Mining Mining Data for Workplace Injuries

Ariana Haidari, Abas Shkembi, Xin Zhang

Introduction

The mining industry is among the most dangerous in the United States [1].Various measures have been implemented over the past century to improve mining-related working conditions, particularly the creation of the Mine Safety and Health Administration (MSHA) in 1977, leading to a dramatic reduction in workplace injuries and fatalities [2]. This trend provides an opportunity to leverage statistical methodology to determine the driving factors of these changes. Additionally, specific occupations within the mining industry are known to have a higher incidence of injury than others. Understanding the differences in these occupations could improve strategies to reduce worker harm.

Investigation of text fields, while data intensive, could provide rich insights into predictors of dangerous job situations and lead to more targeted interventions in improving workplace safety. To do so, text mining techniques, such as natural language processing, can be employed. The aim of this project is to analyze the free-text entries in the Mine Accident, Injuries, and Illness report (MSHA Form 7000-1) from 2000 to 2021 to identify predictors of extreme vs non-extreme injuries. This analysis will identify which jobs are most dangerous.

Methods

Data Source MSHA, under the United States Department of Labor, maintains a database of all mining-related accidents from 1983. Only accidents since January 1, 2000 are currently publicly available on MSHA's Data Retrieval System (https://www.msha.gov/mine-data-retrieval-system)[3]. For each investigation, there are free-text fields that include a narrative of the accident and worker's occupation.

Table 1: Degrees of injuries in MSHA dataset

Degree of injury	N	Degree of injury	N
Days Away From Work Only	80138	Perm Tot Or Perm Prtl Disablty	2213
No Dys Awy Frm Wrk,No Rstr Act	64566	All Other Cases (Incl 1st Aid)	1925
Days Restricted Activity Only	38192	Injuries Due To Natural Causes	1281
Accident Only	28138	Fatality	1052
Dys Awy Frm Wrk & Restrctd Act	18009	No Value Found	910
Occupatnal Illness Not Deg 1-6	9461	Injuries Involvng Nonemployees	510

The original 2000-2021 data has 246,395 entries, with the degree of injury coded into 10 categories (Table 1). Accidents only (n = 28,138 incidents), injuries involving natural causes (n = 1,281 incidents), and injuries without a category (n = 910 incidents) were dropped from analysis. "Extreme injuries" are defined as injuries that resulted in "fatality" or "permanent total/partial disability" categories. Non-extreme injuries are defined as all other injuries.

The final matrix includes 216,066 rows and 426 columns, with each row representing an injury case and each column is an individual occupation descriptor. These descriptors can be considered tokens. A Naive Bayes classifier was used to identify the indicators of "extreme" and "non-extreme" accidents from job descriptions. Naive Bayes classifiers are a family of simple probabilistic classifiers, which assumes all words (or tokens) occur independently. With the term matrix created above, we calculated the likelihood of each token's occurrence in extreme injuries and non-extreme injuries with the following formula:

$$P(token_i|extreme) = \frac{\text{occurrence of } token_i \text{ in extreme injuries}}{\text{total words in extreme injuries}}$$

$$P(token_i|extreme) = \frac{\text{occurrence of } token_i \text{ in non-extreme injuries}}{\text{total words in non-extreme injuries}}$$

In cases where a token never appears in either type of injuries, we use Laplace smoothing method that adds extra occurrence to each token. Because extreme injuries are much more rare than non-extreme injuries, adding the same extra value will result in a bigger increase of likelihood in extreme events than in non-extreme events. To adjust for that effect, we apply weights to the extra values.

$$P(token_i|extreme) = (N_{token_i} + 1 \times N_{token})/(N_{extreme} + N_{token})$$

$$P(token_i|non\text{-}extreme) = (N_{token_i} + k \times N_{token})/(N_{non\text{-}extreme} + k \times N_{token})$$

where $k = N_{non\text{-}extreme}/N_{extreme}$, N is the number of occurrences of words. After calculating the likelihood of each token's occurrence in different degrees of injuries, we calculate the likelihood ratio (LR) comparing the same word's occurrences in extreme injuries versus non-extreme injuries.

$$LR = P(token_i|extreme)/P(token_i|non\text{-}extreme)$$

If the likelihood ratio of a token is bigger than 1, it implies that this token is more indicative of an extreme injury; if it is smaller than 1, then it is more indicative of a non-extreme injury.

The calculation is conducted in RStudio v.4.2.1. Because the dataset contains a huge amount of words and the term document matrix is large, we vectorized our code to improve the efficiency. After calculating the ratio, we picked the top 10 tokens that are most indicative of extreme injuries, then compared the ratio changes over the 22 years to check the trend using Pearson's correlation. We also extracted the top two tokens by their ratios in each of six mining sub-units. Each sub-unit represents a type of mine locations where a collection of jobs are conducted. Therefore, we can get a broad idea of which jobs are most indicative of extreme injuries in various mining locations.

Results

Table 2: Overview of incident records in MSHA dataset

Category	N (%)	Extreme Injuries, N (%)
Total	**216066 (100%)**	**3265 (1.5%)**
Mine Sub-unit		
Dredge	3497 (1.6%)	71 (2%)
Independent shops	716 (0.3%)	12 (1.7%)
Mill operation	59176 (27.4%)	785 (1.3%)
Strip/quarry	69866 (32.3%)	1113 (1.6%)
Surface (underground mine)	7557 (3.5%)	114 (1.5%)
Underground	73203 (33.9%)	1142 (1.6%)
Year		
2000 to 2004	71589 (33.1%)	1039 (1.5%)
2005 to 2009	58497 (27.1%)	905 (1.5%)
2010 to 2014	44508 (20.6%)	689 (1.5%)
2015 to 2021	41472 (19.2%)	632 (1.5%)

Description of the dataset A total of 216,066 injury incidents were included in this analysis, of which 3,265 (1.5%) were considered extreme injury incidents (Table 2). These incidents were associated with a total of 426 tokens that describe the occupation of the miners. Most injury incidents occurred underground (33.9%), in quarries (34.3%), and during mill operations (27.4%). There was a downward trend in the number of injuries since 2000, with 33.1% of incidents arising between 2000 and 2004 and falling to 19.2% by 2015 to 2021. Generally, there was very little difference in the percent of extreme incidents by mine sub-unit or year (1.3% to 2%).

Naive Bayes analysis Of the 426 tokens that describe the occupation, 170 (40%) of words had a likelihood ratio (LR) > 1. Superintendent had the highest LR, with superintendents 3.9 times more likely to experience an extreme injury than a non-extreme injury. Occupations related to auger (LR = 3.8) and stoper (LR = 3.4) made up the top 3 most likely jobs to experience an extreme injury than non-extreme. The rest of the occupations that were the top 10 most likely to experience an extreme injury include work related to "tender" (LR = 3.2), "shaftcrew" (LR = 3.0), "washer" (LR = 2.8), "shaft" (LR = 2.6), "iron" (LR = 2.6), "chute" (LR = 2.5); and "grizzly" (LR = 2.4). Among the jobs which were the most likely to experience a non-extreme injury than an extreme injury, the top 3 were related to security, guard, and watchman. All of these jobs had 0.3 times lower likelihood of a non-extreme injury than an extreme injury.

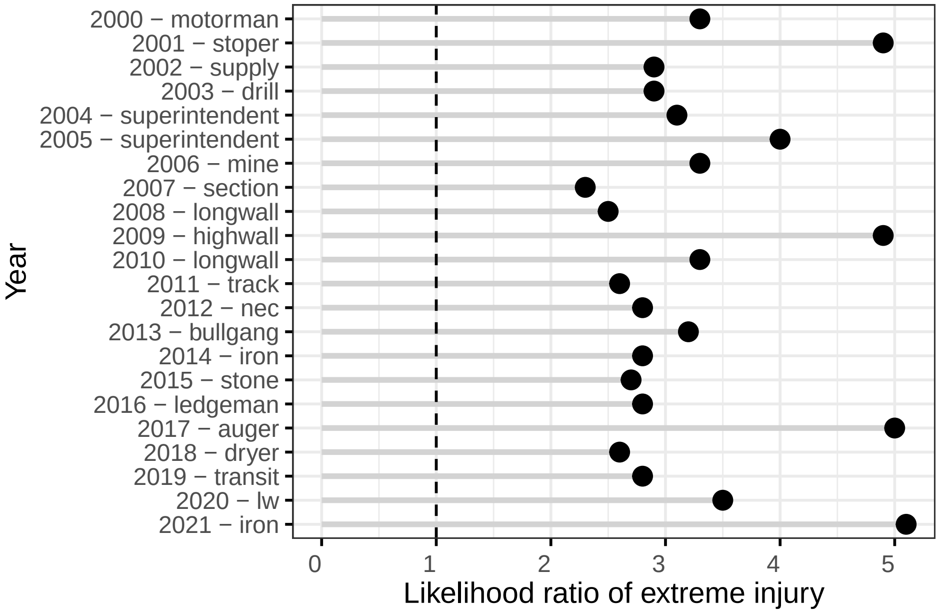

Figure 1: Yearly most indicative tokensof extreme injuries compared to non-extreme injuries

Trends by year Figure 1 presents the most indicative words (highest LR) of extreme injuries by year. In 2021, "iron" work had the highest likelihood ratio, with such workers having 5.1 times higher likelihood of an extreme event than non-extreme event. Work related to augers in 2017 (LR = 5.0), "highwall" drilling in 2009 (LR = 4.9), stopers in 2001 (LR = 4.9), and superintendents in 2005 (LR = 4.0) were the most indicative of extreme injuries. Within the last five years, work related to auger (LR = 5.0), dryer operation (LR = 2.6), transit (LR = 2.8), LW propman (LR = 3.5) and "iron" work (LR = 5.1) were the most likely to experience an extreme injury from 2017 to 2021, respectively.

Yearly trends in the likelihood ratio of the overall top 10 most likely jobs to experience an extreme injury (Table 3) from 2000 to 2021 is shown in Figure 2. Overall, superintendents (r = -0.32) and stopers (r = -0.37) had a decreasing trend in likelihood ratios from year to year. Grizzly (r = 0.59), washer work (r = 0.45), iron work (0.27), tender work (r = 0.24), and chute work (r = 0.22) demonstrated a yearly, increasing trend. Work related to augers, shaft and shaftcrew did not display a yearly trend. None of the top 10 most indicative words consistently had a likelihood ratio > 1 for every year.

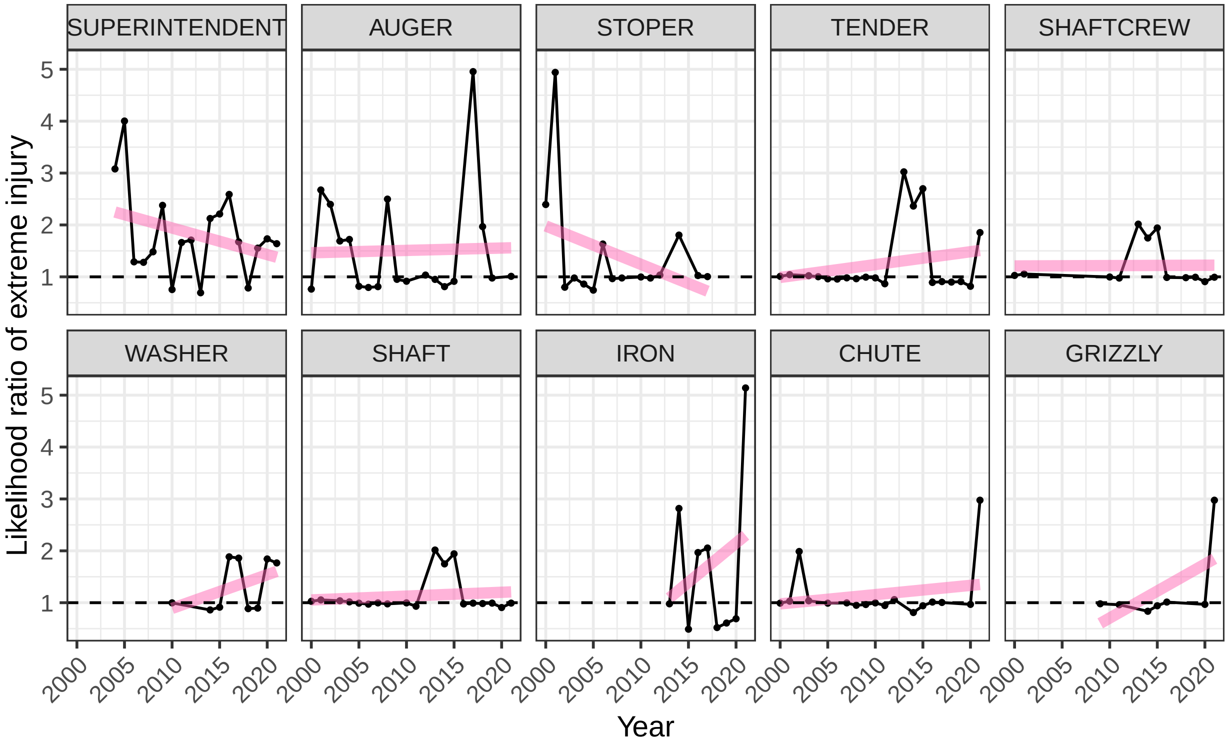

Figure 2: Changes in likelihood ratio from 2000 to 2021 for the top 10 most indicative tokens of extreme injuries compared to non-extreme injuries

Differences by mine sub-unit The top 2 tokens most indicative of extreme injuries compared to non-extreme injuries were examined by mine sub-unit (Figure 3). Among dredge mining, superintendents and drivers were 3.3 and 3.0 times more likely to experience an extreme injury than a non-extreme injury, respectively. Among independent shops, jobs related to "mine" (which were linked with "mine managers" and "mine examiner" in the dataset) were 4.8 times more likely to experience an extreme injury than a non-extreme event. Words related to "manager" and "owner" had the same likelihood ratio (LR = 2.5). Among mill operation, superintendents were 3.1 times more likely to experience an extreme injury than a non-extreme injury, with "motorman", "motor", and "switchman" having similar likelihood ratios (3.1). Among quarries, superintendents were once more the most indicative of an extreme event (LR = 3.4), with augers having 2.9 times higher likelihood of an extreme event than non-extreme. At the surface of underground mines, drivers and truckers were the most likely to experience an extreme event (LRs = 2.9 and LR 2.7, respectively). Lastly, tenders and stopers were the most likely to experience an extreme event at underground mines (LRs = 4.2 and LR = 3.4, respectively). Superintendents were the most indicative of extreme events in three mine sub-units (dredge, mill operations, and quarry) and drivers were the most indicative in dredge mining and at the surface of underground mines.

Discussion

The mining industry is a unique sector of the workforce. It includes many highly specialized sets of tasks, which make it difficult to group jobs into categories that are comparable with other industries. "Grizzly," for example, involves a specific set of tasks not easily interpretable to other contexts. This poses challenges when attempting to determine which roles might be the best targets for injury prevention. Additionally, these job tasks have specific mining-relevant meaning that require a mining context to understand [4]. In this analysis, naive bayes classification with LaPlace smoothing was used to determine the likelihood of an extreme or non-extreme injury event for each job task in the dataset and find that "superintendent"

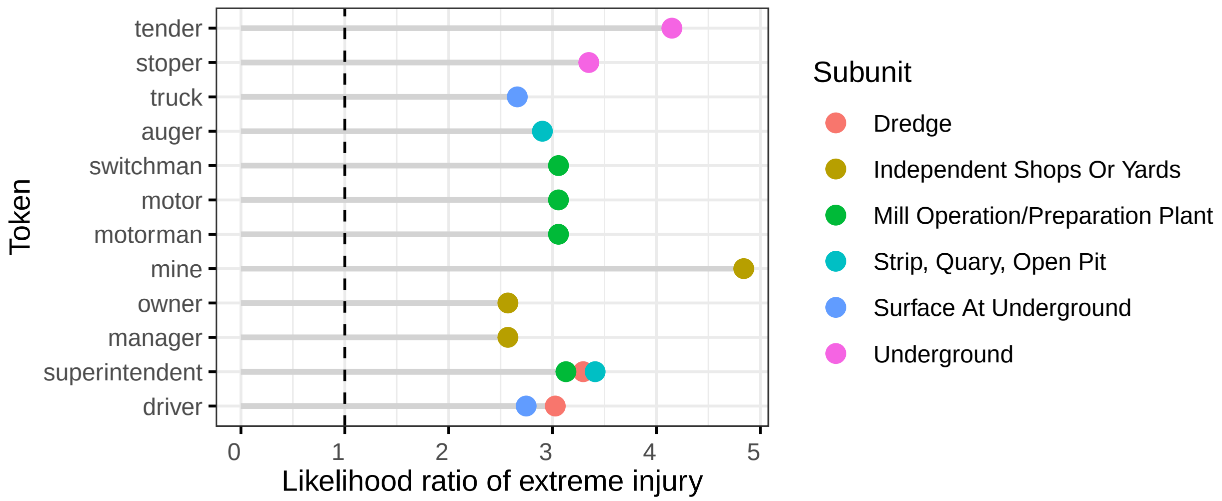

Figure 3: Top 2 most indicative tokens by mine sub-unit of extreme injuries compared to non-extreme injuries

has the highest likelihood of experiencing an extreme injury. This method allows for the job tasks as they appear in the dataset to be used to determine the likelihood of workers experiencing an extreme event. This information can be used by mine leadership and policymakers to identify interventions for improved mine safety for the most dangerous jobs at mine sites.

References

1. U.S. Bureau of Labor Statistics. (n.d.). IIF Home. U.S. Bureau of Labor Statistics. Retrieved December 16, 2022, from https://www.bls.gov/iif/

2. Raj, V. K., and E. K. Tarshizi. “Advanced Application of Text Analytics in MSHA Metal and Nonmetal Fatality Reports.” SME Annual Meeting & Expo: Phoenix, AZ, USA. 2020

3. National Institute for Occupational Safety and Health. Mining Safety and Health Research: MSHA Data File Downloads. Washington, DC: National Institute for Occupational Safety and Health; 2008.

4. National Institute for Occupational Safety and Health. “6 Emerging Issues in Mining Safety and Health.” 2007. The National Academies Press. doi: 10.17226/11850.

Author Contributions **Xin Zhang:** Concept and design, presentation slides, proposal editing, statistical analysis, data interpretation, bibliography, drafting of final report. **Ariana Haidari:** Presentation slides, critical revision, project proposal authorship, data interpretation, final report authorship, bibliography. **Abas Shkembi:** GitHub initiation, drafting of final report, critical revision, statistical analysis, concept and design, project proposal authorship, presentation slides, dataset selection, data interpretation, final report authorship.

Data availability All data and code used in this project analysis is publicly available in the following repository: https://github.com/ashkembi/BIOSTAT625-Final-Project