# Research Questions

**Idea 1:**

**Classify documents as fatality and non-fatality (based on whether they mention death related keywords, e.g. “died”, “fatal”), then find what keywords are associated with each group (e.g. what conditions or causes are more linked to fatal accidents vs. nonfatal) /// predict whether an accident will be fatal or nonfatal based on text (minus death keywords)**

Idea 2:

Compare the accident reports year over year for similarities and, more importantly, differences

Idea 3:

Classify documents based on root cause and try to predict root cause of accident based on description of accident

# 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

Advanced Application of Text Analytics and NLP on MSHA Metal and Nonmetal Fatality Reports

K. V. Raj & E. Tarshizi

1. Introduction
   1. Data about accident volume according to MSHA
   2. About half of all US mining fataliesi were due to accidents involving powered haulage
   3. This paper focuses on accidents that were classified as powered haulage and machinery
2. MSHA Accident Reports
   1. Data is from MSHA reports
3. Methodology
   1. Step 1 acquire text data
   2. Step 2 text and characters were extracted from the text file and put into tabular format
   3. Step 3 apply text mining and NLP techniques on the text data
   4. Text Mining
      1. Tokenization
      2. Convert to lowe case
      3. Remove punctuation
      4. Stop words removed
      5. Removed corpuse stop words (e.g. victim, employee)
      6. Reduced words to root form
      7. Created bi grams
      8. Top modeling
         1. Cluster documents based on co-occurrence of words
         2. Defines a number of “topics” that the texts can be grouped into
      9. LDA
         1. With number of topics decided, assign topics using LDA
   5. Natural Language Processing
      1. POS tagging
      2. Information Extraction
         1. Pulls out features such as name, place, organization, date time
         2. Created an equipment dictionary and could pull that out as well
         3. Makes a chunk that has the TLDR version of the report
4. Conclusions

An Exploration of Text Mining of Narrative Reports of Injury Incidents to Assess Risk

D. Passmore, C. Chae, Y. Kustikova, R. Baker, J. Yim

1. Problem
   1. Coal is big
   2. But it’s hazardous (stats)
   3. Knowledge can help
   4. Text is an underutilized source of said knowledge
2. Focus
   1. Purpose
   2. Exploratory Approach
      1. Goal: 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
3. Method
   1. Data
      1. MSHA Accident Injuries Data Set
   2. Mining of Text
      1. Basic clean up
      2. LDA
   3. Findings
   4. Discussion
   5. Conclusion

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

R. Ganguli, P. Miller, R. Pothina

1. Introduction
   1. Goal: Classify accidents into accident types
2. Importance of This Paper
3. Research Methodology
   1. MSHA Accident Database
   2. Random Forest Classifier
4. Results
   1. Performance within MSHA Data
   2. Performance on Non-MSHA Data
5. Discussion
6. Conclusions