Accident Analysis and Prevention 98 (2017) 359–371

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Classifying injury narratives of large administrative databases for surveillance—A practical approach combining machine learning ensembles and human review

Helen R. Marucci-Wellman [a,*], Helen L. Corns [a], Mark R. Lehto [b]

[a] Center for Injury Epidemiology, Liberty Mutual Research Institute for Safety, 71 Frankland Road, Hopkinton, MA 01748, USA
[b] School of Industrial Engineering, Purdue University, 1287 Grissom Hall, West Lafayette, IN 47907, USA

ARTICLE INFO

Article history:
Received 13 July 2016
Received in revised form 7 October 2016
Accepted 10 October 2016
Available online 15 November 2016

Keywords:
Injury
Narrative text
Injury surveillance
Cause of injury
Machine learning

ABSTRACT

Injury narratives are now available real time and include useful information for injury surveillance and prevention. However, manual classification of the cause or events leading to injury found in large batches of narratives, such as workers compensation claims databases, can be prohibitive. In this study we compare the utility of four machine learning algorithms (Naïve Bayes, Single word and Bi-gram models, Support Vector Machine and Logistic Regression) for classifying narratives into Bureau of Labor Statistics Occupational Injury and Illness event leading to injury classifications for a large workers compensation database. These algorithms are known to do well classifying narrative text and are fairly easy to implement with off-the-shelf software packages such as Python. We propose human-machine learning ensemble approaches which maximize the power and accuracy of the algorithms for machine-assigned codes and allow for strategic filtering of rare, emerging or ambiguous narratives for manual review. We compare human-machine approaches based on filtering on the prediction strength of the classifier vs. agreement between algorithms.

Regularized Logistic Regression (LR) was the best performing algorithm alone. Using this algorithm and filtering out the bottom 30% of predictions for manual review resulted in high accuracy (overall sensitivity/positive predictive value of 0.89) of the final machine-human coded dataset. The best pairings of algorithms included Naïve Bayes with Support Vector Machine whereby the triple ensemble NB_{SW} = $NB_{BI\text{-}GRAM}$ = SVM had very high performance (0.93 overall sensitivity/positive predictive value and high accuracy (i.e. high sensitivity and positive predictive values)) across both large and small categories leaving 41% of the narratives for manual review. Integrating LR into this ensemble mix improved performance only slightly.

For large administrative datasets we propose incorporation of methods based on human-machine pairings such as we have done here, utilizing readily-available off-the-shelf machine learning techniques and resulting in only a fraction of narratives that require manual review. Human-machine ensemble methods are likely to improve performance over total manual coding.

1. Introduction

Advances in information technology in health care over the past two decades have marked a pivotal change in age-old methods for the administration and tracking of medical and other health records. The resulting electronic databases containing real-time human subject data, such as hospital billing records, workers compensation claims or national surveys, create the potential for changes and improvements in public health research and surveillance. Since injuries, a leading cause of death in the United States, have a relatively short latency period, the narratives accompanying structured pre-posited database entries in large administrative data sources are a useful adjunct to pre-coded information on potential causes, prevention and recovery from injury (CDC, 2015; Sorock et al., 1996, 1997; Stutts et al., 2001; Williamson et al., 2001; Lincoln et al., 2004; Lombardi et al., 2005, 2009; Verma et al., 2008; McKenzie et al., 2010; Taylor et al., 2014; Vallmuur, 2015; Vallmuur et al., 2016). However, the number of narratives that can be manually read through is often limited due to resource constraints.

* Corresponding author.
E-mail address: Helen.Wellman@LibertyMutual.com (H.R. Marucci-Wellman).

http://dx.doi.org/10.1016/j.aap.2016.10.014

Over the past two decades we have completed several studies (Lehto and Sorock, 1996; Sorock et al., 1997; Wellman et al., 2004; Lehto et al., 2009; Marucci-Wellman et al., 2011, 2015) on the utilization of computer algorithms to streamline the classification of the event (or causes) leading to injury for surveillance. Our focus has been to create machine learning techniques that can quickly filter through hundreds of thousands of narratives and accurately classify and track high magnitude, high risk and emerging causes of injury, information which can be used to guide the development of interventions for prevention of future injury incidents (Horan and Mallonee, 2003). Our recent work has included classifying workers compensation (WC) injury narratives into BLS Occupational Injury and Illness Classification system (OIICS) event (leading to injury) codes. A recent article published in a special supplement of Injury Prevention, geared toward advancing injury surveillance methods to fit the 21st century, describes the background, growth, value, challenges and future directions of machine learning as applied to injury surveillance. It summarizes our work, as well as that of others, in developing these strategic methods (Vallmuur et al., 2016). We believe what has been learned on these computer-assisted methods could be easily adopted by many other injury surveillance programs nationally and internationally for more timely identification and classification of the circumstances leading to injury.

Our work has included development of human-machine approaches whereby strategic filters are used to identify those weakly predicted by the algorithm to be extracted out and manually reviewed. We have found that the selection of narratives which should be classified by the algorithm vs those which should be classified by a human can be strategically determined by allowing the algorithm to assign the code when two Naïve Bayes machine learning strategies agree on the code or if the code was predicted at a high strength by the Naïve Bayes classifier.

Several studies have shown that a single word version of Naïve Bayes performs quite well for classifying many event categories. (Lehto et al., 2009; Vallmuur, 2015; Nanda et al., 2016). Other studies have shown improvements over Naïve Bayes (NB) for both Support Vector Machine (SVM) (Chen et al., 2015) and Logistic Regression (LR) (Bertke et al., 2016). In what may be currently the most systematic comparison of machine learning methods for injury narrative classification, Chen et al. (2015) found that SVM was overall the best performer on multiple criteria for their classification task. Naïve Bayes, Decision Trees, and Neural Networks tended to perform very similarly to each other at a level of performance a few percentage points lower than SVM. However, even after very extensive data pre-processing was integrated into any of these methods (i.e. correction of misspellings, word stemming, phrase extraction), there were large performance decrements on many of the smaller categories. Alex Measure (2014) and Bertke et al. (2016) demonstrated very good performance using Regularized LR, with approximately a 4-point higher accuracy compared with NB. However, for all of these studies (Chen et al., 2015; Bertke et al., 2016), even the best performing models could not perform well on all categories making the final coded dataset insufficient for surveillance of high risk, emerging risk events.

In our opinion no currently available off-the-shelf machine learning classifier alone is able to achieve high accuracy across all cause of injury classification categories for datasets including many categories of various sizes (Vallmuur et al., 2016). Instead, we believe a human-machine pairing should be optimized. Using agreement between algorithms ***or*** the probability strength of the classifier as a confidence metric can result in high accuracy in the machine classifications and can provide a strategic approach for filtering out narratives (those where the algorithms did not agree or below a certain probability threshold) for manual review (Marucci-Wellman et al., 2011, 2015; Bertke et al., 2016; Nanda et al., 2016).

We have demonstrated the potential for selecting out highly accurate computer-generated codes based on agreement between Naïve Bayesian Models (Marucci-Wellman et al., 2011). We found the two Naïve Bayesian Models (e.g. predictors are single words (NB_{SW}) or predictors are words in sequence ($NB_{BI\text{-}GRAM}$)), offered a practical approach for short narratives resulting in high accuracy across all categories (Marucci-Wellman et al., 2015). These results are also almost identical to what was obtained by Nanda et al. (2016); using the same two models. Bertke and colleagues from the National Institute for Occupation Safety and Health (NIOSH) recently also demonstrated the utility of integrating Logistic Regression (LR) for a similar classification task as ours, classifying WC narratives into OIICS two-digit event classifications (Bertke et al., 2016; narratives were allocated to 19 two-digit BLS OIICS categories). Interestingly, they found very similar results for either: 1) filtering on the probability strength the Logistic Regression model used to make the prediction or 2) filtering using agreement by pairing Logistic Regression (LR) with a Naïve Bayes Single Word Model. They achieved an 85% accuracy overall, and above 80% sensitivity and positive predictive value (PPV) across most large and some small categories, comparing the performance where 25% of the narratives needed to be filtered out for manual review (based on disagreement between the NB and LR predictions).

It is noteworthy that Bertke et al. (2016), with very similar methods and integrating Logistic Regression into the mix of classifiers, but using a very different dataset (workers compensation claims from one state vs. one insurer) obtained results similar to our prior work. The similar methods employed by three separate and distinct research teams have shown that there could be 66%–75% reduction in resources required for the same coding task that historically has been done all manually, yet resulting in a similar level of accuracy. However, the sensitivity for some of the small categories was still limited for the final coded datasets.

The objective of the current study is to test and compare the practicality and performance of a human-machine combined approach for classifying short injury narratives (up to 120 characters) where the selection of computer-generated codes is based on various machine learning ensembles or based on filtering on the prediction strength of each classifier. In this study we use four readily available and easy to integrate machine learning algorithms which have previously been found to be fairly successful for classification of short narratives as described above: 1) Support vector machine (SVM), 2) logistic regression (LR), 3) NB_{SW} and 4) $NB_{BI\text{-}GRAM}$. We demonstrate and compare results of the final coded data using ensemble approaches and alternatively utilizing the strength metric available for each of the classifiers.

We also test the utility of integrating some simple Natural Language Processing (NLP) rules to identify narratives in particular categories (e.g. electrocutions) where we anticipate that some simply-derived indexing rules, based on very strong keywords related to specific exposures, may be able to pull out at least some cases that a machine learning classifier may not be able to find.

2. Methods

Thirty thousand records were randomly extracted from claims filed with a large WC insurance provider between January 1 and December 31, 2007. Four coders, trained on the Bureau of Labor Statistics (BLS) Occupational Injury and Illness Classification system (OIICS) 2012 version, classified records into two-digit event codes using the accident (what happened, 120 character maximum) and injury narratives (type, e.g. strain, fracture, 20 character maximum) as they appeared on the first report of injury. These manual codes served as our "gold standard."

The dataset was then divided into two sets of 15,000 cases: a training set for model development, and a prediction dataset for evaluation. Each record included a unique identifier, a narrative describing how the injury occurred, and a two-digit BLS OIICS event code. The distribution of the two-digit OIICS event codes did not differ between datasets (χ^2 p = 0.87). Further detail on methods are explained in our earlier report (Marucci-Wellman et al., 2015).

The theoretical basis of all four classifiers (NB_{SW}, $NB_{BI\text{-}GRAM}$, SVM and LR) have been previously defined[1] (Lehto et al., 2009; Bertke et al., 2016). Briefly, the Naïve Bayes algorithm calculates the probability of each possible category given the set of words in a narrative (see equation 1 in Lehto et al., 2009). NB determines its estimate by first calculating the probability each word is present in each given category (using the training narratives). These probabilities are then multiplied through, and also multiplied by, the prior probability of the category alone in the training dataset to calculate the un-normalized probability of the category given the words. The category-specific probabilities are then normalized to make the sum of the probability estimates over all categories equal to 1. This estimate is optimal if the words are conditionally independent. The Logistic Regression algorithm assumes that the log likelihood ratio for each category is a linear function of the sum of the weights for each word present in a narrative. Therefore, the assignment of weights for each word in each category is determined by using all the words found in the training dataset as predictors, and optimizing the betas (weights) of the LR model. Using the weights (for each word) generated by the logistic regression model from the training data, the probability of each category can be calculated for subsequent prediction narratives. Logistic regression algorithms normally include a regularization parameter which can be adjusted to prevent over-fitting of the many (thousands of word) predictors. For both LR and NB, the category that is assigned the highest probability using the particular set of words in a narrative is chosen as the algorithm prediction and the corresponding probability provides information about the confidence (strength) of the classification. Finally, Support Vector Machine differs from both LR and NB in that it is a non-probabilistic classifier. However, SVM, similar to LR, attempts to minimize error while penalizing weights assigned to the words, but usually does this by fitting a linear function for each category that optimally discriminates it from the other categories.

Various software packages are now publically available for training (or building) the models based on the training dataset and then making subsequent predictions. For this study, we used the Python software machine learning package (Scikit-learn: Machine Learning in Python, Pedregosa et al., 2011) since it is free to the public, easily downloadable and easily adaptable for development of all four models. The three model routines used in this analyses were: 1) sklearn.linear_model.Logistic Regression, 2) sklearn.naive_bayes.MultinomialNB and 3) sklearn.svm.svc. Default parameters were primarily used for each model with the exception of the regularization (penalty) parameter, set to l1 in the Logistic Regression model, and the Naïve Bayes alpha (smoothing) parameter set to 0.1. The authors' Python code can be made available upon request. The narratives were used in their raw form. Although improved performance can be expected if you clean up misspellings and morph words that have the same meaning into one syntax, we wanted to show what could be achieved with little pre- processing of the narratives. A small list of drop words (following common practice, i.e., A, AN, AND, ETC, HE, HER, HIM, HIS, I, LEFT, LT, MY, OF, RT, RIGHT, SHE, THE, R, L) were globally deleted from the narratives prior to the learning phase.

We then used the predictive models (and probabilities) developed from the training dataset to classify each of the 15,000 prediction narratives into a two-digit BLS OIICS classification. The obtained results were then evaluated, comparing the predictions with the manually-assigned gold standard codes.

Our evaluation metrics were designed to capture the accuracy (i.e. high sensitivity and positive predictive values) that would be required of surveillance, enabling us to compare across many different models. It is important for surveillance that the distribution by category in the final coded dataset is robust (similar to the gold standard) and that small categories or emerging risks can be identified as accurately as large categories. Sensitivity is calculated as the percent of gold standard narratives coded correctly by the algorithm into each category; PPV is the percent of narratives correctly predicted into a category out of the total number of times the algorithm predicted into a category. We did not evaluate specificity and negative predictive value because they were all high (nearing 1.0) with little differentiation across categories (see earlier results in Wellman et al., 2004; Lehto et al., 2009). Summary metrics included the overall sensitivity and PPV of each model. We note that the performance of the larger categories will greatly affect the overall performance of the entire dataset. Therefore, we also provide, as a summary performance statistic, the unweighted values of both sensitivity (mean sensitivity value across all categories) and PPV (mean PPV value across all categories), which consider the performance of each category to have equal weight towards the overall results regardless of size.

Results are presented 1) for each algorithm alone and 2) for human- machine pairings where filtering on the 15,000 prediction narratives occurred to decide which narratives the algorithm would code and which would be manually reviewed. When using agreement between two algorithms (the ensemble approaches, i.e. SVM = LR, NB_{SW} = LR, $NB_{BI\text{-}GRAM}$ = SVM, NB_{SW} = $NB_{BI\text{-}GRAM}$ = SVM, etc.), the filter level is integral to the agreement method; the computer classification is assigned when the algorithms agree and the remainder of the narratives (where the algorithms disagreed on the classification) are filtered out for human review. We compare that to the researcher setting the amount to be manually coded with each algorithm alone, i.e. setting levels of 10–15–30–45% manual coding. This is done by assigning a computer classification for those predicted by the algorithm with the highest prediction strengths, e.g. the top 90–85–70–55% of the narratives and filtering out the remainder, 10–15–30–45% for human review.

2.1. Additional development of NLP rules

We have realized through our work that WC narrative data are very noisy and, since injury narratives can contain many similar words, the algorithms will always make some mistakes and tend to predict the larger categories better than the small. We also realized that some small categories have very unique words or syntax to help with identification. We, therefore, wanted to understand if applying some simple rules (without modifying the structure of the narratives in any way such as including word tagging for nouns vs. verbs, etc.) would help to pull out unique narratives and allow for an accurate computer-assigned code beyond the machine learning strategies for some of the smaller categories. Our strategy was to test out additional methods based on very simple NLP rules using certain unique keywords that would allow for rapid and accurate identification of some narratives (e.g. electrical; explosions; exposure to temperature extremes). One example of an NLP rule set that we designed for explosions includes "explo" AND "ear" or "pressure" or ("inflat" AND 'air' AND "tire" AND ("blew" or "blow or expl")). After developing the rules we then examined whether they

[1] For a good introduction to statistical learning methods such as SVM and LR, see: James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R, available at http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf. Springer, New York.

were capable of identifying additional cases beyond those identified from Logistic Regression alone (the best performing algorithm alone overall) in these categories.

3. Results

3.1. Algorithms alone

Similar to prior results (Lehto et al., 2009; Bertke et al., 2016) we found that each of the four classifiers used on their own had fair performance classifying all of the 15,000 predication narratives (results shown in the top of Table 1). The LR model performed the best overall (sensitivity 73%), second was SVM (sensitivity 70%), third was NB_{SW} (sensitivity 67%) and finally $NB_{BI\text{-}GRAM}$ (sensitivity 66%). All classifiers were within a 7-point accuracy of one another and none of them had consistently high performance across all categories (with very low unweighted sensitivities in the small categories 0.05–0.15). Therefore, the resulting classified datasets would not, in fact, be representative by event category of the population of cases we began with, would include very few of the cases in the small unique categories and would be of limited value for surveillance.

3.1.1. Filtering out narratives for manual review based on the prediction strength of each algorithm

A human-machine pairing approach to coding, based on filtering on the probability strengths used to predict the classification for the three algorithms, demonstrated that LR alone had the highest sensitivity and PPVs across most categories (Table 1). The 70% vs. 30% computer/human assignment (70-30 model) of narratives in Fig. 1 also demonstrate, as an example, that these results offer a

Table 1
Summary Statisics: Comparing methods for complete coding of the entire prediction dataset: Using an algorithm alone and human-machine approaches based on two or three and four model agreement.
[b]Weighted PPV is Positive Predicted Value across all categories: the percentage of narratives correctly coded into a specific category out of all narratives placed into that category by the algorithm.
[c]Unweighted average sensitivity is the average sensitivity across all categories = ∑[each category sensitivities]/(# of categories) (the performance of each category is considered equally).
[d]Unweighted average positive predictive value is the average PPV across all categories = ∑[each category PPVs]/(# of categories with at least one case predicted) (the performance of each category is considered equally) *Note: Some small categories do not contribute to the average PPV given that they are never predicted, PPV can only be calculated for categories that are predicted at least once).
[e]Manually coding cases where there is no agreement of classifications of codes or where the prediction strength is low for a case.
[f]Average of 3 overall metrics: Overall weighted sen/weighted ppv, unweighted average sensitivity, unweighted average PPV.

	Overall Results				Large Categories (n≥100)		Small Categories (n<100)		
	Weighted	Unweighted			Unweighted		Unweighted		
Models	Sen[a]/PPV[b]	Sen[c]	PPV[d]	% of Dataset Manually Coded[e]	Sen[c]	PPV[d]	Sen[c]	PPV[d]	Average of 3 Overall Metrics[f]
Algorithm Alone									
Logistic Regression (LR)	0.73	0.32	0.69	0	0.59	0.68	0.15	0.71	0.58
Support Vector Machine (SVM)	0.70	0.35	0.43	0	0.58	0.62	0.20	0.29	0.49
Naïve Bayes Single Word Model (NB_{SW})	0.67	0.32	0.54	0	0.59	0.59	0.15	0.48	0.51
Naïve Bayes ($NB_{BI\text{-}GRAM}$)	0.66	0.22	0.58	0	0.49	0.58	0.05	0.59	0.49
Two model agreement and manual review									
NB_{SW}=LR	0.86	0.66	0.91	25	0.80	0.84	0.57	0.96	0.81
SVM= LR	0.81	0.52	0.86	14	0.71	0.78	0.41	0.91	0.73
SVM=NB_{SW}	0.86	0.71	0.90	28	0.81	0.84	0.64	0.94	0.82
SVM=$NB_{BI\text{-}GRAM}$	0.89	0.73	0.94	33	0.84	0.90	0.65	0.97	0.85
$NB_{BI\text{-}GRAM}$=LR	0.88	0.67	0.95	29	0.82	0.89	0.57	0.99	0.83
NB_{SW}=$NB_{BI\text{-}GRAM}$	0.86	0.66	0.93	30	0.82	0.86	0.56	0.97	0.81
Three model agreement and manual review									
SVM=NB_{SW}= $NB_{BI\text{-}GRAM}$	0.93	0.80	0.97	41	0.90	0.93	0.74	0.99	0.90
SVM=NB_{SW}= LR	0.89	0.73	0.93	31	0.84	0.88	0.66	0.97	0.85
SVM=$NB_{BI\text{-}GRAM}$=LR	0.91	0.74	0.96	36	0.87	0.92	0.67	0.99	0.87
NB_{SW}= $NB_{BI\text{-}GRAM}$=LR	0.92	0.76	0.96	39	0.88	0.92	0.68	0.99	0.88
Four model agreement and manual review									
SVM=NB_{SW}= $NB_{BI\text{-}GRAM}$=LR	0.93	0.81	0.97	43	0.91	0.94	0.74	1.00	0.90

[a]Weighted Sen is the Sensitivity across all categories: (true positives for the entire dataset) the overall percentage of narratives that had been coded by the experts into each category that were also assigned correctly by the algorithm. The performance of the larger categories will greatly affect the overall performance of the entire dataset.
[b]Weighted PPV is Positive Predicted Value across all categories: the percentage of narratives correctly coded into a specific category out of all narratives placed into that category by the algorithm.
[c]Unweighted average sensitivity is the average sensitivity across all categories = ∑[each category sensitivities]/(# of categories) (the performance of each category is considered equally).
[d]Unweighted average positive predictive value is the average PPV across all categories = ∑[each category PPVs]/(# of categories with at least one case predicted) (the performance of each category is considered equally) *Note: Some small categories do not contribute to the average PPV given that they are never predicted, PPV can only be calculated for categories that are predicted at least once).
[e]Manually coding cases where there is no agreement of classifications of codes or where the prediction strength is low for a case.
[f]Average of 3 overall metrics: Overall weighted sen/weighted ppv, unweighted average sensitivity, unweighted average PPV.

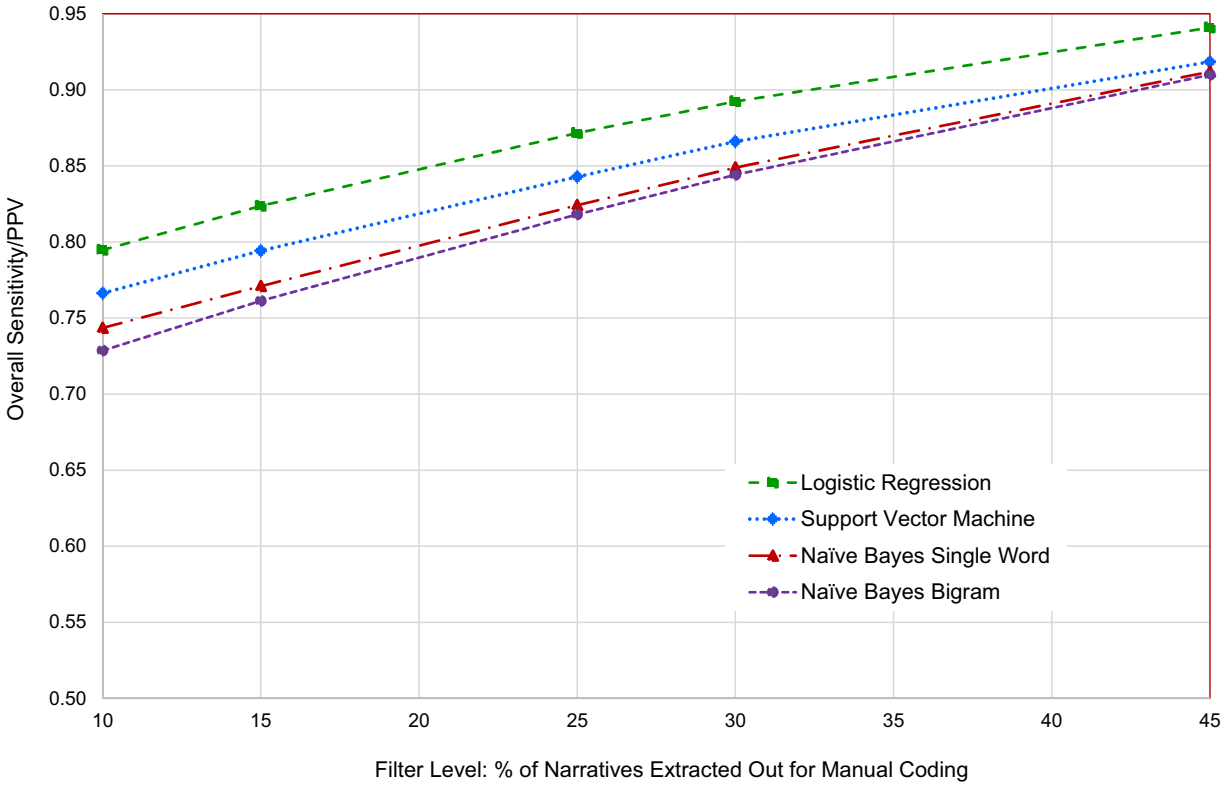


Fig. 1. Overall Sensitivity/PPV of Human-Machine Systems: Logistic Regression (LR), Naive Bayes(NBSW), Support Vector Machines (SVM) and Naive Bayes Bi-Gram (NBBi) with Filters Applied (10%, 15%, 25%, 30% and 45% manual coding).

reasonable accuracy model (0.89 sensitivity/PPV) with less than 1/3 of the narratives requiring manual review.

3.1.2. Filtering out narratives for manual review based on agreement between machine learning algorithms

For the two model agreement, SVM and $NB_{BI\text{-}GRAM}$ paired up well, resulting in an overall accuracy of 0.89, leaving only 33% of the dataset to be manually coded and with surprisingly fairly high accuracy across both large and small categories (Table 1). This model overall performed comparable to the 70-30 LR model just described but required 3% more manual coding (Fig. 2a and b).

When adding the NB_{SW} algorithm to the mix, this agreement model ($SVM = NB_{BI\text{-}GRAM} = NB_{SW}$) had improved performance with an overall accuracy of 0.93, but requiring that 41% of the dataset be manually coded. This model, while again very comparable to the LR model with the same amount of manual coding, had a slight edge over the LR model with regards to the accuracy of the small categories (unweighted sensitivity of the small categories rose to 0.74, Table 1 and Fig. 2a).

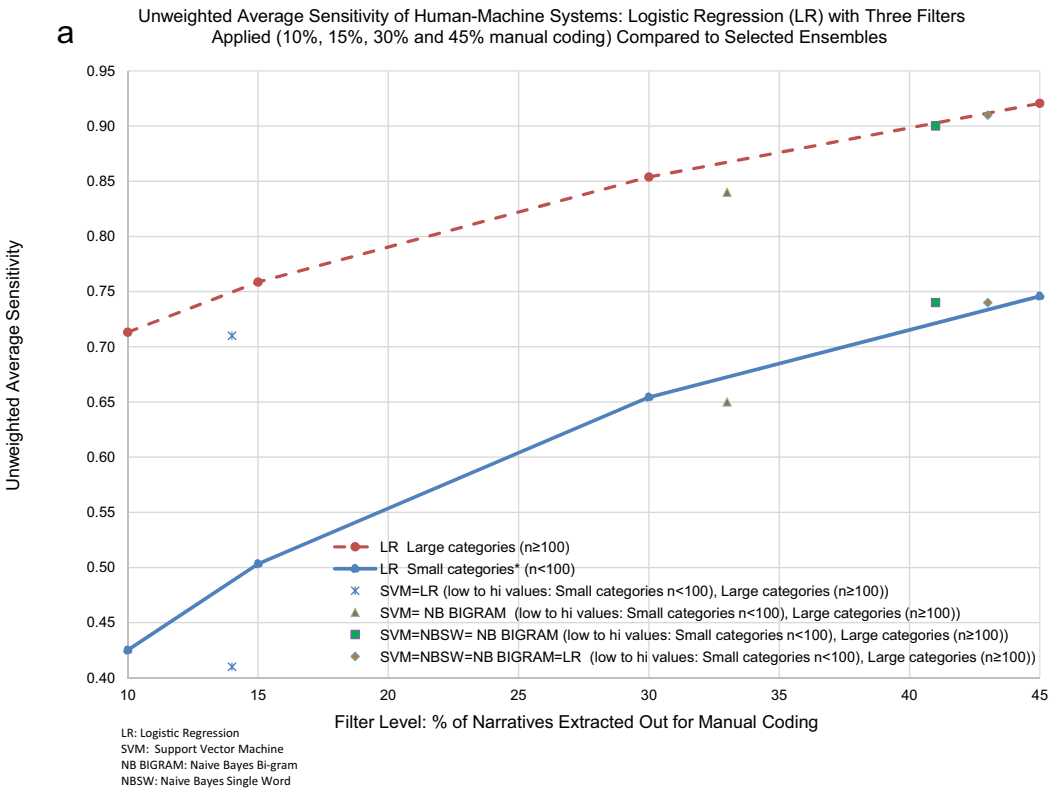
Finally, the agreement model for the four algorithms ($SVM = NB_{SW} = NB_{BI\text{-}GRAM} = LR$), as expected, had the highest accuracy (Tables 1, 2a and 2b) with 93% overall sensitivity and very high sensitivity and PPV across all categories (unweighted sensitivity 0.81, unweighted PPV 0.93) with 43% of the 15,000 narratives left for manual review. Similar to Bertke et al. (2016), however, comparable results occurred when simply removing the bottom 43% of narratives (lowest 43% probabilities used to predict the classifications) using LR alone (Tables 3a and 3b). We did find during this comparison that different small categories improved for the different methods, whereby many of the "exposure to harmful substances" subcategories did better using the ensemble approach while some of the "transportation incidents" subcategories did better with filtering solely on the LR results. In these tables, we also report for comparison the agreement accuracy (and kappa statistic) between each of 2 manual coders out of 4 total manual coders coding a separate dataset with 4000 total narratives.

3.2. Addition of strategic filters for human-machine coded narratives

Because agreement between models creates a fixed filtering amount (labeled in Table 1 as "% of dataset manually coded"), we compare the accuracy of the human-machine ensemble approaches, giving similar importance to large and small categories, with what would be achieved through filtering on the best performing single algorithm, the logistic regression algorithm alone in Fig. 2 (Fig. 2a: unweighted PPV and Fig. 2b: unweighted sensitivity). The results indicate that the logistic regression algorithm performs the best for the lower filters (such as below 35%). As the filtering amount approaches the fixed filtering amounts from the best ensembles based on NB and SVM together, overall performance improves substantially and the ensemble approaches may even surpass the LR model for the small categories. Also, given that NB and SVM are making their assignments in different ways, if both methods assign the same code, we can be more confident that the classification is correct.

3.3. Use of NLP rules for very unique categories

The overall results of LR alone and NLP alone for 5 selected categories are shown in Table 4 to illustrate the utility in pairing some NLP rules with LR to identify more cases from selected small categories. As can be seen from the table the NLP rules alone were able to identify several of the categories with high PPV. For the first four categories the NLP rules also resulted in a higher sensitivity than LR. For example, for electrocution, 12 more cases were found (44% additional cases) by integrating NLP compared with LR alone. However, for other categories, NLP did not do as well as machine



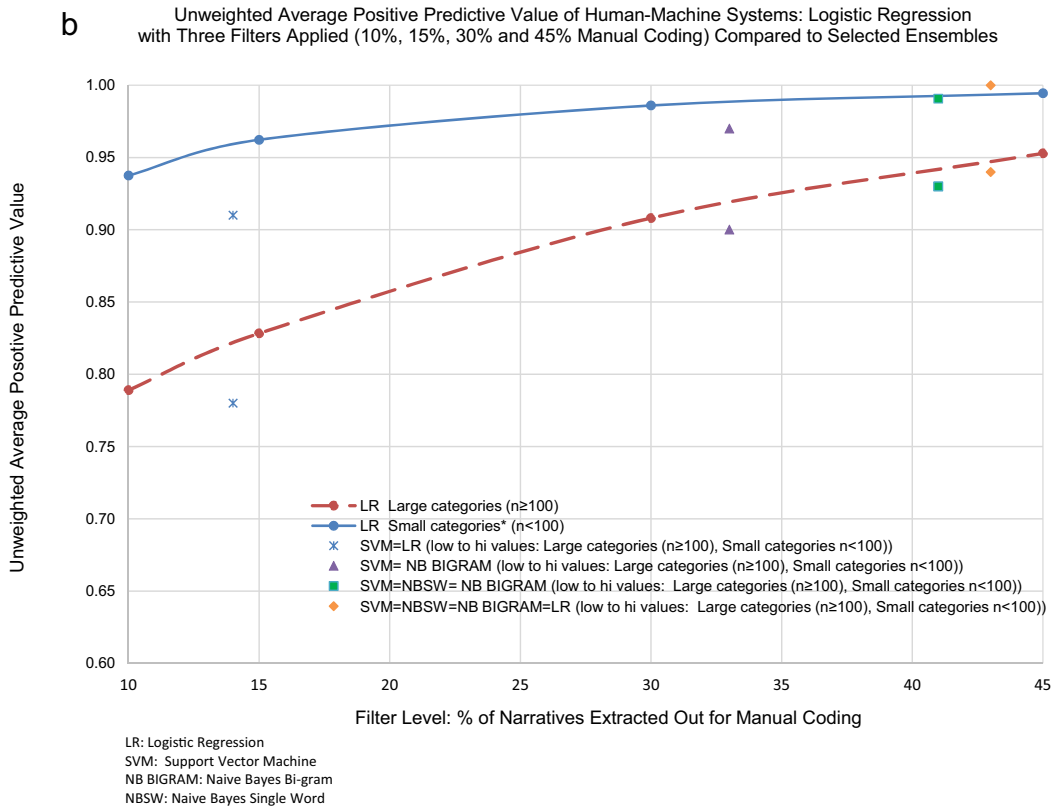


Fig. 2. (a) Unweighted Average Sensitivity of Human-Machine Systems: Logistic Regression (LR) with Three Filters Applied (10%, 15%, 30% and 45% manual coding) Compared to Selected Ensembles. (b) Unweighted Average Positive Predictive Value of Human-Machine Systems: Logistic Regression with Three Filters Applied (10%, 15%, 30% and 45% Manual Coding) Compared to Selected Ensembles.

Table 2a
The accuracy of selective computer coding: four-model ensemble approach, narratives where each algorithm assigned the same classification (57% of the dataset only are classified, n = 8612) vs. those predicted by the Logistic Regression algorithm alone with the highest probabilities (top 57%, n = 8612) (large categories, where ncat ≥ 100).

		Gold Standard[a]		Four-Model Ensemble Filter[b]: Narratives where LR = NB_{SW} = SVM = $NB_{BI-GRAM}$						Narratives where LR Predicted the Classification at a Very High Probability Level (top 57%)					
BLS OIICS 2, Digit Event Code		n	%	n_{pred}[c]	$\%_{pred}$[d,e]	Sen[f]	95% CI	PPV[g]	95% CI	n_{pred}[c]	$\%_{pred}$[d,e]	Sen[f]	95% CI	PPV[g]	95%CI
1* Violence and other injuries by persons or animals															
11	Intentional injury by person	159	1.1	35	0.4	0.73	0.59, 0.83	0.94	0.81, 0.99	21	0.2	0.56	0.38, 0.72	0.95	0.76, 1.00
2* Transportation incidents															
24	Pedestrian vehicular incidents	120	0.8	23	0.3	0.56	0.40, 0.72	0.83	0.61, 0.95	8	0.1	0.38	0.18, 0.62	1.00	0.63, 1.00
26	Roadway incidents involv motorized land vehicle	650	4.3	567	6.6	0.99	0.98, 1.00	0.96	0.94, 0.98	537	6.2	0.99	0.97, 0.99	0.96	0.93, 0.97
27	Nonroadway incidents involv motorized land vehicles	136	0.9	16	0.2	0.32	0.20, 0.47	0.56	0.30, 0.80	17	0.2	0.37	0.20, 0.56	0.65	0.38, 0.86
4* Falls, slips, trips															
41	Slip or trip without fall	806	5.4	271	3.1	0.72	0.67, 0.77	0.81	0.76, 0.85	329	3.8	0.80	0.75, 0.84	0.77	0.72, 0.81
42	Falls on same level	2,148	14.3	1456	16.9	0.92	0.91, 0.94	0.89	0.87, 0.91	1456	16.9	0.91	0.90, 0.93	0.89	0.96, 0.98
43	Falls to lower level	1,065	7.1	640	7.4	0.86	0.84, 0.89	0.84	0.81, 0.87	596	6.9	0.82	0.78, 0.85	0.85	0.82, 0.88
5* Exposure to harmful substances or environments															
53	Exposure to temperature extremes	141	0.9	49	0.6	0.85	0.73, 0.93	0.90	0.78, 0.97	58	0.7	0.78	0.66, 0.87	0.90	0.79, 0.96
55	Exposure to other harmful substances	175	1.2	53	0.6	0.82	0.71, 0.91	0.89	0.77, 0.96	31	0.4	0.77	0.61, 0.89	0.97	0.83, 1.00
6* Contact with objects and equipment															
62	Struck by object or equipment	1,651	11.0	909	10.6	0.89	0.87, 0.91	0.81	0.78, 0.84	775	9.0	0.86	0.84, 0.89	0.83	0.80, 0.85
63	Struck against object or equipment	466	3.1	75	0.9	0.43	0.35, 0.50	0.73	0.62, 0.83	72	0.8	0.45	0.36, 0.54	0.78	0.66, 0.87
64	Caught in or compressed by equipment	505	3.4	264	3.1	0.87	0.83, 0.90	0.82	0.77, 0.87	225	2.6	0.82	0.77, 0.87	0.84	0.79, 0.89
7* Overexertion and bodily reaction															
70	Overexertion and bodily reaction, uns	188	1.3	15	0.2	0.09	0.05, 0.14	0.40	0.16, 0.68	5	0.1	0.00	–	–	–
71	Overexertion involving outside sources[e]	4,189	27.9	3320	38.6	0.97	0.96, 0.97	0.94	0.93, 0.95	3625	42.1	0.98	0.97, 0.98	0.92	0.91, 0.93
72	Repetitive motions involving micro tasks	484	3.2	332	3.9	0.89	0.85, 0.92	0.79	0.74, 0.83	222	2.6	0.87	0.82, 0.91	0.89	0.84, 0.93
73	Other exertions or bodily reactions	916	6.1	370	4.3	0.78	0.74, 0.81	0.82	0.78, 0.86	409	4.7	0.80	0.76, 0.84	0.80	0.76, 0.84
X* All other classifiables n < 100 in training dataset															
xx	Other small n < 100 classifiable categories[h]	632	4.2	43	0.5	0.24	0.18, 0.31	0.88	0.75, 0.96	47	0.5	0.25	0.19, 0.32	0.87	0.74, 0.95
Nonclassifiable															
9999	Nonclassifiable	569	3.8	174	2.0	0.66	0.60, 0.72	0.72	0.65, 0.78	179	2.1	0.74	0.67, 0.80	0.79	0.73, 0.85
Overall		15,000	100.0	8,612	100.0	0.88	0.88, 0.89	0.88	0.88, 0.89	8,612	100.0	0.89	0.88, 0.89	0.89	0.88, 0.89

*Asterisks denote a summary level code not assigned to individual cases.

NB_{SW}: Naïve Bayes Single Word Model, $NB_{BI-GRAM}$: Naïve Bayes Bi-gram Word Model, SVM: Support Vector Machine, LR: Logistic Regression.

[a] Codes assigned by expert manual coders are the Gold Standard. The distribution of the original gold standard dataset, 15,000 are shown for comparison with the distribution resulting from using only the machine coded data.

[b] A filter is just a technique to decide which narratives the computer should classify vs. which should be left for a human to read and classify.

[c] n_{pred} is number predicted into category (this includes both correct and incorrect predictions).

[d] $\%_{pred}$ is percent of cases in whole dataset predicted into category.

[e] The distribution of two-digit classifications will be skewed towards categories with high sensitivity, biasing the finally distribution of the coded datasets.

[f] Sen is Sensitivity: (true positives) the percentage of narratives that had been coded by the experts into each category that were also assigned correctly by the algorithm. Sensitivity values are calculated for the cases contained within the 8,612 (e.g. agree dataset). The sensitivity calculation includes *only* the n_{pred} cases that were *predicted correctly*, i.e. n = 7,569 (data not shown) of the 8,612 cases were correctly predicted by the algorithm resulting in an overall sensitivity of .88. Category specific sensitivities: for example, intentional injury by person, using the ensemble strategy, the sensitivity is calculated as 33 cases predicted correctly out of 45 (those cases from this category present in the 8,612 dataset – data not shown) resulting in a sensitivity of .73.

[g] PPV is Positive Predicted Value: the percentage of narratives correctly coded into a specific category out of all narratives placed into that category by the algorithm.

[h] Overall average results of all small categories.

Table 2b
The accuracy of selective computer coding: four-model ensemble approach, narratives where each algorithm assigned the same classification (57% of the dataset only are classified, n = 8612) vs. those predicted by the Logistic Regression algorithm alone with the highest probabilities (top 57%, n = 8612) (small categories only, where ncat < 100).

		Gold Standard[a]	Four-Model Ensemble Filter[b]: Narratives where LR = NB_{SW} = SVM = $NB_{BI-GRAM}$					Narratives where LR Predicted the Classification at a Very High Probability Level (top 57%)				
BLS OIICS 2-Digit Event Code		(n)	n_{pred}[c]	Sen[d]	95% CI	PPV[e]	95% CI	n_{pred}[c]	Sen[d]	95% CI	PPV[e]	95% CI
1* Violence and other injuries by persons or animals												
12	Injury by person – intentional or intent unknown	96	5	0.10	0.02, 0.27	0.60	0.15, 0.95	3	0.04	0.00, 0.20	0.33	0.01, 0.91
13	Animal and insect related incidents	99	15	0.68	0.45, 0.86	1.00	0.78, 1.00	24	0.77	0.59, 0.90	1.00	0.86, 1.00
2* Transportation incidents												
20	Transportation incident, unspecified	3	0	0.00	–	–	–	0	0.00	–	–	–
21	Aircraft incidents	22	3	0.60	0.15, 0.95	1.00	0.29, 1.00	3	0.43	0.10, 0.82	1.00	0.29, 1.00
22	Rail vehicle incidents	6	0	0.00	–	–	–	0	0.00	–	–	–
23	Animal & other non-motorized vehicle transport incidents	14	1	0.33	0.01, 0.91	1.00	0.03, 1.00	0	0.00	–	–	–
25	Water vehicle incidents	11	0	0.00	–	–	–	0	0.00	–	–	–
3* Fires and explosion												
31	Fires	22	1	0.20	0.01, 0.72	1.00	0.03, 1.00	2	0.20	0.01, 0.72	0.50	0.01, 0.99
32	Explosions	21	1	1.00	0.03, 1.00	1.00	0.03, 1.00	0	0.00	–	–	–
4* Falls, slips, trips												
40	Fall, slip, trip, unspecified	4	0	0.00	–	–	–	0	0.00	–	–	–
44	Jumps to lower level	57	3	0.21	0.05, 0.51	1.00	0.29, 1.00	8	0.31	0.12, 0.62	0.63	0.24, 0.91
45	Fall or jump curtailed by personal fall arrest system	3	0	0.00	–	–	–	0	0.00	–	–	–
5* Exposure to harmful substances or environments												
50	Exposure to harmful substances or environ, unspecified	23	1	0.00	–	–	–	0	0.00	–	–	–
51	Exposure to electricity	27	2	1.00	0.16, 1.00	1.00	0.16, 1.00	1	0.25	0.01, 0.81	1.00	0.03, 1.00
52	Exposure to radiation and noise	38	9	1.00	0.59, 1.00	0.78	0.40, 0.97	5	0.83	0.36, 1.00	1.00	0.48, 1.00
54	Exposure to air and water pressure change	1	0	0.00	–	–	–	0	0.00	–	–	–
57	Exposure to traumatic or stressful even nec	32	1	0.25	0.01, 0.81	1.00	0.03, 1.00	1	0.17	0.00, 0.64	1.00	0.03, 1.00
59	Exposure to harmful substances or environments, nec	1	0	0.00	–	–	–	0	0.00	–	–	–
6* Contact with objects and equipment												
60	Contact with objects and equipment, uns	78	1	0.04	0.00, 0.21	1.00	0.03, 1.00	0	0.00	–	–	–
61	Needle stick	1	0	0.00	–	–	–	0	0.00	–	–	–
65	Struck/caught/crush in collapsing structure, equip or material	5	0	0.00	–	–	–	0	0.00	–	–	–
66	Rubbed or abraded by friction or pressure	16	0	0.00	–	–	–	0	0.00	–	–	–
67	Rubbed abraded or jarred by vibration	7	0	0.00	–	–	–	0	0.00	–	–	
69	Contact with objects and equipment, nec	1	0	0.00	–	–	–	0	0.00	–	–	–
7* Overexertion and bodily reaction												
74	Bodily conditions nec	20	0	0.00	–	–	–	0	0.00	–	–	–
78	Multiple types of overexertions and bodily reactions	23	0	0.00	–	–	–	0	0.00	–	–	–
79	Overexertion and bodily reaction and exertion, nec	1	0	0.00	–	–	–	0	0.00	–	–	–
Overall		632	43	0.24	0.18, 0.32	0.88	0.75, 0.96	47	0.25	0.19, 0.33	0.87	0.74, 0.95

*Asterisks denote a summary level code not assigned to individual cases.
NB_{SW}: Naïve Bayes Single Word Model, $N_{BBI-GRAM}$: Naïve Bayes Bi-gram Word Model, SVM: Support Vector Machine, LR: Logistic Regression.

[a] Codes assigned by expert manual coders are the Gold Standard. The counts of the cases from the small categories in the original gold standard dataset are shown for comparison with what would be found using only the machine coded data.

[b] A filter is just a technique to decide which narratives the computer should classify vs. which should be left for a human to read and classify.

[c] n_{pred} is number predicted into category (this includes both correct and incorrect predictions).

[d] Sen is Sensitivity: (true positives) the percentage of narratives that had been coded by the experts into each category that were also assigned correctly by the algorithm. Sensitivity values are presented in this table for the small category cases contained within the 8,612 (agree dataset). The sensitivity calculation includes *only* the n_{pred} cases that were *predicted correctly*, i.e. n = 38 of the 156 cases (data not shown) that were contained in the agree dataset were correctly predicted by the algorithm resulting in an overall sensitivity of .24. Category specific sensitivities: for example, Injury by person-intentional or intent unknown, using the ensemble 'strategy, the sensitivity is calculated as 3 out of 30 cases (those cases from this category present in the 8,612 dataset – data not shown) correctly predicted resulting in a sensitivity of .10.

[e] PPV is Positive Predicted Value: the percentage of narratives correctly coded into a specific category out of all narratives placed into that category by the algorithm.

Table 3a
The accuracy of the human-machine classification system: implementation of a strategic filter[a] based on a four-model ensemble approach vs. prediction strengths generated by the Logistic Regression algorithm (large categories, where ncat ≥ 100).

BLS OIICS 2, Digit Event Code		Gold Standard[b]		Human-Machine Performance using Four-Model Ensemble Filter for Machine Classifications[c]: LR = NB_{SW} = SVM = $NB_{BI-GRAM}$						Human-Machine Performance Using LR Algorithm Alone for Machine Classifications[c]							
		n	%	n_{pred}[d]	$\%_{pred}$[e,f]	Sen[g]	95% CI	PPV[h]	95% CI	n_{pred}[d]	$\%_{pred}$[e,f]	Sen[g]	95% CI	PPV[h]	95% CI	% Agreement between 2 Manual Coders[i]	Fleiss κ[j] Kappa
1* Violence and other injuries by persons or animals																	
11	Intentional injury by person	159	1.1	149	1.0	0.92	0.87, 0.96	0.99	0.95, 1.00	144	1.0	0.90	0.84, 0.94	0.99	0.96, 1.00	81%, 97%	0.85
2* Transportation incidents																	
24	Pedestrian vehicular incidents	120	0.8	109	0.7	0.88	0.80, 0.93	0.96	0.91, 0.99	107	0.7	0.89	0.82, 0.94	1.00	0.97, 1.00	57%, 78%	0.65
26	Roadway incidents involv motorized land vehicle	650	4.3	668	4.5	0.99	0.98, 1.00	0.97	0.95, 0.98	667	4.4	0.99	0.98, 1.00	0.96	0.95, 0.98	93%, 96%	0.94
27	Nonroadway incidents involv motorized land vehicles	136	0.9	124	0.8	0.86	0.79, 0.92	0.94	0.89, 0.98	123	0.8	0.86	0.79, 0.91	0.95	0.90, 0.98	52%, 84%	0.62
4* Falls, slips, trips																	
41	Slip or trip without fall	806	5.4	774	5.2	0.90	0.87, 0.92	0.93	0.91, 0.95	820	5.5	0.92	0.90, 0.94	0.91	0.88, 0.93	66%, 89%	0.71
42	Falls on same level	2,148	14.3	2202	14.7	0.95	0.94, 0.96	0.93	0.92, 0.94	2179	14.5	0.94	0.93, 0.95	0.93	0.92, 0.94	85%, 93%	0.86
43	Falls to lower level	1,065	7.1	1084	7.2	0.92	0.90, 0.94	0.90	0.88, 0.92	1041	6.9	0.89	0.87, 0.91	0.91	0.89, 0.93	78%, 92%	0.81
5* Exposure to harmful substances or environments																	
53	Exposure to temperature extremes	141	0.9	138	0.9	0.94	0.89, 0.98	0.96	0.92, 0.99	132	0.9	0.89	0.83, 0.94	0.95	0.90, 0.98	82%, 98%	0.88
55	Exposure to other harmful substances	175	1.2	171	1.1	0.94	0.90, 0.97	0.96	0.93, 0.99	167	1.1	0.95	0.90, 0.98	0.99	0.97, 1.00	81%, 96%	0.87
6* Contact with objects and equipment																	
62	Struck by object or equipment	1,651	11.0	1731	11.5	0.94	0.93, 0.95	0.90	0.89, 0.91	1686	11.2	0.94	0.93, 0.95	0.92	0.93, 0.95	82%, 90%	0.82
63	Struck against object or equipment	466	3.1	412	2.7	0.84	0.80, 0.87	0.95	0.93, 0.97	413	2.8	0.85	0.82, 0.88	0.96	0.94, 0.98	66%, 83%	0.68
64	Caught in or compressed by equipment	505	3.4	519	3.5	0.93	0.91, 0.95	0.91	0.88, 0.93	499	3.3	0.92	0.89, 0.94	0.93	0.90, 0.95	72%, 83%	0.75
7* Overexertion and bodily reaction																	
70	Overexertion and bodily reaction, uns	188	1.3	135	0.9	0.67	0.60, 0.74	0.93	0.88, 0.97	153	1.0	0.79	0.72, 0.84	0.97	0.93, 0.99	6%, 48%	0.19
71	Overexertion involving outside sources	4,189	27.9	4295	28.6	0.98	0.97, 0.98	0.95	0.95, 0.96	4403	29.4	0.98	0.98, 0.99	0.93	0.93, 0.94	87%, 95%	0.87
72	Repetitive motions involving micro tasks	484	3.2	522	3.5	0.93	0.91, 0.95	0.87	0.83, 0.89	479	3.2	0.94	0.92, 0.96	0.95	0.93, 0.97	71%, 83%	0.75
73	Other exertions or bodily reactions	916	6.1	895	6.0	0.91	0.88, 0.92	0.93	0.91, 0.94	916	6.1	0.91	0.89, 0.93	0.91	0.89, 0.93	56%, 85%	0.64
X* All other classifiables n < 100 in training dataset																	
xx	Other small n < 100 classifiable categories[k]	632	4.2	519	3.5	0.81	0.78, 0.84	0.99	0.99, 1.00	515	3.4	0.81	0.77, 0.84	0.99	0.97, 1.00	–	–
Nonclassifiable																	
9999	Nonclassifiable	569	3.8	553	3.7	0.89	0.86, 0.91	0.91	0.88, 0.93	556	3.7	0.91	0.89, 0.93	0.93	0.91, 0.95	69%, 84%	0.72
Overall		15,000	100.0	15,000	100.0	0.93	0.93, 0.94	0.93	0.93, 0.94	15,000	100.0	0.93	0.93, 0.94	0.93	0.93, 0.94	77%, 90%	0.78

*Asterisks denote a summary level code not assigned to individual cases.
NB_{SW}: Naïve Bayes Single Word Model, $NB_{BI-GRAM}$: Naïve Bayes Bi-gram Word Model, SVM: Support Vector Machine, LR: Logistic Regression.

[a] A filter is just a technique to decide which narratives the computer should classify vs. which should be left for a human to read and classify.

[b] Codes assigned by expert manual coders are the Gold Standard.

[c] Machine Classifications: For ensemble method include where the algorithms agreed on the code, for LR include narratives classified at a very high probability. In both Human-Machine approaches, 57% of the dataset is machine coded, 43% is manually coded.

[d] n_{pred} is number predicted into category (this includes both correct and incorrect predictions).

[e] $\%_{pred}$ is percent of cases in whole dataset predicted into category.

[f] The distribution of two-digit classifications will be skewed towards categories with high sensitivity, biasing the finally distribution of the coded datasets.

[g] Sen is Sensitivity: (true positives) the percentage of narratives that had been coded by the experts into each category that were also assigned correctly by the algorithm.

[h] PPV is Positive Predicted Value: the percentage of narratives correctly coded into a specific category out of all narratives placed into that category by the algorithm.

[i] Inter-rater agreement between 4 expert manual coders classifying a separate set of narratives (n = 4000). Two-coder agreement, for example, 6 total comparisons, coder 1 compared with 2, 3, 4, coder 2 compared with 3, 4, coder 3 compared with 4.

[j] Fleiss κ between 0 and 1, >0.6 considered good agreement, >0.8 considered very good agreement.

[k] Results when grouping all of the small categories (overall average results of small categories).

Table 3b
The accuracy of the human-machine classification system: implementation of a strategic filter[a] based on a four-model ensemble approach vs. prediction strengths generated by the Logistic Regression algorithm (small categories only, where ncat < 100).

		Gold Standard[b]	Human-Machine Performance using Four-Model Ensemble Filter for Machine Classifications[c]: LR = NB_{SW} = SVM = $NB_{BI-GRAM}$					Human-Machine Performance Using LR Algorithm Alone for Machine Classifications[c]					% Agreement between 2 Manual Coders[g]	Fleiss κ[h] Kappa
BLS OIICS 2-Digit Event Code		(n)	n_{pred}[d]	Sen[e]	95% CI	PPV[f]	95% CI	n_{pred}[d]	Sen[e]	95% CI	PPV[f]	95% CI		
1* Violence and other injuries by persons or animals														
12	Injury by person – intentional or intent unknown	96	71	0.72	0.62, 0.81	0.97	0.90, 1.00	72	0.73	0.64, 0.82	0.97	0.90, 1.00	47%–78%	0.57
13	Animal and insect related incidents	99	92	0.93	0.86, 0.97	1.00	0.96, 1.00	92	0.93	0.86, 0.97	1.00	0.96, 1.00	79%–94%	0.87
2* Transportation incidents														
20	Transportation incident, unspecified	3	2	0.67	0.09, 0.99	1.00	0.16, 1.00	3	1.00	0.29, 1.00	1.00	0.29, 1.00	0%–0%	0.00
21	Aircraft incidents	22	20	0.91	0.71, 0.99	1.00	0.83, 1.00	18	0.82	0.60, 0.95	1.00	0.81, 1.00	0%–75%	0.17
22	Rail vehicle incidents	6	3	0.50	0.12, 0.88	1.00	0.29, 1.00	4	0.67	0.22, 0.96	1.00	0.40, 1.00	0%–100%	0.67
23	Animal & other non-motorized vehicle transport incidents	14	12	0.86	0.57, 0.98	1.00	0.74, 1.00	10	0.71	0.42, 0.92	1.00	0.69, 1.00	0%–0%	0.00
25	Water vehicle incidents	11	4	0.36	0.11, 0.69	1.00	0.40, 1.00	3	0.27	0.06, 0.61	1.00	0.29, 1.00	0%–88%	0.25
3* Fires and explosion														
31	Fires	22	18	0.82	0.60, 0.95	1.00	0.81, 1.00	19	0.82	0.60, 0.95	0.95	0.74, 1.00	55%–88%	0.58
32	Explosions	21	21	1.00	0.84, 1.00	1.00	0.84, 1.00	17	0.81	0.58, 0.95	1.00	0.80, 1.00	44%–83%	0.46
4* Falls, slips, trips														
40	Fall, slip, trip, unspecified	4	3	0.75	0.19, 0.99	1.00	0.29, 1.00	3	0.75	0.19, 0.99	1.00	0.29, 1.00	0%–0%	0.00
44	Jumps to lower level	57	46	0.81	0.68, 0.90	1.00	0.92, 1.00	49	0.81	0.70, 0.91	0.94	0.83, 0.99	51%–90%	0.65
45	Fall or jump curtailed by personal fall arrest system	3	2	0.67	0.09, 0.99	1.00	0.16, 1.00	0	0.00	–	–	–	0%–0%	0.00
5* Exposure to harmful substances or environments														
50	Exposure to harmful substances or environ, unspecified	23	22	0.91	0.72, 0.99	0.95	0.77, 1.00	21	0.91	0.72, 0.99	1.00	0.84, 1.00	21%–88%	0.33
51	Exposure to electricity	27	27	1.00	0.87, 1.00	1.00	0.87, 1.00	24	0.89	0.71, 0.98	1.00	0.86, 1.00	65%–88%	0.81
52	Exposure to radiation and noise	38	40	1.00	0.91, 1.00	0.95	0.83, 0.99	37	0.97	0.86, 1.00	1.00	0.91, 1.00	54%–100%	0.80
54	Exposure to air and water pressure change	1	1	1.00	0.03, 1.00	1.00	0.03, 1.00	0	0.00	–	–	–	0%–100%	0.40
57	Exposure to traumatic or stressful even nec	32	29	0.91	0.75, 0.98	1.00	0.88, 1.00	27	0.84	0.67, 0.95	1.00	0.87, 1.00	73%–85%	0.80
59	Exposure to harmful substances or environments, nec	1	1	1.00	0.03, 1.00	1.00	0.03, 1.00	1	1.00	0.03, 1.00	1.00	0.03, 1.00	0%–100%	0.12
6* Contact with objects and equipment														
60	Contact with objects and equipment, uns	78	55	0.71	0.59, 0.80	1.00	0.94, 1.00	64	0.82	0.72, 0.90	1.00	0.94, 1.00	12%–63%	0.25
61	Needle stick	1	1	1.00	0.03, 1.00	1.00	0.03, 1.00	1	1.00	0.03, 1.00	1.00	0.03, 1.00	–	–
65	Struck/caught/crush in collapsing structure, equip or material	5	5	1.00	0.48, 1.00	1.00	0.48, 1.00	5	1.00	0.48, 1.00	1.00	0.48, 1.00	0%–0%	0.33
66	Rubbed or abraded by friction or pressure	16	14	0.88	0.62, 0.98	1.00	0.77, 1.00	14	0.88	0.62, 0.98	1.00	0.77, 1.00	0%–50%	0.11

67	Rubbed abraded or jarred by vibration	7	3	0.43	0.10, 0.82	1.00	0.29, 1.00	4	0.57	0.18, 0.90	1.00	0.40, 1.00	0%–67%		0.14
69	Contact with objects and equipment, nec	1	0	0.00	–	–	–	1	1.00	0.03, 1.00	1.00	0.03, 1.00	–		–
7* Overexertion and bodily reaction															
74	Bodily conditions nec	20	16	0.80	0.56, 0.94	1.00	0.79, 1.00	17	0.85	0.62, 0.97	1.00	0.80, 1.00	0%–75%		0.33
78	Multiple types of overexertions and bodily reactions	23	11	0.48	0.27, 0.69	1.00	0.72, 1.00	9	0.39	0.20, 0.61	1.00	0.66, 1.00	0%–0%		0.00
79	Overexertion and bodily reaction and exertion, nec	1	0	0.00	–	–	–	0	0.00	–	–	–	–		–
Overall		632	519	0.81	0.78, 0.84	0.99	0.98, 1.00	515	0.81	0.78, 0.84	0.99	0.97, 1.00			

*Asterisks denote a summary level code not assigned to individual cases.
NB_{SW}: Naïve Bayes Single Word Model, $NB_{BI\text{-}GRAM}$: Naïve Bayes Bi-gram Word Model, SVM: Support Vector Machine, LR: Logistic Regression.
[a] A filter is just a technique to decide which narratives the computer should classify vs. which should be left for a human to read and classify.
[b] Codes assigned by expert manual coders are the Gold Standard.
[c] Machine Classifications: For ensemble method include where the algorithms agreed on the code, for LR include narratives classified at a very high probability. In both Human-Machine approaches, 57% of the dataset is machine coded, 43% is manually coded.
[d] n_{pred} is number predicted into category (this includes both correct and incorrect predictions).
[e] Sen is Sensitivity: (true positives the percentage of narratives that had been coded by the experts into each category that were also assigned correctly by the algorithm.
[f] PPV is Positive Predicted Value: the percentage of narratives correctly coded into a specific category out of all narratives placed into that category by the algorithm.
[g] Inter-rater agreement between 4 expert manual coders classifying a separate set of narratives (n = 4000). Two-coder agreement, for example, 6 total comparisons, coder 1 compared with 2, 3, 4, coder 2 compared with 3, 4, coder 3 compared with 4.
[h] Fleiss κ between 0 and 1, >0.6 considered good agreement, >0.8 considered very good agreement.

learning (i.e. "slip or trip without a fall") and the low PPV indicates there were many false positives using this approach. For this category, the NLP rules performed poorly compared with LR and adding in the additional cases identified from the NLP would lower the PPV of this category substantially.

4. Discussion

For this study we compare the utility of four classification algorithms for classifying the event leading to injury using injury narratives of a large WC dataset. One advantage of using "off-the-shelf" approaches is that they can sometimes be quickly and easily combined to yield results quite competitive with modern state-of-the-art classifiers, yet with minimal cost. For example, Wang and Manning (2012) found that a simple model relying more heavily on NB to classify shorter narratives and SVM for the longer ones outperformed several state-of-the-art classifiers for classification of short text snippets. This was accomplished without requiring the development of ontologies or complex preprocessing of the data (Wang and Manning, 2012).

These results easily show that, if resources are constrained at a specific low level (e.g. you only have human resources to classify 15–30% of the dataset), a simple approach with very good accuracy would be to apply a probability strength as the filter based on the LR algorithm alone. It is noteworthy that LR achieved the highest performance of the individual models and was comparable to the best ensemble approaches. We found that filtering out the 30% of narratives predicted at the lowest probabilities, allowing the LR algorithm to code 70% of the narratives and leaving 30% for manual review, resulted in fairly high accuracy (0.89).

We found, as expected, that the most conservative and most accurate ensemble approach would be to filter out cases for the algorithm to code based on agreement between all four models. Alternatively, the use of fewer models or models that operate in similar ways results in less filtering. This may be good if you cannot afford to manually classify a large portion of the narratives; however, you will sacrifice some performance. Noteworthy was the higher accuracy found when matching the NB algorithms with SVM, whereby the amount of manual review was also only about 1/3 of the original dataset. Previous studies have consistently shown that filtering on agreement between models can lead to large improvements in performance for small and hard-to-predict categories (Marucci-Wellman et al., 2011, 2015; Bertke et al., 2016; Nanda et al., 2016). We believe, there are several advantages of filtering in this way (vs on the strength metric of the classifier). If two or more algorithms agree, especially if they are making their assignment in different ways (i.e. optimization vs non optimization models) we can have more confidence that when used together the results in the long run will be more robust, than using one algorithm alone. We think this may become more evident as these methods are attempted on even larger datasets or to predict more refined codes (at the three or four digit levels). We know from our experience that finding the rare categories continues to be a challenge with any algorithm and even with manual coding; When two or more algorithms disagree on a code this appears to highlight that the narrative contains something unique which separates it apart from the larger categories. Finally, these methods automatically determine a fixed level of filtering where you can be confident that the resultant human-machine coded system will be fairly accurate without adding an additional level of analysis of the predicted results (i.e. trying to determine what would be a good level of filtering for overall results on large and small categories).

Using this approach to filtering builds on the idea that models making predictions in fundamentally different ways are less likely to agree for hard-to-predict categories and more likely to agree for

Table 4
Extracting out very unique types of injury narratives from the prediction dataset using Logistic Regression vs. Natural Language Processing alone.

	Categorys of Interest	Gold Std	Logisic Regression Alone			Natural Language Processing Rules Alone			# Extra Identified Correctly (% Category)	
		n	n_{pred}[a]	Sen[b]	PPV[c]	n_{pred}[a]	Sen[b]	PPV[c]		
Sample Categories with very unique characteristics										
32	Explosions	21	7	0.19	0.57	18	0.71	0.83	11	(52.4)
51	Exposure to electricity	27	5	0.15	0.80	17	0.59	0.94	12	(44.4)
53	Exposure to temperature extremes	141	104	0.60	0.82	113	0.65	0.81	6	(4.3)
55	Exposure to other harmful substances	175	102	0.47	0.81	132	0.60	0.80	22	(12.6)
Sample category without very unique characteristics										
41	Slip or trip without a fall	806	870	0.70	0.65	1069	0.58	0.58	199	(24.7)

[a] n_{pred} is number predicted into category.
[b] Sen is Sensitivity: (true positives) the percentage of narratives that had been coded by the experts into each category that were also assigned correctly by the algorithm.
[c] PPV is Positive Predicted Value: the percentage of narratives correctly coded into a specific category out of all narratives placed into that category by the algorithm.

easy categories. These are the ones that the machine should code since humans are also prone to errors and the computer can be more systematic. One requirement for this approach to be effective is that each model must perform reasonably well on its own. The second is that the predictions must be independent. As noted above, previous studies have shown that all four algorithms meet the first requirement. For the second requirement there are fundamental differences in these models which should be considered. NB differs from both SVM and LR in that it does not directly attempt to minimize an error function. SVM differs from NB and LR in that it does not estimate the probability of the category given the words. The use of bi-gram predictions might differ from single word-based predictions when bi-grams have very category-specific meanings.

Our results support and complement the results by Bertke and colleagues yet using an entirely different set of data and expanding the number of categories from 19 (used in Bertke et al., 2016) that needed to be differentiated to 44. We included in our dataset any classification that was made by two separate coders in classifying the 15,000 unique narratives. Some classifications were assigned only a very few times (even as few as once). Also, since WC narratives are often composed of short incomplete sentences with a lot of ambiguity, sometimes the specific cause of injury cannot be discriminated between two categories. One example would be the narrative "EMPLOYEE WAS CLEANING A CONCRETE PUMP AND RIGHT HAND WAS SOMEHOW SEVERED IN THE PROCESS" Here it can be easily seen that, based on interpretation of the narrative, this was some sort of contact injury, but it is difficult to determine from this narrative alone if it was a struck by, struck against or caught incident. Since we know at a minimum that the injury was caused by some form of contact, we still can code this with that degree of specificity as contact non-specified and this provides information that can be used for surveillance. The words from the narratives in the non-specified categories, therefore, will always be very similar to the words in the narratives from the more specific categories (struck by and struck against) and it will be very difficult for a machine learning algorithm to figure this out. During manual coding, coders often will disagree on these types of classifications, some coders believing the information is implied in the narrative, others feeling the information is not in the narrative. Given that these limitations exist in administrative narratives, we believe the accuracy results of even these categories to be quite good as compared with the agreement between manual coders (see Tables 3a and 3b). Therefore, the human-machine methods presented here can be used not only to find large, easily defined categories (such as overexertion or fall on the same level) but also to classify to a lower level of specificity often required of these types of narratives. We note that it is very important that your training narratives come from the same data source with the same level of detail that you plan to use to predict codes through machine learning (i.e. workers compensation data and Emergency Department data would require different training data sets)

It will always be difficult for an algorithm on its own to be able to assign classifications in all categories with the same level of confidence and very difficult to improve the accuracy of computer-generated codes for the small categories. Getting the rarer events coded accurately requires either sophisticated filtering or integration of highly tailored resource-intensive methods such as natural language processing. It is interesting to speculate that such methods could be tailored to help find emerging risks.

However, applying some very simple rules may be a way of complementing results of the machine classifiers when using an algorithm alone to classify all narratives. This provides an alternative strategy particularly for some of the small categories. It may also be the best approach for extracting out very specific narratives such as injuries caused by electrical contact from a large dataset if the purpose of your research or surveillance effort is to look specifically at one type of event or outcome.

However, if you want better results and have the resources to do some manual coding, the filtering approach will do better for these same categories (see Table 3b). While developing NLP rules can be tedious, we limited our development to those categories where we knew some unique words would be able to easily find some categories. These simple rules could be shared among people using the same coding protocol for their work (i.e. once someone has developed a rule to extract electrical injuries, this same rule can be used by many others in the field with very good results). However, while a finite set of rules can be developed for some categories where there are very specific words (i.e. electrocutions) it would very difficult to come up with a set of NLP rules that would do well for other categories like 'slip or trip without a fall' which has very similar words to those used in the larger 'fall on same level' category. More research is necessary in integrating speech tagging (lift as a verb vs. noun) into both the utility of the algorithms for making predictions and for generating a more complex enhanced set of rules for specific categories to be shared. There is a potential for improving on both the ensemble and LR-alone approaches by integrating speech tagging and NLP rules.

We have demonstrated with this work that there are many alternatives to manually coding all narratives from administrative datasets for surveillance. The human-machine methods we are suggesting here result in high accuracy of the machine coded narratives and allow for a much smaller subset of narratives to be manually reviewed. We believe this may result in higher accuracy of the final coded dataset given it is well known there are inconsistencies and errors in human codes. It has been shown that devoting more expertise and time to a subset of unique cases, that are extracted out because algorithms disagreed on the code, can result in better performance than when coders are faced with all cases (Nanda et al., 2016) mixed together which was the case here for our gold

standard codes; Finding and categorizing a rare event without first reducing the dataset to a manageable size creates a situation similar to finding a needle in a haystack. We also know that human coders can become bored with or just inattentive to repetitive and mundane narratives which may lead to coding inconsistencies over time. Conversely, an algorithm can code systematically and consistently for a limitless amount of repetitive, mundane narratives without experiencing fatigue.

5. Conclusion

We stated from the beginning (Wellman et al., 2004) that we believe that manual coding should never be completely replaced for such short noisy injury narratives as would be found in many administrative datasets. A best practice approach should incorporate some manual coding, assigning a computer classification only for more repetitive events where the models are able to confidently predict the correct classification. To classify injury narratives contained in large administrative datasets for surveillance, we recommend incorporation of methods based on human-machine pairings such as we have done here, utilizing readily available off-the-shelf machine learning techniques in order to maximize accuracy across many categories while minimizing manual review (e.g. to apply the correct filter for any particular dataset). These methods build off our prior results by integrating Logistic Regression and Support Vector Machine algorithms with Naïve Bayes and result in high accuracy, potentially higher than manual review alone, while significantly reducing the human resources required to accomplish the task. Finding the rare categories continues to be a challenge with any algorithm and even with manual coding. It is clear that finding these categories will take enhanced strategies such as integration of NLP or well thought out ensemble approaches.

Acknowledgements

The Liberty Mutual Research Institute for Safety sponsored the study. The parent company, Liberty Mutual Insurance, had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

The authors would like to express their gratitude to Dr. Tin-Chi Lin, Dr. Yulan Liang and Ted Courtney for reviewing the manuscript and Ms Margaret Rothwell for editing on the final manuscript. We would also like to thank Alex Measure from the Bureau of Labor Statistics for making us aware of the utility of the logistic regression algorithm and Python for this work.

References

Bertke, S.J., Meyers, A.R., Wurzelbacher, S.J., Measure, A.C., Lampl, M.P., Robins, D., 2016. Comparison of methods for auto-coding causation of injury narratives. Accid. Anal. Prev. 88, 117–123.

CDC, 2015. Leading Causes of Death Reports, National and Regional, 1999–2014. Centers for Disease Control and Prevention, Available at: http://webappa.cdc.gov/sasweb/ncipc/leadcaus10_us.html (accessed 27.08.15).

Chen, L., Vallmuur, K., Nayak, R., 2015. Injury narrative text classification using factorization model. BMC Med. Inform. Decis. Mak. 15 (1), S5.

Horan, J.M., Mallonee, S., 2003. Injury surveillance. Epidemiol. Rev. 25, 24–42.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning with Applications in R. Springer, New York, Available at: http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf.

Lehto, M.R., Sorock, G., 1996. Machine learning of motor vehicle accident categories from narrative data. Methods Inf. Med. 35 (4–5), 309–316.

Lehto, M., Marucci-Wellman, H., Corns, H., 2009. Bayesian methods: a useful tool for classifying injury narratives into cause groups. Inj. Prev. 15 (4), 259–265.

Lincoln, A.E., Sorock, G.S., Courtney, T.K., Wellman, H.M., Smith, G.S., Amoroso, P.J., 2004. Using narrative text and coded data to develop hazard scenarios for occupational injury interventions. Inj. Prev. 10 (4), 249–254.

Lombardi, D.A., Pannala, R., Sorock, G.S., Wellman, H., Courtney, T.K., Verma, S., Smith, G.S., 2005. Welding related occupational eye injuries: a narrative analysis. Inj. Prev. 11 (3), 174–179.

Lombardi, D.A., Matz, S., Brennan, M.J., Smith, G.S., Courtney, T.K., 2009. Etiology of work-related electrical injuries: a narrative analysis of workers' compensation claims. J. Occup. Environ. Hyg. 6 (10), 612–623.

Marucci-Wellman, H., Lehto, M., Corns, H., 2011. A combined Fuzzy and Näive Bayesian strategy can be used to assign event codes to injury narratives. Inj. Prev. 17 (6), 407–414.

Marucci-Wellman, H., Lehto, M., Corns, H., 2015. A practical tool for public health surveillance: semi-automated coding of short injury narratives from large administrative databases using Naïve Bayes algorithms. Accid. Anal. Prev. 84, 165–176.

McKenzie, K., Scott, D.A., Campbell, M.A., McClure, R.J., 2010. The use of narrative text for injury surveillance research: a systematic review. Accid. Anal. Prev. 42, 354–363.

Measure, A.C., 2014. Automated Coding of Worker Injury Narratives. JSM 2014 – Government Statistics Section. U.S. Bureau of Labor Statistics, Washington, D.C, Available at: http://www.bls.gov/osmr/pdf/st140040.pdf.

Nanda, G., Grattan, K.M., Chu, M.T., Davis, L.K., Lehto, M.R., 2016. Bayesian decision support for coding occupational injury data. J. Saf. Res. 57, 71–82.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Sorock, G., Ranney, T., Lehto, M., 1996. Motor vehicle crashes in roadway construction work zones: an analysis using narrative text from insurance claims. Accid. Anal. Prev. 28, 131–138.

Sorock, G.S., Smith, G.S., Reeve, G.R., Dement, J., Stout, N., Layne, L., Pastula, S., 1997. Three perspectives on work-related Injury surveillance systems. Am. J. Ind. Med. 32, 116–128.

Stutts, J.C., Reinfurt, D.W., Staplin, L., Rodgman, E.A., 2001. The Role of Driver Distraction in Traffic Crashes. AAA Foundation for Traffic Safety, Washington, D.C.

Taylor, J., Lacovaraa, A.V., Smith, G.S., Pandiana, R., Lehto, M., 2014. Near-miss narratives from the fire service: a Bayesian analysis. Accid. Anal. Prev. 62, 119–129.

Vallmuur, K., Marucci-Wellman, H.R., Taylor, J.A., Lehto, M., Corns, H.L., Smith, G.S., 2016. Harnessing information from injury narratives in the 'big data' era: understanding and applying machine learning for injury surveillance. Inj. Prev. 22 (Suppl. (1)), i34–i42.

Vallmuur, K., 2015. Machine learning approaches to analyzing textual injury surveillance data: a systematic review. Accid. Anal. Prev. 79, 41–49.

Verma, S.K., Lombardi, D.A., Chang, W.R., Courtney, T.K., Brennan, M.J., 2008. A matched case-control study of circumstances of occupational same-level falls and risk of wrist, ankle and hip fracture in women over 45 years of age. Ergonomics 51 (12), 1960–1972.

Wang, S., Manning, C.D., 2012. Baselines and bigrams: simple, good sentiment and topic classification. In: Proceedings of ACL Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, Jeju, Republic of Korea, Jul 8–14, pp. 90–94.

Wellman, H.M., Lehto, M.R., Sorock, G.S., 2004. Computerized coding of injury narrative data from the National Health Interview Survey. Accid. Anal. Prev. 36 (2), 165–171.

Williamson, A., Feyer, A.M., Stout, N., Driscoll, T., Usher, H., 2001. Use of narrative analysis for comparisons of the causes of fatal accidents in three countries: new Zealand, Australia, and the United States. Inj. Prev. 7 (Suppl. (1)), i15–i20.