alone can be misleading, as it tends to favor the majority class, potentially overlooking critical but less frequent accident types. To improve reliability, we developed a semi-automated framework where high-confidence predictions were classified automatically, while low-confidence cases were flagged for manual review. This approach strikes a balance between automation and expert oversight, ensuring both classification accuracy and reduced manual effort in mining safety interventions.

## 4. Results and Discussion

The classification models were evaluated using Accuracy, Precision, Recall, and F1-Score to assess their effectiveness in identifying accident causes from DGMS reports. Among the models tested, Logistic Regression with Adjusted Weights outperformed others, achieving an accuracy of 84%, recall of 85%, and F1-score of 82%, making it the most robust classifier for mining fatality narratives. Naïve Bayes and Support Vector Machines (SVM) also demonstrated stable performance, with recall values of 81% and 76%, respectively. Random Forest, in contrast, exhibited lower recall (63%), indicating challenges in correctly classifying underrepresented accident categories.

Table 1. BOW (uni-gram and bi-gram)

| Models | Accuracy | Precision | Recall | F-1 Score |
|---|---|---|---|---|
| Naïve Bayes | 0.79 | 0.76 | 0.81 | 0.77 |
| Logistic Regression | 0.79 | 0.80 | 0.80 | 0.77 |
| Logistic Regression (Weights Adjusted) | **0.84** | **0.83** | **0.85** | **0.82** |
| Support Vector Machines | 0.79 | 0.80 | 0.76 | 0.76 |
| Random Forest | 0.70 | 0.78 | 0.63 | 0.64 |

To balance classification accuracy and automation efficiency, a semi-automated classification framework was implemented, wherein predictions exceeding a predefined probability threshold were auto-coded, while lower-confidence cases were flagged for manual review. The impact of this threshold on classification performance is illustrated in Figure 1, which demonstrates that as the confidence threshold increases, precision improves, but recall declines. This trade-off occurs because higher thresholds ensure that only the most confident predictions are auto-coded, reducing misclassifications, whereas, lower thresholds maximize auto-coding efficiency but increase the likelihood of errors. The graph further indicates that maintaining a threshold of around 0.4 to 0.5 provides an optimal balance, ensuring high classification accuracy while keeping manual review efforts minimal.

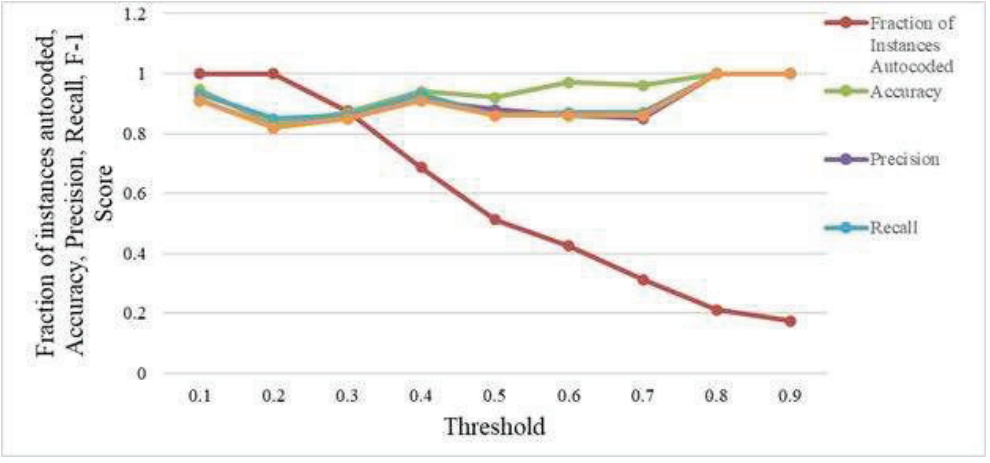


Figure 1. Effect of classification threshold on model performance and fraction of autocoded instances

The misclassification trends observed in the analysis reveal that accident causes related to "Ground Movement" and "Falls" (Other than Falls of Ground) were frequently confused, likely due to overlapping terminology and similarities in textual descriptions. This suggests that further refinement of feature extraction methods, such as context-aware word embeddings, may enhance classification accuracy. Additionally, while BoW with bigram representation improved recall, it also introduced some noise, leading to minor reductions in precision. Alternate embeddings such as TF-IDF, Word2Vec, BERT could enhance classification performance and future work could explore these advanced embeddings to assess their impact on accident text classifications and compare their effectiveness with BoW.

Overall, the results demonstrate that machine learning models, particularly Logistic Regression with Adjusted Weights, can effectively classify accident causes, reducing reliance on manual coding and improving efficiency in mining safety analysis. However, implementing a semi-automated approach with a tunable probability threshold is essential to mitigate misclassifications and maintain the reliability of automated classification in real-world applications.

The classification structure introduced in this study possesses significant promise for enhancing safety management in mining. Nevertheless, its practical implementation presents obstacles. A significant challenge is guaranteeing the model's functionality across various mining conditions. Accident reports differ significantly based on the type of mine, reporting methodology, and terminology employed. The reporting style must adhere to established guidelines for this framework to gain widespread acceptance. Obtaining regulatory approval and ensuring smooth integration into the current safety workflows will be crucial for the effective implementation of the framework.

## 5. Conclusion

Mining remains one of the most hazardous industries globally, necessitating efficient and accurate analysis of accident data to improve safety interventions. Traditionally, manual coding of accident reports has been labor-intensive, inconsistent, and prone to human error, highlighting the need for automated solutions. This study demonstrates the feasibility of machine learning-driven classification of fatal mining accident causes, using DGMS reports from India as a case study. Among the models tested, Logistic Regression with Adjusted Weights outperformed others, achieving high recall and F1-score, making it the most suitable for handling imbalanced accident data.

To ensure scalability and reliability, a semi-automated classification framework was implemented, where high-confidence cases were auto-coded, while low-confidence cases were flagged for manual review. This approach balances automation efficiency with expert verification, offering a scalable solution for mining safety agencies worldwide. The findings contribute to the growing field of Natural Language Processing (NLP) applications in occupational safety, emphasizing the potential of machine learning in mining accident analysis. Future research should explore context-aware language models and hybrid NLP techniques to further enhance classification accuracy and extend this framework to global mining datasets, improving data-driven risk mitigation strategies in hazardous industries.

**References**


Abdat, Fazia, Sylvie Leclercq, Xavier Cuny, and Claire Tissot. "Extracting recurrent scenarios from narrative texts using a Bayesian network: application to serious occupational accidents with movement disturbance." *Accident Analysis & Prevention* 70 (2014): 155-166.

Breiman, Leo. "Random forests." *Machine learning* 45 (2001): 5-32.

Das, Souvik, Dhruva Rajesh Khanwelkar, and J. Maiti. "A semi-automated coding scheme for occupational injury data: An approach using Bayesian decision support system." *Expert Systems with Applications* 237 (2024): 121610.

Gupta, Aryan Kumar, Chunduru Geetha Venkata Sai Pardheev, Sinjana Choudhuri, Souvik Das, Ashish Garg, and Jhareswar Maiti. "A novel classification approach based on context connotative network (CCNet): A case of construction site accidents." *Expert Systems with Applications* 202 (2022): 117281.

Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

Liu, Chang, and Shiwu Yang. "Using text mining to establish knowledge graph from accident/incident reports in risk assessment." *Expert Systems with Applications* 207 (2022): 117991.

Lombardi, David A., Simon Matz, Melanye J. Brennan, Gordon S. Smith, and Theodore K. Courtney. "Etiology of work-related electrical injuries: a narrative analysis of workers' compensation claims." *Journal of occupational and environmental hygiene* 6, no. 10 (2009): 612-623.

Murty, M. Narasimha, and V. Susheela Devi. *Pattern recognition: An algorithmic approach*. Springer Science & Business Media, 2011.

Sharma, Amit, and Bibhuti Bhusan Mandal. "Attenuation of mechanical vibration during transmission to human body through mining vehicle seats." *Mining, Metallurgy & Exploration* 38, no. 3 (2021): 1449-1461.

Verma, Abhishek, Sudha Das Khan, J. Maiti, and O. B. Krishna. "Identifying patterns of safety related incidents in a steel plant using association rule mining of incident investigation reports." *Safety science* 70 (2014): 89-98.