Accident Analysis and Prevention 88 (2016) 117–123

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Comparison of methods for auto-coding causation of injury narratives

S.J. Bertke[a,*], A.R. Meyers[b], S.J. Wurzelbacher[b], A. Measure[c], M.P. Lampl[d], D. Robins[d]

[a] National Institute for Occupational Safety and Health, Division of Surveillance, Hazard Evaluations, and Field Studies, Industrywide Studies Branch, 1090 Tusculum Ave, Cincinnati, OH 45226, United States
[b] National Institute for Occupational Safety and Health, Division of Surveillance, Hazard Evaluations, and Field Studies, Industrywide Studies Branch, Center for Workers' Compensation Studies, 1090 Tusculum Ave, Cincinnati, OH 45226, United States
[c] Bureau of Labor Statistics, Occupational Safety and Health Statistics, 2 Massachusetts Avenue, Washington, DC 20212, United States
[d] Ohio Bureau of Workers' Compensation, Division of Safety & Hygiene, 13430 Yarmouth Drive, Pickerington, OH 43147, United States

ARTICLE INFO

Article history:
Received 27 August 2015
Received in revised form
13 November 2015
Accepted 7 December 2015
Available online 30 December 2015

Keywords:
Auto-coding
Naïve Bayes
Regularized logistic regression
Injury narratives
Workers' compensation

ABSTRACT

Manually reading free-text narratives in large databases to identify the cause of an injury can be very time consuming and recently, there has been much work in automating this process. In particular, the variations of the naïve Bayes model have been used to successfully auto-code free text narratives describing the event/exposure leading to the injury of a workers' compensation claim. This paper compares the naïve Bayes model with an alternative logistic model and found that this new model outperformed the naïve Bayesian model. Further modest improvements were found through the addition of sequences of keywords in the models as opposed to consideration of only single keywords. The programs and weights used in this paper are available upon request to researchers without a training set wishing to automatically assign event codes to large data-sets of text narratives. The utility of sharing this program was tested on an outside set of injury narratives provided by the Bureau of Labor Statistics with promising results.

Published by Elsevier Ltd.

1. Introduction

The National Institute for Occupational Safety and Health (NIOSH) maintains a database from the Ohio Bureau of Workers' Compensation (OHBWC) containing over 1 million workers' compensation (WC) claims from 2001 to 2011. For tracking, trending and prevention purposes, it is crucial to identify the event or exposure leading to the injury for each claim. For example, an intervention program attempting to prevent back strains would benefit from the knowledge of the leading cause of this injury (i.e. overexertion, bodily reaction to slip/trip/fall, etc.). In the OHBWC database however, event/exposure was captured in a free-text field, usually filled out by the injured worker, describing the events leading to the accident. Categorizing these claims into standardized event/exposure categories, such as the Occupational Injury and Illness Classification System (OIICS) developed by the Bureau of Labor Statistics (BLS), would require manually reading each claim and assigning an event/exposure code.

Recently, researchers (Wellman et al., 2004; Lehto et al., 2009; Marucci-Wellman et al., 2011; Bertke et al., 2012; Taylor et al., 2014) demonstrated that computer learning algorithms using variations of the naïve Bayes model can auto-code injury narratives into different causation or event/exposure groups efficiently and accurately. In addition, a webinar (CWCS, 2014) was held by the NIOSH Center for Workers' Compensation Studies with participation by experts from the Liberty Mutual Research Institute for Safety (Helen L. Corns and Helen Marucci-Wellman), NIOSH (Stephen J. Bertke), Bureau of Labor Statistics (Alexander Measure), and Purdue University (Mark R. Lehto) presenting work on the topic of auto-coding workers' compensation narratives. The presenters demonstrated that the algorithms could code thousands of claims in a matter of minutes or hours with a high degree of accuracy by "learning" from claims previously coded by experts, referred to as a training set. Furthermore, these algorithms provided a score for each claim that reflected the algorithm's confidence in the prediction and, therefore, claims with low confidence scores could be flagged for manual review.

The majority of recent research into auto-coding injury narratives has focused on variations of the naïve Bayes models (Vallmuur, 2015) and while these models have been shown to be highly effective and intuitive, alternative machine learning approaches have been shown to out-perform them in many applications (Measure, 2014). One method in particular is referred to as regularized logistic regression and evaluating its performance in comparison to the naïve Bayes model is one focus of this study.

* Corresponding author.
E-mail address: inh4@cdc.gov (S.J. Bertke).

http://dx.doi.org/10.1016/j.aap.2015.12.006
0001-4575/Published by Elsevier Ltd.

Another purpose of this study is to explore the features used by these auto-coders. Previously, the main features considered were the occurrence or nonoccurrence of certain individual words. However, in natural language, words do not generally occur individually and often sequences of words commonly appear together. For example, common key words of interest for coding event/exposure are "FELL" and "OFF" and these words are helpful in identifying an injury caused by a *slip, trip, or fall*. However, the occurrence of the sequence "FELL OFF" is also common and could provide further evidence of a *slip, trip, or fall.* An example of the utility of considering two word sequences can be seen in the claim narrative "DRIVER FELL ASLEEP WENT OFF RIGHT SIDE OF ROAD INTO DITCH." This narrative contains both "FELL" and "OFF" but does not contain the sequence "FELL OFF," so identification of (or lack of) this feature could provide more evidence for a non-fall event/exposure.

The use of two-word sequences is not a new concept in the computational linguistics field. In fact, within the field of coding injury narratives, Lehto et al. (2009) and Marucci-Wellman et al. (2011) have considered two-word (and longer) sequences in a separate model referred to as "Fuzzy Bayes." Also, Grattan et al. (2014) and Marucci-Wellman et al. (2015) used two-word sequences within the Naïve Bayes framework, however, single-word and two-word sequences were used in separate models, not in a single model. Measure (2014) provides a more exhaustive investigation into which features optimize various auto-coder models and found that both the Naïve Bayes and logistic event auto-coders benefit from including single word and two-word features along with the North American Industry Classification System (NAICS) code of the employing establishment in a single model.

Finally, not all researchers or public health practitioners have access to a set of previously coded records to use as a training set on their un-coded data and most privacy agreements would prohibit providing/publishing workers' compensation claims. However, each of these auto-coding methods involve calculating a table of "weights" (coefficients) associated with each feature by event/exposure code. The weights table has all the necessary information from the training set needed to auto-code additional claims and can be easily be constructed in a way that has all personally identifiable information removed. As a result, the tables from this study are available upon request to the public (email cwcs@cdc.gov). Since this table of weights has been optimized on the data from this study (OHBWC claims), we tested the feasibility of using these weights to auto-code other injury narratives by sharing it with BLS and asking them to evaluate its ability to assign event/exposure codes to Survey of Occupational Injuries and Illnesses (SOII) cases that had been previously manually coded.

In short, this paper will: (1) investigate the performance of a naïve Bayes model vs a logistic model, (2) investigate the performance of adding two word sequences into a single model, (3) demonstrate the feasibility of sharing an auto-coder pre-trained with OHBWC claims with an outside researcher.

2. Methods

2.1. Auto-coding procedures

Two general auto-coding procedures were compared for this study: Naïve Bayes and regularized logistic regression. Details of these procedures can be found in the Appendix. In short, both procedures attempt to calculate the probability a given claim is the result of a particular injury or illness event/exposure by considering the relevant features of the claim. The event/exposure with the highest probability is assigned to the claim and the associated calculated probability is retained as a score value representing the confidence that the auto-coder assigned the correct category.

For this study, relevant features included: (1) the occurrence/nonoccurrence of a list of keywords in the narrative, (2) the occurrence/nonoccurrence of a sequence of two keywords in the narrative, (3) the resulting injury diagnoses categorized into 57 groupings. We defined keywords as any word that occurred in at least 3 claims of the training set and did not appear in a list of so-called "stop-words" (common, less informative words such as "the", "a", "an", etc.). A sequence of two keywords was defined as any two keywords that occurred consecutively in a given narrative, after stop-words were removed. Finally, the 57 resulting injury categories were based on the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) code for the "optimal return to work" (i.e. most severe) diagnosis (Beery et al., 2014) listed on the claim, which OHBWC defines as the injury that most likely will keep the injured worker off of work for the longest period of time and is assigned via a proprietary OHBWC algorithm. Details for the injury category variable have been previously described (Bertke et al., 2012) and inclusion of this additional field previously showed a substantial improvement on the auto-coding performance, namely raising the overall accuracy by about 5%.

2.2. Event/exposure categories

The auto-coding methods used in this study were used to code claims to a 2-digit OIICS event/exposure category. The OIICS system is a hierarchical sequence of numbers, where each digit indicates a further level of detail describing the event leading to the injury. For example, a claim coded as a 4 represents a *Slip, Trip, or Fall* and this can be further specified with a second digit as a *slip or trip without fall* (41), *falls on same level* (42) or *falls to lower level* (43). The full list of event/exposures can be found at: http://www.bls.gov/iif/oiics_manual_2010.pdf.

The OIICS coding system has a code of 9 indicating a claim that is un-classifiable and this is either due to a vague narrative or a narrative that is completely missing. In addition, when coding to the 2-digit event/exposure level, sometimes it was possible to identify the first digit (division) but there was not enough information to assign a more detailed category. In this instance, a zero is used as the second digit to signify "unspecified" claims within a specific division.

2.3. Evaluation

The test data used for this study consist of 7200 manually coded claims from a stratified random sample of allowed claims from 2001 to 2009 in the OHBWC database. The database contains a narrative for each claim answering the following the question: Description of accident (Describe the sequence of events that directly injured the employee, or caused the disease or death.) Claims were stratified to produce an equal number of medical-only and lost-time claims and equal numbers of claims per calendar month. All claims were manually coded by an experienced coder and coded to the 2-digit event/exposure OIICS code level. To estimate inter-coder reliability, one third of the claims were randomly assigned to a second experienced coder and manually coded.

To evaluate each method, the 7200 claims were randomly split into a training set consisting of 6200 claims and a prediction set of the remaining 1000 manually coded claims. All claims with a code of 9 and codes with a second digit of 0 were removed from the training set, since these claims were determined to be un-classifiable or not further classifiable beyond the first digit (division). These claims were *not* removed from the prediction set, however, so that the prediction set would be a representative sample of un-coded claims. As a result, the auto-coder will assign its "best guess" to a two-digit level and hopefully claims with a manual classification