Advanced Application of Text Analytics in MSHA Metal and Nonmetal Fatality Reports

Conference Paper · February 2020

CITATIONS
2

READS
576

2 authors, including:

Vaibhav Raj
Centers for Disease Control and Prevention
28 PUBLICATIONS **72** CITATIONS

SEE PROFILE

SME Annual Meeting
Feb. 23 - 26, 2020, Phoenix, AZ

Preprint 20-111

ADVANCED APPLICATION OF TEXT ANALYTICS IN MSHA METAL AND NONMETAL FATALITY REPORTS

K. V. Raj, CDC NIOSH, Spokane WA
E. K. Tarshizi, National Univ., San Diego CA

ABSTRACT

Advanced data science techniques have been applied in a verity of fields to improve health and safety and reduce accident and fatality rates. Similar to other industries, the mining industry requires the adoption and implementation of modern research methods to continue boosting safety in mine operations. Text analytics, in general, is one of the essential methods used to analyze unstructured data that is in text format. Using text mining and Natural Language Processing (NLP) techniques to extract patterns with incidents and identify insightful information from contents and reports is extremely valuable to produce actionable recommendations and strategies for applying text mining and NLP to accident text data.

In this research investigation, advanced text mining techniques and NLP were applied to the U.S. Mine Safety and Health Administration (MSHA) final fatality reports in metal and nonmetal operations (surface and underground) to perform preprocessing, Exploratory Data Analysis (EDA), and various statistics to discover insightful word relation patterns in the reports. For this purpose, the final fatality reports from 2010 through 2017 were collected and cleaned. In addition, topic modeling was done to group the reports with similar underlying themes or topics.

INTRODUCTION

Worker safety is of paramount importance to any organization, especially in high-risk sectors such as mining, as any fatality or injury adds an economic and social burden not only to the organization but also to the family of the worker. The number of fatalities and injuries in the mining industry has seen a decreasing trend over the years (with some spikes) but are still at unacceptable levels. According to the Mine Safety and Health Administration (MSHA) Accidents and Injuries dataset, from 2010 through 2017 a total of 306 fatalities occurred in all mining sectors, and of these, 147 were reported to have occurred in the metal/nonmetal (MNM) and stone, sand, and gravel (SSG) mining sectors. MSHA data also shows that 45% of all the MNM and SSG mining sector fatalities were attributed to machinery and powered haulage [1, 2]. Researchers in the past have pointed out that human/machine interactions lead to most of the fatalities and injuries [3-6]. With persistent fatal incidents involving powered haulage, MSHA has focused on reducing fatalities via their Powered Haulage Safety Initiative: "*About half of all U.S. mining fatalities in recent years – including 13 of the 27 fatalities in 2018 – were due to accidents involving powered haulage. That classification includes mobile equipment, conveyor systems, and anything else under power that hauls people or materials. MSHA has made the prevention of powered haulage accidents a priority, with an initial focus on three areas: mobile equipment at surface mines, seat belt usage, and conveyor belt safety*" [7]. Through this initiative, MSHA is planning to reduce powered haulage fatalities by focusing on mobile equipment at surface mines, seat belt usage, and conveyor safety.

In response to this, researchers from the National Institute for Occupational Safety and Health (NIOSH) are focusing their efforts on investigating powered haulage and machinery fatalities [8]. NIOSH investigations include going through the accidents and injuries dataset as well as the final fatality reports. The final fatality reports are published by MSHA after a thorough investigation of the fatal accidents. Analysis of the accidents and injuries dataset and the final fatality reports can give meaningful insight into the accident, which can eventually be helpful in reducing the fatal incidents. The study described in this paper focused on fatal accidents from 2010 through 2017 that were classified as powered haulage and machinery for the purpose of applying advanced text mining and Natural language processing (NLP) tools.

Text mining and NLP are data science techniques used for extracting meaningful information from unstructured text data, which in this case comes from MSHA final fatality reports. Text mining is broadly referred to as the process of extracting useful information from data sources through the use of statistics, identification, and exploration of interesting patterns [9]. NLP is a branch of computational linguistics for information extraction and uses part-of-speech (POS) tagging and named-entity extraction for revealing useful information about the text data [9, 10]. Applying these techniques for accident investigation is not new, as several researchers [11-16] have used such techniques in the past. Chokor et al. [11] applied unsupervised machine learning for analyzing Arizona Occupational Safety and Health Administration (OSHA) injury reports and used machine learning for clustering of accident attributes such as falls, struck by object, etc. Tixier et al. devised a ruled-based approach with the creation of a keyword dictionary for content analysis for construction safety [12]. Tirunagari conducted a study to investigate maritime accidents using text mining and two methods, namely the Pattern Classification method and the Connectives method [13]. Brown discussed the role and application of text mining algorithms to determine accident characteristics developed from the narrative reports submitted on rail accidents to the Federal Railroad Administration. He preferred the combination of text mining techniques with ensemble methods like Latent Dirichlet Allocation (LDA) [14]. Nakata [15] proposed a solution to the challenge of analyzing unstructured textual reports of accidents to identify the flow of events by focusing on two adjacent events in a large corpus of text documents while applying text mining techniques on NASA's Aviation Safety reports. More recently, Zhang et al. [16] applied text mining and NLP techniques on construction site accidents. They used classification algorithms for classifying accidents and an NLP rule-based chunking approach for identifying common objects that cause an accident.

The application of text mining for accidents related to the mining industry is rarely considered. Tarshizi et al. [17] did text mining analysis on MSHA fatal accident narratives for coal mining. They applied text mining on narratives of the fatal accidents, which involved looking into word frequency and word correlation plotting to identify factors that were the possible cause of an accident. The work being presented here will be the first attempt to apply text mining and NLP to the MSHA final fatality reports in the MNM mining sector, specifically as related to powered haulage and machinery accidents.

MSHA ACCIDENT REPORTS

The data used in this research were the MSHA final fatality reports, downloaded from the MSHA website. There are PDF versions of the reports as well as text versions on the website. The PDF versions have pictures and other information in the reports; however, for the purpose of this study only text versions were extracted from the website. The 69 reports selected for this study were from 2010 through 2017, and all were from the MNM sector with accident classifications of powered haulage or machinery.

METHODOLOGY

The process of applying text mining and NLP to the MSHA fatality reports is illustrated in Figure 1. The first step (data acquisition) is to acquire text data from the MSHA site. The text versions of reports are first extracted from the website and text data is subsequently extracted from individual reports. In this step, MSHA final fatality reports were collected from the website and text files were created based on accident dates. Next, the actual text/characters were extracted from the text file and put in tabular format for ease of accessing data in text mining and NLP programs. This process was time consuming as not all reports were formatted exactly the same even though the reports were from similar MSHA accident classifications. Figure 2 shows a sample of the tabular data created based on the data extracted from the reports. The tabulated data include report identification number, facility type, MSHA accident classification, overview, general information about the mine, description of the accident, discussion, and MSHA root cause analysis.

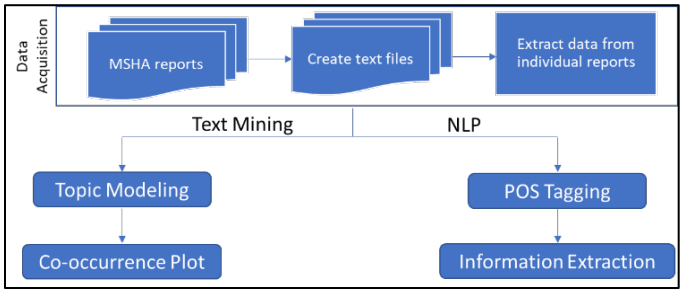

Figure 1. Process followed for applying text mining and NLP techniques.

1 REPORT_ID	2 FACILITY_TYPE	3 CLASSIFICATION	4 ACCIDENT_DT	5 OVERVIEW	6 GENERAL_INFO	7 DESCRIPTION	8 DISCUSSION	9 ROOT_CAUSE
'MAI-2010-07'	'Surface'	'Machinery'	'05/24/2010'	'On May 24, 2...	'U. S. Lime Compa...	'On the day of th...	'Location of the...	'A root cause an...
'MAI-2010-05'	'Surface'	'Machinery'	'05/05/2010'	'Cody Dean, ...	'Brierfield Quarry, ...	'On the day of th...	'1x11287 char'	'A root cause an...
'MAI-2011-16'	'Surface'	'Machinery'	'12/15/2011'	'Wesley J. She...	'Damascus 535 Cr...	'the day of the a...	'Location of the...	'A root cause an...
'MAI-2011-13'	'Surface'	'Machinery'	'11/07/2011'	'Bruce A. And...	'Anderson Sand an...	'the day of the a...	'Location of the...	'A root cause an...
'MAI-2011-01'	'Surface'	'Machinery'	'02/12/2011'	'Jose A. Soto, ...	'Hardee Phosphat...	'n the day of the ...	'Location of the...	'A root cause an...
'MAI-2012-17'	'Surface'	'Machinery'	'11/01/2012'	'Stephen J. Wi...	'Mt. Marion Pit an...	'phen J. Wickha...	'Location of the...	'The investigator...
'MAI-2012-14'	'Surface'	'Machinery'	'09/26/2012'	'On Septemb...	'North Pit, a sand ...	'On the day of th...	'Location of the...	'Investigators co...
'MAI-2012-07'	'Surface'	'Machinery'	'05/23/2012'	'On May 23, 2...	'Broken Bow Sand ...	'On the day of th...	'Location of the...	'Investigators co...
'MAI-2013-10'	'Surface'	'Machinery'	'08/05/2013'	'On August 5,...	'Mill, Crushing LLC...	'On August 5, 20...	'Location of the...	'The investigator...
'MAI-2013-23'	'Surface'	'Machinery'	'07/10/2013'	'Joe Donald T...	'Troy Plant, a cons...	'the day of the a...	'Location of the...	'null'
'MAI-2013-02'	'Surface'	'Machinery'	'01/21/2013'	'On January 2...	'Apex Quarry and ...	'the day of the a...	'Location of the...	'A root cause an...
'MAI-2014-24'	'Surface'	'Machinery'	'12/29/2014'	'On Decembe...	'Tilden Plant, a sur...	'the day of the a...	'Location of the...	'The investigator...
'MAI-2014-25'	'Surface'	'Machinery'	'11/10/2014'	'On Novemb...	'R&S Stone, I...	'Friday Novemb...	'Location of the...	'The investigator...
'MAI-2014-10'	'Underground'	'Machinery'	'04/28/2014'	'On April 28, ...	'Midas Mine, a gol...	'the day of the a...	'Location of the...	'The investigator...
'MAI-2014-08'	'Surface'	'Machinery'	'04/24/2014'	'On April 24, ...	'Gravelite Division,...	'the day of the a...	'Location of the...	'A root cause an...
'MAI-2014-07'	'Surface'	'Machinery'	'04/17/2014'	'Harold Steeg...	'Hafenstine Plant ...	'the day of the a...	'Location of the...	'A root cause an...
'MAI-2015-16'	'Surface'	'Machinery'	'12/15/2015'	'Bernard L. Ge...	'The Davenport Pla...	'On the day of th...	'Location of the...	'null'
'MAI-2015-14'	'Underground'	'Machinery'	'08/03/2015'	'On August 3,...	'SSX Mine, a multi...	'On the day of th...	'Location of the...	'Investigators co...
'MAI-2015-12'	'Surface'	'Machinery'	'07/10/2015'	'On July 10, 2...	'The Dry Fork Sand...	'the day of the a...	'1x11157 char'	'The investigator...
'MAI-2015-06'	'Surface'	'Machinery'	'03/23/2015'	'On March 23...	'Kane's Quarry, a s...	'the day of the a...	'GeologyThe mi...	'The investigator...
'MAI-2015-04'	'Surface'	'Machinery'	'01/26/2015'	'William K. St...	'Swift Creek Mine, ...	'the day of the a...	'WeatherThe w...	'The investigator...

Figure 2. Tabular data created from MSHA reports.

The next step was to apply text mining and NLP techniques on the text data extracted. For text mining, topic modeling and co-occurrence plots were generated; whereas, for NLP part-of-speech (POS) tagging and information extraction were done. Details of each technique will be provided in the following sections.

Text Mining

After the tabulation of reports, text mining techniques such as topic modeling was applied to identify and assign underlying topics for the reports. Figure 3 shows the steps involving text mining. Before applying the topic modeling technique to the reports, texts are broken into individual units, called tokens, and cleaned. The process of breaking up the text units into words and certain characters, such as punctuation, in a sequence is called tokenization, and the documents containing a series of tokens are called tokenized documents. After tokenization of text, the entire body of text is converted to lower case, followed by removal of punctuation. Apart from that, the most frequently occurring words such as a, and, to, an, the etc. are removed from the tokenized documents. The most frequently occurring words are termed "stopwords." Apart from stopwords, a set of words, such as victim, employee, occur etc. which are present in text but do not add value to the analysis and occur frequently, were also removed. These words were manually cleaned after going through word frequencies. Further, the texts in the tokenized documents were normalized by reducing different forms of the same word to its root form, e.g. kill, killed, killing can be expressed as kill. After cleaning and normalization of the tokenized document, words in the document can be analyzed as one unit called a "unigram," in two units called a "bigram," and so on. This process is termed data preprocessing.

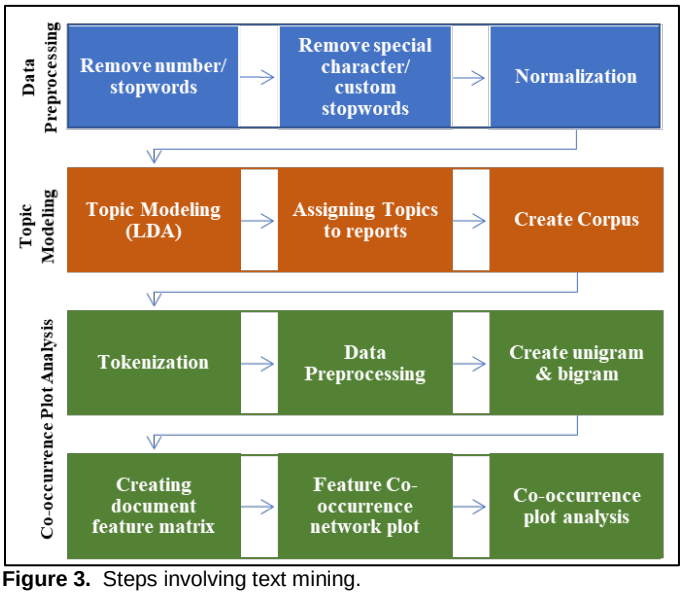

Figure 3. Steps involving text mining.

Topic modeling of a group of documents is done to cluster the documents based on the distribution of co-occurring of words among the documents. The unigram unit was used for topic modeling and the Latent Dirichlet Allocation (LDA) model [18] was selected for topic modeling. Initially, topic modeling was done for the entire collection of 69 reports; however, since the reports were a collection for two MSHA accident classifications, namely powered haulage and machinery, the reports were sub-divided into two sets based on their MSHA classification. There were 43 reports for powered haulage accidents and 26 for machinery-related accidents. At first, the LDA model was run to decide a suitable number of topics for a set of documents. Figures 4 and 5 show the results for the LDA model aimed at determining a suitable number of topics for powered haulage and machinery reports. Figure 4 shows that 18 is the number of topics that is suitable for 43 powered haulage reports, whereas Figure 5 shows that for the 26 machinery reports a suitable number of topics is 8. The suitable number of topics was determined where the model showed that validation perplexity was minimized.

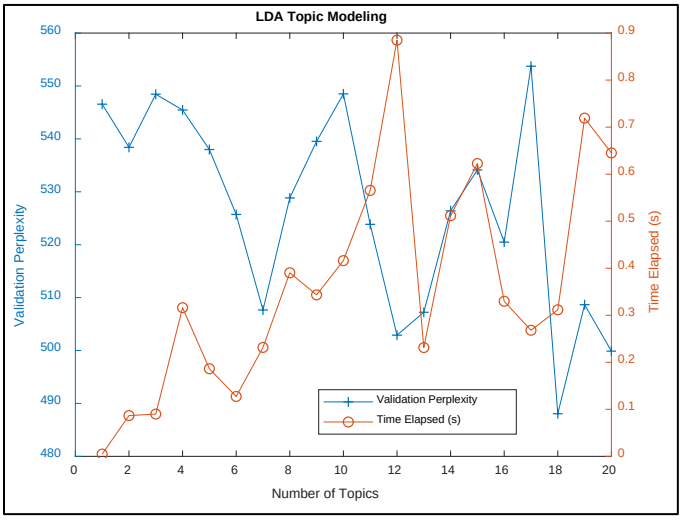

Figure 4. LDA topic modeling to find a suitable number of topics for powered haulage reports.

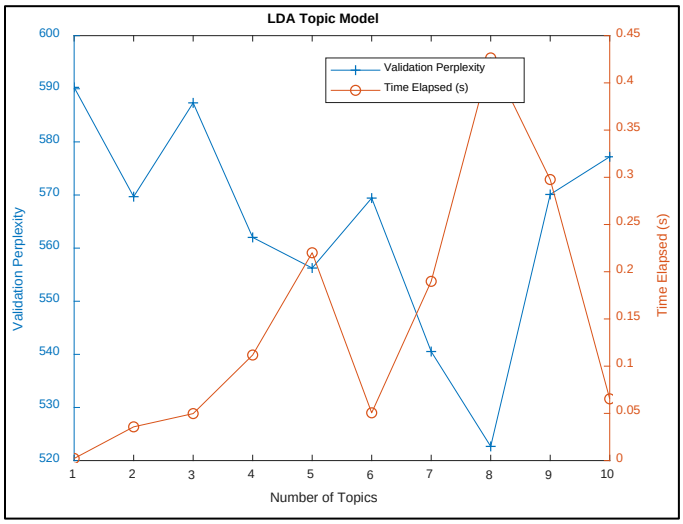

Figure 5. LDA topic modeling to find a suitable number of topics for machinery reports.

Once the number of suitable topics was determined, topics were assigned to the reports by LDA topic modeling. In this step, LDA modeling is done to cluster different reports to topic number. Some of the powered haulage reports clustered to their assigned topics are presented in Figure 6. It can be seen from Figure 6 that Topic Number 3 has the maximum number of reports assigned to it (15) and Topic Number 12 has 5 reports assigned. There were 9 topics with only 1 assigned report, 4 topics with 2 reports, and 2 topics with 3 reports. On further investigation, it was found that Topic Number 3 is associated with haul truck accidents, whereas Topic Number 12 is associated with incidents involving conveyor belts. Topics with only one report have been found to have accidents unique in nature, such as victim got pinned while performing maintenance on forklift or victim was pinned between two loaded ore cars. Similarly, LDA topic modeling was done for machinery reports (Figure 7). However, there are some issues with LDA topic modeling as there were a total of 9 conveyor-related incidents, but LDA topic modeling clustered 5 incidents together and other incidents were clustered individually or with other types of accident topics. This could be due to the lower number of reports used to train the LDA model. This was similar to the case with machinery accidents LDA topic modeling where unrelated accidents were clustered together. It should be noted that there were only 26 machinery accidents for topic modeling.

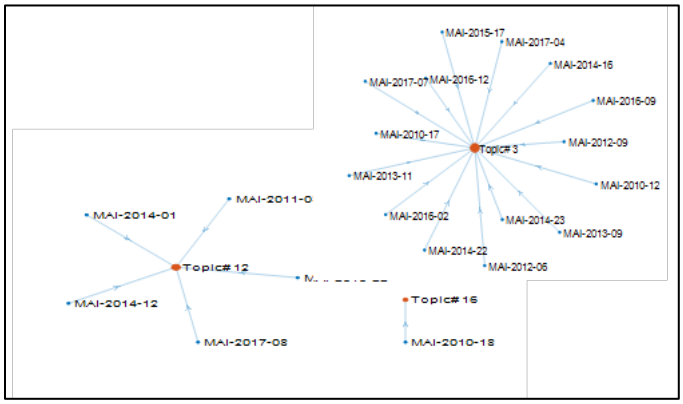

Figure 6. Examples of powered haulage reports assigned to different topic numbers.

The MATLAB® Text Analytics Toolbox was used as the programming language for extracting data from the reports and for LDA topic modeling. After the topic modeling using LDA was done, the text data was modified to be used in R statistical computing programming language for feature co-occurrence plot analysis. The quanteda package within R was used during this investigation. R is an open-source program that has been used for a variety of analytics and text mining projects. Quanteda is a toolkit for managing textual data and applying natural language processing (NLP) tasks [19].

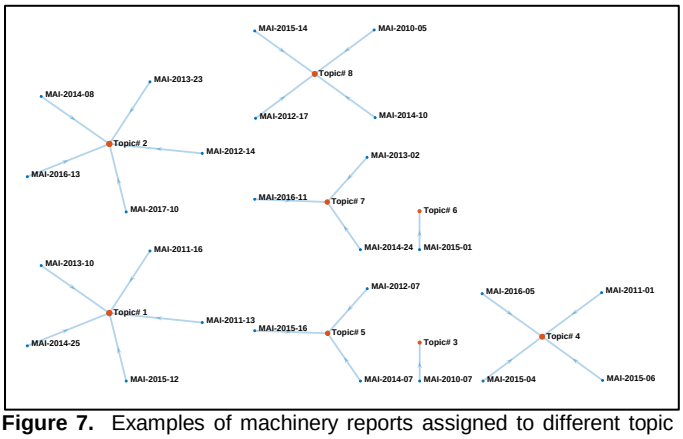

Figure 7. Examples of machinery reports assigned to different topic numbers.

Technically, after modeling the topics, which is an unsupervised machine learning method, a new set of incident reports were selected for the feature co-occurrence matrix and plot analysis. Using this new/refined set of topics, the associated reports as text files were imported into R using the readtext package before the preprocessing stages. For most text mining techniques, it is essential to tokenize the created corpus into smaller text features. This technique, along with other preprocessing methods, such as removal of punctuation, numbers, separators, symbols and hyphens, enhance the accuracy and computational analysis process [20]. After filtering the general English stopwords, custom-built lists of stopwords, particularly for both unigrams and bigrams of the incident reports, were developed. Finally, the features of the tokens were converted to lower cases as part of normalization.

A document-term matrix (DTM), known as document-feature matrix (DFM) as used in the quanteda package, is a mathematical matrix of the “bag-of-words” and one of the most common numerical representations of text corpus or tokens. After developing a DFM of the preprocessed tokens of the incident reports for each individual new topic, feature co-occurrence matrices (FCM) of the tokens (unigrams and bigrams) were created, and the most frequently co-occurring words were plotted to visualize semantic network analyses. Figure 8 shows the bigram co-occurrence plot of Topic Number 1 for powered haulage accidents. It can be seen from the figure that there are two distinct clusters of words. Topic Number 1 comprises two reports, wherein one report incident involves an all-terrain vehicle (ATV) with the victim driving on difficult terrain, and in another incident, a trailer truck was involved.

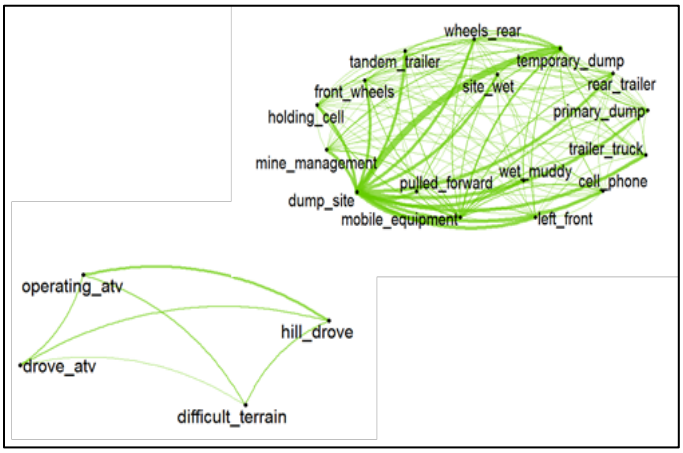

Figure 8. Co-occurrence plot for Topic Number 1 for powered haulage accidents.

Given the nature of the accidents presented in Figure 8, the co-occurrence plot was able to make a distinction between the two accidents. The co-occurrence plot for Topic Number 3 for powered haulage accidents is presented in Figure 9. Topic Number 3 is associated with 15 haul truck accidents and unigram co-occurrence shows the word "truck" at the center of the plot. The word truck is associated with slope, control, berm, crane, etc., which points to the accident's conditions, i.e. where the truck might have lost control, traveled down the ramp, or went off the berm. This co-occurrence plot gives some insight into the conditions surrounding the incident. Similarly, co-occurrence plots for other topics were also plotted and analyzed.

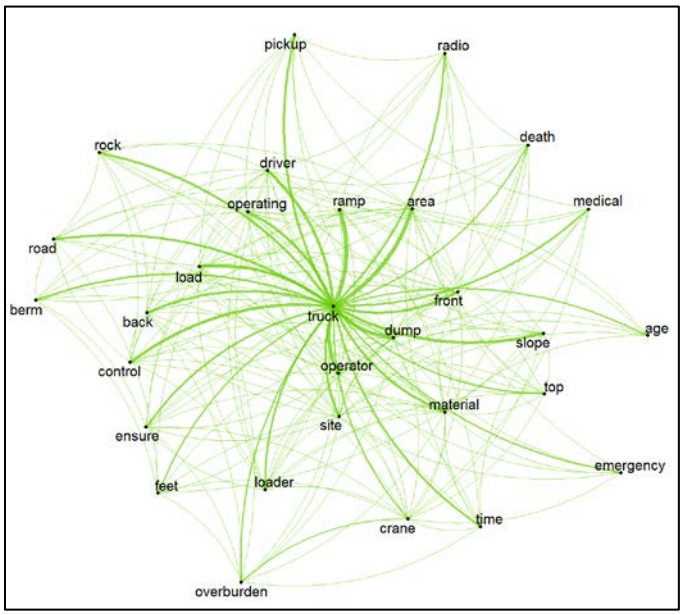

Figure 9. Co-occurrence plot for Topic Number 3 for powered haulage accidents.

The co-occurrence plot for Topic Number 7 for machinery accidents is shown in Figure 10. From the figure, it can be seen that there are three distinct incidents associated with Topic Number 7 where one of the incidents is associated with a forklift, another one with a kiln, and the last one with a chute. It is interesting to note that forklift accidents are classified as machinery as well as powered haulage. In the accident involving a kiln, it appears that the victim was performing maintenance. The co-occurrence plot has been helpful in identifying the co-occurrence of tokens in the same report as well as in the combination of reports. This gives an insight into what was in the text without thoroughly reading it.

Natural Language Processing

NLP was used after the text data was extracted from the fatality reports. This was done in the Python programming language using the spaCy package. The first step in this process is tokenization. However, after the tokenization, stopwords are not removed as the NLP looks into full text and assigns part-of-speech (POS) and word interdependency, which is the grammatical relationships in a sentence. The next step is POS tagging, which is the assigning of POS to each token of the text and finding grammatical relationships in the sentence [21]. Figure 11 shows an example sentence with POS assigned to each word and their grammatical relationship. It shows that victim (proper noun) and died (verb) is related to a dependency label "nsubj" which is a nominal subject. Here victim is nominal subject to the verb, died, and this verb is also related to another verb, rolled, by a dependency label "advcl" (adverbial clause modifier). This is done by a machine learning model based in the spaCy package.

Another important feature of NLP is information extraction, which is done by recognizing entities such as name, place (geo-location), organization, date-time, etc. This is also done by using a machine learning algorithm with training data coming from a corpus inbuilt in the spaCy program. This process was able to recognize the entities such as person, date-time, and places; however, there were some mismatches in recognizing organization, ordinal number, and quantities [21]. Moreover, for accident analysis, it was important to recognize relevant entities such as equipment type. Therefore, a dictionary was created for the equipment used in those reports, and the model was trained to recognize equipment in the text. Figure 12 shows some of the entities recognized during the process. The model was able to recognize persons in the text; however, their names are not shown here. Apart from the person's name, the model was able to recognize the time where time was mentioned. With inclusion of equipment in the model, the model was able to recognize haul truck, front-end loader, and dozer as equipment in the text.

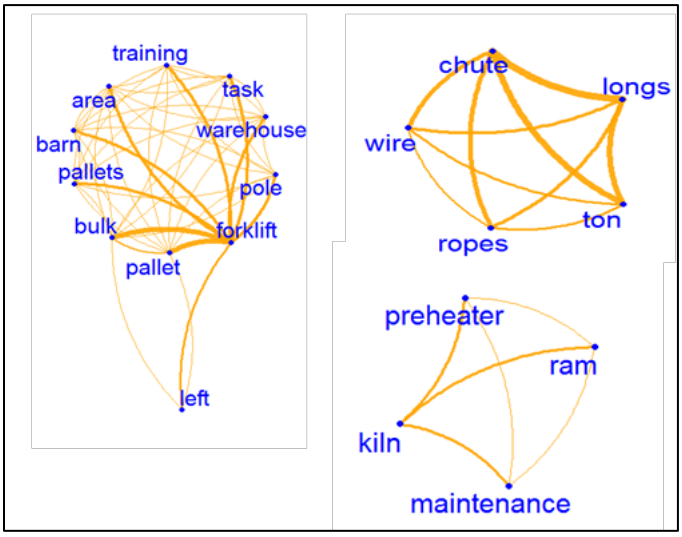

Figure 10. Co-occurrence plot for Topic Number 7 for machinery accidents.

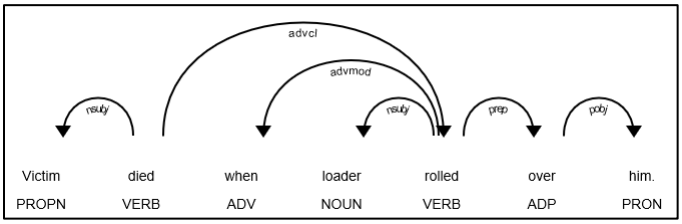

Figure 11. Assigned POS and grammatical relationship within an example sentence.

After entity recognition, information extraction was done to extract some useful information from the reports. For this purpose, the chunking method was used to extract nouns and words around them as well as verbs and words around them. The information extracted for one of the reports after chunking is presented below.

> "*Chunk - October On, victim died, he cleaning, a belt conveyor tail pulley on*"

After examining the chunk, it can be said that the victim was killed when he was doing some cleaning work on a belt conveyor. The chunk extraction was done by manually looking at the chunks and then selecting them. This process reduces the effort of an investigator since they do have to go over the entire text of the report but rather can simply skim through the chunk to make meaningful conclusions. Also looking at the wordcloud of all the chunks can provide a clear perspective of the incident. Figure 13 shows the wordcloud of the chunks extracted from the report, and it is more evident from the wordcloud that the victim was cleaning the belt conveyor and the conveyor belt energized. Similarly, if we look at the wordcloud of chunks of information extracted from Topic Number 12 for powered haulage accidents presented in Figure 14, it can be inferred that most of the accidents in Topic Number 12 happened around belt conveyors where the victim got entangled or fell on it. Figure 15 presents the

wordcloud from Topic Number 2 for machinery accidents, which shows that most of the accidents happened around a dozer or drill machine. On further investigation, it was found that 4 out of 5 accidents associated with Topic Number 2 were due to dozers and 1 was due to a drill machine. Injuries in the cases of dozer incidents were attributed to falling into a pond, not wearing a seat belt or loss of control.

Figure 12. Example of entity recognition done on one of the accident reports.

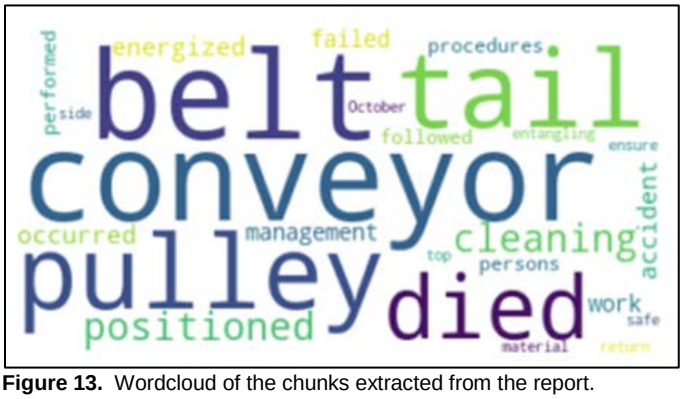

Figure 13. Wordcloud of the chunks extracted from the report.

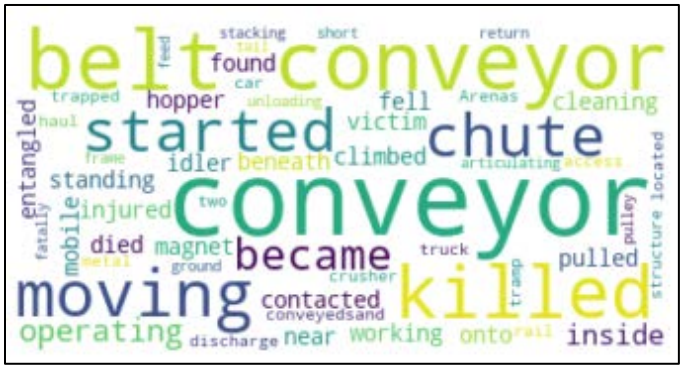

Figure 14. Wordcloud of chunks from Topic Number 12 for powered haulage accidents.

CONCLUSIONS

Text mining and natural language processing (NLP) has been used for accident analysis in the construction industry and other industry sectors for quite some time, and these methods have the potential to identify factors that lead to fatalities and injuries. In the mining sector, this study is the first attempt to apply text mining and NLP for analyzing MSHA final fatality reports. It was found that the application of topic modeling along with the co-occurrence plot can provide a quick insight into a cluster of text, which can be helpful in clustering reports based on similar underlying themes or topics; whereas, the application of NLP proved better for the understanding of sentence structure and is helpful in information extraction. One of the important aspects of using these techniques is their ability to produce meaningful results in a relatively short amount of time. Topic modeling can cluster similar incidents together and save time in manual grouping. NLP can be useful in chunking and extracting entities can give a user some meaningful information regarding the text.

Figure 15. Wordcloud of chunks from Topic Number 2 from machinery accidents.

The focus during topic modeling should be on training the model for better clustering of text and removing stopwords so that the output is not biased due to the occurrence of the most frequent words in text. In the case of NLP, most of the training data comes from work done in the field of literature and social sciences, which lacks many of the terms used in the mining industry. However, creating your own dictionary for mining terms can be helpful as was done in this study where the authors created a dictionary called “equipment” for identifying mining equipment for entity recognition.

DISCLAIMER

The findings and conclusions in this report are those of the author(s) and do not necessarily represent the position of the National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention. Mention of any company or product does not constitute endorsement by NIOSH.

REFERENCES

1. MSHA (2018). "*MSHA Accident Data Sets*". [cited 2018 1/24]; available from: https://arlweb.msha.gov/OpenGovernmentData/OGIMSHA.asp.
2. NIOSH (2018). "*MSHA Data File Downloads*". [cited 2018 3/14]; available from: https://www.cdc.gov/niosh/mining/data/default.html.
3. Ruff, T., P. Coleman, and L. Martini (2011), "*Machine-Related Injuries in the US Mining Industry and Priorities for Safety Research*". International Journal of Injury Control and Safety Promotion, **18**(1): pp. 11-20.
4. Kecojevic, V., D. Komljenovic, W. Groves, and M. Radomsky (2007), "*An Analysis of Equipment-related Fatal Accidents in U.S. Mining Operations: 1995–2005*". Safety Science, **45**(8): pp. 864-874.
5. Zhang, M., V. Kecojevic, and D. Komljenovic (2014), "*Investigation of Haul Truck-related Fatal Accidents in Surface Mining using Fault Tree Analysis*". Safety Science, **65**: pp. 106-117.
6. Groves, W., V. Kecojevic, and D. Komljenovic (2007), "*Analysis of Fatalities and Injuries involving Mining Equipment*". Journal of Safety Research, **38**(4): pp. 461-470.
7. MSHA (2018). "*Powered Haulage Safety Initiative*". [cited 2018 06/15]; available from: https://www.msha.gov/news-

media/special-initiatives/2018/05/31/powered-haulage-safety-initiative.

8. NIOSH (2018). "*Identification of Key Factors Affecting Machine-related Fatalities and Injuries in MNM Mining Sectors*". available from: https://www.cdc.gov/niosh/mining/researchprogram/projects/fatalitiesinjuriesMNMmining.html.
9. Feldman, R. and J. Sanger (2007), "*The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*": Cambridge University Press.
10. Clark, A., C. Fox, and S. Lappin (2013), "*The Handbook of Computational Linguistics and Natural Language Processing*": John Wiley & Sons.
11. Chokor, A., H. Naganathan, W.K. Chong, and M.E. Asmar (2016), "*Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning".* Procedia Engineering, **145**: pp. 1588-1593.
12. Tixier, A.J.P., M.R. Hallowell, B. Rajagopalan, and D. Bowman (2016), "*Automated Content Analysis for Construction Safety: A Natural Language Processing System to Extract Precursors and Outcomes from Unstructured Injury Reports".* Automation in Construction, **62**: pp. 45-56.
13. Tirunagari, S. (2015), "*Data Mining of Causal Relations from Text: Analysing Maritime Accident Investigation Reports"*, in *arXiv*.
14. Brown, D.E. (2015), "*Text Mining the Contributors to Rail Accidents".* IEEE Transactions on Intelligent Transportation Systems, **17**(2): pp. 346-355.
15. Nakata, T. (2017). "*Text-Mining on Incident Reports to Find Knowledge on Industrial Safety*". *In Proceedings of the 2017 Annual Reliability and Maintainability Symposium (RAMS)*, Orlando, FL: IEEE.
16. Zhang, F., H. Fleyeh, X. Wang, and M. Lu (2019), "*Construction Site Accident Analysis using Text Mining and Natural Language Processing Techniques".* Automation in Construction, **99**: pp. 238-248.
17. Tarshizi, E., M.W. Buche, B. Inti, and R. Chappidi (2018), "*Text Mining Analysis of U.S. Department of Labor's MSHA Fatal Accident Reports for Coal Mining".* Mining Engineering, **70**(4): pp. 43-48.
18. Blei, D.M., A.Y. Ng, and M.I. Jordan (2003), "*Latent Dirichlet Allocation".* Journal of Machine Learning Research, **3**: pp. 993-1022.
19. Benoit, K., K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo (2018), "*quanteda: An R Package for the Quantitative Analysis of Textual Data".* Journal of Open Source Software, **3**(30): pp. 774.
20. Welbers, K., W. Van Atteveldt, and K. Benoit (2017), "*Text Analysis in R ".* Communication Methods and Measures, **11**(4): pp. 245–265.
21. Explosion (2019). "*spaCy*". Annotation Specifications [cited 2019 10/21/2019]; available from: https://spacy.io/api/annotation#pos-tagging.