Computer Classification of Injury Narratives Using a Fuzzy Bayes Approach: Improving the Model

Helen R. Marucci[1,3], Mark R. Lehto[2], and Helen L. Corns[1]

[1] Liberty Mutual Research Institute for Safety, 71 Frankland Road, Hopkinton, MA 01748, USA
[2] School of Industrial Engineering, Purdue University, 1287 Grissom Hall, West Lafayette, In 47907, USA
[3] Department of Work Environment, UMASS Lowell, One University Avenue, Lowell, MA 01854, USA

Abstract. This paper summarizes improvements to an earlier developed Fuzzy Bayes approach for assigning coding categories to injury narratives randomly extracted from a large U.S. insurer. Improvements to the model included: adding sequenced words as predictors and removing common subsets prior to calculation of word strengths. Removing subsets and adding word sequences improved prediction strengths for sequences found frequently in the training dataset, and resulted in more intuitive predictions and increased prediction strengths. Improved accuracy was found for several categories that had proved difficult to code in the past. This study also examined the effectiveness of a two-tiered approach, in which narratives were first categorized at the broad level (such as [falls]), before classification at a more refined level (such as [falls from heights].) The overall sensitivity following a two-tiered approach was 79% for predicting classifications at the broad category level and 66% for the more refined prediction categories.

Keywords: Textmining, Occupational Safety, Workers Compensation, BLS, Textminer.

1 Introduction

Injury narratives can provide useful information for selection of interventions for the prevention of injuries [1], [2] [6]. Using the computer to help classify narratives has the potential for reducing the burden implicit in manually reviewing and classifying large numbers of narratives from administrative injury databases.

A computerized approach based on Fuzzy Bayes logic was introduced in previous studies [3], [4], [5]. This study furthers that investigation by: 1) using separate prediction and training datasets, 2) adding sequence words as predictors, 3) reducing the strength of common words, found in more than one category, by removing common subsets, and 4) determining the value of using this method to assign more specific classifications.

M.J. Smith, G. Salvendy (Eds.): Human Interface, Part I, HCII 2007, LNCS 4557, pp. 500–506, 2007.

2 Methods

Workers' compensation claims case narratives were randomly extracted from a large US insurer. A trained, three person coding panel classified the narratives into two digit event categories (leading to injury) according to the Bureau of Labor Statistics (BLS) Standard Occupational Injury and Illness Classification (SOIIC) coding scheme [8]. Only narratives where at least 2 out of 3 coders agreed on the code were used (n=10,389), these codes were considered as the gold standard for comparing computerized coding. Three thousand narratives were extracted as the prediction data set, while the Fuzzy Bayes classifier "trained' from the probabilities of categories given word, word combinations or word sequences of 7,389 narratives. A previous study [5] found that using multiple word predictors, (the MW model) when compared with using single words alone, improved both the sensitivity and specificity of the computer generated codes. In this study all models were compared with the MW model.

As reported earlier [5] and repeated here using a new classification strategy and narratives, this method is capable of predicting injury event categories with a high level of accuracy (MW model, 78%). The new strategy included adding two word sequences to the wordlist and removing subsets (MW2SEQSUB model), this marginally improved the overall accuracy of the model at the category level (79%). Results for each category (for training and prediction datasets combined and the prediction data set only) are given in Tables 1 and 2 respectively. However, the prediction of the category when including word sequences were determined at higher strengths and were clearly more intuitive (Table 3). The optimistic bias [7] of predicting training and prediction narratives combined as compared with prediction narratives alone ranged from 6 to 10%.

3 Results

If *sequenced* words are found frequently in a category of the training set, the strength of the sequence as a predictor can be much higher than using words not in sequence. The MW2SEQ model correctly predicted 33 additional narratives, which had been poorly predicted in the MW model. The mean prediction strength for those 33 narratives was 7% higher with the sequenced words. Overall the sequence model predicted classifications at higher strengths than word combinations (MW mean strength .82, MW2SEQ mean strength .87). Examples of the improved predictions from the MW model to the MWM2SEQ model are shown in table 3.

There was a larger improvement when adding word sequences at the two digit level with the largest improvement in the contact category 'struck by' (8%). This may be important because the sensitivity of contact categories have been consistently low [5]. Results for two digit categories (with greater than 100 claim narratives) are provided in Table 4. The larger categories maintained a high level of accuracy ('Fall from same level' had sensitivity of 84% and 'highway incident' had sensitivity of 97%).

Finally, although the model with word sequences predicted more specific classifications, some of the categories had too few training narratives (i.e. [fires and explosions], n=42) for an adequate number of unique predictors to be generated.

Table 1. Sensitivity, Specificuity and Positive Predicitive Value of Bureau of Labour Statistics (BLS) Event Categories Comparison to Manual Coding of two Different Fuzzy Bayes Models: Training and Prediction Cases Combined (N=10,389)

		Multiple word model			Multiple & 2 word sequences model (including removing subsets)		
BLS Event Category	N	Sensitivity	Specificity	Positive Predictive	Sensitivity	Specificity	Positive Predictive
Contact with objects & equipment	1,890	0.73	0.98	0.90	0.76	0.98	0.91
Falls	1,617	0.93	0.94	0.76	0.93	0.95	0.78
Bodily Reaction & Exertion	3,315	0.89	0.92	0.84	0.91	0.93	0.85
Exposure to Harmful Substances or Env	1,290	0.90	0.98	0.87	0.90	0.98	0.88
Transportation Accidents	1,351	0.95	0.94	0.71	0.95	0.97	0.82
Fires & Explosions	63	0.25	1.00	0.94	0.30	1.00	0.86
Assaults & Violent Acts	382	0.59	1.00	0.90	0.59	1.00	0.92
Other Events or Exposures	481	0.22	1.00	0.88	0.31	1.00	0.88
Overall	**10,389**	**0.83**	**0.95**	**0.83**	**0.85**	**0.96**	**0.85**

Table 2. Sensitivity, Specificuity and Positive Predicitive Value of Bureau of Labour Statistics (BLS Event Categories Comparison to Manual Coding of two Different Fuzzy Bayes Models: Prediction cases Only (N=3,000)

		Multiple word model			Multiple word and 2 Word Sequences Model (including removing subsets)		
BLS Event Category	N	Sensitivity	Specificity	Positive Predictive	Sensitivity	Specificity	Positive Predictive
Contact with Objects & Equipment	557	0.65	0.97	0.84	0.69	0.97	0.84
Falls	470	0.90	0.92	0.68	0.89	0.93	0.69
Bodily reaction & Exertion	964	0.84	0.90	0.80	0.85	0.91	0.81
Exposure to Harmful Substances or Environment	361	0.88	0.98	0.85	0.87	0.98	0.85
Transportation Accidents	370	0.92	0.96	0.75	0.91	0.96	0.77
Fires & Explosions	21	0.19	1.00	1.00	0.19	1.00	1.00
Assaults & Violent Acts	117	0.49	1.00	0.88	0.47	1.00	0.90
Other Events or Exposures	140	0.17	1.00	0.71	0.25	1.00	0.73
Overall	**3,000**	**0.78**	**0.94**	**0.79**	**0.79**	**0.95**	**0.80**

Table 3. Examples of Improved Predictions for Multiple and 2 Word Sequence Model

		Multiple Word Model (Predicted incorrectly)		Multiple and 2 word Sequences Model (Predicted correctly)	
Claim Narrative	BLS event category	Prediction Strength	Predictor	Prediction Strength	Predictor
Employee was en route to job site and was in a motor vehicle accident and lacerated his face.	Transportation	0.78	LACERATED	0.94	IN-MOTOR
Battery fell on foot.	Contact w Obj & Equip	0.78	FELL&ON	0.92	FELL-ON&ON-FOOT
Hit knee on floor	Contact w Obj & Equip	0.57	FLOOR &HIT&KNEE	0.70	KNEE-ON
On the phone when the phone zapped employee in the ear then computer blew up and started smoking.	Expos to Harmful Sub or Env	0.78	COMPUTER &IN&ON	0.89	IN-EAR
Emp was struck on left foot by a lamp when the lamp fell offthe counter	Contact w Obj & Equip	0.78	FELL&ON	0.88	FELL&STRUCK-ON
Order selecting food products, balancing full. Box fell against his leg.	Contact w Obj & Equip	0.77	FELL	0.83	BOX-FELL
Emp was putting items away above her head on the retail floor causing a strain to her rt shoulder blade	Bodily Reaction & Exertion	0.80	FLOOR &HEAD	0.86	SHOULDER&TO-RT
Emp in back of vehicle man with gun jumped in vehicle emp jumpedout hitting lt knee on steering wheel	Assaults & Violent Acts	0.80	STEERING &WITH	0.86	WITH-GUN
Emp unloading trailer, tried to move bundle of rebar, 1400 lb. Pipe fell on left foot	Contact w Obj & Equip	0.87	FELL&TO &TRIED	0.92	FELL-ON&ON-FOOT
Clt wa putting things on shelf and a co-wprker struck clt with forklift	Transportation	0.86	SHELF &STRUCK	0.91	WITH-FORKLIFT
Trying to prevent student from falling	Bodily Reaction & Exertion	0.75	STUDENT	0.80	FROM-FALLING&TO

Table 4. Sensitivity, Specificuity and Positive Predicitive Value of Bureau of Labour Statistics (BLS) Event Categories Comparison to Manual Coding of two Different Fuzzy Bayes Models: Prediction Cases Only (N=3,000)

		Multiple Word Model				Multiple and 2 word Sequences Model (including removing subsets)			
Two Digit BLS Event Code	N	Sensitivity	Specificity	Positive Predictive	Negative predictive	Sensitivity	Specificity	Positive Predictive	Negative predictive
Contact:									
Struck against	163	0.35	0.99	0.67	0.96	0.42	0.99	0.71	0.97
Struck by	272	0.42	0.97	0.56	0.94	0.50	0.97	0.62	0.95
Caught in-between	100	0.79	0.99	0.65	0.99	0.81	0.99	0.67	0.99
Falls:									
Fall to lower level	159	0.57	0.98	0.56	0.98	0.59	0.98	0.63	0.98
Fall on same level	299	0.84	0.93	0.58	0.98	0.84	0.94	0.60	0.98
Bodily reaction & exertion									
Bodily Motion	218	0.36	0.98	0.62	0.95	0.36	0.98	0.60	0.95
Slip,Trip without fall	162	0.36	0.99	0.63	0.96	0.43	0.99	0.64	0.97
Overexertion	460	0.85	0.90	0.61	0.97	0.85	0.91	0.64	0.97
Repetative Motion	98	0.77	0.99	0.68	0.99	0.80	0.99	0.72	0.99
Exposure to Harmful sub or Env									
Exposure to temperature extremes	108	0.86	0.99	0.70	0.99	0.85	0.99	0.71	0.99
Exposure to caustic substances	123	0.79	0.99	0.85	0.99	0.81	0.99	0.79	0.99
Transportation									
Highway accident	205	0.97	0.96	0.64	1.00	0.97	0.96	0.67	1.00
Assaults and Violent Acts									
Assaults & violent acts by persons	111	0.59	0.99	0.77	0.98	0.59	0.99	0.78	0.98
Unclassifiable	140	0.29	0.99	0.69	0.97	0.35	0.99	0.71	0.97
Overall	**3,000**	**0.63**	**0.97**	**0.63**	**0.98**	**0.66**	**0.97**	**0.66**	**0.98**

4 Conclusions

From this study it can be seen that using a Fuzzy Bayes approach, threshold values can be used to filter out more difficult narratives for manual review. Examination of Tables 1 and 2 suggest that the mean strengths of the [falls] category compared with the [contact] category would result in more of the latter narratives filtered out for manual review, a process that would improve the final accuracy substantially. Some additional improvements are being considered for improving the accuracy of smaller categories.

In summary, a more refined computerized approach for classifying narratives using a Fuzzy Bayes approach has been developed. Sequenced words can be stronger predictors, if found sufficiently frequently in the training dataset, and their results use more intuitive predictors. Removing subsets is an important strategy to consider when common single word predictors are found in more than one category.

References

1. Sorock, G., Smith, G., Reeve, G., et al.: Three perspectives on work-related injury surveillance systems. Am J. Ind. Med. 32, 116–128 (1997)
2. Smith, G.S.: Public health approaches to occupational injury prevention: do they work? Inj. Prev. 7(suppl. I), i3–i10 (2001)
3. Lehto, M., Sorock, G.: Machine learning of motor vehicle accident categories from narrative data. Methods Info Med. 35(4-5), 309–316 (1996)
4. Sorock, G., Ranney, T., Lehto, M.: Motor vehicle crashes in roadway construction work zones: an analysis using narrative text from insurance claims. Accid. Anal. Prev. 28, 131–138 (1996)
5. Wellman, H.M., Lehto, M.R., Sorock, G.S.: Computerized coding of injury narrative data from the National Health Interview Survey. Accid Anal. Prev. 36, 165–171 (2004)
6. Lincoln, A.E., Sorock, G.S., Courteney, T.K., Wellman, H.M., Smith, G.S., Amoroso, P.J.: Using narrative text and coded data to develop hazard scenarios for occupational injury interventions. Inj. Prev. 10, 249–254 (2004)
7. Clancy, E.A.: Factors Influencing the Resubstitution Accuracy. Multivariate Classification Analysis: Implications for Study Design in Ergonomics, Ergonomics 40(4), 417–427 (1997)
8. Bureay of Labor Statistics. Occupational injury and illness classification manual. Us Department of Labor, Washington, DC (December 1992)