Research Questions:

Idea 1: Classify documents as fatality and non-fatality (based on whether they mention death related keywords), then find what keywords are associated with each group (e.g. what conditions or causes are more linked to fatal accidents vs. nonfatal)

Idea 2:

Literature Review:

Deep Learning and Network Analysis – Classifying and Visualizing Accident Narratives in Construction

Botao Zhong, Xing Pan, Peter Love, Lieyun Ding, Weili Fang

1. Introduction
   1. Significance of Study
      1. “More often than not managers are not provided with timely and fact-based information about accident causation as it is typically in an unstructured or semi-structure format.”
      2. “Managers often glimpse over reports as they can be rich in content and in some cases lengthy. As consequence valuable information that describes circumstances and conditions may be overlooked.”
   2. Goal of the Paper
      1. “Not to provide new insights into the causes of accidences per se, but demonstrate that deep learning can be used to extract unstructured safety data from accident text narratives automatically.”
2. Related Work
   1. Classify workers’ comp claims into categories
   2. K-means-based clustering approach to accident texts to support safety inspections
   3. NLP rule-based automated content analysis to extract precursors and outcomes from injury texts
3. Research Approach
   1. Data Material and Preprocessing
      1. Data source: Unlabeled OSHA reports
      2. Manually labeled half of the sample size
         1. Primarily labeling the primary cause of the incident
      3. Tokenization
      4. Stop word removal
      5. Train/test/validation split
      6. N-grams
4. CNN-based Classification of Accident Narratives
   1. “CNN can automatically determine discriminative phrases in text using a max-pooling layer, instead of through manual feature engineering with domain knowledge”
   2. CNN-based Deep Learning Model
      1. Word Embedding
         1. The process of converting words to a vector matrix
         2. They did a special word vectorization process that gets over word vectorization challenges such as loss of word order and oversize of dimensionality
      2. Convolution Kernel and MLP Classifier
         1. Created a neural network model that takes in text and spits out a classification
         2. Basic cross validation and hyper parameter tuning
      3. Model Testing and Evaluation
         1. Compare to some other paper’s “shallow learning models”
         2. CNN outperformed
         3. Confusion matrix
5. Topic mining and LDA-based network analysis
   1. LDA model finds topics and their corresponding keywords
   2. They minimize time spent manually doing this
   3. Insight: caught in between is often associated with body parts
   4. Insight: falls is often associated with things that fall e.g. towers
   5. LDA-based Network Analysis
      1. Word Co-occurrence Network uses graphs as a means to represent words as nodes and identify their relationships with one another
6. Discussion
7. Limitations
   1. Some accidents might fit multiple categories, but we just put it into one. Need to develop a multi-label classifier.
8. Conclusion

Text Mining Analysis of U.S. Department of Labor’s MSHA Fatal Accident Reports for Coal Mining

E. Tarshizi, M. W. Buche, B. Inti and R. Chappidi

1. Introduction
   1. Accidents are bad
   2. MSHA monitors accidents for miners’ safety
   3. Objective: “Identify opportunities resulting from previously unexplored directions in order to provide additional insights into potential safety recommendations”
2. Literature Review
   1. Tirunagari conducted a study investigating maritime accidents using text mining; he identified causes for accidents
      1. Used:
         1. Naïve Bayes (classifier)
         2. SVM (classifier)
         3. Connectives Method (cause and effect terms)
   2. Panthi and Ahmed identified the factors that led to the accidents under consideration and devised necessary measures to prevent repetition of those accidents
   3. Nakata also tried to do cause and effect
3. Data Description
   1. Source: MSHA reports from 2010-2018, totaling 119
   2. Consists of date of accident, age and experience of the victim, brief explanation of hwo the accident happened, describing the cause and conditions resulting in the mishap
   3. Report is a paragraph in length (3-5 lines of text)
4. Text Mining Methodology
   1. Text mining is good
5. Text Mining Analysis and Techniques
   1. Data Importing
      1. Summary report of each accident was stored in an individual text file, all files were imported into an R dataframe with col 1 being doc ID and second col being report text
   2. Data Preprocessing
      1. Remove numbers (dates, age, years of experience)
         1. removeNumbers in R
      2. Remove punctuation
      3. Remove whitespace
   3. Data Structuring
      1. Word Vectorization
         1. Text2vec package
      2. Tidytext package
      3. Tokenization
      4. Remove stopwords
   4. Exploratory Analysis
      1. Most Frequent Words
         1. Categorize your words (vehicles, actions)
      2. TF-IDF
      3. Correlation Network Plot
   5. Conclusion and Future Study
      1. Considerable portion of fatalities happen due to worker misalignment with equipment and vehicles in the work environment