minerals

MDPI

Article

Effectiveness of Natural Language Processing Based Machine Learning in Analyzing Incident Narratives at a Mine

Rajive Ganguli [1,*], Preston Miller [2] and Rambabu Pothina [1]

[1] Department of Mining Engineering, University of Utah, Salt Lake City, UT 84112, USA; rambabu.pothina@utah.edu
[2] Teck Red Dog Operations, Anchorage, AK 99503, USA; Preston.Miller@teck.com
* Correspondence: rajive.ganguli@utah.edu

Abstract: To achieve the goal of preventing serious injuries and fatalities, it is important for a mine site to analyze site specific mine safety data. The advances in natural language processing (NLP) create an opportunity to develop machine learning (ML) tools to automate analysis of mine health and safety management systems (HSMS) data without requiring experts at every mine site. As a demonstration, nine random forest (RF) models were developed to classify narratives from the Mine Safety and Health Administration (MSHA) database into nine accident types. MSHA accident categories are quite descriptive and are, thus, a proxy for high level understanding of the incidents. A single model developed to classify narratives into a single category was more effective than a single model that classified narratives into different categories. The developed models were then applied to narratives taken from a mine HSMS (non-MSHA), to classify them into MSHA accident categories. About two thirds of the non-MSHA narratives were automatically classified by the RF models. The automatically classified narratives were then evaluated manually. The evaluation showed an accuracy of 96% for automated classifications. The near perfect classification of non-MSHA narratives by MSHA based machine learning models demonstrates that NLP can be a powerful tool to analyze HSMS data.

Keywords: mine safety and health; accidents; narratives; machine learning; natural language processing; random forest classification

Citation: Ganguli, R.; Miller, P.; Pothina, R. Effectiveness of Natural Language Processing Based Machine Learning in Analyzing Incident Narratives at a Mine. *Minerals* **2021**, *11*, 776. https://doi.org/10.3390/min11070776

Academic Editor: Yosoon Choi

Received: 22 May 2021
Accepted: 14 July 2021
Published: 17 July 2021

1. Introduction

Workers' health and safety is of utmost priority for the sustainability of any industry. Unfortunately, occupational accidents are still reported in high numbers globally. According to the recent estimates published by the International Labour Organization (ILO), 2.78 million workers die from occupational accidents and diseases worldwide [1]. In addition, 374 million workers suffer from non-fatal accidents, and lost work days represent approximately 4% of the world's gross domestic product [2,3]. It is, therefore, not surprising that researchers are constantly investigating factors that impact safety [4,5], or finding innovations and technology to improve safety [6,7].

As to the U.S. mining industry, for years 2016–2019, the National Institute for Occupational Safety and Health (NIOSH), a division of the US Centers for Disease Control and Prevention (CDC) reports 105 fatal accidents and 15,803 non-fatal lost-time injuries [8]. To bring down the rate of serious injuries and fatalities, the industry analyzes incident reports to conduct root cause analysis and identify leading indicators. Unfortunately, as noted by the International Council on Mining and Metals, a global organization of some of the largest mining companies of the world, the vast trove of incident data is not analyzed as much as it could be due to lack of analytics expertise at mine sites [9]. With the advances in natural language processing (NLP), there is now an opportunity to create NLP-based tools to process and analyze such textual data without requiring human experts at the mine site.

Natural language processing (NLP) has been explored as a tool to analyze safety reports since the 1990s [10,11]. This paper, intended for a mining industry audience, presents in this section, a brief history of NLP and its use in analyzing safety reports. NLP is the automated ability to extract useful information out of written or spoken words of a language. Exploring its application to safety is logical, as safety reports are valuable information. If causation and associated details can be automatically extracted from the safety reports, NLP can be used to quickly gain insight into safety incidents from historical reports that are filed away in the safety management databases. Additionally, with smartphone-based work site observations apps becoming popular, NLP tools can be useful in providing real time insights as incidents and observations are reported in real time. For example, in a confidential project, one of the authors of this paper advised an industrial site about a hazardous practice at the operation using an NLP analysis of data collected using a smartphone-based application. This hazard became apparent after evaluating the data because several employees had noted the practice in their worksite observations.

The efforts to apply NLP to extract causation from safety reports received a major boost when the Pacific Northwest National Laboratory (PNNL) put together a large team in the early 2000s to apply NLP and analyze aviation safety reports from the National Aeronautics and Space Administration's (NASA) aviation safety program [12]. The "meaning" of a sentence depends not just on the words, but also on the context. Therefore, PNNL used a variety of human experts to develop algorithms to extract human performance factors (HPF) from report narratives. HPF definitions were adopted from NASA [13]. The PNNL approach consisted of artificial intelligence (AI) after the text was preprocessed using linguistic rules. The linguistic rules, developed by human experts, considered specific phrases and sentence structures common in aviation reports. When automated, these rules were able to identify causes of safety incidents on par with human experts. The PNNL team, however, noted the reliance of the algorithms on human experts with domain-specific knowledge.

New developments have reduced human involvement in text analysis [14]. These developments include identifying linguistic features such as parts of speech, word dependencies, and lemmas. A million-sentence database (or "corpus" to use NLP terminology) may only contain 50,000 unique words once words such as 'buy' and 'bought' (one is a lemma of the other) are compressed into one; though that is also a choice for the human expert. After vectorization, each sentence in the database is a vector of length 50,000, with most elements being zero (a twelve-word sentence will only have ones in twelve places). When the relative order of words in a sentence is taken into account, common phrases can be identified easily. Thus, after preprocessing with NLP techniques, classical statistics and machine learning techniques can be applied to classify text. Baker et al., 2020 [15] used a variety of NLP and machine learning techniques to classify incident reports and predict safety outcomes in the construction industry. Tixier et al., 2016 developed a rule based NLP algorithm that depends on a library of accident related keywords to extract precursors and outcomes from unstructured injury reports in the construction industry [16]. In another study that was conducted on narratives from Aviation Safety Reporting System (ASRS), NLP-based text preprocessing techniques along with k-means clustering classification were used to identify various safety events of interest [17]. Baillargeon et al., 2021 [18] used NLP and machine learning techniques to extract features of importance to the insurance industry from public domain highway accident data. In an analysis conducted on infraction history of certain mine categories, ML-based classification and regression tree (CART) and random forest (RF) models were used on Mine Safety and Health Administration (MSHA) database narratives in predicting the likely occurrence of serious injuries in near future (the following 12-month period) [19].

The application of NLP-based machine learning to mining industry safety data is relatively new. Yedla et al., 2020 [20] used the public domain (MSHA) database to test the utility of narratives in predicting accident attributes. They found that vectorized forms of narratives could improve the predictability of factors such as days away from work.

Other researchers used NLP to analyze fatality reports in the MSHA database [21]. Using co-occurrence matrices for key phrases, they were able to identify some of the common causes of accidents for specific equipment.

2. Importance of this Paper

In safety-related research, it is typical to demonstrate NLP and machine learning capabilities on public domain databases. Models are first developed on a public domain database, after which its capabilities are demonstrated on an independent subset of the same database. Since modeling and subsequent demonstration of model capabilities happen on the same dataset, there is no certainty that these approaches or models would be effective on databases created by other sources. For example, every entry in an MSHA database is made by a federal employee. Would a federal employee describe an incident the same way as a mining company employee? If yes, then there exists a specific language for mine safety that is shared by safety professionals. This 'language', if it exists, can be leveraged to make NLP-based machine learning of mine safety data very effective.

This paper advances the use and application of NLP to analyze mine safety incident reports by demonstrating that machine learning models developed on public domain mine safety databases can be applied effectively on private sector safety datasets. Therefore, it demonstrates that there is a language of safety that spans organizations. Furthermore, this paper identifies key attributes of specific categories of incidents. This knowledge can be used to improve algorithms and/or understand their performance.

More generally, the paper advances the field of mine safety research. Currently, data-mining-based mine safety researchers focus only on categorical or numerical data. Therefore, gained insights are limited to statistical characterization of data (such as average age, or work experience) or models based on these data [4]. If narratives are available with incident data (as they often are), this paper will encourage researchers to evaluate them to glean more insights into the underlying causes.

3. Research Methodology

3.1. MSHA Accident Database

The MSHA accident database [22] has 57 fields used to describe safety incidents including meta-data (mine identification, date of incident, etc.), narrative description of the incident, and various attributes of the incidents. Some of the data is categorical such as body part injured and accident type. More than eighty-one thousand (81,298) records spanning the years 2011 to early 2021 were used in this research. Any operating mine in the United States that had a reportable injury is in the database. Thus, the database reflects many types of mines, jobs, and accidents.

Accidents are classified in the database as belonging to one of 45 accident types. Examples include "Absorption of radiations, caustics, toxic and noxious substances", "Caught in, under or between a moving and a stationary object", and "Over-exertion in wielding or throwing objects". Looking at these definitions, it appears that MSHA defined them to almost answer the question "What happened?" Thus, the category is simply the high level human summary of the narrative, i.e., the category is the "meaning" of the narrative. In this paper, the MSHA accident type is considered a proxy for the meaning of the narrative. Narratives are typically five sentences or less.

3.2. Random Forest Classifier

The random forest (RF) technique was used to classify the narratives based on accident types. Random forests are simply a group of decision trees. Though described here briefly, those unfamiliar with decision trees are referred to Mitchell, 1997 [23], a good textbook on the topic and the source for the description below. A decision tree is essentially a series of yes or no questions applied to a particular column ("feature") of the input data. The decision from the question (for example, miner experience > 10, where miner experience is a feature in the data set) segments the data. Each question is, thus, a "boundary" splitting

the data into two subsets of different sizes. The segmented data may be further segmented by applying another boundary, though the next boundary may be on another feature. Applying several boundaries one after the other results in numerous small subsets of data, with data between boundaries ideally belonging to a single category. The maximum number of decision trees applied in the longest pathway is called the "tree depth". The method works by applying the sequence of boundaries to a sample, with the final boundary determining its class. Note that while one boundary (also called "node") makes the final decision on the class for one sample, some other boundary may make the decision for another sample. It all depends on the path taken by a particular sample as it travels through the tree. When the final boundary does not result in a unanimous class, the most popular class in the subset is used as the final decision of the class.

Boundaries are set to minimize the error on either side of the boundaries. The combination of a given data set and given boundary criteria will always result in a specific tree. In an RF, a decision tree is formed by randomly selecting (with replacement) the data. Thus, while a traditional decision tree will use the entire modeling subset for forming the tree, a decision tree in an RF will use the same amount of data, but with some samples occurring multiple times, and some not occurring at all. Thus, the same data set can yield multiple trees. In the RF technique, multiple trees formed with a random selection of data are used to classify the data. One can then use any method of choice to combine predictions from the different trees. This method of using a group of trees is superior to using a single decision tree.

In this paper, an RF classifier was applied to model the relationship between a narrative and its accident type. A non-MSHA database would contain narratives, but not any of the other fields populated by MSHA staff. Since the goal of the project is to test it on non-MSHA data, no other field in the database was used to strengthen the model. Half of the records were randomly selected to develop the model. It was tested on the remaining half of the records to evaluate its performance on the MSHA data. In the final step, the model was tested on non-MSHA data. There is no standard for what proportion of data to use for training and testing subsets, though it is expected that the subsets be similar [24]. A 50–50 split is a common practice [25,26]. RF models were developed using the function RandomForestClassifier () in the SCIKIT-LEARN [27] toolkit. As is common practice in machine learning [28], the authors did not code the RF but used a popular tool instead.

Modeling starts by making a list of non-trivial words in the narratives. As is typical in NLP, the narratives were pre-processed before the list of non-trivial words is made. Pre-processing consisted of:

- Changing case to lower case.
- Removal of specific words: This consisted of the removal of acronyms common in MSHA databases, and a custom list of "stop words". Stop words are words such as stray characters, punctuation marks, and common words that may not add value. These are available from several toolkits. The stop words list available from NLTK [29] was modified and used in this paper.
- Lemmatizing: This was done using the lemmatizer in the spacy [30] toolkit. Lemmatizing is the grouping of similar words, or rather, identifying the foundational word. This is done so that related words are not considered separately. For example, consider the two sentences, "He was pushing a cart when he got hurt" and "He got hurt as he pushed a cart". The lemmatizer would provide "push" as a lemma for both pushing and pushed, and push would replace pushed and pushing in the narrative.

The combined length of all narratives was 1.72 million words, consisting of 31,995 unique words or "features". The list of unique features is called the vocabulary. The input data set is then prepared by selecting the top 300 most frequently occurring words ("max features"). Essentially, the vocabulary is cut from its full length to just the words occurring most frequently. These words are used to vectorize each narrative such that each narrative is represented as a vector of size 300. The value at a given location in the vector would

represent the number of occurrences of that word in that narrative. The top 5 words were: fall, right, left, back, and cause.

The output for the narrative consisted of a 1 or a 0, indicating whether it belonged ("1") to a particular category of accident or not ("0"). "Max features" is a parameter in RF modeling, and was set to 300 after trial and error exercises. Similarly, the number of trees ("n_estimators") was set to 100. Another parameter is "max_depth" (maximum depth of tree). This parameter was not set. Whenever a parameter is not specified, the tool uses default values. In the default setting for tree depth, data is continually segmented till the final group is all from the same class. According to the user guide of the tool, the main parameters are the number of trees, and max features. The rest of the parameters were not set, i.e., default values were used. The interested reader can visit the provided links for technical details about the toolkits in the footnotes, including the default values. The tool combines the outputs of the various trees by averaging them to obtain the final classification.

Among the 45 accident types are some whose names start with the same phrase. For example, there are four over-exertion (OE) types, all of which start with the phrase over-exertion. They are (verbatim): Over-exertion in lifting objects, over-exertion in pulling or pushing objects, over-exertion in wielding or throwing objects, and over-exertion NEC. Accident categories whose names begin with the same phrase are considered to belong to the same "type group", with the phrase defining the grouping.

NEC stands for "not elsewhere classified," and is used within some type groups. When it exists, it is often the largest sub-group as it is for everything that is not easily defined. There are 11 types that start with "Fall", including two that start with "Fall to". Five types start with "Caught in". Six start with "Struck by". These accident type groups contain 26 of the 45 accident types, but 86% of all incidents (35,170 out of 81,298). Table 1 shows the four type groups that were modeled in this paper. Separate models were developed for some of the sub-groups to get an understanding of these narrowly defined accidents. These were:

- Over-exertion in lifting objects (OEL).
- Over-exertion in pulling or pushing objects (OEP).
- Fall to the walkway or working surface (FWW).
- Caught in, under or between a moving and a stationary object (CIMS), and
- Struck by flying object (SFO).

Table 1. The four type groups of accidents modeled in the paper.

Type Group: Caught in	Type Group: Fall	Type Group: Over-Exertion	Type Group: Struck
Caught in, under, or between a moving and a stationary object	Fall down raise, shaft or manway	Over-exertion in lifting objects	Struck by concussion
Caught in, under, or between collapsing material or buildings	Fall down stairs	Over-exertion in pulling or pushing objects	Struck by falling object
Caught in, under, or between NEC	Fall from headframe, derrick, or tower	Over-exertion in wielding or throwing objects	Struck by flying object
Caught in, under, or between running or meshing objects	Fall from ladders	Over-exertion NEC	Struck by powered moving object
Caught in, under, or between two or more moving objects	Fall from machine		Struck by rolling or sliding object
	Fall from piled material		Struck by... NEC
	Fall from scaffolds, walkways, platforms		
	Fall on same level, NEC		
	Fall onto or against objects		
	Fall to lower level, NEC		
	Fall to the walkway or working surface		

Thus, a total of nine RF models were developed; four for the four type groups, and five for the specific types. Table 2 shows the characterization of the training and testing subsets

that went into developing the models. It is apparent that each category was represented about the same in the two subsets.

Table 2. Various accident categories in the training and testing subsets. Each subset has 40,649 samples.

Subset	Type Group: OE	Type Group: Caught in	Type Group: Struck by	Type Group: Fall	OEP	OEL	FWW	CIMS	SFO
Training	8909	4563	10,216	4802	1290	2838	2130	3337	1586
Testing	8979	4524	10,226	4926	1275	2961	2130	3310	1590

In classification exercises, it is common to develop a single model to classify a data set into multiple categories, rather than develop models for each category individually. The reason for developing nine models instead of one is discussed in the next section.

4. Results

4.1. Performance within MSHA Data

Table 3 shows a summary of the modeling within the MSHA test set. To understand the table, consider the OE type group. Of the 40,649 records in the test set, 8979 records were from this type. The success of an RF model can be determined by identifying the OE type as OE type and/or by classifying a non-OE type (31,670 records) as not belonging to OE. This is shown below through a simple computation.

Table 3. Results of RF models in the MSHA test set.

Metrics	Type Group: OE	Type Group: Caught in	Type Group: Struck by	Type Group: Fall	OEP	OEL	FWW	CIMS	SFO
Records from Category	8979	4524	10,226	4926	1275	2961	2130	3310	1590
Overall Success % from Category	92%	96%	90%	95%	98%	96%	96%	95%	97%
Accurately Predicted	81%	71%	75%	71%	37%	59%	34%	55%	25%
False Positive	4%	1%	5%	2%	<1%	<1%	<1%	2%	<1%

- Total samples (n_samples): 40,649
- Total samples in target category (n_target): 8979
- Total samples in other categories (n_other): n_samples − n_target = 31,670
- Samples from target category predicted accurately (n_target_accurate): 7248
- Samples from other category predicted wrongly as target (false_predicts): 1331
- Samples from other category predicted correctly as other (other_accurate): 31,670 − 1331 = 30,339
- Percentage of targets accurately predicted: 100 × n_target_accurate/n_target = 100 × 7248/8979 = 81%
- False positive rate: false_predicts/n_other = 1331/31,670 = 4%
- Total correct predictions (total_correct): n_target_accurate + other_accurate = 7248 + 30,339 = 37,587
- Overall success rate (%) = 100 × total_correct/_samples = 100 × 37,587/40,649 = 92%

The overall success was 92%, i.e., a very high proportion of narratives were classified correctly as belonging to OE type group, or as not belonging to OE type group. Though it is an indicator of overall success, this type of evaluation is not particularly useful, as classifying a narrative as "not belonging to OE" is not helpful to the user. It is more useful to look at how successful RFs were in correctly identifying narratives from the accident type in question (OE type group in this example). As shown in the table and in the example computation, 81% of these 8918 (7248) were accurately identified. The false positive rate was 4%, i.e., 1331 of the 31,670 non-OE records were identified as OE. The low positive rate

implies that if a narrative was classified as belonging to the OE type group, it was highly likely to belong to that type. The success in the other type groups was lower, and ranged from 71% to 75%, with false positives ranging from 1% to 5%. Thus, one could expect RF to accurately identify about 75% of the narratives in the MSHA database from the four type groups, with a good false positive rate.

The success rate takes a dramatic downturn with the individual models. Only 25% to 59% of narratives belonging to the individual types are correctly classified though with a negligible false positive rate. The negligible false positive implies that when the model classifies the narrative as belonging to a specific category, it is almost guaranteed to be in that category. The low number of records in the individual categories is one part of the explanation of the poor performance, as models would be less powerful if they are trained on fewer records. For example, only about 3% of the records were from the OEP category. This means that 97% of the data seen by the OEP model was not relevant to identifying OEP. An additional explanation is obtained from trigram analysis of the narratives that belong to these accident types. Trigrams explore the sets of three words that occur consecutively the most. Trigram analysis was conducted using the NLTK collocations toolkit.

Table 4 shows the tri-word sequences that occur the most frequently in the OE accident types. They are listed in order of frequency. The overlap between the tri-words is immediately apparent. Back, shoulders, knee, abdomen, and groin are injured most in these types of accidents. The overlap between OEP and OEL would cause accidents to be misclassified as belonging to the other category. This issue is also evident in the Fall accident types (Table 5), where losing balance, slipping, and falling seem to be the major attributes. Even the two types "Caught in" and "Struck by" have some overlap (Table 6). Caught in makes it apparent that it is the fingers that are predominantly injured in this type of accident. SFO highlights that eyes and safety glasses are impacted when someone is struck by a flying object.

Table 4. Results of trigram analysis on OE accident types.

Type Group: OE	OE Lifting	OE Pulling
feel pain back	feel pain back	feel pain back
pain low back	pain low back	feel pain shoulder
feel pain low	feel pain low	feel pain right
feel pain right	feel low back	feel pain low
feel pain shoulder	feel pain shoulder	feel pain left
feel pain left	feel pain right	feel pain groin
feel pain knee	feel pain left	feel pain abdomen

Table 5. Results of trigram analysis on Fall accident types.

Fall	FWW
lose balance fall	lose balance fall
slip fall ground	slip fall right
cause lose balance	slip fall left
foot slip fall	slip fall ground
slip fall backward	cause lose balance
step lose balance	place restrict duty
lose balance cause	slip fall ice

Table 6. Results of trigram analysis on Caught in and Struck by accident types.

Caught in . . .	**CIMS**	**Struck by** . . .	**SFO**
right index finger	right index finger	piece rock fell	wear safety glass
left index finger	left index finger	rock fall strike	safety glass eye
right middle finger	left ring finger	cause laceration require	eye safety glass
left ring finger	right middle finger	left index finder	behind safety glass
right ring finger	right ring finger	strike left hand	go safety glass
left middle finger	pinch index finger	right index finger	safety glass face
pinch index finger	left middle finger	wear safety glasses	safety glass left

The success rate for classification was dramatically lower when a single RF model was developed to classify the narratives into separate categories. OEP, OEL, FWW, CIMS, SFO had success rates of only 23%, 33%, 19%, 29%, and 17% respectively compared to 37%, 59%, 34%, 55%, 25% respectively. Multiple models for multiple categories would require that multiple models be applied to the same data, resulting in multiple predictions of category. It would be possible then for a particular narrative to be categorized differently by the different models. In such situations, one could determine the similarity between the narrative and the narratives from the multiple categories in the training set to resolve the conflicting classifications. The features (words) of the category within the training set are the foundation behind the model for the category. For example, the words in the "Struck by" category in the training set play a key role in what RF trees are formed in the "Struck by" model. Thus, when a test narrative is classified as "Struck by" by one model, and "Caught in" by another, one could find the similarity between words in the test narrative, and the words in the two categories of the training data, "Struck by" and "Caught in", to resolve the conflict. This is demonstrated in the next section.

4.2. Performance on Non-MSHA Data

The nine RF models were applied to data from a surface metallic mine in the United States that partnered in this project. The data consisted of narratives that described various safety incidents. Injury severity ranged from very minor incidents to lost time accidents. Narratives were typically longer than MSHA narratives (about twice the length), and formats were sometimes different (such as using a bulleted list). They usually had more details about the incident. The narratives were written by a staff member from the safety department. Narratives from the 119 unique incidents logged in 2019 and 2020 were analyzed. Some narratives were duplicated in the database. Duplicates of narratives were ignored. Each model was applied to the 119 narratives separately.

The RF models classified 76 out of the 119 narratives (Table 7) with a high degree of success. 17 narratives were classified by multiple models, but not misclassified (explained later). Forty-three (43) narratives were ignored by all nine models, i.e., they were not classified as belonging to a particular category. The classifications were manually evaluated by the authors to see if they would match the MSHA Accident Types. In many cases, the MSHA database contained an accident that was not only similar to the narrative being manually evaluated but was also classified into the same accident type as the narrative in question. Therefore, the manual validation was easy. A narrative was deemed as accurately classified if it was also classified as such by the authors. The 43 narratives that were not classified by any of the nine models could possibly belong to one of the 19 MSHA accident types not modeled in this paper. The overall success rate was 96%.

Table 7. Performance of RF models on non-MSHA data.

Metrics	**OE**	**OEP**	**OEL**	**Fall**	**FWW**	**Caught in**	**CIMS**	**Struck by**	**SFO**	**Overall**
Number	26	1	4	14	3	9	7	27	2	93
Validation	85%	100%	100%	100%	100%	100%	100%	100%	100%	96%

The OE category is quite broad and, therefore, one would expect some narratives to be wrongly classified as OE. Therefore, it is not surprising that 4 out of the 26 classified as OE did not belong in that category. One narrative involved an employee who had a pre-existing soreness in the wrist. The 'incident' was simply the employee reporting to the clinic. Two incidents involved employees backing into or walking into a wall or object while working. The fourth incident involved chafing of the calves from new boots. Some of these incidents would perhaps have been also classified differently had models been developed for the other accident types.

Table 8 shows examples of some of the narratives and the automated classifications. Examples are shown for the narrowest categories as they would normally be the most challenging to identify. Table 9 shows how the overlapping occurred in the 17 narratives. Three narratives were classified as both Fall and FWW, while seven were categorized as both "Caught in" and CIMS. Since nine models were used in parallel, it was possible for each narrative to be categorized into nine different categories. Yet, no narrative was categorized as belonging to three or more different categories. Except for one, these overlaps should be expected. For example, OEL is a subset of OE. Therefore, a narrative classified as OEL by the OEL model is expected to be also classified as OE by the OE model. The overlap between a type group and one of its sub-type is a confirmation that models are working properly. It is good that there was no overlap between OEL and OEP. The overlap between "Caught in" and "Struck by" was surprising as they are different categories. The narrative that was classified as both "Caught in" and "Struck by" is (verbatim): "while installing a new motor/pump assy. using portable a cherry picker, the cherry picker tipped over and the assembly caught the employee leg and ankle between the piping and the motor assembly." Tools and equipment that tip over and cause injury have been reported in the "Struck by" category in the MSHA database. A limb caught in between two objects is reported in the "Caught in" category in the MSHA database. Thus, the RF models were correct in their classification of the narrative. However, the overlap in classification presents a good opportunity to demonstrate how one could use "similarity scores" to resolve the overlap. The steps of the process, to resolve conflicting classifications of "Caught in" and "Struck by" are:

1. Consider the non-trivial words in the problem narrative: "instal new motor/pump assy.use portable cherry picker cherry picker tip assembly catch leg ankle piping motor assembly". This list of non-trivial words was obtained after pre-processing. Note that "instal" is not a typo but a product of the lemmatizer.
2. Consider the word frequencies of the training set when the accident category was "Caught in". There were 4894 unique words in the 4563 narratives from that category. The top 5 words were finger (0.036), hand (0.021), right (0.015), pinch (0.0148), and catch (0.0143) with the number in parenthesis indicating the proportion of times the word occurred within that category of narratives.
3. Similarly, consider the list of words in the "Struck by" category. There were 7758 unique words in the 10,216 narratives. The top 5 words were strike (0.019), left (0.014), right (0.014), cut (0.013), and fall (0.012).
4. Now obtain the similarity score between the narrative and a category by weighing each word of the narrative by the proportion of occurrence within the category. This makes sense as the frequency of occurrence of a word in a category is an indicator of its importance to the category. For example, if "leg" gets "Caught in" less frequently than "Struck by", it will occur in lower proportion in "Caught in" than in "Struck by". The words in the "Struck by" list occurred 16 times in the narrative for a total similarity score of 0.0168. There are 13 unique words in the 16 occurrences. The top 3 contributors were "leg", "/" and "install" with scores of 0.004, 0.0027, and 0.0023 for each occurrence in the narrative.
5. Similarly, obtain the total similarity score for all the other categories. For "Caught in", the score is 0.0338. The top 3 contributors in the narrative were "catch" (0.014), "tip" (0.0045), and "install". It is insightful to note how much more "catch" contributed as

a top word than "leg" did as a top word. Clearly, "catch" is a bigger determiner of "Caught in" than leg is of "Struck by".

6. The decision as to which category the narrative belongs is the one with the highest similarity score. In this case, the narrative is deemed to be of the category "Caught in".

Table 8. Examples from the partner mine HSMS, and the automated classifications. Narratives are shown verbatim, but some text has been deleted (identified by . . .) to not disclose sensitive information.

Accident Type	Narrative
OEP	Employee pulled a heavy bag with helper and felt sharp pain in mid back area
OEL	. . . Employee strained lumbar back while carrying a portable generator...
FWW	The operator . . . . began the pre-shift walk around, but did not notice the slick ground conditions. The operator was not wearing any type of traction device, and slipped and landed on their side/back.
SFO	.. While doing so a small piece of shrapnel from shank guard struck mechanic in the left inner thigh and was lodged into skin . . .
CIMS	While moving a turbo charger rotor, employee pinched finger between the rotor shaft and the crate . . .

Table 9. Counts of overlapping accident types.

Overlapping Types	Count
Fall, FWW	3
Caught in, Struck by	1
OEL, OE	3
OEP, OE	1
Struck by, SFO	2
Caught in, CIMS	7

5. Discussion

Two thirds of the narratives in the partner database could be successfully classified (96% accuracy) without any human intervention. The narratives that are not automatically classified could belong to categories not modeled in this paper. At this time, they were not manually analyzed to determine their nature. The nearly absent overlap in predictions for distinct accident types is encouraging as that allows the multiple-model-for-multiple-category approach to work. That is further strengthened by the low false positive rates for the distinct categories, i.e., when a particular model for a distinct category (say OEP) claims that a narrative belongs to that category, the classification is most likely valid. The similarity score approach is presented to resolve cases where a narrative is classified into multiple categories due to the use of multiple models.

The classifications done in the paper were not an empty computational exercise thanks to how MSHA classified the accidents. An increase in narratives being classified as SFO would tell management that foreign matter was entering the eyes of their employees. This is the same as humans reading the narratives, understanding them, and reaching that conclusion. Thus, in some sense, the RF models picked up what the narratives "meant". The high classification success rate also meant that there were specific ways safety professionals describe incidents and that NLP tools can extract that language.

These tools have excellent applicability to help the mining industry reach the industry goal of preventing serious injury and fatalities. On noting an increase in SFO classifications, management can deploy eye protection related interventions. An increase in OEL incidents could result in more training about safe lifting. The safety "department" in most mines means a single person with no mandate or expertise to analyze data. These types of tools can assist mines to analyze data without human intervention. As mines deploy smartphone-based apps to collect employee reports on worksites, the volume of information will

explode. However, these tools will help mines process that data and identify hazards before they become incidents.

The detection rate for the narrowest of categories needs to be improved. Improving this would be the most logical next step for this research. A reason why NLP tools were not always effective may be how incidents are described in the narratives. A limitation of the approach is that it is dependent on the terminology and the writing style. For example, "roof bolter" related incidents may not be detected by NLP in narratives when the writer uses the term "pinner" to refer to a bolter (though the diligent NLP developer would notice the frequent occurrence of "pinner" in narratives involving "roof"). "Pinner" is a common term for roof bolters in certain parts of the US. Terminology aside, writing style can vary dramatically depending on the region and the English language abilities of the writer. Considering all of these, the MSHA database may not be a great resource for English based NLP tools in other parts of the world. Regardless, organizations (or nations) developing their own NLP tools could provide training to standardize the writing of safety narratives, so that data is generated to assist automation.

The extremely low false positive rate for the narrowest accident types is a wonderful argument for considering these tools. The overall false positive rate across all accident types is quite low, which is good.

6. Conclusions

Natural language processing based random forest models were developed to classify narratives in the MSHA database depending on accident types. Nine models were developed. Four of the models, i.e., Over-exertion, Fall, "Caught in" and "Struck by", looked at type groups, i.e., groups of particular accident types. Five models looked at specific accident types within these broad groups. They were: Over-exertion in lifting objects, Over-exertion in pulling or pushing objects, Fall to the walkway or working surface, "Caught in", under or between a moving and a stationary object, and Struck by flying object. All models had high overall success rates (typically 95% or higher) in classification on MSHA data when considering both false positive and false negative rates. The success in detecting an accident type within a narrative was higher for type groups (71–81%) than for individual categories (25–59%). Detection was done with low false positive rates for type groups (1–5%), and extremely low false positive rate (<1%) for individual categories.

When a single model was developed to classify narratives into multiple categories, it did not perform as well as when a separate model was developed for each category. A similarity score based method was developed to resolve situations where a particular narrative may be classified differently according to different models.

When applied to non-MSHA data, the developed models were successful in classifying about two-thirds of the narratives in a non-MSHA database with 96% accuracy. The narratives that are not classified by the models could belong to accident types not modeled in this paper. In classifying the non-MSHA narratives with near perfect accuracy, the paper demonstrates the utility of NLP-based machine learning in mine safety research. It also demonstrates that there exists a language for mine safety, as models developed on narratives written by MSHA personnel apply to narratives written by non-MSHA professionals. They also demonstrate that natural language processing tools can help understand this language automatically.

Author Contributions: Conceptualization, R.G.; data curation, P.M. and R.P.; formal analysis, R.G., P.M., and R.P.; funding acquisition, R.G.; investigation, R.G., P.M., and R.P.; methodology, R.G., P.M., and R.P.; validation, R.G., P.M., and R.P.; visualization, R.G. and P.M.; writing—original draft, R.G.; writing—review & editing, P.M. and R.P. All authors have read and agreed to the published version of the manuscript.

Funding: This project was partially supported by the National Institute of Occupational Safety and Health.

Data Availability Statement: Not Applicable.

Acknowledgments: This project was partially supported by the National Institute of Occupational Safety and Health. Their support is gratefully acknowledged.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. ILO. Safety and Health at the Heart of the Future of Work: Building on 100 Years of Experience. International Labour Organization, April 2019 Issue. Available online: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/publication/wcms_686645.pdf (accessed on 10 April 2021).
2. Hämäläinen, P.; Takala, J.; Boon, K.T. Global estimates of occupational accidents and work-related illnesses. In Proceedings of the XXI World Congress on Safety and Health at Work, Marina Bay Sands, Singapore, Workplace Safety and Health Institute, Marina Bay Sands, Singapore, 3–6 September 2017.
3. Takala, J.; Hämäläinen, P.; Saarela, K.; Yun, L.; Manickam, K.; Jin, T.; Heng, P.; Tjong, C.; Kheng, L.; Lim, S.; et al. Global estimates of the burden of injury and illness at work in 2012. *J. Occup. Environ. Hyg.* **2014**, *11*, 326–337. [CrossRef] [PubMed]
4. Jiskani, I.M.; Cai, Q.; Zhou, W. Distinctive model of mine safety for sustainable mining in Pakistan. *Min. Metall. Explor.* **2020**, *37*, 1023–1037. [CrossRef]
5. Talebi, E.; Rogers, W.P.; Morgan, T.; Drews, F.A. Modeling Mine Workforce Fatigue: Finding Leading Indicators of Fatigue in Operational Data Sets. *Minerals* **2021**, *11*, 621. [CrossRef]
6. Basu, A.J.; Kumar, U. Innovation and technology driven sustainability performance management framework (ITSPM) for the mining and minerals sector. *Int. J. Surf. Min. Reclam. Environ.* **2004**, *18*, 135–149. [CrossRef]
7. Aznar-Sáncheza, J.A.; Velasco-Muñoza, J.F.; Belmonte-Ureña, L.J.; Manzano-Agugliaro, F. Innovation and technology for sustainable mining activity: A worldwide research assessment. *J. Clean. Prod.* **2019**, *221*, 38–54. [CrossRef]
8. NIOSH. NIOSH Mine and Mine Worker Charts. Available online: https://wwwn.cdc.gov/NIOSH-Mining/MMWC (accessed on 15 June 2021).
9. ICMM. 2012. Available online: http://www.icmm.com/en-gb/guidance/health-safety/indicators-ohs (accessed on 5 April 2021).
10. Garcia, D. COATIS, an NLP system to locate expressions of actions connected by causality links. In *Knowledge Acquisition, Modeling and Management. EKAW 1997*; Plaza, E., Benjamins, R., Eds.; Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence); Springer: Berlin/Heidelberg, Germany, 1997; Volume 1319, pp. 347–352. [CrossRef]
11. Kaplan, R.; Berry-Rogghe, G. Knowledge-based acquisition of causal relationships in text. *Knowl. Acquis.* **1991**, *3*, 317–337. [CrossRef]
12. Posse, C.; Matzke, B.; Anderson, C.; Brothers, A.; Matzke, M.; Ferryman, T. Extracting information from narratives: An application to aviation safety reports. In Proceedings of the IEEE Aerospace Conference Proceedings, Big Sky, MT, USA, 5–12 March 2005. [CrossRef]
13. Maille, N.P.; Ferryman, T.A.; Rosenthal, L.J.; Shafto, M.G.; Statler, I.C. What Happened, and Why: Towards an Understanding of Human Error Based on Automated Analyses of Incident Reports—Volume I. NASA ONERA, 2015. Available online: https://ntrs.nasa.gov/api/citations/20060023334/downloads/20060023334.pdf?attachment=true (accessed on 15 April 2021).
14. Jurafsky, D.; Martin, J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2020. Available online: https://web.stanford.edu/~{}jurafsky/slp3/ed3book_dec302020.pdf (accessed on 18 January 2021).
15. Baker, H.; Hallowell, M.R.; Tixier, A.J.-P. AI-based prediction of independent construction safety outcomes from universal attributes. *Autom. Constr.* **2020**, *118*, 103146. [CrossRef]
16. Tixier, A.J.-P.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. *Autom. Constr.* **2016**, *62*, 45–56. [CrossRef]
17. Rose, R.; Puranik, T.G.; Mavris, D.N. Natural language processing based method for clustering and analysis of aviation safety narratives. *Aerosp.* **2020**, *7*, 143. [CrossRef]
18. Baillargeon, J.T.; Lamontagne, L.; Marceau, E. Mining actuarial risk predictors in accident descriptions using recurrent neural networks. *Risks* **2021**, *9*, 7. [CrossRef]
19. Gernard, J.M. Machine learning classification models for more effective mine safety inspections. In Proceedings of the 2014 International Mechanical Engineering Congress and Exposition IMECE2014, Montreal, QC, Canada, 14–20 November 2014.
20. Yedla, A.; Kakhki, F.D.; Jannesar, A. Predictive modeling for occupational safety outcomes and days away from work analysis in mining operations. *Int. J. Environ. Res. Public Health* **2020**, *17*, 1–17.
21. Raj, V.K.; Tarshizi, E.K. *Advanced Application of Text Analytics in MSHA Metal and Nonmetal Fatality Reports*; SME Annual Meeting & Expo: Phoenix, AZ, USA, 2020.
22. MSHA (Mine Safety and Health Administration). Mine Data Retrieval System. Available online: https://www.msha.gov/mine-data-retrieval-system (accessed on 31 January 2021).
23. Mitchell, T.M. Machine Learning. In *Machine Learning*; McGraw-Hill: New York City, NY, USA, 1997; Volume 45.

24. Ganguli, R.; Dagdelen, K.; Grygiel, E. *Systems Engineering. Mining Engineering Handbook*; Darling, P., Ed.; Society for Mining, Metallurgy and Exploration, Inc.: Englewood, CO, USA, 2011.
25. Röger, C.; Ismayilova, I. Predicting ambient traffic of a vehicle from road abrasion measurements using random forest. In Proceedings of the Conference 13th International Workshop on Computational Transportation Science (IWCTS'20), Seattle, WA, USA, 3 November 2020; pp. 1–7.
26. Weedon, M.; Tsaptsinos, D.; Denholm-Price, J. Random forest explorations for URL classification. In Proceedings of the 2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), London, UK, 19–20 June 2017; Institute of Electrical and Electronics Engineers, Inc.: New York City, NY, USA, 20 June 2017. ISBN 9781509050604. [CrossRef]
27. Scikit-Learn. sklearn.ensemble.RandomForestClassifier. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed on 15 January 2021).
28. Humphries, G.R.W.; Magness, D.R.; Huettmann, F. *Machine Learning for Ecology and Sustainable Natural Resources Management*; Springer: Berlin/Heidelberg, Germany, 2018. [CrossRef]
29. NLTK. Natural Language Tool Kit. Available online: https://www.nltk.org/ (accessed on 15 January 2021).
30. Explosion Spacy. Industrial-Strength Natural Language Processing. Available online: https://spacy.io/ (accessed on 15 January 2021).