Accident Analysis and Prevention 110 (2018) 115–127

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Full length article

Improving autocoding performance of rare categories in injury classification: Is more training data or filtering the solution?

Gaurav Nanda[a,*], Kirsten Vallmuur[b,c], Mark Lehto[a]

[a] School of Industrial Engineering, Purdue University, USA
[b] Current: Australian Centre for Health Services Innovation, School of Public Health and Social Work, Queensland University of Technology, Australia
[c] Formerly: Centre for Accident Research and Road Safety-Queensland, School of Psychology and Counselling, Queensland University of Technology, Australia

ARTICLE INFO

Keywords:
Injury autocoding
Rare categories
Text classification
Machine learning
More training data vs filtering
Human-machine systems

ABSTRACT

Introduction: Classical Machine Learning (ML) models have been found to assign the external-cause-of-injury codes (E-codes) based on injury narratives with good overall accuracy but often struggle with rare categories, primarily due to lack of enough training cases and heavily skewed nature of injurdata. In this paper, we have: a) studied the effect of increasing the size of training data on the prediction performance of three classical ML models: Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM) and Logistic Regression (LR), and b) studied the effect of filtering based on prediction strength of LR model when the model is trained on very-small (10,000 cases) and very-large (450,000 cases) training sets.

Method: Data from Queensland Injury Surveillance Unit from years 2002–2012, which was categorized into 20 broad E-codes was used for this study. Eleven randomly chosen training sets of size ranging from 10,000 to 450,000 cases were used to train the ML models, and the prediction performance was analyzed on a prediction set of 50,150 cases. Filtering approach was tested on LR models trained on smallest and largest training datasets. Sensitivity was used as the performance measure for individual categories. Weighted average sensitivity (WAvg) and Unweighted average sensitivity (UAvg) were used as the measures of overall performance. Filtering approach was also tested for estimating category counts and was compared with approaches of summing prediction probabilities and counting direct predictions by ML model.

Results: The overall performance of all three ML models improved with increase in the size of training data. The overall sensitivities with maximum training size for LR and SVM models were similar (~82%), and higher than MNB (76%). For all the ML models, the sensitivities of rare categories improved with increasing training data but they were considerably less than sensitivities of larger categories. With increasing training data size, LR and SVM exhibited diminishing improvement in UAvg whereas the improvement was relatively steady in case of MNB. Filtering based on prediction strength of LR model (and manual review of filtered cases) helped in improving the sensitivities of rare categories. A sizeable portion of cases still needed to be filtered even when the LR model was trained on very large training set. For estimating category counts, filtering approach provided best estimates for most E-codes and summing prediction probabilities approach provided better estimates for rare categories.

Conclusions: Increasing the size of training data alone cannot solve the problem of poor classification performance on rare categories by ML models. Filtering could be an effective strategy to improve classification performance of rare categories when large training data is not available.

1. Introduction and background

Injury surveillance data which includes external cause of injury codes (E-codes) are valuable in facilitating analyses to understand the primary causes of injuries to direct prevention efforts. Previous work in the area of safety research have analyzed injury data from various sources, such as, injury surveys conducted by government- National Health Injury Survey (NHIS) NHIS, 2017), workplace injuries- Workers Compensation Data and Survey of Occupational Injuries and Illnesses (SOII) (Wiatrowski, n.d.), and data collected in hospital emergency departments (Queensland Injury Surveillance Unit, n.d). Most of these injury databases contain basic fields for reporting details of injury, including injury diagnoses (nature and body region of injuries), narratives on reasons for presentation/how the injury occurred, and sometimes E-codes. If present, E-codes are typically assigned manually by either treating doctors/nurses or by clinical coders based on the

* Corresponding author.
E-mail address: gnanda@purdue.edu (G. Nanda).

https://doi.org/10.1016/j.aap.2017.10.020
Received 6 February 2017; Received in revised form 13 August 2017; Accepted 21 October 2017
Available online 08 November 2017
0001-4575/

narrative and other fields. Some data sources only collect injury narratives and do not routinely assign E-codes, limiting the utility of these data for injury surveillance purposes. Manually assigning E-codes to injury data is very time and resource consuming. As described by (Measure, 2014), the estimated effort for initial coding of about 300,000 incidents reported in SOII each year is around 25,000 man-hours apart from several additional man-hours to find and fix errors in manual coding.

Machine Learning (ML) methods that can learn from previously coded data can be used for assigning codes (autocoding) to injury data based on the injury narrative and other available fields. Various studies on autocoding the injury data have shown that classical machine learning algorithms such as Naïve Bayes (NB) (Marucci-Wellman et al., 2011) (Bertke et al., 2012) (Marucci-Wellman et al., 2015), Support Vector Machine (SVM) (Marucci-Wellman et al., 2017)(Chen et al., 2015), and Logistic Regression (LR) (Marucci-Wellman et al., 2017) (Bertke et al., 2016) yield good overall classification performance but often struggle to classify rare categories (i.e., categories with very few cases in the training and prediction datasets). Some of the possible reasons are: severely unbalanced datasets (i.e. large portion of cases belong to very few (typically 3–4) categories), ambiguous narratives, and inconsistency in manual coding of training data to name a few.

For most of the classical ML algorithms, the goal is to maximize the overall accuracy and the assumption is that distribution of data in the training set and prediction set is similar (Provost, 2000). For a severely unbalanced dataset (such as the injury dataset), most of the classical ML methods that aim to maximize accuracy over the entire dataset tend to over-predict the majority classes. This adversely affects the probability of prediction of rare categories and we often observe that the rare categories are under-predicted– resulting in lower sensitivity for rare categories. This issue is often referred to as the 'Class Imbalance' problem in machine learning research (Chawla et al., 2004) and is a fairly well-recognized and studied issue as most of real-world data is imbalanced in nature. Various data-sampling and algorithm-level approaches have been proposed to tackle this problem (Provost, 2000), (Chawla et al., 2004) (Phua et al., 2004) (Prati et al., 2015) (Sun et al., 2009).

Previous experimental studies on these approaches have shown that there is no single approach that works best, and the effectiveness of an approach depends on the classifier and the performance evaluation measure being used (Van Hulse et al., 2007).

1.1. Weighted average sensitivity vs unweighted average sensitivity

Most of the earlier studies in the area of autocoding injury data have reported the overall sensitivity (i.e., the weighted average sensitivity (WAvg) where weights are based on number of cases in each E-code category) of machine learning algorithms and some studies have reported the sensitivities of individual causation categories. While WAvg or overall sensitivity is a good measure for studying the overall prediction performance of a ML model (as it clearly indicates the proportion of cases correctly predicted), it can be misleading when the dataset is heavily skewed.

For example, if a dataset has two categories- A and B, where the category A accounts for 99% of cases in prediction and training sets, and a ML model predicts category A with considerably high sensitivity of 0.9 and category B with extremely low sensitivity of 0.01. In this situation, the overall sensitivity of the ML model will still be considerably high (0.89) ignoring the fact that the ML model performed poorly in predicting category B accurately. Given the unbalanced or skewed nature of the injury data, where most of the cases belong to very few E-code categories such as Falls, Struck, etc., using overall sensitivity as a performance measure does not give the true picture of the ML model's performance.

It has also been argued by many machine-learning researchers that overall sensitivity is not the best measure of performance of a ML model when dealing with a heavily skewed dataset because overall accuracy overweighs performance for the larger classes and does not reflect the performance for smaller classes (which have typically low recall and precision). Therefore, other measures of performance, such as F-measure, ROC-curve, etc. have been suggested for use when dealing with heavily skewed datasets (Guo et al., 2008).

The performance measure 'Unweighted Average Sensitivity' (UAvg), which is a simple average of sensitivities of each E-code category, does not consider the size of the category and is hence a relatively unbiased performance measure as compared to WAvg. In our earlier example of a heavily skewed dataset with two categories − A (99% cases, sensitivity = 0.9), and B (1% cases, sensitivity = 0.01), the UAvg of ML model will be 0.455 −which reflects the fact that the ML model performs well only for half (one out of two) categories. UAvg has been used as a reported measure of overall classification performance of ML models in a recent journal article by (Marucci-Wellman et al., 2017).

In this paper, we have also used UAvg as a measure of overall classification performance as our focus is on the prediction performance of ML models on rare categories. In studies where the main focus is on studying the total number of cases correctly predicted by a ML model, WAvg would be an appropriate performance measure. A good ML model should be able to classify all the categories accurately, which would mean a high value of both WAvg and UAvg.

We have summarized the results from previous work in autocoding injury data in Table 1, where the weighted average sensitivities and unweighted average sensitivities are presented as available in the papers.

As shown in Table 1, there is a considerable gap between WAvg and UAvg for most of the studies. These results re-iterate that most of the machine-learning methods do not perform well in classifying rare categories.

One of the approaches often suggested to improve the classification performance for rare categories is to increase the size of training set (Vallmuur et al., 2016). Previous studies on evaluating the impact of the size of training set on prediction performance have reported that increasing the size of training set resulted in improved overall sensitivity of autocoding, but the level of improvement progressively diminishes as the size of training set grows (Bertke et al., 2012) (Taylor et al., 2014) (Chen et al., 2015). In this paper, we examine if using a very large training dataset (up to 450,000 cases) to train the machine learning model can solve the problem of low sensitivity of autocoding rare categories.

It is to be noted that in all of the previous studies on autocoding injury narratives, the size of training dataset has been significantly less (maximum = 40,000 cases) than the size of training datasets used in this study (up to 450,000 cases). One of the previous studies in injury autocoding (Measure, 2014) used a relatively large training set of about 195,000 cases from SOII 2011 but since the sensitivities for individual E-code categories and number of E-code categories were not reported in the article, it is not included in Table 1.

1.2. Filtering

Filtering cases for manual review based on different approaches such as prediction strength and agreement in prediction results of different ML models, has been tested as an effective strategy to improve the classification performance of rare E-code categories (Wellman et al., 2004) (Corns et al., 2007) (Marucci-Wellman et al., 2011) (Marucci-Wellman et al., 2015) (Nanda et al., 2016) (Bertke et al., 2016) (Marucci-Wellman et al., 2017). Such filtering approaches would be particularly helpful when large training dataset (with enough cases of rare categories) is not available for the ML model to learn from. However, it is unknown whether filtering approaches on smaller data sets provide similar, better or poorer classification results to increasing training set sizes as there has been no research published comparing these approaches in this field.

Table 1
Weighted and Unweighted Average Sensitivity Reported in Previous Studies.

Reference	No. of Training Cases	No. of Prediction Cases	No. of Categories	Model Used	Weighted Average Sensitivity	Unweighted Average Sensitivity
Wellman et al. (2004)	5677	Same as Training Set	13	Single Word Fuzzy Bayes Multiple Word Fuzzy Bayes	71.3% 82.7%	55% 72.3%
Marucci et al. (2007)	7389	3000	8	Multiple Word Fuzzy Bayes Multiple Word and Two-Word Sequence Fuzzy Bayes	78% 79%	63% 64%
Marucci-Wellman et al., 201 Marucci-Wellman et al. (2011))	11000	3000	19	Fuzzy Bayes Naïve Bayes	64% 70%	– –
(Bertke et al., 2012)	2240	800	8 8	Naïve Bayes (Text Only) Naïve Bayes (Text + Injury)	84.1% 86%	57.7% 64.7%
Taylor et al. (2014)	764	Same as Training Set	12	Naïve Bayes Fuzzy Bayes	67.8% 74%	76.6% 52.5%
	764	293	12	Naïve Bayes and Fuzzy Bayes Agree	60.2%	–
Chen et al. (2015)	10,000	5000	20	Non Negative Matrix Factorization + SVM	92%	78.6%
Marucci-Wellman et al. (2015)	15,000	15,000	18	Single Word Naïve Bayes Two-Word Sequence Naïve Bayes	67% 65%	54.6% 49.6%
Nanda et al. (2016)	40,000	10,000	48	Single Word Naïve Bayes Two-Word Sequence Naïve Bayes	66% 69%	29% 30%
Bertke et al. (2016)	6200	1000	19	Naïve Bayes Single Word + Injury Naïve Bayes Single Word + Injury + Sequences Logistic Regression Single Word + Injury Logistic Regression Single Word + Injury + Sequence	65.3% 65.6% 69.4% 70.8%	57.7% 59.04% 56.68% 57.48%

1.3. Objectives of study

The main objectives of the current study were to: a) examine the effect of increasing the training data to a very large size on the prediction performance of three classical ML models: Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM) and Logistic Regression (LR), and b) compare the prediction results of increased training size with prediction strength based filtering on the LR model.

2. Methods

In this section, we discuss about the methodology of studying the prediction performance of different ML models with increasing training data and comparing it with filtering approaches.

2.1. Multiple training sets

We used the QISU data collected from years 2002–2012 which contained 500,150 injury-related emergency department patient records from a sample of hospitals in Queensland, Australia. QISU data is collected by the triage nurse when a patient presents to the emergency department and nurses assign a series of codes to describe the external cause of the injury, objects/substances involved, activity and place of occurrence, as well as a short injury description narrative to describe how the injury occurred. These data are compiled by the central QISU team, and data are reviewed and cleaned by QISU coders to improve the accuracy of data. These data are available in a de-identified format for researchers upon request.

The data was divided into multiple training sets of different sizes and a single prediction set which was a randomly chosen sample of 50,015 cases (10% of the entire QISU dataset). There was no overlap between the prediction set and any of the training sets. We used eleven training sets of different sizes with the smallest one having about 10,000 cases (2% of QISU dataset) and the largest training set having 450,135 cases (90% of QISU dataset). Each of the training sets are referred to as '%nT' further in the paper, where n is the percentage of total QISU cases, for example, the training set of 25,007 cases is referred to as 5%T. The names of training sets and prediction set along with the number of cases are provided in Table 2.

The QISU data was originally manually coded at triage with 30 distinct E-codes–codes which correspond with the ICD external cause of injury categories(National Data Standards for Injury Surveillance Version, 1998). Ten of these codes were 'children' of 'parent' categories and for the purposes of this study the parent categories were used for training and prediction, e.g., E-code "9 Fall-low" (on same level or less than 1m) was combined with E-code "10 Fall-high" (drop of 1 m or more) to form E-code "910 Fall-low and high".

The list of E-codes with the percentage of cases for each E-code in each of the training sets and the prediction set is shown in Table 3. It is to be noted that: a) more than 50% cases in any training and prediction

Table 2
Number of Cases in Training Sets and Prediction Set.

Dataset Name	2%T	5%T	10%T	20%T	30%T	40%T	50%T	60%T	70%T	80%T	90%T	PRED
Number of Cases	10001	25007	50015	100030	150045	200060	250075	300090	350105	400120	450134	50015

Table 3
List of E-Codes and their Percentage of Cases in Each Training and Prediction Dataset.

		2%T	5%T	10%T	20%T	30%T	40%T	50%T	60%T	70%T	80%T	90%T	PRED
	Number of Cases in each dataset (in thousands)	10	25	50	100	150	200	250	300	350	400	450	50
E-CODE	E-CODE DESCRIPTION	Percentage of Cases of the E-code Category in Each Training and Prediction Set											
120	Motor vehicle − driver and passenger (1 Motor vehicle −Driver, 2 Motor vehicle −Passenger)	3.26	3.37	3.06	3.02	3.13	3.16	3.11	3.1	3.05	2.99	2.99	2.94
340	Motorcycle driver and passenger (3 Motorcycle −Driver, 4 Motorcycle −Passenger)	3.14	1.95	2	2.04	2.22	2.31	2.47	2.42	2.39	2.34	2.35	2.52
5	Pedal cyclist or pedal cycle passenger	3.09	3.05	3.15	3.2	3.27	3.22	3.12	2.98	2.85	2.74	2.67	2.27
6	Pedestrian	0.34	0.49	0.53	0.47	0.48	0.45	0.43	0.42	0.4	0.39	0.39	0.33
7	Other or unspecified transport	0.55	0.39	0.44	0.57	0.59	0.58	0.57	0.55	0.55	0.54	0.57	0.47
8	Horse Related	0.65	0.83	0.87	0.92	0.88	0.81	0.8	0.75	0.74	0.75	0.75	0.81
910	Fall- low and high (9 Fall − Low (on same level, or < 1 m drop or no info), 10 Fall − High (drop of 1 m or more)	30.72	30.54	30.67	30.49	30.35	31.04	31.01	31.66	31.77	31.86	31.79	30.95
112	Drowning, submersion (11 Drowning, submersion − in swimming pool, 12 Drowning, submersion − other than in swimming pool)	0.01	0.06	0.09	0.09	0.1	0.09	0.09	0.1	0.09	0.09	0.09	0.07
13	Other threat to breathing	0.35	0.28	0.38	0.37	0.36	0.38	0.39	0.37	0.37	0.4	0.39	0.34
14	Fire, flames, smoke	0.26	0.24	0.26	0.29	0.3	0.3	0.32	0.32	0.32	0.32	0.32	0.31
156	Exposure to hot fluid, gas, or solid (15 Hot drink, food, water, other fluid, steam, gas, 16 Exposure to hot object or solid substance (incl. contact burn)	1.83	2.01	1.98	1.94	1.96	1.97	1.97	1.92	1.89	1.89	1.88	1.62
178	Poisoning −drug or others (17 Poisoning − Drug or medicinal substance, 18 Poisoning − Other or unspecified substance	1.53	2.71	2.75	2.6	2.43	2.26	2.21	2.18	2.18	2.15	2.1	1.68
19	Firearm	0.00	0.04	0.03	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02
20	Cutting, piercing object	8.21	9.46	9.02	8.67	8.44	8.27	8.21	8.09	7.97	7.73	7.43	4.28
212	Dog and animal related excluding horse (21 Dog related (Incl. Bitten, struck by), 22 Animal related (Excl. Horse or dog)	3.43	4.37	4.48	4.32	4.25	4.04	4.01	3.93	3.95	3.97	3.96	3.99
24	Machinery	3.89	3.49	3.49	3.72	4.08	3.97	3.91	3.76	3.77	3.72	3.6	2.49
25	Electricity	0.24	0.23	0.26	0.26	0.25	0.25	0.27	0.25	0.25	0.25	0.26	0.31
267	Hot and cold conditions (26 Hot conditions (natural origin), sunlight, 27 Cold conditions (natural origin)	0.03	0.06	0.09	0.07	0.08	0.08	0.09	0.08	0.09	0.09	0.1	0.07
289	Other and unspecified external cause (28 Other specified external cause, 29 Unspecified external cause)	11.50	12.63	12.59	12.96	13.01	12.82	12.59	12.45	12.62	12.68	12.94	17.13
310	Struck by or collision with person or object (30 Struck by or collision with person, 31 Struck by or collision with object)	26.98	23.79	23.89	23.98	23.8	23.97	24.44	24.66	24.72	25.07	25.41	27.39

set belong to only two categories − 910 (Fall-low and high) and 310 (Struck by or collision with person or object), b) there are nine categories that have less than 1% cases in all the training and prediction sets (E-codes 112, 13, 14, 19, 25, 267, 6, 7 and 8), and c) the distribution of cases among individual E-code categories varies, but does not change significantly between different training sets.

We tested three classical Machine Learning models: Multinomial Naïve Bayes (MNB), Support Vector Machine (SVM), and Logistic Regression (LR) for this study.

Each of the ML models was trained on different training sets and their classification performance on the prediction set was analyzed. A good overview of these models can be found in various data mining textbooks (Aggarwal and Zhai, 2012; Witten and Frank, 2005) and previous works in autocoding of injury data (Marucci-Wellman et al., 2011) (Chen et al., 2015), (Measure, 2014). One of the key differences between these models is that while Naive Bayes is a generative model, LR and SVM are discriminative classifiers. For the given data $\boldsymbol{x}$ and class labels $\boldsymbol{y}$, Naïve Bayes learns the joint probability $\boldsymbol{p(x,y)}$ from the training set and uses Bayes rule to calculate $\boldsymbol{p(y|x)}$. On the other hand, LR and SVM models estimate $\boldsymbol{p(y|x)}$ directly from the training data. (Ng and Jordan, 2002)

WEKA (Hall et al., 2009), a popular open-source machine learning package was used for evaluating the prediction performance of different ML algorithms trained on different training sets. In WEKA, the classifier "NaiveBayesMultinomial" was used for MNB model, and "L2-Regularized Logistic Regression model (dual)" and "L2-Regularized L2-Loss Support Vector Model (dual)" classifiers were used from the LIBLINEAR library(Fan et al., 2008) in WEKA for LR and SVM models respectively. Default hyper-parameter values were used for all the classifiers. WEKA uses a "text to word-vector" function to transform the words in the narrative into a vector form. This function uses the "maximum number of words for each category" as an input parameter which was set to 12,000 for each run with different training datasets to use most of the words in the narrative for making predictions. The number of words used for prediction increased with increase in training size. All three ML algorithms were trained on different training sets and their prediction performance was studied using the performance measures discussed in the next sub-section.

2.2. Performance measures

We used three performance measures for this study: Category-sensitivity for studying the prediction performance on individual categories, and weighted average sensitivity and unweighted average sensitivity for studying the overall prediction performance.

Sensitivity (also called Recall), which signifies the capability of a classifier to identify a particular category correctly, is a common measure used for evaluating classification performance in machine learning tasks. Since our focus is to evaluate if the ML algorithm can identify rare categories with increased or improved training data, we used sensitivity as the primary measure of performance for individual E-code categories. The 'Category Sensitivity' can be defined as:

$$\text{Category Sensivity} = \frac{True\ Positives}{True\ Positive + False\ Negative} = \frac{Number\ of\ Correctly\ Predicted\ Cases\ for\ a\ Category}{Total\ Number\ of\ Cases\ in\ Prediction\ Set\ for\ that\ Category}$$

Weighted Average Sensitivity (WAvg) and Unweighted Average Sensitivity (UAvg) were used as the measures of overall classification performance as defined below:

$$\text{Weighted Avg. Sensivity} = \frac{\Sigma\{(\text{Sensitivity of Ecode Category})^* (\text{Number of Cases in the Category})\}}{\text{Total Number of Cases in Prediction Set}}$$

$$\text{Unweighted Average Sensivity} = \frac{\Sigma(\text{Sensivityof Ecodecategory})}{\text{Total Number of Category}}$$

The weighted average sensitivity has also been referred to as 'overall sensitivity' and reported in most of the previous works in autocoding injury data. As discussed previously, the main issue with weighted average sensitivity is that due to the heavily-skewed nature of injury data, the sensitivities of most common categories greatly influence the overall sensitivity. The unweighted average sensitivity does not consider the size of each E-code category, and therefore, presents a more balanced picture of the classification performance across all categories. A robust unbiased classifier is expected to yield similar values for weighted and unweighted average sensitivities.

2.3. More training data vs filtering approach

ng a small portion of cases for manual review has been tested to be an effective strategy to assign E-codes to injury data with good accuracy for rare categories. In recent papers by (Marucci-Wellman et al., 2017) and (Bertke et al., 2016), it has been discussed that among classical ML models − SVM, MNB, and LR, the prediction strength based filtering approach on the LR model yields largest improvement in sensitivity for all categories. Therefore, we compared the results obtained from increasing the training data size with the approach of filtering cases for expert review based on prediction strength (probability of prediction) outputted by the LR model. For this approach of prediction strength based filtering, the prediction results of the LR model were sorted according to the probability of prediction and cases with the least probability were filtered for expert review. Two levels of filtering based on prediction strength were tested − a) filtering the bottom 10% cases, and b) filtering the bottom 25% cases. These filtering approaches were examined on the LR models trained on the smallest training set (2%T) and the largest training set (90%T).

Assuming the human review to be consistent with the original E-code assigned to the cases filtered, the correction-after-filtering prediction results were used to calculate the sensitivities of each E-code category, the weighted average sensitivity and the unweighted average sensitivity. The prediction results using filtering approach were compared with the prediction results when a very large training set such as 80%T or 90%T was used for training the ML models. We also studied the prediction performance of filtering approach when different prediction strength thresholds (0.5, 0.8 and 0.9) were applied on LR model trained on very-small (2%T) and very-large (90%T) training sets.

3. Results and discussion

In this section, we first discuss how increasing the training data size impacts the overall prediction performance and sensitivities of individual E-code categories for different ML algorithms. Then, we discuss the effect of using a balanced training set on prediction performance of various ML algorithms.

The sensitivities for each E-code categories, overall weighted average sensitivity (WAvg) and overall unweighted average sensitivity (UAvg) for different training sets are reported in: Table 4 for Logistic Regression, Table 5 for Support Vector Machine and Table 6 for Multinomial Naïve Bayes.

As shown in Tables 4, 5 and 6, all the machine learning models showed improvement in the overall prediction performance with increasing size of the training set. The improvements in weighted average sensitivities (WAvg) with increase in training size from 2%T to 90%T for different models were somewhat different, with LR- 8%, SVM- 10%

Table 4
Sensitivity of E-code Categories for Different Training Sets Using LR.

LR	120	340	5	6	7	8	910	112	13	14	156	178	19	20	212	24	25	267	289	310	UAvg	WAvg
2%T	0.737	0.775	0.848	0.44	0.047	0.936	0.903	0	0.012	0.301	0.841	0.531	0	0.856	0.758	0.608	0.419	0	0.452	0.76	0.487	0.742
5%T	0.723	0.673	0.889	0.452	0.034	0.951	0.906	0.162	0.082	0.32	0.863	0.697	0.111	0.794	0.827	0.672	0.632	0	0.52	0.736	0.526	0.752
10%T	0.743	0.69	0.906	0.53	0.112	0.956	0.919	0.405	0.146	0.392	0.861	0.73	0.111	0.833	0.853	0.669	0.748	0.143	0.541	0.741	0.573	0.767
20%T	0.787	0.746	0.913	0.596	0.167	0.958	0.921	0.486	0.181	0.497	0.872	0.742	0.222	0.835	0.865	0.725	0.8	0.286	0.536	0.752	0.614	0.776
30%T	0.832	0.813	0.919	0.663	0.193	0.958	0.911	0.622	0.193	0.575	0.873	0.745	0.111	0.87	0.869	0.762	0.806	0.371	0.534	0.761	0.637	0.782
40%T	0.9	0.825	0.905	0.651	0.18	0.96	0.921	0.649	0.24	0.614	0.887	0.742	0.222	0.867	0.867	0.73	0.845	0.371	0.528	0.767	0.651	0.787
50%T	0.913	0.884	0.901	0.62	0.176	0.96	0.922	0.703	0.281	0.667	0.9	0.768	0.222	0.898	0.862	0.605	0.858	0.457	0.535	0.772	0.662	0.791
60%T	0.919	0.896	0.905	0.627	0.185	0.963	0.925	0.703	0.287	0.712	0.897	0.777	0.222	0.895	0.86	0.68	0.877	0.457	0.522	0.783	0.671	0.795
70%T	0.926	0.901	0.912	0.645	0.18	0.968	0.929	0.676	0.304	0.725	0.89	0.77	0.333	0.863	0.864	0.763	0.865	0.486	0.529	0.796	0.682	0.802
80%T	0.922	0.905	0.918	0.645	0.21	0.965	0.927	0.676	0.339	0.706	0.894	0.78	0.222	0.832	0.87	0.761	0.884	0.486	0.539	0.818	0.681	0.809
90%T	0.929	0.915	0.915	0.675	0.253	0.965	0.921	0.676	0.368	0.686	0.9	0.815	0.333	0.795	0.879	0.757	0.89	0.514	0.601	0.828	0.696	0.821

Table 5
Sensitivity of E-code Categories for Different Training Sets Using SVM.

SVM	120	340	5	6	7	8	910	112	13	14	156	178	19	20	212	24	25	267	289	310	UAvg	WAvg
2%T	0.763	0.798	0.832	0.488	0.155	0.948	0.873	0	0.111	0.444	0.84	0.53	0	0.833	0.8	0.602	0.665	0.057	0.408	0.699	0.516	0.712
5%T	0.714	0.68	0.893	0.482	0.12	0.948	0.9	0.189	0.129	0.373	0.86	0.71	0.111	0.763	0.843	0.73	0.774	0.2	0.461	0.711	0.552	0.735
10%T	0.741	0.704	0.905	0.566	0.167	0.948	0.916	0.541	0.175	0.51	0.861	0.732	0.222	0.789	0.86	0.678	0.761	0.429	0.454	0.7	0.603	0.74
20%T	0.827	0.778	0.903	0.639	0.193	0.96	0.923	0.622	0.211	0.523	0.873	0.757	0.333	0.797	0.873	0.728	0.832	0.429	0.415	0.723	0.635	0.75
30%T	0.875	0.817	0.919	0.62	0.21	0.963	0.908	0.622	0.251	0.647	0.862	0.758	0.333	0.794	0.877	0.768	0.852	0.514	0.456	0.762	0.658	0.767
40%T	0.899	0.831	0.849	0.614	0.172	0.96	0.916	0.676	0.316	0.667	0.893	0.766	0.444	0.823	0.873	0.73	0.89	0.486	0.494	0.755	0.669	0.775
50%T	0.907	0.892	0.898	0.566	0.189	0.965	0.916	0.73	0.327	0.745	0.882	0.767	0.444	0.884	0.862	0.606	0.903	0.457	0.512	0.761	0.677	0.782
60%T	0.911	0.904	0.91	0.59	0.223	0.968	0.92	0.73	0.322	0.752	0.888	0.768	0.556	0.882	0.868	0.685	0.923	0.571	0.504	0.776	0.698	0.789
70%T	0.92	0.912	0.909	0.62	0.215	0.973	0.924	0.703	0.357	0.752	0.887	0.777	0.556	0.856	0.867	0.772	0.903	0.543	0.51	0.788	0.702	0.796
80%T	0.92	0.905	0.919	0.633	0.232	0.97	0.924	0.676	0.345	0.739	0.883	0.798	0.333	0.833	0.874	0.773	0.935	0.514	0.518	0.81	0.692	0.803
90%T	0.926	0.912	0.927	0.681	0.27	0.975	0.921	0.676	0.357	0.706	0.893	0.817	0.333	0.798	0.886	0.772	0.935	0.514	0.578	0.822	0.700	0.816

and MNB- 6%. For the improvement in unweighted average sensitivity (UAvg) with increase in training size from 2%T to 90%T, MNB showed the largest improvement (~ 31%) followed by LR (~21%) and SVM (~18%). Fig. 1 shows the unweighted average sensitivity and weighted average sensitivity of each ML method for different training sets.

3.1. Weighted average sensitivity

As shown in Fig. 1, the weighted average sensitivity (WAvg) for each ML model is considerably higher than its unweighted average sensitivity (UAvg). This highlights the biased performance of the classifiers, i.e. all the ML models yield substantially higher sensitivity for bigger categories as compared to rare categories. It should be noted that the difference between the UAvg and WAvg for each ML model is relatively larger for the smaller training set size and reduces as the training set size is increased. This indicates that with increasing number of examples of rare categories in the training dataset, the model learns to better classify rare categories.

Among classifiers, the weighted average sensitivities for LR and SVM for different training sizes were very similar. The WAvg of LR model was somewhat better than SVM for smaller training sets (till 30%T). This indicates that with increasing training data, both SVM and LR yield a similar classification performance. The WAvg of MNB was relatively lower than SVM and LR. It could be observed in Fig. 1 that a) the difference in WAvg of MNB and LR remained about the same with increasing training set size, and b) the difference in WAvg of MNB and SVM was smaller for smaller training size (till 30%T) and then it remained about the same for larger training sets.

3.2. Unweighted average sensitivity

As shown in Fig. 1, while MNB showed steady improvement in UAvg consistently with increasing training data size, for LR and SVM models, the improvement in UAvg plateaued with increasing training data size–particularly after 70%T (which seemed to be the asymptotic performance for LR and SVM models). The unweighted average sensitivity of SVM was consistently better than LR, followed by MNB.

It is to be noted that while the unweighted average sensitivity improved a lot with increased training set size for all the models, the improvement in weighted average sensitivity was to a lesser extent. This indicates that with increasing the training set size, the sensitivities of rare categories improved.

We can also observe from Fig. 1 that the improvement in unweighted average sensitivity (UAvg) with increased training set size is steep and steady for MNB. However, for LR and SVM, the improvement in unweighted average sensitivity slows down with increase in training size, particularly after 60%T, where the UAvg of SVM and LR do not change with increase in training data. In Fig. 1, by inspecting the UAvg-MNB and WAvg-MNB curves and extrapolating them, it seems likely that more training data the UAvg-MNB will further increase and may converge with WAvg-MNB. This is in contrast with the other two ML models, SVM and LR, for which the UAvg and WAvg curves do not seem to converge with extrapolation and seem to reach an asymptotic difference between WAvg and UAvg.

In one of the widely-discussed papers on generative vs. discriminative classifiers by (Ng and Jordan, 2002), the authors mentioned that the generative MNB model would yield better classification performance than discriminative LR model with lesser training data and as the size of training data increases, the LR model would yield better classification performance. However, our results do not show such a pattern in the results. Our results show that the discriminative models LR and SVM yield better classification performance even with smaller training data. This might be because of the large number of categories and the heavily-skewed nature of injury data, due to which the MNB model did not have enough training cases for the smaller categories to be able to predict them correctly.

Table 6
Sensitivity of E-code Categories for Different Training Sets Using MNB.

MNB	120	340	5	6	7	8	910	112	13	14	156	178	19	20	212	24	25	267	289	310	UAvg	WAvg
2%T	0.882	0.658	0.443	0.006	0.004	0.012	0.918	0	0	0	0.757	0.506	0	0.851	0.596	0.363	0.045	0	0.345	0.758	0.340	0.694
5%T	0.897	0.556	0.713	0.048	0.004	0.101	0.905	0	0.006	0.02	0.853	0.763	0	0.861	0.729	0.484	0.148	0	0.402	0.694	0.390	0.702
10%T	0.905	0.67	0.772	0.163	0.009	0.237	0.902	0.027	0.047	0.026	0.895	0.81	0	0.872	0.766	0.534	0.265	0	0.421	0.691	0.429	0.715
20%T	0.924	0.744	0.818	0.229	0.034	0.343	0.898	0.108	0.076	0.059	0.914	0.841	0	0.868	0.792	0.588	0.432	0	0.432	0.698	0.467	0.725
30%T	0.937	0.786	0.82	0.259	0.052	0.385	0.893	0.189	0.088	0.085	0.925	0.842	0	0.88	0.792	0.592	0.465	0.057	0.412	0.708	0.484	0.726
40%T	0.942	0.839	0.807	0.446	0.112	0.595	0.885	0.405	0.263	0.216	0.938	0.85	0	0.888	0.801	0.592	0.71	0.114	0.426	0.714	0.550	0.735
50%T	0.944	0.866	0.817	0.38	0.116	0.595	0.883	0.405	0.287	0.333	0.926	0.862	0	0.903	0.792	0.508	0.742	0.057	0.438	0.715	0.551	0.736
60%T	0.946	0.882	0.828	0.458	0.155	0.637	0.88	0.568	0.368	0.458	0.93	0.859	0	0.904	0.793	0.564	0.794	0.114	0.426	0.713	0.585	0.737
70%T	0.943	0.885	0.84	0.536	0.197	0.706	0.883	0.595	0.398	0.536	0.93	0.857	0	0.896	0.804	0.635	0.8	0.171	0.419	0.714	0.607	0.74
80%T	0.945	0.891	0.843	0.566	0.219	0.763	0.884	0.649	0.468	0.575	0.931	0.852	0	0.888	0.809	0.662	0.839	0.143	0.415	0.725	0.622	0.744
90%T	0.944	0.897	0.846	0.62	0.296	0.802	0.88	0.703	0.538	0.601	0.935	0.858	0	0.875	0.812	0.674	0.877	0.229	0.443	0.732	0.646	0.752

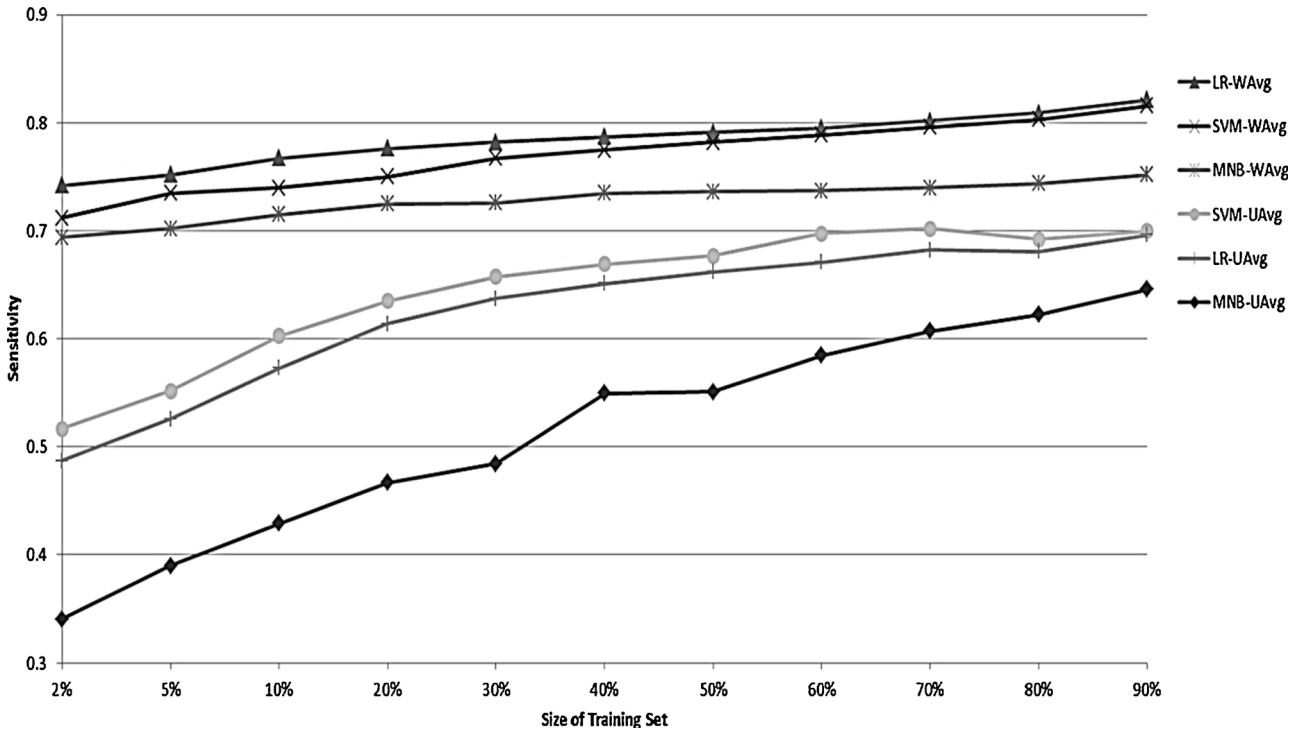

Fig. 1. Unweighted Average and Weighted Average Sensitivities of ML Models for Different Training Sets.

3.3. Sensitivity of individual categories

The sensitivity of most E-code categories improved steadily with increased size of training set for each of the ML methods. There were few categories for which the sensitivity decreased by a small value with increasing training size − this may be due to the ML model itself or the quality of data added to the training set with increased size. In Figs. 2–4, the sensitivities of four representative rare categories (7, 13, 112, and 267 − these together constitute about 1.1% of total cases) and three largest categories (289, 310 and 910–these together constitute about 67% of total cases) are plotted against the size of training set for the classifiers LR, SVM and NB respectively.

As shown in Figs. 2, 3, and 4, the sensitivity of the largest category 910 (Falls, with 30% of total cases) does not change much with increasing size of training set for any of the ML models. This is probably because there are already enough training examples of this category even in the smallest training set for each of the models to be able to distinguish it from other categories. It is also to be noted that with 5%T, all the ML models yield almost the same sensitivity (0.9) for category 910.

The sensitivity of category 310 (Struck by/against object/person, with 24% of total cases) increases around 10% (0.71–0.81) with increase in training data from 5%T to 90%T for LR and SVM, but for MNB, the increase in sensitivity is only about 4% (0.69–0.73) with increase in training data from 5%T to 90%T. Category 310 is closely related to other categories with closely-related terms to terms present in 24 (Machinery) (e.g., hit hand on, was hit with) and 20 (Cut, Pierce) (e.g., cut, laceration). Hence, with increased training data, the ML models seem to better predict category 310.

Category 289 (Other and Unspecified (which captures all causes of injuries not described by the other specific codes), with 13% of total cases), although a relatively large category, is difficult to predict because it is very open ended and has a lot of overlap with various other categories. There is improvement in the sensitivity of category 289 with increased training size but it varies with the ML model. The increment between 5%T and 90%T for different models are: 8% for LR, 11% for SVM, and 4% for MNB.

The sensitivity improves significantly with increased training data for all the smaller categories shown in Figs. 2, 3, and 4. It is to be noted that all the ML models showed the most improvement in individual category sensitivity with increase in training data size from 5%T to 90%T for the category 112 (Drowning and Submersion, with about

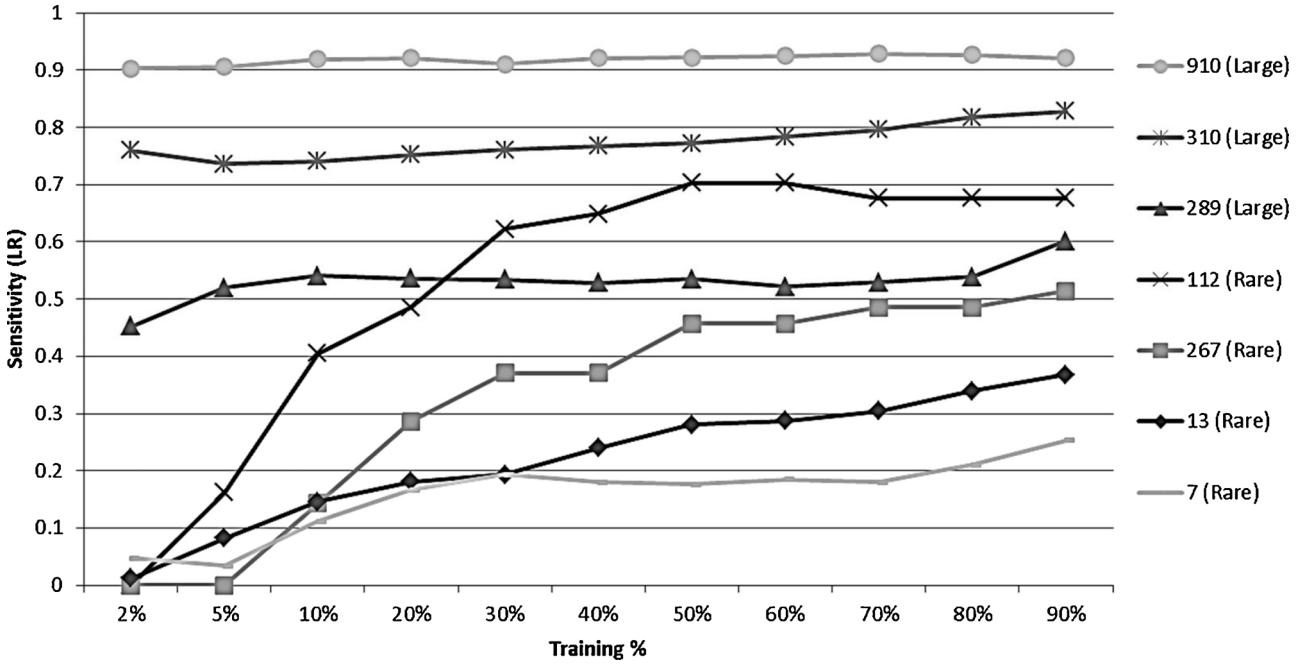

Fig. 2. Plot of sensitivity of selected E-codes vs size of training set for LR model.

0.1% cases). SVM showed an improvement of 49%, LR 51%, and MNB showed the largest improvement (~70%). Category 112 is a rare category but has a specific definition and is not closely related to other categories except 289. Hence, it seems intuitive that with increased training data size, all the ML models are able to classify the narratives for this category.

Category 7 (Other or unspecified transport (which captures all transport other than motor vehicle/motorcycles/pedal cycles/horse-related and pedestrian), with about 0.6% cases) is a rare category which does not have a specific definition and has narrative words that are closely related to larger transportation-related categories – 120 (Motor-vehicle – driver and passenger), 340 (Motorcycle-driver and passenger), and 5 (Pedal-cycle). We did not observe a large improvement in the sensitivity of category 7 (around 20% for all ML models). It might be because even with increased training data, the distribution was still skewed, and given the non-specific definition of the category, it would be difficult for the ML models to distinguish it from related larger categories.

Category 13 (Other threat to breathing, with about 0.4% cases) is another rare category with not a very specific definition and has overlapping definition with category 289(Other Unspecified); our observation from the data is that category 13 is often miscoded and assigned to the non-specific code 289 (other and unspecified external cause) by human coders. MNB yields the highest improvement (about 53%) in sensitivity of category 13 with increment in training data from 5%T to 90%T, while LR and SVM showed modest improvements

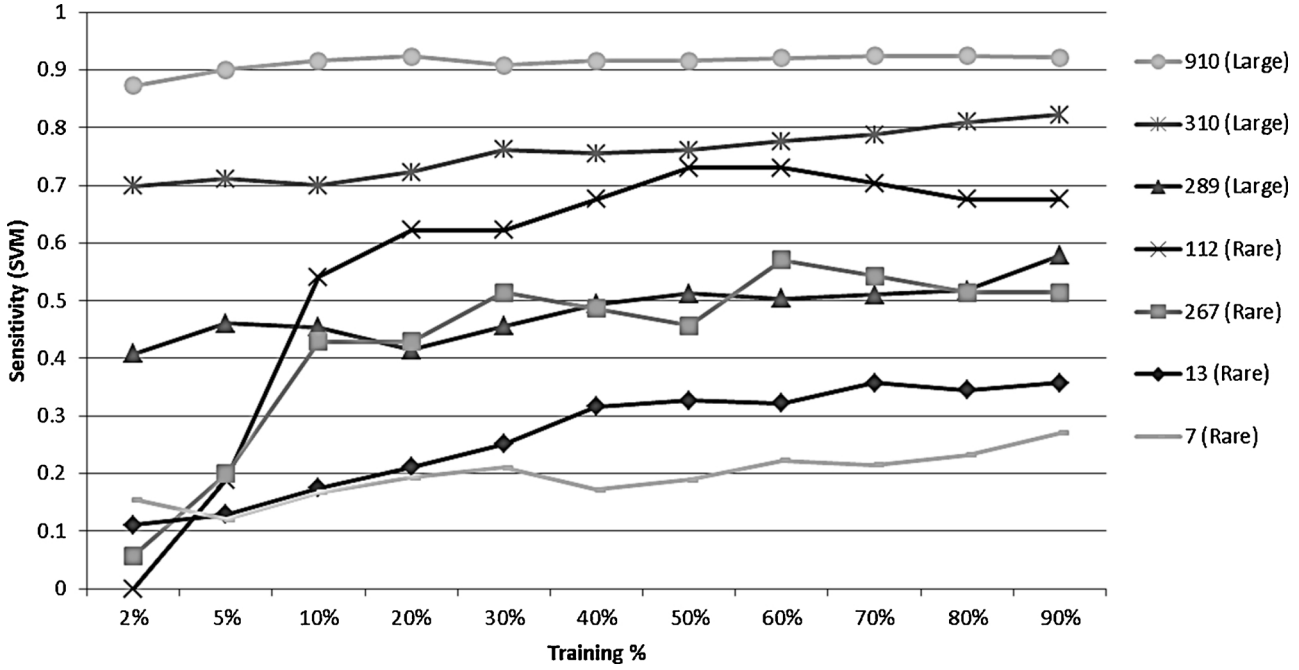

Fig. 3. Plot of sensitivity of selected E-codes vs size of training set for SVM Model.

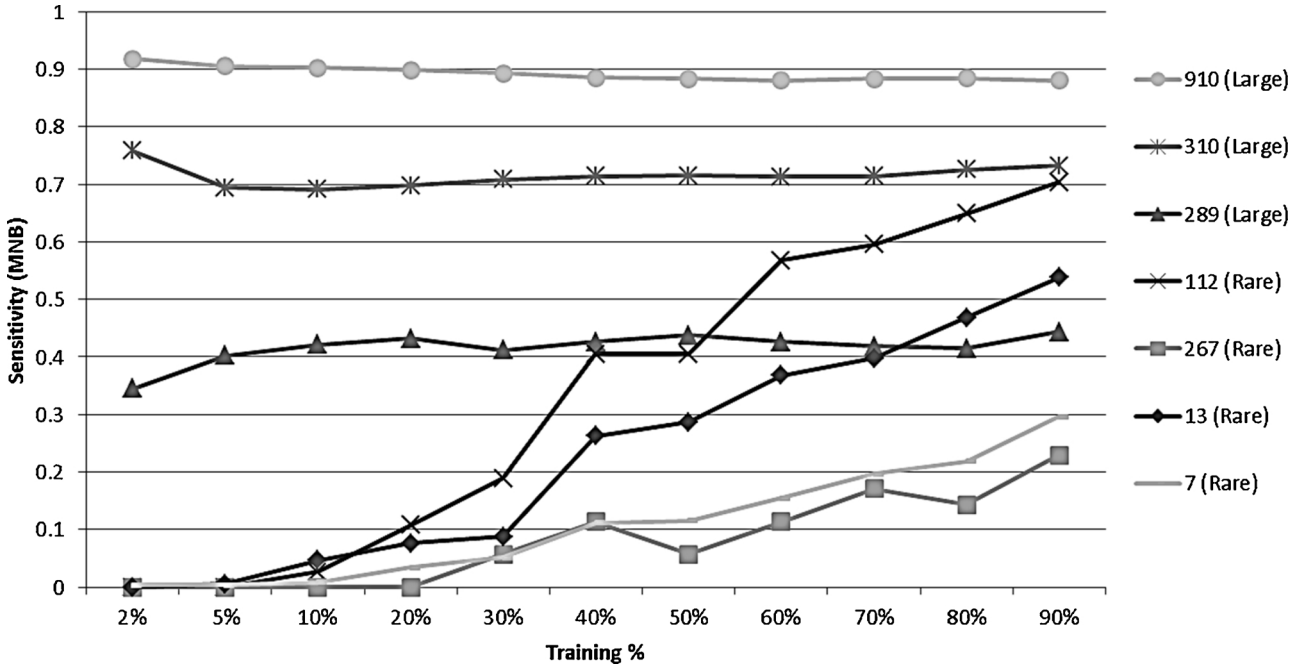

Fig. 4. Plot of sensitivity of selected E-codes vs size of training set for MNB Model.

(around 22%).

Category 267 (Hot and Cold Conditions, with about 0.08% cases) has narrative words that are closely related to category 156 (Exposure to hot substances) but has some specific indicators as per the definition, such as, sun-stroke, sunburn, etc. The improvement in sensitivity of category 267 with increased training data (from 5%T to 90%T) varied for each ML method, for SVM it is about 30%, for LR about 51%, and for MNB the improvement is about 20%.

3.4. Filtering approach

The prediction results of LR models trained on 2%T and 90%T training sets were sorted in descending order of the probability of prediction, and the bottom-10% and the bottom-25% cases were filtered for expert review. For the cases filtered for expert review, it was assumed that the expert-assigned E-code would be consistent with the original E-code assigned to each case, and the sensitivities were calculated for each E-code category. The results for different filtering approaches for the models trained on 90%T and 2%T training sets are presented in Table 7.

As shown in Table 7, for the LR model trained on 2%T, the bottom10% filtering and correction-after-filtering resulted in WAvg of 81% and UAvg of 69%, which is close to the prediction performance of LR model trained on a large dataset 90%T, that has a WAvg of 82.1% and UAvg of 69.6%. The results indicate that for the LR model, even with small amount of training data and relatively small level of filtering (only 10% data), prediction performance similar to the LR model trained on a very large training set can be obtained. For one of the extremely rare categories, "7-Other Transport", the sensitivity with 2%T and 10% filtering is 0.23- which is similar to the sensitivity for that for the LR model trained on 90%T without filtering, 0.253.

The sensitivities of each E-code for different filtering approaches are also plotted in Fig. 5.

As shown in Fig. 5, the 25% filtering on LR model trained on 90%T results in largest sensitivities for all the E-codes. It is to be noted that 10% filtering on LR model trained on 90%T and 25% filtering on LR model trained on 2%T yield similar level of sensitivities for most of the E-codes and also in terms of overall performance. The WAvg of both approaches is 0.88 and the UAvg of 10% filtering-90%T training is slightly higher, 0.82, as compared to 25%filtering-2%T training, 0.80. The similar level of performances for these two approaches again shows that the lack of large amount of training data can be compensated by a higher level of filtering. The highest level of prediction performance is achieved when both large training data and higher level of filtering are applied.

It is to be noted that the UAvg and WAvg calculations for the filtering approach are based on the assumption that the cases filtered for human review would be coded consistently with the original code. Both the approaches – obtaining good quality training data in large amount and getting a portion of injury data to be coded by experts – are expensive and difficult to achieve. Depending on the availability of either resource, appropriate approaches can be used.

3.5. Prediction strength threshold

We can observe from Table 7 that the sensitivity for the extremely rare categories- ("7-Other Transport", "13- Other Threat to Breathing", "14- Fire, Flames, Smoke", and "19–Firearms") is relatively lower as compared to other categories, even with a very large training set (90%T) and 25% filtering. Other studies have also reported lower sensitivity of prediction by ML models ((Marucci-Wellman et al., 2017) (Nanda et al., 2016) (Marucci-Wellman et al., 2015) for some of the rare categories such as "14-Fire, Flames, Smoke" and "7-Other Transport". The study by (Chen et al., 2015) that used the same QISU dataset as ours and applied elaborate preprocessing in addition to machine learning, reported very high overall sensitivity but relatively lower sensitivities for the extremely rare categories – "7-Other Transport", "13- Other Threat to Breathing", "14- Fire, Flames, Smoke", and "19–Firearms". Some of the possible reasons behind the poor performance of these categories include: very few cases in training dataset, closely related larger categories (e.g., 14- Fire and 156-Hot Objects and Liquids), and inconsistent manual coding of training data (similar narratives assigned different categories by different coders).

Analyzing the E-codes predicted for rare categories by different ML models trained on 90%T dataset, we observed that the category "7-Other Transport" gets frequently misclassified as the larger categories "910-Falls" and "310-Struck" by the model. We further examined the narrative of cases from category "7-Other Transport" and identified

Table 7
Sensitivity of E-codes after Manual Correction of Filtered Cases for Different Approaches.

Size of Training Set		2%T		90%T	
% of Cases Filtered based on Prediction Probability of LR Model		10%	25%	10%	25%
E-code	E-Code Description	Sensitivity After Manual Correction of Filtered Cases			
120	Motor vehicle − driver and passenger (1 Motor vehicle −Driver, 2 Motor vehicle −Passenger)	0.84	0.91	0.96	0.97
340	Motorcycle driver and passenger (3 Motorcycle −Driver, 4 Motorcycle −Passenger)	0.82	0.89	0.93	0.96
5	Pedal cyclist or pedal cycle passenger	0.88	0.92	0.94	0.97
6	Pedestrian	0.62	0.78	0.77	0.86
7	Other or unspecified transport	0.23	0.48	0.47	0.70
8	Horse Related	0.95	0.98	0.98	0.99
910	Fall- low and high (9 Fall − Low (on same level, or < 1 m drop or no info), 10 Fall − High (drop of 1 m or more)	0.92	0.96	0.94	0.97
112	Drowning, submersion (11 Drowning, submersion − in swimming pool, 12 Drowning, submersion − other than in swimming pool)	0.46	0.76	0.81	0.92
13	Other threat to breathing	0.41	0.62	0.54	0.77
14	Fire, flames, smoke	0.55	0.71	0.78	0.89
156	Exposure to hot fluid, gas, or solid (15 Hot drink, food, water, other fluid, steam, gas, 16 Exposure to hot object or solid substance (incl. contact burn)	0.91	0.95	0.92	0.95
178	Poisoning −drug or others (17 Poisoning − Drug or medicinal substance, 18 Poisoning − Other or unspecified substance	0.76	0.87	0.89	0.95
19	Firearm	0.33	0.33	0.44	0.78
20	Cutting, piercing object	0.91	0.95	0.89	0.94
212	Dog and animal related excluding horse (21 Dog related (Incl. Bitten, struck by), 22 Animal related (Excl. Horse or dog)	0.83	0.88	0.91	0.95
24	Machinery	0.72	0.84	0.83	0.91
25	Electricity	0.68	0.84	0.94	0.97
267	Hot and cold conditions (26 Hot conditions (natural origin), sunlight, 27 Cold conditions (natural origin)	0.60	0.80	0.74	0.86
289	Other specified and unspecified external cause (28 Other specified external cause, 29 Unspecified external cause)	0.59	0.75	0.74	0.87
310	Struck by or collision with person or object (30 Struck by or collision with person, 31 Struck by or collision with object)	0.80	0.88	0.88	0.94
Weighted Average Sensitivity (Wavg)		0.81	0.88	0.88	0.94
Unweighted Average Sensitivity (Uavg)		0.69	0.80	0.82	0.91

some of the prominent word combinations in the narrative such as “Boat + Fell”. By filtering cases from the QISU dataset that had both the words − “Boat” and “Fell” and examining the E-codes assigned to them by human coders, we observed that such cases were inconsistently coded in categories 7- Other Transport and 910-Falls. Some of the other examples of inconsistent human coding identified in the QISU dataset were: a) cases involving choking were inconsistently coded into categories “13–Other Threat to Breathing” and “289- Other Specified and Unspecified”, b) cases involving contact with hot coal were inconsistently coded by human coders in categories “14-Fire” and “156-Hot object and Liquids”.

To examine the extent of filtering required to accurately code these extremely rare categories with small training data (2%T) and large training data (90%T), we applied prediction strength thresholds of 0.5, 0.8 and 0.9 on the prediction output by the LR model trained on 2%T and 90%T datasets and analyzed the prediction performance and extent of filtering required. As mentioned in the Background section, the approach of applying prediction strength thresholds on LR model output has been tested in some of the previous papers such as (Marucci-Wellman et al., 2017). If the prediction probability for a case in the prediction set as outputted by the LR model was less than the threshold, it was filtered for expert review and it was assumed that the expert-assigned code would be consistent with the originally assigned E-code. The sensitivities for the rarest categories (7, 13, 14, and 19), WAvg,

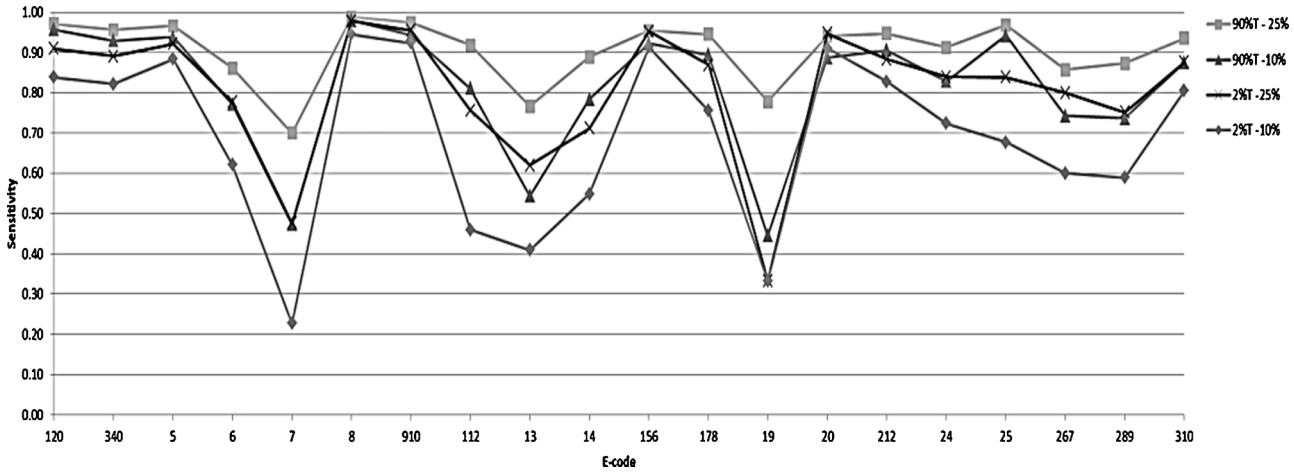

Fig. 5. Sensitivities of E-codes for Different Filtering Approaches.

Table 8
Sensitivity of Rare Categories, WAvg, UAvg, and Percentage of Cases Filtered for Different Levels of Prediction Strength Thresholds on LR Model trained on 2%T and 90%T.

	Sensitivity after Manual Review of Filtered Cases					
Training Set	2%T			90%T		
Prediction Strength Threshold on LR Model	0.5	0.8	0.9	0.5	0.8	0.9
ECODE						
7- Other or unspecified transport	0.33	0.76	0.91	0.48	0.86	0.94
13-Other Threat to Breathing	0.49	0.95	0.98	0.56	0.91	0.98
14–Fire, Flames, Smoke	0.63	0.95	0.99	0.78	0.94	0.97
19 – Firearm	0.33	0.67	1	0.44	1	1
WAvg	0.84	0.97	0.99	0.88	0.98	0.99
UAvg	0.74	0.94	0.99	0.82	0.97	0.99
%Cases Filtered in Prediction Set (Total 50,000 cases)	16%	52%	71%	11%	42%	57%

UAvg, and percentage of cases filtered after applying the threshold on the output of LR model trained on 2%T and 90%T are presented in Table 8.

As shown in Table 8, applying a prediction-strength threshold of 0.5 on 90%T-LR model resulted in good overall performance (WAvg and UAvg) and filters relatively smaller portion (11%) of cases, but the sensitivities of extremely rare categories were still low. This means that a threshold of 0.5 is not able to filter cases from extremely rare categories even when the LR model is trained on a very large dataset. When a threshold of 0.8 was applied on 90%T-LR model, it resulted in extremely good overall prediction performance and relatively high sensitivities for the extremely rare categories and filtered about 42% cases. When the prediction-strength threshold of 0.8 was applied on 2%T-LR model, it filtered about 52% cases, and resulted in very good overall prediction performance (WAvg = 0.97, UAvg = 0.94) and improved sensitivities of the extremely rare categories. It is to be noted that even after applying a prediction-strength threshold of 0.8 on the 2%T-LR model, the sensitivities of two extremely rare categories – 7 and 19, were relatively lower. When a higher threshold of 0.9 is applied on the 2%T-LR model, it resulted in extremely high sensitivities for all the rare categories, but it comes at a cost of filtering a considerably large (71%) number of cases. Comparing the prediction performances of: a) 0.8 threshold on 90%T-LR model, and b) 0.9 threshold on 2%T-LR model, we can infer that having a larger training set for the LR model can reduce the number of cases to be manually reviewed to achieve same prediction performance for rare categories (as well as overall). It is to be noted that even with a very large training set, a sizeable portion of cases need to be filtered based on prediction strength to code extremely rare categories with high accuracy.

3.6. Estimating category counts

In many situations, estimating the total number of cases for E-codes (referred to as 'category counts' in this paper) is sufficient for injury surveillance purposes, not requiring prediction of E-code for each individual narrative (Bertke et al., 2016). One way to accomplish this is the approach discussed in the previous section – to predict the E-code for each narrative in the dataset using trained machine learning models and then calculate the count for each E-code category. An alternative approach discussed by (Bertke et al., 2016) is summing the category-specific prediction probabilities outputted by the LR model across all narratives in the prediction set to estimate the overall category counts of E-codes.

We tested the approach of summing prediction probabilities for estimating category counts in the prediction set for LR and MNB models trained on three different training sizes: 5%T (small), 50%T (medium), and 90%T (large). We compared these estimations of the category counts with the number of predictions made by LR and MNB models for each E-code category (where the model selects the E-code with highest prediction probability). The results showed that approach of summing prediction probabilities resulted in relatively better estimate of category count for some of the rare categories – 7, 19, 112 and 267 for LR model trained on 5%T, 50%T, and 90%T. On the other hand, we did not observe any considerable difference in the category counts estimated using summing prediction probabilities approach and counting the direct predictions made by the MNB model for any of the three training sets.

For the LR model trained on 90%T dataset, we then compared the category count estimated using following approaches: a) summing prediction probability (referred to as LR 90%T Prob-Sum), b) direct predictions made the LR model (referred to as LR 90%T), and c) filtering bottom 25% cases based on prediction strength for expert review

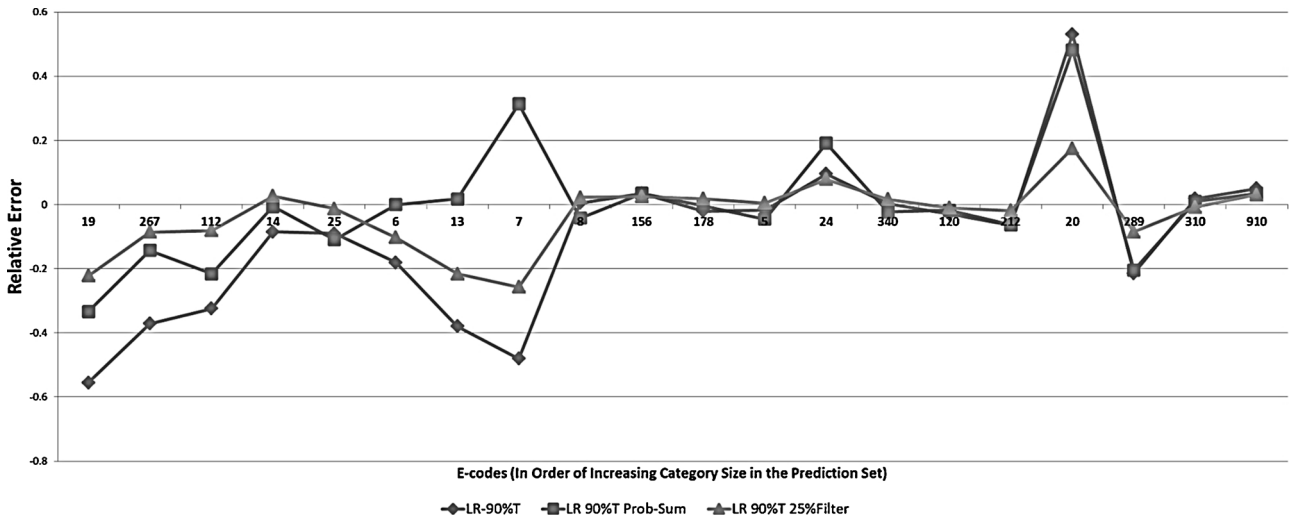

Fig. 6. Comparison of category counts estimated for E-codes in the prediction set using different approaches.

(referred to as LR 90%T 25%Filter). The results are presented in Fig. 6 in form of a plot where the X-axis represents different E-codes sorted by increasing number of cases and the Y-axis represents the 'relative error' in category count which is calculated for each E-code as:

$$Relative\ Error = \frac{Estimated\ Count - Actual\ Count}{Actual\ Count}$$

In Fig. 6, a positive relative error implies overestimation of the category count for that particular E-code by the approach used. Similarly, a negative relative error implies underestimation of category count for that E-code by the approach used.

As shown in Fig. 6, the category count estimate using the filtering approach (LR 90%T 25%Filter) corresponds to the lowest Relative Error (among the three approaches) for all except two E-codes: 6 and 13 −where the probability summing approach (LR 90%T Prob-Sum) resulted in a better estimate of category counts as compared to other approaches. It is to be noted that the category counts for most of the rare categories such as 19, 267, 112, 14 and 25 were underestimated using both probability summing (LR 90%T Prob-Sum) and direct prediction (LR 90%T) approaches.

These results demonstrate that: a) the filtering approach leads to more accurate category counts as compared to the probability summing and counting direct prediction approaches, and b) the probability summing approach leads to better category count estimates for most of the rare categories as compared to the direct prediction approach.

4. Conclusions and future work

In this study, we examined the variation in prediction performance of various ML methods (LR, SVM and MNB) with increases in the size of the training set and compared this to results obtained from filtering approaches. We increased the size of the training set from 10,000 cases (about 1/5 of the size of prediction set, 50,150) to 450,000 cases (about 10 times the size of the prediction set) and studied the effect on each ML method's overall sensitivity, unweighted average sensitivity and individual category sensitivities. Given the similar level of overall sensitivity (WAvg) of LR and SVM models, and relatively higher level of UAvg for SVM as compared to LR, SVM seem to be more robust for working with skewed data such as the injury data.

With increasing size of the training dataset, the sensitivities of smaller categories tend to improve more than larger categories for each ML method. We also noticed that the sensitivity of smaller E-code categories with specific and non-overlapping definitions improved to a larger extent with increasing training data size, as compared to other smaller E-code categories which had relatively non-specific definitions. Even the improved sensitivity values of smaller categories were significantly less than the sensitivities of larger categories. This may be due to the heavily skewed distribution of the data, overlap in definitions of certain categories, and inconsistent manual coding of the narratives.

The larger categories such as 910 (Falls), 289 (Other Unspecified), and 310 (Struck) constitute more than 70% of the data, and the ML methods yielded good prediction sensitivity for these categories even with the smallest training dataset. Therefore, we did not observe a large increment in the overall sensitivity of any of the ML methods with increase in size of training set. It can be inferred that although the prediction ability of ML models improved with increased size of the training set, the gap between the sensitivity of larger categories and smaller categories still remained considerably wide. This means that more training data alone cannot solve the problem of low sensitivity of rare categories yielded by ML methods. The quality of training data and level of skewness in distribution of categories play an important role in the performance of ML models.

Filtering approaches offer promising results and can be utilized in absence of good quality and quantity of training data. Our results showed that filtering approaches even with smaller training size can bring similar level of improvement in prediction performance as large training data. The best classification results were obtained when higher level of filtering (25%) was applied with LR model trained on a very large dataset but given the cost associated with obtaining training data and manual filtering, both may not be available in practical settings.

Various advanced data-based approaches such as "Under-sampling" (under-sample the high frequency categories in the training set) and "Over-sampling" (adding duplicate examples of rare category cases in the training set) have been proposed in machine learning research to tackle the class imbalance problem (Van Hulse et al., 2007). We have been working on using these techniques for injury classification. Future research could be directed towards testing if these sophisticated methods could be useful for classifying injury data.

Apart from the challenges posed by the injury data, the inherent inability of ML models to use semantic and syntactic information beyond the word occurrence and co-occurrence is also one of the main reasons for limited improvement in classification performance by different ML models. Identifying the semantic relationship between the words (Subject, Verb, and Object) can help in better interpretation of injury narrative and therefore, better classification performance. In one of our working papers, we have tested the use of statistical grammatical parsers such as the Stanford Parser (De Marneffe et al., 2006) (which are trained on large textual datasets), but due to the short and noisy nature of injury narrative, the results of parsing are not very accurate. On the other hand, domain-knowledge-based word-sequence and word co-occurrence rules can help the ML models by providing the additional information about deeper concepts present in the narrative which the ML models are not able to extract. We have observed that some of the rare E-code categories may be better predicted with incorporation of linguistic rule-based approaches in the modelling process. Future research could be directed towards how to efficiently extract deeper concepts from the narrative and develop a framework to utilize the knowledge of these concepts in addition to the prediction by ML models.

Vallmuur et al. (2016) discussed the potential for researchers to work collaboratively to build a shared knowledge database to provide larger training datasets to facilitate model development, and Purdue University is currently developing an open source framework for this purpose. As the data collection increases, we will have access to more training data for the rare categories which can be used to build balanced training sets to train the ML models.

However, more training data is not the only solution needed as it results in a marginal improvement in the prediction of rare categories. Filtering approaches can be used to improve the prediction performance of ML models in case large training datasets are not available. Filtering approaches targeted towards rare categories should be explored in order to achieve larger improvement in UAvg with lesser manual review. In addition to the filtering approaches tested, other sophisticated filtering approaches should be explored, such as rule-based filtering, or filtering based on the agreement in prediction results of ML models trained on balanced and unbalanced datasets.

Machine learning is a valuable method to facilitate coding of injury narratives and studies of approaches to improve the accuracy of these methods are important for advancing the use of machine learning in the injury field.

References

Aggarwal, C.C., Zhai, C., 2012. In: Aggarwal, C.C., Zhai, C. (Eds.), A Survey of Text Classification Algorithms. Springer, US, pp. 163–222. (Retrieved from). http://link.springer.com/chapter/10.1007/978-1-4614-3223-4_6.

Bertke, S.J., Meyers, A.R., Wurzelbacher, S.J., Bell, J., Lampl, M.L., Robins, D., 2012. Development and evaluation of a Naïve Bayesian model for coding causation of workers' compensation claims. J. Safety Res. 43 (5-6), 327–332. http://dx.doi.org/10.1016/j.jsr.2012.10.012.

Bertke, S.J., Meyers, A.R., Wurzelbacher, S.J., Measure, A., Lampl, M.P., Robins, D., 2016. Comparison of methods for auto-coding causation of injury narratives. Accid. Anal. Prev. 88, 117–123. http://dx.doi.org/10.1016/j.aap.2015.12.006.

Chawla, N.V., Japkowicz, N., Kotcz, A., 2004. Editorial: special issue on learning from imbalanced data sets. ACM Sigkdd Explor. Newsl. 6 (1), 1–6. (Retrieved from). http://dl.acm.org/citation.cfm?id=1007733.

Chen, L., Vallmuur, K., Nayak, R., 2015. Injury narrative text classification using factorization model. BMC Med. Inform. Decis. Mak. 15 (1), S5. http://dx.doi.org/10.1186/1472-6947-15-S1-S5.

Corns, H.L., Marucci, H.R., Lehto, M.R., 2007. In: Smith, M.J., Salvendy, G. (Eds.), Development of an Approach for Optimizing the Accuracy of Classifying Claims Narratives Using a Machine Learning Tool (TEXTMINER[4]). Springer, Berlin Heidelberg, pp. 411–416. (Retrieved from). http://link.springer.com/chapter/10.1007/978-3-540-73345-4_47.

De Marneffe, M.-C., MacCartney, B., Manning, C.D., 2006. Generating Typed Dependency Parses from Phrase Structure Parses 6. pp. 449–454. (Retrieved from). http://t3-1.yum2.net/index/nlp.stanford.edu/manning/papers/LREC_2.pdf.

Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C., 2008. LIBLINEAR: A Library for Large Linear Classification. (Retrieved from). http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.140.9959.

Guo, X., Yin, Y., Dong, C., Yang, G., Zhou, G., 2008. On the class imbalance problem. In: IEEE. In 2008 Fourth International Conference on Natural Computation 4. pp. 192–201. http://dx.doi.org/10.1109/ICNC.2008.871.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Explor. Newsl. 11 (1), 10–18. http://dx.doi.org/10.1145/1656274.1656278.

Marucci, H.R., Lehto, M.R., Corns, H.L., 2007. Computer Classification of Injury Narratives Using a Fuzzy Bayes Approach: Improving the Model. In: Smith, M.J., Salvendy, G. (Eds.), Springer, Berlin Heidelberg, pp. 500–506. (Retrieved from). http://link.springer.com/chapter/10.1007/978-3-540-73345-4_57.

Marucci-Wellman, H., Lehto, M., Corns, H., 2011. A combined fuzzy and Naive bayesian strategy can be used to assign event codes to injury narratives. Injury prevention. J. Int. Soc. Child Adolesc. Inj. Prev. 17 (6), 407–414. http://dx.doi.org/10.1136/ip.2010.030593.

Marucci-Wellman, H.R., Lehto, M.R., Corns, H.L., 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. http://dx.doi.org/10.1016/j.aap.2015.06.014.

Marucci-Wellman, H.R., Corns, H.L., Lehto, M.R., 2017. Classifying injury narratives of large administrative databases for surveillance—a practical approach combining machine learning ensembles and human review. Accid. Anal. Prev. 98, 359–371. http://dx.doi.org/10.1016/j.aap.2016.10.014.

Measure, A.C., 2014. Automated Coding of Worker Injury Narratives. (Boston, MA, USA. Retrieved from). http://www.bls.gov/osmr/pdf/st140040. pdf.

NHIS, 2017. National Health Interview Survey Homepage. Retrieved April 15, 2014, from. http://www.cdc.gov/nchs/nhis.htm.

Nanda, G., Grattan, K.M., Chu, M.T., Davis, L.K., Lehto, M.R., 2016. Bayesian decision support for coding occupational injury data. J. Safety Res. 57, 71–82. http://dx.doi.org/10.1016/j.jsr.2016.03.001.

National Data Standards for Injury Surveillance Version, 1998. National Data Standards for Injury Surveillance Version 2.1. AIHW National Injury Surveillance Unit Cat. no. NISU 6638., Canberra (Retrieved from). http://www.aihw.gov.au/publication-detail/?id=6442466997.

Ng, A.Y., Jordan, M.I., 2002. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (Eds.), On Discriminative Vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes. MIT Press, pp. 841–848. (Retrieved from). http://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-and-naive-bayes.pdf.

Phua, C., Alahakoon, D., Lee, V., 2004. Minority report in fraud detection. ACM SIGKDD Explor. Newsl. 6 (1), 50. http://dx.doi.org/10.1145/1007730.1007738.

Prati, R.C., Batista, G.E.A.P.A., Silva, D.F., 2015. Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. Knowl. Inf. Syst. 45 (1), 247–270. http://dx.doi.org/10.1007/s10115-014-0794-3.

Provost, F., 2000. Machine Learning from Imbalanced Data Sets 101. pp. 1–3. (Retrieved from). http://www.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-001.pdf.

Queensland Injury Surveillance Unit. (n.d.). Retrieved July 6, 2016, from http://www.qisu.org.au/ModCoreFrontEnd/index.asp?pageid=109.

Sun, A., Lim, E.-P., Liu, Y., 2009. On strategies for imbalanced text classification using SVM: a comparative study. Decis. Support Syst. 48 (1), 191–201. http://dx.doi.org/10.1016/j.dss.2009.07.011.

Taylor, J.A., Lacovara, A.V., Smith, G.S., Pandian, R., Lehto, M., 2014. Near-miss narratives from the fire service: a Bayesian analysis. Accid. Anal. Prev. 62, 119–129. http://dx.doi.org/10.1016/j.aap.2013.09.012.

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. http://dx.doi.org/10.1136/injuryprev-2015-041813.

Van Hulse, J., Khoshgoftaar, T.M., Napolitano, A., 2007. Experimental Perspectives on Learning from Imbalanced Data. ACM, New York, NY, USA, pp. 935–942. http://dx.doi.org/10.1145/1273496.1273614.

Wellman, H.M., Lehto, M.R., Sorock, G.S., Smith, G.S., 2004. Computerized coding of injury narrative data from the National Health Interview Survey. Accid. Anal. Prev. 36 (2), 165–171. http://dx.doi.org/10.1016/S0001-4575(02)00146-X.

Wiatrowski, W.J., (n.d.). The BLS Survey of Occupational Injuries and Illnesses: A Primer. Retrieved from https://www.bls.gov/iif/soii_primer.pdf.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, second edition. Morgan Kaufmann (Retrieved from). https://books.google.com/books?id=QTnOcZJzlUoC.