Accident Analysis and Prevention 108 (2017) 122–130

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Construction accident narrative classification: An evaluation of text mining techniques

Yang Miang Goh*, C.U. Ubeynarayana

Safety and Resilience Research Unit (SaRRU), Dept. of Building, School of Design and Environment, National Univ. of Singapore, 4 Architecture Dr., 117566, Singapore

ARTICLE INFO

Keywords:
Accident classification
Construction safety
Data mining
Support vector machine
Text mining

ABSTRACT

Learning from past accidents is fundamental to accident prevention. Thus, accident and near miss reporting are encouraged by organizations and regulators. However, for organizations managing large safety databases, the time taken to accurately classify accident and near miss narratives will be very significant. This study aims to evaluate the utility of various text mining classification techniques in classifying 1000 publicly available construction accident narratives obtained from the US OSHA website. The study evaluated six machine learning algorithms, including support vector machine (SVM), linear regression (LR), random forest (RF), k-nearest neighbor (KNN), decision tree (DT) and Naive Bayes (NB), and found that SVM produced the best performance in classifying the test set of 251 cases. Further experimentation with tokenization of the processed text and non-linear SVM were also conducted. In addition, a grid search was conducted on the hyperparameters of the SVM models. It was found that the best performing classifiers were linear SVM with unigram tokenization and radial basis function (RBF) SVM with uni-gram tokenization. In view of its relative simplicity, the linear SVM is recommended. Across the 11 labels of accident causes or types, the precision of the linear SVM ranged from 0.5 to 1, recall ranged from 0.36 to 0.9 and F1 score was between 0.45 and 0.92. The reasons for misclassification were discussed and suggestions on ways to improve the performance were provided.

1. Introduction

Workplace safety and health is a major concern in the construction industry in many countries (Zhou et al., 2015). To improve the industry's safety and health performance, the industry needs to learn from past accidents effectively (Chua and Goh 2004). However, accident reports are typically unstructured or semi-structured free-text data that require significant manual classification before statistical analyses can be conducted to facilitate interventions. These classification tasks are typically conducted at organizational and national levels. Due to the resource-intensiveness of the classification process, significant amount of resources need to be spent on classifying accident narratives, but the consistency of the classification is hard to be ascertained. On the other hand, organizations that choose not to classify accident narratives, will suffer loss of precious data for learning and accident prevention.

There had been an increased interest in automatic classification or auto-coding of accident narratives through the application of text mining techniques. These studies typically aim to improve the consistency, productivity and efficiency of accident narrative classification (e.g. Chen et al., 2015b; Marucci-Wellman et al., 2011; McKenzie et al., 2010b; Taylor et al., 2014; Tixier et al., 2016; Vallmuur 2015; Vallmuur et al., 2016). The results appear to be promising, but there are concerns that the success of automatic classification of accident narratives is very sensitive to the dataset and the effectiveness of classification algorithms may not be consistent across different datasets. There is also a wide range of text mining techniques and the usefulness of different techniques in the context of accident narrative classification need to be evaluated. Even though automatic classification of accident narratives does not generate new knowledge per se, it may be argued that with higher efficiency, more incident data can be collected and more detailed analytics can be conducted to produce useful insights that would not be available when fewer incidents were classified by human coders.

This study aims to evaluate the utility of various text mining classification techniques in classifying publicly available accident narratives obtained from the US OSHA website (Occupational Safety and Health Administration, 2016). This study also contributes to future studies on accident narrative classification by making available a dataset of 4470 construction accident narratives to other researchers (see Appendix A). The dataset includes 1000 narratives labelled in this study and 3470 narratives that were not labelled. The subsequent sections provide an overview of current text mining research on accident narratives, the text data that were used in this study, an overview of the

* Corresponding author.
E-mail address: bdggym@nus.edu.sg (Y.M. Goh).

http://dx.doi.org/10.1016/j.aap.2017.08.026
Received 15 May 2017; Received in revised form 24 July 2017; Accepted 26 August 2017

text mining techniques implemented in this study, the results of the evaluation, and discussion and recommendations for future research on text mining of accident narratives.

2. Literature review

2.1. Text mining techniques

Text mining is a well-researched field. One of the common tasks in text mining is classification of text data (Sebastiani 2002). Text classification is the task of assigning one or more class labels to a document using a predefined set of classes or labels. The supervised machine learning approach to text classification relies on an initial set of corpus (or collection of documents) with known class labels. This corpus is split into training and testing datasets in order to train and then ascertain the performance of the classifier. The classifier is trained by observing the characteristics of the training dataset through different machine learning algorithms.

Data for text classification is typically represented using vector space model. In this model, each document is represented as a vector of terms. Another way to look at these terms is that they are essentially a bag of words (Bird et al., 2009). Terms are features that represent a document, which could be a single word, phrase, or string. To distinguish between documents in a corpus, each feature for each document is given numeric values to show the importance of that term to the document (Keikha et al., 2008).A commonly used vector space model is the term frequency–inverse document frequency (tf-idf) representation (Peng et al., 2014). In the tf-idf representation, values, or x_{ik} weights, reflecting the importance of each given feature of a document is given by

$$x_{ik} = f_{ik} \times \log\left(\frac{N}{n_i}\right) \tag{1}$$

where f_{ik} is the frequency of feature i in document k, N is the number of documents in the corpus, and n_i is the number of documents where feature i occurs. Once the document is represented using a suitable vector space representation model, the data can be trained and classified using typical data mining techniques such as decision tree, neural network, support vector machine and Bayesian network (Raschka 2015; Witten 2011).

2.2. Performance metrics

This study adopts the use of recall, precision and F1 score (or F-measure) (Buckland and Gey 1994) to evaluate the performance of the machine learning algorithms experimented. Table 1 and Equations (2) to (4) define these metrics. Essentially, precision is a measure of how accurate the positive predictions are and recall is a measure of how many of the actual positives the model can identify (Williams 2011). F1 score combines precision and recall to provide an overall assessment of performance of the classifier. As these metrics are widely used and discussed in the literature, readers can refer to text mining or machine learning textbooks (e.g. Bird et al., 2009; Witten 2011) for their detailed description.

$$Precision = \frac{TP}{(TP + FP)} \tag{2}$$

Table 1
True and false positives and negatives (adapted from Bird et al. (2009)).

	Relevant	Irrelevant
Retrieved	True Positives (TP)	False Positives (FP) (Type I error)
Not retrieved	False Negatives (FN) (Type II error)	True Negatives (TN)

$$Recall = \frac{TP}{(TP + FN)} \tag{3}$$

$$F1Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

2.3. Past studies on accident narrative classification

There were several other studies that applied text mining techniques in the analysis of injury narratives. In Chen et al. (2015a), the study aimed to automatically classify narratives in emergency room medical reports into common injury cause codes. The authors argued that injury narratives have unique characteristics that make them different from general documents and a detailed experiment is needed to evaluate the usefulness of different text mining techniques for their dataset. It was found that the use of matrix factorization coupled with support vector machine (SVM), gave the best classification performance. The authors reported recall ranging from 0.48 to 0.94 and precision ranging from 0.18 to 0.95 for different classification labels. McKenzie et al., (2010a) also attempted to classify emergency department injury narratives for the purpose of injury surveillance to support an evidence-based public health response. The study compared keyword search, index search, and text mining. Text mining was conducted using a content text mining software, Leximancer (Leximancer Pty Ltd, 2016), and it was found that text mining approach provided the best performance. Bertke et al. (2012) made use of Naïve Bayesian classifiers to classify workers' medical compensation claims into three "claim causation" categories, i.e. musculoskeletal disorder (MSD), slip trip fall (STF), or others (OTH). The study found that the Naïve Bayesian classifier was able to achieve "approximately 90% accuracy" for MSD, STF and OTH classifications. However, it was observed that when OTH was being broken up into lower level classifications, the performance of the classifier dropped significantly.

Tanguy et al. (2015) evaluated an aviation safety report corpus, which contains 136,861 documents, using support vector machine. As part of the pre-processing and experimental design, numerous forms of text units were created. Some of the text units explored include word, word stems (e.g. "falling" is converted to its word stem, "fall"), and N-grams of words and stems. An N-gram refers to a set of N adjacent words (Bird et al., 2009). The study found that use of bi-gram and tri-gram of stemmed narratives produced the best results in their preliminary classifications. They constructed a binary classifier for each target label (e.g. air traffic management, bird strike, runway excursion and glider towing related event) and that means 37 classifiers were trained. However, the authors only reported the results for seven of the classifiers, the precision ranged from 0.6 to 0.96, recall was 0.36–0.93, and F1 score was 0.45–0.95. The authors highlighted that the performance for each classifier is dependent on issues such as "rarity, difficulty and inconsistency" of the text data.

Taylor et al. (2014) trained Fuzzy and Naïve Bayesian models to assign mechanism of injury and injury outcome for a set of fire-related near miss narratives obtained from the National Firefighter Near-Miss Reporting System. Their algorithms achieved sensitivity (same as recall) of between 0.602 and 0.74. Taylor et al. (2014) also made a comparison with five other studies and claimed that their findings are "are comparable with the growing body of seminal studies on narrative autocoding".

For the construction industry, there were several studies that utilized text mining approaches in areas such as dispute resolution (Fan and Li 2013), cost overrun (Williams and Gong 2014) document retrieval (Yu and Hsu 2013) and classification of field inspection records (Chi et al., 2016). Specifically in the domain of construction accident narrative classification, Tixier et al. (2016) made use of a customized term lexicon (keyword dictionary) as well as a set of rules to automatically classify construction incident narratives. The study was conducted on a dataset with 2201 accident narratives. The lexicons were