An exploration of text mining of narrative reports of injury incidents to assess risk

David Passmore[1,*], *Chungil* Chae[2], *Yulia* Kustikova[3], *Rose* Baker[4], and *Jeong-Ha* Yim[5]

[1]Penn State University, Workforce Education and Development, 305D Keller, University Park, Pennsylvania, USA 16802
[2]Penn State University, Applied Cognitive Science Lab, E365 IST Building, University Park, Pennsylvania, USA 16802
[3]Federal State Educational Institution of Higher Education, National Research Moscow State University of Civil Engineering, Yaroslavskoye Shosse 26, Moscow, 129337 Russian Federation
[4]University of North Texas, Learning Technologies, Discovery Park, G150, 3940 North Elm Street, Denton, Texas USA 76207
[5]University of Georgia, Lifelong Education, Administration, and Policy, 850 College Station Road, Athens, Georgia, USA 30602

Abstract. A topic model was explored using unsupervised machine learning to summarized free-text narrative reports of 77,215 injuries that occurred in coal mines in the USA between 2000 and 2015. Latent Dirichlet Allocation modeling processes identified six topics from the free-text data. One topic, a theme describing primarily injury incidents resulting in strains and sprains of musculoskeletal systems, revealed differences in topic emphasis by the location of the mine property at which injuries occurred, the degree of injury, and the year of injury occurrence. Text narratives clustered around this topic refer most frequently to surface or other locations rather than underground locations that resulted in disability and that, also, increased secularly over time. The modeling success enjoyed in this exploratory effort suggests that additional topic mining of these injury text narratives is justified, especially using a broad set of covariates to explain variations in topic emphasis and for comparison of surface mining injuries with injuries occurring during site preparation for construction.

1 Problem

Coal is an abundant natural and economic resource in the United States of America (USA). The USA holds a demonstrated reserve base totalling 476 billion short tons of coal [1]. Coal is mined in the USA primarily using two methods: underground mining and surface mining. Underground mining technology and mechanization is suited uniquely to mining coal from rock strata in restricted subterranean workspaces that feature continuously changing geological conditions and pervasive environmental hazards. However, the

[*] Corresponding author: passmore.david@gmail.com

equipment and processes applied in surface mining are quite similar to the technologies and techniques used for site preparation in most construction projects.

Coal mining is a relatively hazardous industry in the USA. Coal miners are more likely to be killed or to incur severe non-fatal injuries than are workers in other industries [2]. In the USA during 2017, 3.18 injuries were reported to the U.S. Mine Safety and Health Administration (MSHA) for every 200,000 hours that coal miners worked, including 0.02 fatalities per 200,000 hours worked [3]. Historically, surface coal mines have produced approximately 1.5 to 2 times the tonnage of coal extracted from underground mines [4]. Yet, underground coal miners have suffered rates of injury roughly two times higher than miners in surface coal mines [5].

Knowledge about the conditions and causes of coal mine injuries is required to improve the safety of coal miners and to reduce the risk of general losses associated with disastrous coal mine events. Quantitative measures of the correlates, causes, and consequences of injury incidents are recorded in administrative databases of the MSHA and often are analysed to understand the epidemiology of injuries of coal miners [6]. Also available, although rarely used to understand coal miner injuries except for their anecdotal value, are written narrative accounts of injury incidents. Deriving useful knowledge from these narratives is difficult because natural human language is nuanced, ambiguous, and relatively unstructured. Also, these narratives about injury incidents typically are not written by injured miners, but often by safety officers or supervisors without formal training in the development of narratives that include critical, clear, and organized details about the incidents.

The term, *text mining*, refers to the use of automated methods for exploiting the enormous amount of knowledge that can be embedded in text [7]. Text mining could glean knowledge from written narratives about injury incidents to aid in the design of policies, programs, and work environments to promote the prevention of injuries to coal miners. Text mining of narrative reports about industrial injury incidents is not a new approach. For example, text mining has been applied to narrative descriptions of railroad accidents [8], farm tractor fatalities [9], nail gun injuries in construction [10], and in many other studies of industrial injury epidemiology [11]. In some cases, narratives are referenced in reports to add contextual information to explain injury correlates that initially were identified through quantitative relationships among characteristics of miner injuries [12].

2 Focus

2.1 Purpose

In this article we document the mining of text narratives written to describe 77,215 coal mine injury incidents in the USA between 2000 and 2015 that were reported to the MSHA. A form of machine learning [13] called *topic modeling* [14] reduced the large body of text in the narratives to a smaller set of topics believed to be latent in the narratives. The type of machine learning applied in our research is classified as *unsupervised* machine learning because the modeling of the injury narratives does not start from an established ontology of injury types or from a base of a priori assumptions made about the topical structure of the text narrative of injuries [15].

Computational topic models are developed as mathematical algorithms, not as a result of manual/clerical classifications of bodies of text narratives written by humans. A topic model discovers "the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes” [14]. Topic modeling does not “require any prior annotations or labeling of documents—the topics emerge from the original texts” [14].

2.2 Exploratory approach

Although our research uncovered fundamental, formerly unexplored knowledge about topics that can be derived from USA coal mine injury narratives, our approach to this research primarily is exploratory. We wish to determine whether methods for topics modeling were sensitive enough to discriminate among topics extracted from the technical, industry-specific, contextually-situated, and often informal language used to describe coal mining injuries.

Also examined were whether emphases among topics modeled from the injury narratives could be differentiated by the degree of injury the miner experienced, the location in the mine where the injury occurred, and the year of injury occurrence. Understanding such variations can add to the salience and interpretability of models of injury topics embedded in free text.

Topic models not only are descriptive, but also are useful for prediction [16]. Topic models reduce the dimensionality of free text to aid in summarizing large document collections. In addition, algorithms developed from models of topics that are latent in a large body of injury narratives also can serve as engineering features in structured machine learning models [17] that could promote loss control by predicting future injury outcomes for coal miners, a possibility we discuss after presenting the method and findings of this research.

3 Method

3.1 Data

We accessed the *MSHA Accident Injuries Data Set* [18] containing information about all 213,047 injuries reported by metal and non-metal mine operators in the USA between 2000 and 2015 as required under Part 50, Title 30, of the Code of Federal Regulations [19]. These regulations stipulate that injuries must be reported within 10 days after their occurrence on “Mine Accident, Illness, and Injury, Report (MSHA Form 7000-1)” [20]. Only injury reports from coal mines were selected for our analysis. Eliminated were reports of incidents involving: illnesses rather than injuries; fatalities and injuries due to natural, not work, causes; non-employees; and missing information about the degree of injury to the miner and the location of the injury in the mine. As a result of these restrictions, 77,215 injury reports remained from which we modeled topics from injury narratives. The reshaped data set that we analysed is downloadable from an online repository [21].

Writers of text narratives of miner injury incidents are directed to describe fully the conditions contributing to an injury, including involvement of: mine equipment; jobs skills and miner proficiency, training, and attitude; protective devices or clothing; and compliance with work rules and regulations. One information limitation is that these narratives are restricted to no more than 401 text characters. Differences in topics embedded in text narratives were examined using metadata identifying the degree of injury the miner experienced (ranging in order of severity from fatality, disability, lost work time, and down to no days lost from work or work restrictions), the location in the mine where the injury occurred (underground, surface, other areas on the mine property), and the year of injury occurrence.

3.2 Mining of text

Injury narratives were prepared for topic modeling using techniques common in text mining practice [22]. In brief, text in narratives was transformed to: maintain lowercase only; remove punctuation, double whitespaces, numbers, equations, and stop words (i.e., prepositions, articles, and other English words that have contextual function, but have trivial meaning); delete special characters and identifiable non-English words; and eliminate HTML. Every word was lemmatized, also, by eliminating inflectional affixes of words (e.g., "practices," "practicing," and "practiced" were reduced to "practic").

Topics were modeled from injury text narratives using a technique called Latent Dirichlet Allocation (LDA) [14, 16]. This technique assumes that a document is composed of a distribution of topics. The theory behind LDA is that writers prepare text about a topic by drawing words from distributions of words that are typical for the topic. Different topics are characterized by more or less unique sets of words. LDA of text narratives about injuries proceeds by determining how words cluster over documents to form topical threads that run throughout documents. Of course, words are not unique to single topics. Some words are common to many topics. Therefore, LDA is a process in which associations between documents and topics and topics and words are represented using Bayesian estimation of the joint probability distributions of documents, words, and topics. Formally, a Dirichlet distribution is a family of distributions of values over multiple categories, which is ideal for representing mixed membership distributions [23]. A Dirichlet distribution is, then, a distribution of distributions. To allow reproduction of our analysis by interested researchers, the annotated R programming code [24] applied to MSHA injury narrative text in the MSHA data that we reshaped [21] is downloadable from an online repository [25]. Solution of equations necessary to model the topics in text injury narratives required approximately 24 straight hours of computing time on a personal computer configured with 16 GB of RAM.

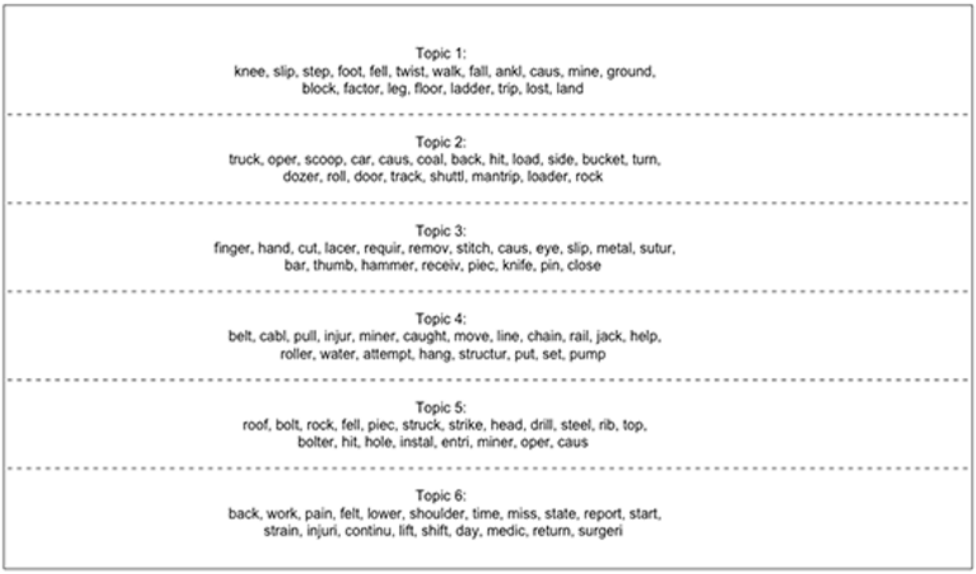
4 Findings

Space limitations require truncation of our findings merely to report narrow features of the topic model estimated, a reporting approach that fits the exploratory aims of this study. In brief, a coherent model of topics was estimated form the MSHA injury text narratives, Topics modeled varied by location of the injury on the mine property, degree of injury, and over years.

Six topics were distilled from modeling of the MSHA injury narrative text. Shown in Figure 1 are the lemmatized words associated most frequently with each of the six topics. Review of the words associated with each topic, along with inspection of the full text of a sample of the narratives closely associated with each topic (narratives not shown here), yields a sense of the meaning that can summarize each topic.

Topic 6 included the highest proportion of words from all of the MSHA injury narrative text that was clustered on a single topic and to reduce the length of this account of our work, we focus the remainder of our findings on Topic 6. Similar displays of findings for all six topics are provided in an online repository [26]. One way to generally describe Topic 6 is that it covers a theme involving incidents resulting in strains and sprains of musculoskeletal systems.

Shown in Figure 2, Figure 3, and Figure 4 are variations in Topic 6 by, respectively, location of the injury on the mine property, degree of injury to the miner, and the year of injury occurrence. Each figure plots point estimates and interval estimates of variation around plotted points. For instance, overlap of interval estimates in Figure 2 indicate that injuries in underground mines are relatively less frequently described by Topic 6 than injuries occurring in surface or other locations. Injuries resulting disability are described more strongly by Topic 6 than are other degrees of injury (Figure 3). Displayed in Figure 4 is a rising assignment of injuries associated with Topic 6 between 2000 and 2015.



Topic 1:
knee, slip, step, foot, fell, twist, walk, fall, ankl, caus, mine, ground, block, factor, leg, floor, ladder, trip, lost, land

Topic 2:
truck, oper, scoop, car, caus, coal, back, hit, load, side, bucket, turn, dozer, roll, door, track, shuttl, mantrip, loader, rock

Topic 3:
finger, hand, cut, lacer, requir, remov, stitch, caus, eye, slip, metal, sutur, bar, thumb, hammer, receiv, piec, knife, pin, close

Topic 4:
belt, cabl, pull, injur, miner, caught, move, line, chain, rail, jack, help, roller, water, attempt, hang, structur, put, set, pump

Topic 5:
roof, bolt, rock, fell, piec, struck, strike, head, drill, steel, rib, top, bolter, hit, hole, instal, entri, miner, oper, caus

Topic 6:
back, work, pain, felt, lower, shoulder, time, miss, state, report, start, strain, injuri, continu, lift, shift, day, medic, return, surgeri

Fig. 1. Words associated most frequently with each of the six topics modeled.

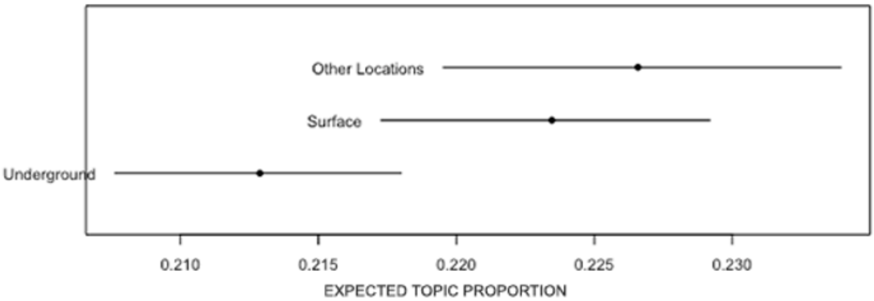


Fig. 2. Variation in Topic 6 by location of injury on the mine property.

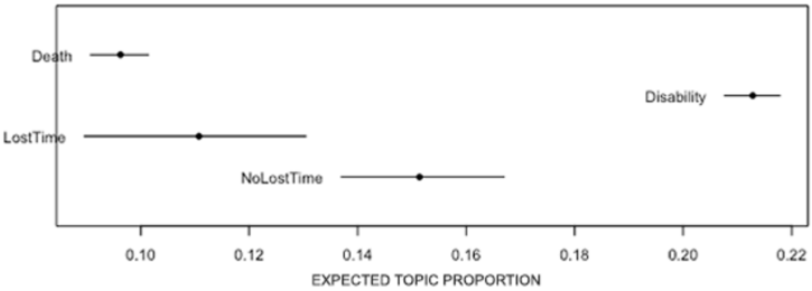


Fig. 3. Variation in Topic 6 by degree of injury to miner.

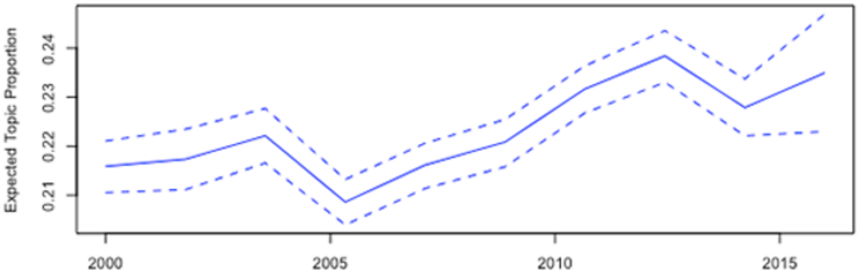


Fig. 4. Variation in Topic 6, 2000-2015.

4 Discussion

Modeling of topics latent in the MSHA text narratives of injuries was accomplished using unsupervised machine learning methods that rely on LDA analytics. One topic, Topic 6, displayed explainable and coherent variations over three metadata elements, adding to our confidence that these text narratives can help describe and explain correlates of injuries to coal miners. In a nutshell, the narratives clustered around Topic 6 seem to refer most frequently to surface or other locations rather than underground locations that resulted in disability and that, also, increased secularly over time.

Through this study, we constructed a brief, verbal, human-understandable statement of meaning for Topic 6 based on our review of words associated with the topic and on our review of a sampling of narratives that are representative of the topic. This meaning helps to describe the topic in general ways. Depending on analytical aims, assigning a meaning to a topic is not necessary because the mathematical algorithm generated to model topics can have greater analytical use than can verbal descriptions. For instance, supervised machine learning models can use the proportions of total words in all injury narratives assigned to each topic as features (i.e., independent variables) to predict an outcome (e.g., injury incidence or severity) in a supervised machine learning effort.

A mathematical algorithm derived for a topic model could help identify miners at risk for injury as well as inform decisions about allocation of injury prevention resources to high risk groups of miners, mining techniques, equipment, or personal protective devices and clothing based on knowledge identified from supervised machine learning. An illustrative use of probabilistic assignments of words to topics is an algorithm constructed from 1 million customer reviews about video games to predict the helpfulness of new customer reviews to improve the games [27].

The balance necessary to judge the relative usefulness of verbal or algorithmic results of topic models is difficult to weigh. On one hand, complex algorithmic models almost are impossible for humans to examine and understand and, therefore, are nothing more than “black box” processes that require blind trust from users. On the other hand, use of verbal accounts of topic models requires human intuition that is naively impressionistic, not reproducible over examiners of the same model, and prone to confirmation and other biases [28].

5 Conclusions

We conclude that a more complete topic modeling of MSHA injury narratives is justified, given the modeling success enjoyed in this exploratory effort. Future efforts could repair the body of narratives delivered by MSHA by correcting obvious spelling errors made by the reporters of the narratives. Also, the sensitivity of a topic model could be tested by adding stop words to remove terms that are common in mining (e.g., such as "mine," "miner," and "mining"), but that perhaps are uninformative in discriminating among distinct and separate topics.

The *MSHA Accident Injuries Data Set* contains other metadata significant for use as covariates to identify variations in topics modeled. For example, the MSHA data set includes information, especially for underground mining, about the specific workplace location of an injury, method of mining applied when an injury happened, a detailed accounting of specialized equipment in use, part of body injured, gender of a miner, time of day of an injury, and seasonal period when an injury occurred. In addition, these metadata are linkable with other quantitative and qualitative MSHA data about size of the workforce and the level of production at a mine as well as citations by government inspectors and regulators of legal culpability for violations of safety, health, and other requirements at a mine.

Because surface mining of coal shares similar processes, equipment, and technologies with work to prepare construction sites, launching a line of inquiry about topic model differences between on MSHA text narratives and construction injury narratives assembled by the U.S. Occupational Safety and Health Administration might prove fruitful for injury prevention efforts for both industries. Additional risk analysis approaches might also prove advantageous for analysis of the risks of coal mine injuries [29, 30].

References

1. U.S. Energy Information Administration. How large are U.S. coal reserves? (2018) https://www.eia.gov/coal/reserves/
2. U.S. Bureau of Labor Statistics. Injuries, illnesses, and fatalities in the coal mining industry. (2010) https://www.bls.gov/iif/oshwc/osh/os/osar0012.htm
3. U.S. Mine Safety and Health Administration. Mine safety and health at a glance. (2017) https://www.msha.gov/data-reports/statistics/mine-safety-and-health-glance
4. U.S. Mine Safety and Health Administration. Coal employment and production. (2017) https://dol-msha-peir-mshagov-prod.s3.amazonaws.com/s3fs-public/Data_Reports/DEC_15_2016_Historical_MIWQ_Employment_and_Production.pdf
5. U.S. Bureau of Labor Statistics. Coal mining injuries, illnesses, and fatalities fact sheet. (2010) https://www.bls.gov/iif/oshwc/osh/os/osar0012.pdf
6. B. Nowrouzi, M. Rojkova, J. Casole, B. Nowrouzi-Kia. A bibliometric review of the most cited literature related to mining injuries. Int. J. Min. Reclam. Env. **31**, 276-285 (2016) doi:10.1080/17480930.2016.1138850
7. S, Inzalkar, J Sharma. A survey on text mining techniques and application. Int. J. Res. In Sci. & Eng. **24**, 1-14 (2015)
8. D. Brown. Text mining the contributors to rail accidents. IEEE T. Intell. Transp. Sys. **17**, 346-355 (2016). doi:10.1109/tits.2015.2472580
9. T. Bunn, S. Slavova, L. Hall. Narrative text analysis of Kentucky tractor fatality reports. Accident Anal. Prev. **40**, 419-425 (2008) doi:10.1016/j.aap.2007.07.010
10. J. Dement, H. Liscomb, L. Li, C. Epling, T. Desai. Nail gun injuries among construction workers. Appl. Occup. Environ. Hyg. **18**, 374-383 (2010) doi:10.1080/10473220301365
11. K. McKensie, D. Scott, M. Campbell, R. McClure. The use of narrative text for injury surveillance research: A systematic review. Accident Anal. Prev. **42**, 354-363 (2010) doi:10.1016/j.aap.2009.09.020

12. K. Biswas, R. Zipf. Root causes of groundfall related incidents in U.S. mining industry. 22nd Int. Conf. on Ground Contr. in Min. (2003) https://www.cdc.gov/niosh/mining/userfiles/works/pdfs/rcogr.pdf
13. A. Samuel. Some studies in machine learning using the game of checkers. IBM J. Res. Dev., **3**, 211-229 (1959)
14. D. Blei. Probabilistic topic models. Commun. ACM. **55**, 77-84 (2012) doi:10.1145/2133806.2133826
15. T. Hastie, R. Tibshirani, J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). (Springer, 2009)
16. S. Debortoli, O. Müller, I. Junglas, J. Brocke. Text mining for information systems researchers: An annotated topic modeling tutorial. Comm. Assoc. Infor. Sys. **39**, 7 (2016) doi:10.17705/1cais.03907
17. S. Kotsiantis, I. Supervised machine learning: A review of classification techniques. Informatica. **31**, 249-268 (2007) https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf
18. U.S. Mine Safety and Health Administration. Accident Injuries Data Set (issued annually). https://arlweb.msha.gov/OpenGovernmentData/DataSets/Accidents.zip
19. Section 103 of the Federal Mine Safety and Health Amendments Act of 1977, Public Law 95-164. https://arlweb.msha.gov/regs/30cfr/
20. Mine Accident, Injury, and Illness Report MSHA Form 7000-1. https://arlweb.msha.gov/forms/70001INB.HTM
21. C. Chae, D. Passmore. *MSHA data analyzed for topic modeling*. (2018). https://osf.io/ab8rk/
22. L. Kurgan, P. Musilek. A survey of knowledge discovery and data mining process models. Knowl. Eng. Rev. **21**, 1-24 (2008).
23. B. Frigyik, A. Kapila, M. Gupta. Introduction to the Dirichlet distribution and related processes. (2010) https://goo.gl/LKk79j
24. R Core Team. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria, 2013) http://www.R-project.org/.
25. C. Chae, D. Passmore. *R code for topic modeling MSHA data*. (2018). https://osf.io/ab8rk/
26. C. Chae, D. Passmore. *Supplementary figures that report findings for topic modeling MSHA data*. (2018). https://osf.io/ab8rk/
27. O. Müller, J. vom Brocke, S. Debortoli. Utilizing big data for information systems research: Challenges, promises and guidelines. European Journal of Information Systems, **25**, 289–302 (2016) doi:10.1057/ejis.2016.2
28. J. Grimmer, B. Stewart. Text as data: The promise and pitfalls of automated content analysis for political texts. Political Analysis, **21**(3), 267–297 (2013) doi:10.1093/pan/mps028.
29. BorkovskayaV.G., Bardenwerper W., Roe R. Interactive teaching of risk management in the Russian construction industry. IOP Conf. Series: Materials Science and Engineering, **365** (2018) 062030 doi:10.1088/1757-899X/365/6/062030
30. Borkovskaya V.G., Passmore D. Behavioral engineering model to identify risks of losses in the construction industry. Advances in Economics, Business and Management Research (France-Netherlands). Atlantis Press. In press.