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

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