Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 145 (2016) 1588 – 1593

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Sustainable Design, Engineering and Construction

Analyzing Arizona OSHA Injury Reports Using Unsupervised Machine Learning

Abbas Chokor[a]*, Hariharan Naganathan[a], Wai K. Chong[a], Mounir El Asmar[a]

[a]*School of Sustainable Engineering and the Built Environment, Arizona State University, 660 South College Ave., Tempe, AZ 85287, United States*

Abstract

As the construction continue to be a leading industry in the number of injuries and fatalities annually, several organizations and agencies are working avidly to ensure the number of injuries and fatalities is minimized. The Occupational Safety and Health Administration (OSHA) is one such effort to assure safe and healthful working conditions for working men and women by setting and enforcing standards and by providing training, outreach, education and assistance. Given the large databases of OSHA historical events and reports, a manual analysis of the fatality and catastrophe investigations content is a time consuming and expensive process. This paper aims to evaluate the strength of unsupervised machine learning and Natural Language Processing (NLP) in supporting safety inspections and reorganizing accidents database on a state level. After collecting construction accident reports from the OSHA Arizona office, the methodology consists of preprocessing the accident reports and weighting terms in order to apply a data-driven unsupervised K-Means-based clustering approach. The proposed method classifies the collected reports in four clusters, each reporting a type of accident. The results show the construction accidents in the state of Arizona to be caused by falls (42.9%), struck by objects (34.3%), electrocutions (12.5%), and trenches collapse (10.3%). The findings of this research empower state and local agencies with a customized presentation of the accidents fitting their regulations and weather conditions. What is applicable to one climate might not be suitable for another; therefore, such rearrangement of the accidents database on a state based level is a necessary prerequisite to enhance the local safety applications and standards.

Keywords: Accident; Construction; Data Analysis; Injury; Natural Language Processing; OSHA; Safety.

* Corresponding author. Tel.: +1-480-758-1427
E-mail address: achokor@asu.edu

doi:10.1016/j.proeng.2016.04.200

1. Motivation

Construction safety issues and its related injuries and fatalities have significantly reduced after implementing the rules and standards of the Occupational Safety and Health Administration (OSHA). These standards involve injury prevention strategies such as job hazard analysis or substance abuse programs that are primarily tied back to the management practices [1]. The analysis is performed with the data collected by OSHA every year from different construction firms and has made a mandatory rule where all the incidents must be reported to OSHA. Construction fatalities and injuries result in immense societal costs, totalling approximately $15 billion in lost revenue every year [2]. Despite the abundant research that has been motivated by the afore-mentioned alarming injury and fatality rates, safety performance in construction has been plateauing in recent years, and the implementation of effective injury prevention practices has reached saturation [3].

The focus of this paper is driven by [4], that implements Natural Language Processing (NLP) into the construction safety incidents. NLP as a widespread analytical approach is commonly used in the information technology and retail product development and marketing team to identify the products that attracts public, which can bring in more people and profit. While the previous studies have used supervised machine learning algorithms that are time consuming, this paper investigates the strength of unsupervised machine learning accompanied with NLP in grouping the accident reports. The study applies a data-driven unsupervised K-Means-based clustering on the OSHA construction accident reports, collected from the state of Arizona. The results of the conducted analysis showcase the power of the proposed methodology in defining a customized state-based classification of the accidents. While a similar analysis can be applied for other states, the findings of this paper helps a better understand for construction accidents, from where the decision makers and rescue leaders can take real time decision to minimize the number injuries and fatalities.

2. Related Work

Safety communication has been a practice in the construction industry which is sharing of safety knowledge that occurs through channels that are pre-established specifically for safety [5]. Others showed that the most successful supervisors tend to have open discussions with workers from different trades about safety issues and provide necessary advice [6], [7]. While communication with the supervisors is critical to reduce safety issue, it cannot stop the accidents if the workers are not aware of the danger. In order to overcome this issue, standards are established along with safety programs to learn, identify the risks and to perform quick mitigation of the issue. There are accidents and deaths happening every year even after establishing OSHA standards which effectively means that the administration has a need to look at the safety concerns with different approaches than traditional way. Thus, database on accidents, injuries and other incidents are collected by OSHA from 1970 from all parts of United States. Using this database, the types of injuries and their impact can be known by which the project managers can alert their workers on safety precautions.

There are numerous approaches adopted by different researchers in handling and converting the database into useful information. From basic statistics to high level computerized automated techniques, researchers have implemented numerous techniques striving to reduce the accidents, by identifying the patterns and predicting the parent factors (weather, location and time) for the causality. For example, Hallowell and Gambatese identified the risk factors involved in concrete framework [8]. Shin et al. identified the factors that affect safety in tower crane installation and dismantling in the construction industry [9] while Shapira et al [10] developed a integrative model for evaluating the tower crane safety factors.

The disadvantage of these models developed by researchers are that they are not based on empirical data and has limited scope of application[3], [4]. Thus, developing models for each accident is laborious and time consuming. Also, the dynamics of construction work are not well captured [11] and thus there[4] is a need to learn the patterns from the data. To overcome these limitations, Esmaeili and Hallowell [3], [11] proposed a unified attribute-based framework that allows standard risk factor and outcome variables to be extracted from naturally occurring accident reports. However, formal analysis of the database is always time consuming and requires high skilled labor which is again an expensive methodology for the administration to perform [1], [4]. To overcome this, [4] proposed an automated model that can perform feature extraction from unstructured injury reports using NLP as a base processing technique.

While this methodology of addressing construction worker safety is successful in reducing the fatality over the years, there is still a need to a better understanding of the incidents. Moreover, the large databases of OSHA alongside the difficulty in reading and interpreting the reports code are a challenge for decision makers. Although construction accidents reasons are well identified and recognized, they vary from a state to another based the weather conditions and local regulations. While previous studies have done supervised machine learning algorithms, none of the reviewed paper has integrated NLP and clustering to address the construction accidents on a state level.

3. Objectives and Methodology

Given the big data of OSHA historical events and reports, a manual analysis of the fatality and catastrophe investigations content is a time consuming and expensive process [12]. This paper aims to evaluate the strength of unsupervised machine learning and Natural Language Processing (NLP) in supporting safety inspections and reorganizing accidents database on state level. The methodology used to collect data and analyze accident reports is detailed next and entails four steps: (1) collecting construction accident reports from OSHA Phoenix, AZ office; (2) preprocessing the accident reports and weighting terms to implement Machine Learning (ML) algorithms; (3) applying a data-driven unsupervised K-Means-based clustering approach to segregate the collected reports; (4) analyzing the different identified clusters and discussing the potential uses of the presented methodology in emphasizing data patterns.

3.1. Accident Reports Collection

Fatality and catastrophe investigation summaries, (OSHA 170 form), are developed after OSHA conducts an inspection in response to a fatality or catastrophe. The summaries provide a complete description of the incident, generally including events leading to the incident and causal factors. Summaries currently available include completed investigations from 1984 through 2015 [13]. These summaries undergo a process for screening and revising the information and adding keywords. These accident reports can be easily searched by keyword, text in the summary description, event date, and industry (SIC). First, to meet the objectives of this study, the authors selected OSHA Phoenix office, covering the accidents that occurred in the state of Arizona. Second, all the summaries, including the accident description, event date, and SIC were collected. Third, the construction summaries were filtered by searching for *Division C: Construction* reports having SIC codes ranging between 1521 and 1799. Out of the 1044 collected reports, 513 accidents or 49.1% were in construction and will be considered for this study.

3.2. Data Preprocessing and Terms Weighting

This phase intends to remove meaningless data from accident descriptions and retrieve relevant features from raw data records. In the proposed methodology, data preprocessing consists of three steps:

A. Tokenization: It is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens [14]. The list of tokens becomes input for further processing such as parsing or text mining.

B. Stemming (lemmatization): English words like ‘accident’ can be inflected with a morphological suffix to produce *accidents, accident’*, *accident’s*. . Since the meaning of different words could be the same but their form different, it is necessary to identify each word from using its stem form. As stemming reduces the number of unique vocabulary items that need to be tracked and speeds up a variety of computational operations, the Snowball Stemmer has been applied on the collected accident descriptions [15].

C. Stopwords removal: Stopwords are nuisance and often do not carry much meaning [16]. For many NLP purposes, stopwords removal is a common preprocessing step, some of these words are: *the*, *a*, *of*, *in*, etc...

After preprocessing the accidents description, a term frequency-inverse document frequency (TFIDF) analysis is implemented to weight terms for information retrieval. The TFIDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general [17]. Thus, the following forumla [18] was applied to calculate the weight of terms and therefore create the TFIDF matrix:

$$TFIDF = TF \times IDF = \frac{t_j}{f_j} \times log\frac{N}{n_t}$$

where t_j is the frequency of term t in report R_j, f_j is the total instances of terms in report R_j, N is the total reports number in the set, and n_t is the total number of reports with term t_j.

3.3. K-Means Documents Clustering

K-means clustering is a method commonly used to automatically partition a data set into K groups [19]. Using the TFIDF matrix, the study approach consists of modifying the traditional K-Means algorithm to avoid predetermining the number of clusters. The model development comprises including the cluster validation in the loop for a specified interval in order to decide the K, the suitable number of clusters. This task involves four steps as follows:

A. Choose the top K reports, having the highest Euclidean length L_j, to be the initial centroids of K clusters. If t_j is the frequency of term t in R_j, the Euclidean length of report j, L_j, is calculated as follows:

$$L_j = \frac{1}{\sqrt{t_j^2}}$$

B. Assign each report to the centroid with the highest cosine similarity value. The similarity value, $Sim(R_j,R_c)$ between the terms vector of report j, $V(R_j)$, and the terms vector of the centroid report c, $V(R_c)$, is equivalent to:

$$Sim(R_j, R_c) = cos\,\langle\, V(R_j), V(R_c)\,\rangle$$

C. Compute the similarity value, $S(R_j,C)$ between report R_j and cluster C of length i, C={R_1, R_2, ..., R_{i-1}, R_i}as follows:

$$S(R_j, C) = \frac{1}{i} \times \sum_{k=1}^{i} Sim(R_j, R_k)$$

Choose report R_h with the highest similarity value, $S(R_h, C)$, to be the centroid of C.

D. If all new centroids are the same as the old ones, exit the loop. Otherwise, return to step B.

3.4. Clusters Analysis and Discussion

This study attempts to explore new patterns behind the unsupervised clustering of accident description reports. For each clusters, the key terms are extracted. Moreover, multidimensional scaling is used to reduce the dimensionality within the corpus and visualize levels of similarity of the dataset individual reports. An analysis of the key terms and similarities paves the way for a deep understanding of the clusters characteristics. Later, a discussion of the clustering potential in in supporting and retrieving data patterns is presented.

4. Preliminary Results and Discussion

This section presents the results of K-Means documents clustering, applied on accident description reports. Without predetermining the number of clusters, the presented algorithm found K to be equal to four clusters. A two-dimensional visualization of the clusters is shown in Figure 1. As shown in the figure, the accident description reports segregate into four distinct and well-defined clusters, containing reports of a single class. Table 1 below lists the top ten key terms within each identified clusters.

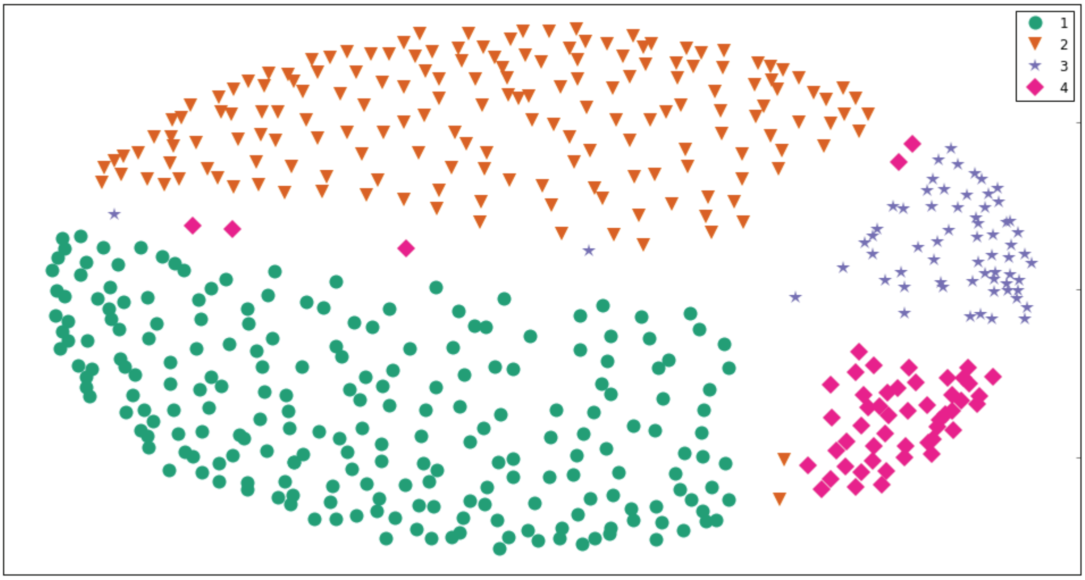

Figure 1. Two-dimensional visualization of the clusters

A deep examination of the clusters shows the accident description reports to be clustered based on the accident attributes. An analysis of key terms within each cluster each shown below:

- Cluster 1 – “falls”: involves the accidents of falling due to the loss of balance, excessive heat, or the absence of fall protection. Such accidents include the fall from roof, ladder, scaffold, or truss. Cluster 1 comprehends 42.9% of the total reports.
- Cluster 2 – “struck by object”: includes the accidents of workers who are strucked and crashed by falling objects and equipment. Cranes, walls, and trucks are causing some of these injuries. Cluster 2 contains 34.3% of the total reports.
- Cluster 3 – “electrocutions”: covers the accidents of workers who were shocked or electrocuted. It also comprises the burns and fires caused by electrocution. Cluster 3 is found in 12.5% of the reports.
- Cluster 4 – “trenches collapse”: comprises the accident of being caught or buried in trenches. It covers the accidents happening during excavation due to backhoes and collapse of unstable soils. Cluster 4 covers 10.3% of the total reports.

Table 1. Top ten key terms for the identified clusters

Falls	Struck by Object	Electrocutions	Trenches Collapse
Balance	Amputated	Burn	Backhoe
Falling	Crane	Electrical	Buried
Fracture	Crushed	Electrocuted	Caught
Heat	Equipment	Fire	Cave-in
Ladder	Falling	Grounding	Collapse
Opening	Head	Line	Excavation
Protection	Run	Lockout	Pipe
Roof	Struck	Overhead	Shoring
Scaffold	Truck	Power	Trench
Truss	Wall	Shock	Unstable

While the importance and contribution of the proposed methodology is underlined in the savings of time and resources to manually labeling the OSHA database, the new method introduced in this paper can be replicated for another state. For instance, the construction accidents in Arizona are most probably to be different than those of Wisconsin and Michigan due to the difference in weather conditions. Yet, the results of construction accidents in Arizona are comparable to the percentages reported by the Bureau of Labor Statistics (BLS) [20]. In 2014, the BLS reports the leading causes of construction accidents to be falls (39.9%), electrocutions (8.5%), struck by Object (8.4%), and caught-in/between (1.4%).

5. Conclusions and Future Work

This paper showcases the strength of unsupervised machine learning and NLP in supporting safety inspections, and reorganizing accidents database on a state based level. By taking the construction accident description that occurred in the state of Arizona, as an example, the proposed K-Means-based clustering approach was able to rearrange the accidents based on the type of accidents. The results show the construction accidents to be occurring because of four main reasons: falls, struck, electrocutions, and trenches collapse. The study highlights the need to customize the presentation of the accidents to fit the conditions and characteristics of construction within each state. Customizing the presentation of the accidents is applicable not only to the accidents in the state of Arizona but also to any other state or country around the world. Although this paper highlights the capacity of unsupervised machine learning in grouping similar accidents and discovering the main types, this preliminary study has limitations that need to be addressed in a larger research effort focusing on the topic: the sample size used in this study is limited to 513 reports and the paper investigates the reports from only one specific geographical area. The preliminary findings of this study are currently being improved by investigating a larger sample of accident reports and states, and also considering testing additional types of unsupervised machine learning algorithms.

References

[1] M. Zeynalian, B. Trigunarsyah, and H. R. Ronagh, “Modification of Advanced Programmatic Risk Analysis and Management Model for the Whole Project Life Cycle ’ s Risks,” *J. Constr. Eng. Manag.*, vol. 138, no. January, pp. 51–60, 2013.

[2] U.S. Bureau of Labor Statistics, “Census of fatal occupational injuries,” 2015.

[3] B. Esmaeili and M. Hallowell, “Attribute-Based Risk Model for Measuring Safety Risk of Struck-by Accidents,” in *Construction*, 2012, no. 2012, pp. 289–298.

[4] A. J. P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, “Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports,” *Autom. Constr.*, vol. 62, pp. 45–56, 2016.

[5] R. Alsamadani, M. Hallowell, and A. N. Javernick-will, “Measuring and modelling safety communication in small work crews in the US using social network analysis,” *Constr. Manag. Econ.*, vol. 31, no. 6, pp. 568–579, 2013.

[6] R. S. J. Wyer, “Effects of incentive to perform well, group attraction, and group acceptance on conformity in a judgmental task.,” *J. Pers. Soc. Psychol.*, vol. 4, no. 1, pp. 21–26, 1966.

[7] M. J. Smith, H. H. Cohen, A. Cohen, and R. J. Cleveland, “Reprint of ‘Characteristics of Successful Safety Programs,’” *Journal of Safety Research*, 2013.

[8] M. R. Hallowell and J. A. Gambatese, “Construction Safety Risk Mitigation,” no. December, pp. 1316–1323, 2009.

[9] I. J. Shin, “Factors that affect safety of tower crane installation/dismantling in construction industry,” *Saf. Sci.*, vol. 72, pp. 379–390, 2015.

[10] A. Shapira, M. Simcha, and M. Goldenberg, “Integrative model for quantitative evaluation of safety on construction sites with tower cranes,” *J. Constr. Eng. Manag.*, vol. 138, no. 11, pp. 1281–1293, 2012.

[11] B. Esmaeili, M. R. Hallowell, and B. Rajagopalan, “Attribute-Based Safety Risk Assessment . I : Analysis at the Fundamental Level,” *J. Constr. Eng. Manag.*, vol. 141, no. 8, pp. 1–15, 2015.

[12] M. V. Prades, "Attribute-based Risk Model for Assessing Risk to Industrial Construction Tasks," University of Colorado at Boulder, 2014.

[13] Occupational Safety & Health Administration, "OSHA Data & Statistics," [Online]. Available: https://www.osha.gov/oshstats/. [Accessed 9 March 2016].

[14] C. Silva and B. Ribeiro, "The Importance of Stop Word Removal on Recall Values in Text Categorization," in The international joint conference on Neural Networks, 2003.

[15] M. F. Porter, Snowball: A language for stemming algorithms, 2001.

[16] S. Porkodi, "Rule based approach for constructing gene/protein names dictionary from medline abstract," International Journal of Advances in Computing and Information Technology , 2012.

[17] N. Jain, P. Agarwal and J. Pruthi, "HashJacker-Detection and Analysis of Hashtag Hijacking on Twitter," International Journal of Computer Applications, vol. 114, no. 19, 2015.

[18] G. Salton and C. Buckley, "On the use of spreading activation methods in automatic information," in The 11th annual international ACM SIGIR conference on Research and development in information retrieval, 1988.

[19] K. Wagstaff, C. Cardie, S. Rogers and S. Schrödl, "Constrained k-means clustering with background knowledge," in International Conference on Machine Learning, 2001.

[20] Bureau of Labor Statistics, "Census of Fatal Occupational Injuries Summary," 2014.