Transportation Safety and Environment, 2020, Vol. 0, No. 0 1–14

doi: 10.1093/tse/tdaa010
Research Article

RESEARCH ARTICLE

Application of machine learning tools in classifying pedestrian crash types: A case study

Subasish Das[1,*], Minh Le[2] and Boya Dai[3]

[1]Texas A&M Transportation Institute, Road Safety, 1111 Rellis Parkway, Suite 4414, Bryan, TX 77807, United States, [2]Texas A&M Transportation Institute, Research and Implementation Division, 12700 Park Central, Suite 1000, Dallas, TX 75251, United States and [3]Texas A&M Transportation Institute, Planning and Engagement, 505 East Huntland Drive, Suite 455, Austin, TX 78752, United States

*Corresponding author. E-mail: subasishsn@gmail.com

Abstract

Crash occurrence is a complex phenomenon, and crashes associated with pedestrians and bicyclists are even more complex. Furthermore, pedestrian- and bicyclist-involved crashes are typically not reported in detail in state or national crash databases. To address this issue, developers created the Pedestrian and Bicycle Crash Analysis Tool (PBCAT). However, it is labour-intensive to manually identify the types of pedestrian and bicycle crash from crash-narrative reports and to classify different crash attributes from the textual content of police reports. Therefore, there is a need for a supporting tool that can assist practitioners in using PBCAT more efficiently and accurately. The objective of this study is to develop a framework for applying machine-learning models to classify crash types from unstructured textual content. In this study, the research team collected pedestrian crash-typing data from two locations in Texas. The XGBoost model was found to be the best classifier. The high prediction power of the XGBoost classifiers indicates that this machine-learning technique was able to classify pedestrian crash types with the highest accuracy rate (up to 77% for training data and 72% for test data). The findings demonstrate that advanced machine-learning models can extract underlying patterns and trends of crash mechanisms. This provides the basis for applying machine-learning techniques in addressing the crash typing issues associated with non-motorist crashes.

Keywords: pedestrian crash; crash typing; machine learning; precision; accuracy

Received: 5 December 2019; **Revised:** 16 April 2020; **Accepted:** 6 May 2020

1. Introduction

Every year in the United States, the number of pedestrian deaths in collisions grows. In 2014, 4910 pedestrians were killed in US road accidents, followed by 5495 and 5987, respectively, in 2015 and 2016. On average, in 2016, a pedestrian was killed every 1.5 hours in a traffic collision [1]. These figures are comparable with those involving bicyclists, with 840 fatal bicycle crashes in the USA in 2016, up from 829 in 2015. This accounted for 2.2% of all traffic deaths during that year [2].

The development of effective countermeasures to prevent bicyclist and pedestrian crashes is typically inhibited by state crash databases that include inadequate information about these types of crash. The Pedestrian and Bicycle Crash Analysis Tool (PBCAT) was developed in order to resolve these issues. PBCAT is a stand-alone crash-typing application [4]. With the intention of helping local and state pedestrian and bicycle coordinators, technicians and designers by addressing collision issues, PBCAT permits users to construct an information database associated with crashes between pedestrians or bicyclists and motor vehicles. This option offers more precise crash details that explain the pre-crash actions of the involved parties. Using this description, PBCAT users are able to evaluate the data and construct reports, and then choose appropriate countermeasures to address the safety concerns [3]. In 2010, portions of PBCAT were adopted into the National Highway Traffic Safety Administration (NHTSA) records-based data-collection systems, the National Automotive Sampling System (NASS), the General Estimates System (GES) and the Fatality Analysis Reporting System (FARS). The legacy NASSGES was replaced by the Crash Report Sampling System (CRSS) in 2016.

The traditional approach has been to establish relationships between crash frequency and environmental conditions, traffic characteristics and roadway geometry. Recently, more attention has been directed toward the identification of factors that significantly influence various crash characteristics. The traditional data-analysis procedures use crash data structured similarly to police reports to perform injury-severity or crash-frequency analysis. In most police crash reports, a textual description of the crash event is included, but these textual crash narratives are not usually stored electronically. The narratives are generally unstructured or semi-structured textual data, and considerable manual effort is required to obtain information from them. Through the exploration of crash narratives, the possibility of losing specific details from these textual reports is high.

This study was designed to mitigate the current research gap by identifying pedestrian crash types (in regard to the intention of the pedestrians) with the use of machine-learning algorithms. This study aims to evaluate the efficiency of various machine-learning classification techniques in classifying crash narratives obtained from seven years of crash data in Texas. The study evaluated three machine-learning algorithms: support vector machines (SVMs), random forests (RFs) and XGBoost.

The structure of this paper is as follows. The following section is the literature review. The concepts of the modelling tools are then briefly introduced, after which the data preparation and model developments are demonstrated. Finally, the results of this evaluation are described, followed by the final conclusions and discussion.

2. Literature review

The literature review is divided into two major sections: (i) studies associated with pedestrian and bicycle crash typing, and (ii) studies associated with crash-narrative analysis.

2.1 Pedestrian and bicycle crash-typing analysis

Over the past 30 years, attempts made to reduce the number of bicycle–vehicle crashes and related deaths have been successful in Wisconsin. On this subject, Amsden and Huber [5] analysed bicycle–vehicle crashes in more detail and distinguished common attributes among crashes, specifically associated with roadway characteristics, traffic conditions and the users involved in the crashes, using PBCAT (version 2.0b) and a geographic information system. Recent statistics suggest about 20% (1002 out of 5376) of all pedestrian deaths in 2015 involved pedestrians over 65 years of age. Using joint correspondence analysis, Das et al. [6] used three years (2014 to 2016) of US FARS information to identify crucial relations between contributing variables.

Additionally, Das et al. [7] identified important relations between variables contributing to elderly pedestrian crashes. Using empirical Bayes (EB) information data mining, the authors analysed three years (2014 to 2016) of fatal older pedestrian crashes from FARS. Ernst [8] suggested that high-speed highways present the greatest threat to pedestrians.

Schneider and Stefanich [9] developed a new method called the location–movement classification method for classifying pedestrian and bicycle crashes. This new method provided useful information that was not captured by the well-established NHTSA crash typology when applied to a sample of 296 pedestrian and 229 bicycle crashes reported in Wisconsin between 2011 and 2013. Berkow et al. [10] conducted a critical evaluation of all types of state crash report in the USA to determine how each type captured location data on crashes involving bicyclists in report areas. The findings revealed that many types of state crash report did not provide sufficient detail for bicycle crash identification, especially for crashes linked to driving on sidewalks.

This section of the literature review reveals that Schneider and Stefanich [9] and Berkow et al. [10] used innovative approaches in improving non-motorist crash typing. Crash-narrative analysis was not explored in previous studies associated with pedestrian/bicyclist crash typing. The following section provides a brief overview of the crash-narrative analysis approaches used for different traffic crash-related research problems.

2.2 Crash-narrative analysis

Text-mining methods have proven to be useful by extracting valuable information from large text-based data sets. Text mining is used primarily to discover trends and anomalies, identify contributing factors, and develop predictive models that can act as a reinforcing guide to solve real-world problems [11, 12]. Previous studies have implemented in-depth text mining for crash and injury analysis, primarily to gain insights from occupational crash reports [13–20], health care reports [21–23], automobile crash reports [24–28] and other resources [11, 12].

2.2.1 Concept chain queries and Haddon's matrix. Researchers have developed a special form of text mining called concept chain queries for discovering essential evidence trails across documents that can be used to explain relationships between given factors [12]. This text-search technique was later developed to research the utility of crash-narrative text analysis for producing codes for injury mechanisms [20]. Researchers have utilized Haddon's matrix as the conceptual framework to code text from work-related injury reports to determine contributing factors of the injuries [18]. Furthermore, Haddon's matrix provides a coded data set and coding rules that divide fatal incidents into three event phases: pre-event, event and post-event [19].

2.2.2 Machine-learning algorithms. In crash-narrative analysis, two major approaches are widely used: (i) determining hidden trends from unstructured texts, and (ii) classifying crash types from crash-narrative reports.

Although small and less diverse data can make it difficult to identify recurrent scenarios from narrative text, a combined naive-fuzzy Bayesian approach allows for greater accuracy in narrative classification, and also selects the most pertinent data for manual review to reduce the workload for human coders [13, 22, 29]. Researchers have also used DUALIST, an online interactive program that allows novice users to organize thousands of narratives efficiently (within a few minutes) after an hour of training [11]. There has been further research to evaluate the effectiveness of the Bayesian-based model in comparison to other machine-learning algorithms, including neural networks [11], logistic regression models [31, 32] and SVM models [32]. These models all produce relatively accurate classifications for the emerging causes of occupational injury. A semi-supervised set covering machine (S3CM) learning algorithm has also been developed to identify coronary angiogram results and ovarian cancer diagnoses from electronic health-record narratives. The S3CM performance has been compared with that of the transudative support vector machine (TSVM), the original fully supervised set covering machine (SCM) and the 'freetext matching algorithm' natural-language processor; researchers found that the S3CM performed better than the TSVM and the fully supervised SCM after training with pre-classified test sets. Furthermore, this model does not depend on linguistic rules, but it does require further studies utilizing other electronic health-record data sets [33].

Recent research has applied text mining to crash analysis in the field of transportation research [24]. This research used a connectionist-based model for classifying free-text crash descriptions, and researchers used singular value decomposition for feature extraction and network training. Using human-classified data through both a fuzzy Bayes and a keyword-based model, the researchers analysed the performance and found that the connectionist and fuzzy Bayes model outperformed the keyword model. Another study conducted exploratory text mining and

EB data mining to determine the relationship between vehicle condition and automotive safety [25]. This study highlighted a number of vehicular manufacturing imperfections as critical factors, such as the 'brake system', 'airbags', 'seat belts' and 'speed control'. Latent Dirichlet allocation (LDA), a three-level hierarchical Bayesian model, was developed to determine major recurring crash factors from the text in Federal Railroad Administration reports. Further analysis was conducted using the Jigsaw text-visualization software and the text-clustering method, and an equivalent effect was found. A combination of random forest, LDA and partial least squares techniques was applied to accurately estimate the cost of railroad crashes and identify contributing factors [31]. Earlier researchers also used logistic regression [28, 29, 33], clustering [33] and other computational tools. For example, the Statistical Analysis System and Leximancer [30, 32] were used to determine the factors contributing to vehicle crashes. This section of the literature review reveals that computerized approaches and predictive models have greater potential to develop standardized crash-narrative text analysis and reduce human error in crash and injury surveillance. As predictive accuracies are emphasized in these studies, a wide variety of machine-learning algorithms were used to determine the best predictive method. One of the key limitations of machine-learning models is model interpretability. Although interpretable machine learning (IML) has been widely used in other research domains, none of the above-mentioned studies used IML in their crash-narrative investigation.

The review of the literature indicates that there is a need for an in-depth investigation of pedestrian crash-typing analysis using innovative methods such as crash-narrative analysis. This study therefore used three machine-learning tools to classify crash types from pedestrian crash-typing data from two locations in Texas. In addition, an odds-ratio analysis was performed to provide context for model interpretation.

3. Methodology

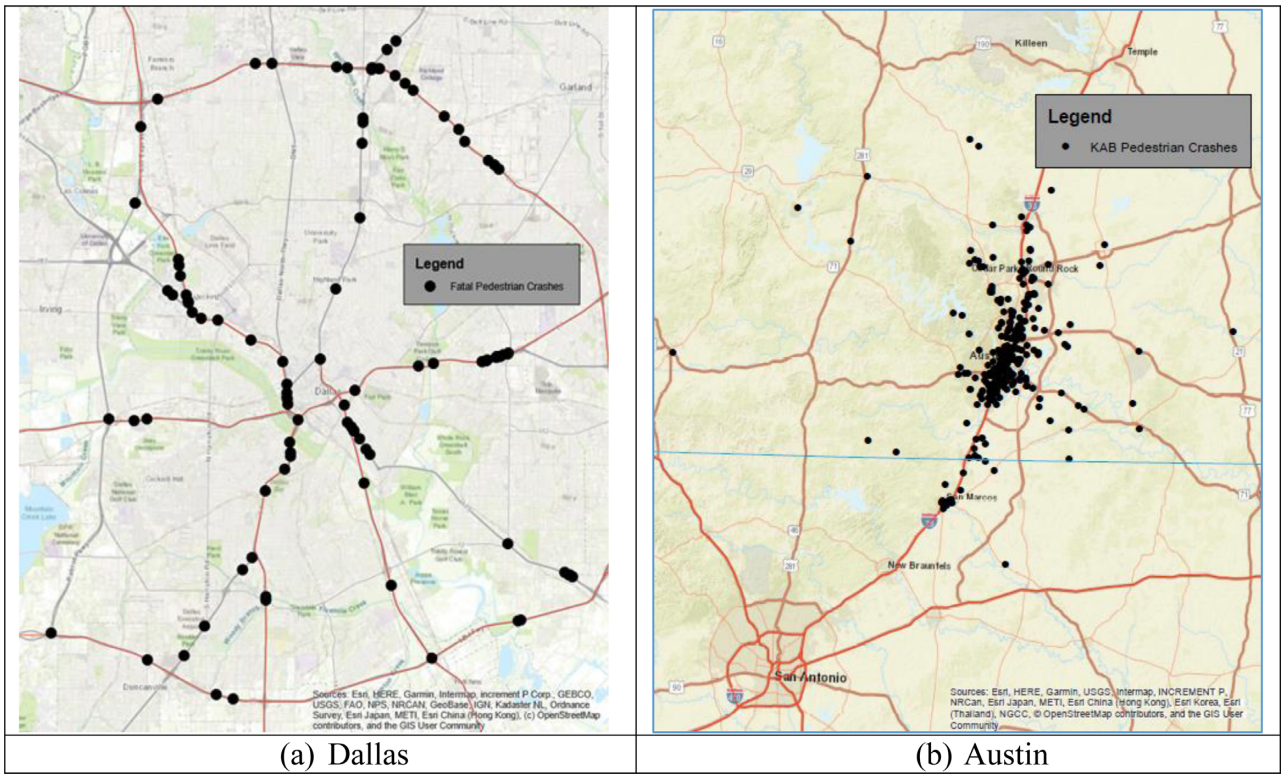
3.1 Machine-learning models

One of the central qualities of artificial intelligence (AI) is the ability to learn. Machine learning spans multiple AI disciplines; it is a method used to train tools to make connections and learn patterns in order to make precise estimations. Machine-learning models are typically used to extract information from raw data. There are two types of machine learning: supervised learning and unsupervised learning. A machine-learning algorithm creates a set of rules for the computational tool to follow in order to learn how to complete a specific task. The outcome of the machine-learning algorithm is a machine-learning model. Models can estimate, categorize or achieve other goals based on the problem type. Unlike machine learning, conventional statistical modelling identifies associations between variables using mathematical equations. Ease of interpretation is the greatest advantage of conventional statistical modelling. One significant drawback of this method, however, is that predetermined assumptions must be made at the outset.

Natural-language processing, text mining and machine learning are useful data science tools. These methods can be used to collect and analyse data in order to discover hidden trends from huge text corpora, like crash-narrative report databases. This study includes longitudinal studies designed to find significant patterns in the data that can be used to improve classification accuracy. The research team used three different machine-learning tools to perform the analysis: SVMs, RFs and XGBoost. A description of these tools is provided below. Interested readers can consult Bishop's book [35] for more detail.

3.1.1 Random forests. RFs are based on two main principles: the bagging principle [36] and the random subspace method [37]. The random subspace method constructs a collection of decision trees with random predictors. The general architecture of an RF includes several steps: (i) generating clusters to grow a tree by randomly selecting the explanatory variables, (ii) using the explanatory variables at the node of the tree to classify labels at this node, (iii) running the out-of-bag (OOB) data to determine misclassification, (iv) repeating the first three steps until minimum OOB is achieved, and (v) assigning each observation to a final class by majority classification estimates.

3.1.2 Support vector machines. SVMs have proven successful in various real-world learning tasks [38]. The SVM framework is explicitly defined by a separate hyperplane that acts as a discriminatory classifier. To explain further, an algorithm that sorts fresh examples creates an ideal hyperplane by considering the marked training data



(a) Dallas (b) Austin

Fig. 1: Locations of pedestrian crashes in the Dallas and Austin data sets

(controlled learning). The hyperplane is a row that separates a plane into two halves in a two-dimensional space where each category lies on either side [39]. The SVM framework works similarly to other machine-learning algorithms: it randomly selects a training set. The generalized portrait algorithm was introduced in 1963 by Vapnik and Lerner; it has a core algorithm that produces SVMs, which are statistical learning-theory algorithms that implement the structural risk-minimization inductive principle to achieve good generalization on a small number of learning patterns. In 1974, Vapnik created the research field of statistical learning theory [38]. Vapnik et al. developed the current SVM framework based on a separable bipartition problem at the AT&T Bell Laboratories in 1992 [39]. An SVM works by mapping the data x into a high-dimensional feature space F via non-linear mapping and performing linear regression in this space. SVMs are more capable of achieving significant gains than other current, high-performing techniques, and they can accomplish many distinct learning tasks.

3.1.3 XGBoost. Extreme Gradient Boosting (XGBoost) is a gradient-boosting library algorithm based on gradient-boosted decision trees. It is generally used for improving model accuracy and robustness. Gradient boosting is an ensemble technique that connects predictors sequentially and corrects prior designs. Instead of assigning distinct weights to the classifiers after each iteration, this technique fits the fresh model to the past prediction's fresh residuals and then minimizes the loss by incorporating the recent estimate [34]. XGBoost is based on machine-learning algorithms within a gradient-boosting framework. It provides a parallel tree-boosting algorithm that reaches the optimized point in a fast and accurate way.

3.2 Data description

3.2.1 Data preparation. The data set of the current study is police crash reports from two locations (the City of Dallas and Austin District) in Texas (see Fig. 1). The Dallas data involves freeway facilities, where pedestrians are not expected because they are prohibited. The Austin data also contains non-freeways where pedestrians are provided with a large number of access points. The crash data set for Austin covers severe crashes (K = fatal, A = incapacitating injury, B = non-incapacitating injury) on all roadways and the corresponding crash-narrative report in text format for each crash in the year 2018. Meanwhile, the

crash data set for Dallas covers fatal crashes (K) on freeways between the years 2008 and 2017. The data set consists of 341 and 101 crash narratives for Austin and Dallas, respectively. These crash narratives create two text corpora), consisting of 30 000 and 14 000 words, respectively.

3.2.2 Descriptive statistics. In a study conducted by Le et al. [41], crash data from the Dallas District of the Texas Department of Transportation (TxDOT) was analysed; the data included 8332 pedestrian crashes (KAB) from 2008 to 2017. Of these crashes, 4696 crashes (56%) occurred within the City of Dallas, of which 327 crashes (7%) occurred on freeways (access-controlled highways). Of these freeway crashes, 128 (39%) were fatal. The crashes that resulted in fatalities were reviewed to gain a better understanding of the fatal crashes that occurred on Dallas freeways. It is important to note that 25 of these crash reports were not available.

A key objective of the researchers was to determine if the pedestrians involved in the crashes were on the freeway intentionally or not. To achieve this, researchers carefully reviewed each crash narrative and diagram to determine why the pedestrian was at the crash location. The crashes were then coded by selecting one of the plausible reasons listed below:

(i) Changing seats in vehicle,
(ii) Commuting/moving from one place to another,
(iii) Crossing roadway,
(iv) Fleeing police,
(v) Jumping from bridge,
(vi) Jumping from car,
(vii) Previous crash,
(viii) Retrieving items from road,
(ix) Stalled vehicle,
(x) Standing in traffic,
(xi) Standing on median, on shoulder or off road,
(xii) Suicide,
(xiii) Taking pictures,
(xiv) Unconscious,
(xv) Walking along the sidewalk,
(xvi) Walking or lying down in traffic,
(xvii) Walking or lying down on median, on shoulder or off road,
(xviii) Working,
(xix) Unknown/other

Some cases could fall into multiple categories; for example, a pedestrian could have been leaving a 'stalled vehicle' and also 'standing on median, on shoulder or off road' when they were fatally struck. The research team defined an 'unintended' pedestrian as a person who was struck and (i) associated with a vehicle at the scene (stalled or otherwise), or (ii) a worker actively performing their duty at the scene. On the other hand, 'intended' indicates a person who did not meet the criteria of 'unintended'. For consistency, the researchers generally coded 'stalled vehicle' even though other reasons could have explained why the pedestrian was at the crash location. Table 1 shows a list of measures or reasons and their association with the classifications 'intended' and 'unintended' for the Dallas data set.

It is important to classify the intent of the pedestrians on the freeway to better design policies and treatments for reducing such crashes. As shown in Table 1, a majority of the fatal crashes involved pedestrians intentionally walking on Dallas freeways. These results were unexpected because pedestrians are legally prohibited from walking on freeways. It should be noted that the unintended pedestrian coded as 'fleeing police' first exited his car before being fatally struck as he was crossing the freeway. This finding did not support the statewide study by Fitzpatrick et al. [42], which found that 5% (24 of 474) of fatal freeway pedestrian crashes were not associated with a vehicle. But the study also found that 68% of crash reports analysed did not include the reason why the pedestrian was at the crash location. The researchers suspect that the Dallas sample may need to be expanded beyond Dallas and include other crash severity types to get a truer picture of how many pedestrians are intended vs. unintended.

In another study, Hudson and Boya [43] collected and analysed over 2500 KAB pedestrian crashes that occurred on on- and off-system roadways within the TxDOT Austin District over a seven-year period between 2011 and 2018. The researchers aimed to identify the events prior to the crashes in order to assign fault. They analysed information from crash reports involving pedestrians and bicyclists from the TxDOT Crash Records Information System (CRIS). However, CRIS does not provide the level of detail required for this analysis, so crash narratives were obtained from police reports (CR-3s) and entered into PBCAT, which classified crashes by type. Fig. 2 presents example illustrations of crash types involving typical pedestrian actions. As shown in the figure, the crash-type images guide the user

Table 1: Reporting on intended vs. unintended pedestrians in fatal freeway crashes (Dallas data)

Reason	Intended	Unintended	Undefined	Total
Crossing roadway	38	–	–	38
Walking or lying down in traffic	11	–	–	11
Standing in traffic	6	–	–	6
Walking or lying down on median, on shoulder or off road	3	–	–	3
Fleeing police	2	1	–	3
Suicide	2	–	–	2
Commuting/moving from one place to another	1	–	–	1
Standing on median, on shoulder or off road	1	–	–	1
Previous crash	–	10	–	10
Retrieving items from road	–	1	–	1
Stalled vehicle	–	20	–	20
Unconscious	–	–	1	1
Working	–	3	–	3
Unknown	1	–	2	3
Missing reports (blank)	–	–	25	25
Total	65	35	28	128

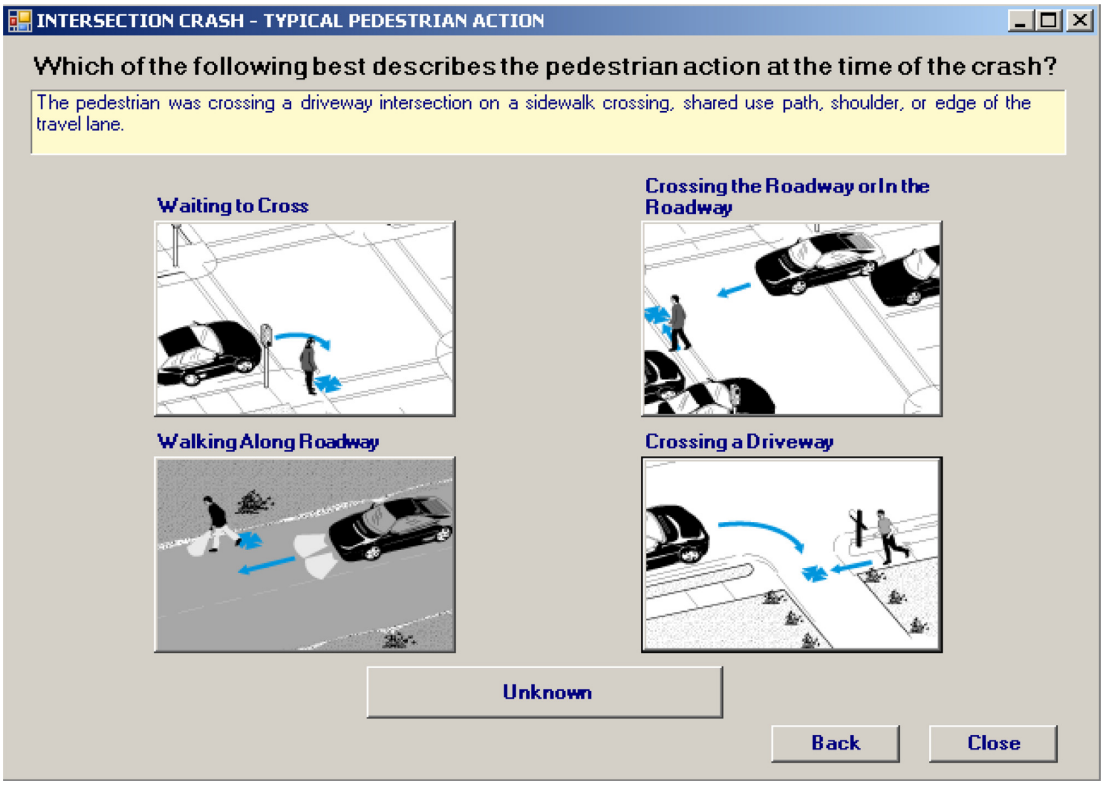


Fig. 2: Examples of the intersection crash type in PBCAT

in defining the correct circumstances. When the mouse pointer hovers over an image, the corresponding action description appears in the narration box.

The answer selected determines the follow-up questions presented. For instance, if the user selects 'waiting to cross', the next screen will ask whether the motor vehicle was turning or not turning at the time of the crash, or if there is insufficient information. If 'crossing the roadway or in the roadway' is selected, a different set of options will appear, as shown in Fig. 3.

In the data-collection phase of this study, the researchers determined the cause of each crash by compiling information from police reports, satellite images, pedestrian laws, and guidelines from the City of Austin and TxDOT. In crash-typing research efforts, it is also important to evaluate 'at-fault' scenarios to determine better policies and intervention designs. The police reports con-

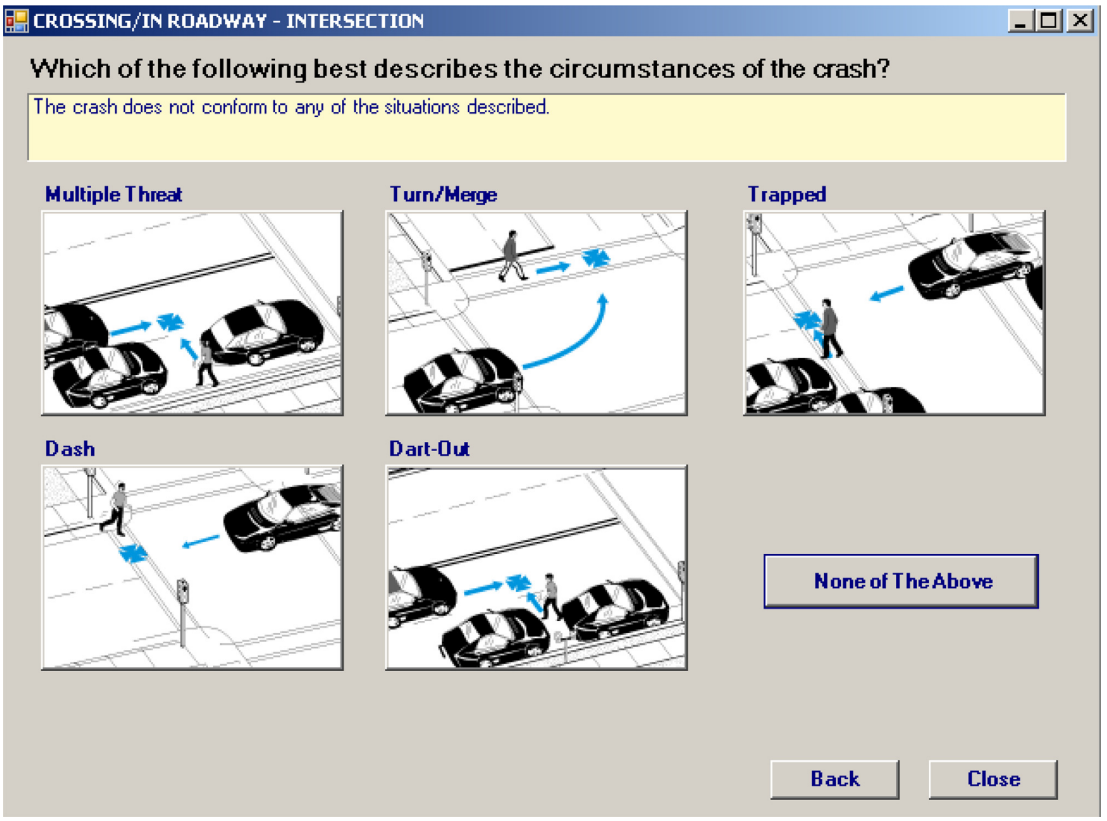


Fig. 3: Examples of the crossing/in roadway crash type in PBCAT

tained information on the crash location, contributing factors, citation and crash narrative, as well as a crash diagram. The satellite images allowed the researchers to obtain an accurate understanding of the street layout and roadway facilities, as they were sometimes portrayed inaccurately in or missing from the police reports, and assisted with the 'at-fault' determination for the crash. The final decision was affected by related laws and guidance, such as the Texas Transportation Code [44] and the guidance leaflets *Pedestrian Safety: A Guide to Applicable Laws in Austin* [45] and *Pedestrian Safety and the Law* [46].

The analysis results of pedestrian crashes occurring in Austin in 2018 suggested that, on average, motorists (45%) and pedestrians (42%) were almost equally at fault (see Table 2). Over half (59%) of crashes occurred when pedestrians were walking/running into the road (dash/dart-out) or crossing the roadways/driveways/expressways. The two most common crash groups involved a pedestrian crossing the roadway (non-expressway) in the following situations: (i) when a vehicle was not turning (29%) and (ii) when a vehicle was turning (21%). These were followed by unusual circumstances (16%) and dash/dart-out (9%). In the event that a pedestrian crossed the roadway and they were struck by a vehicle that was turning, the motorist was more likely to be at fault (79%); if the vehicle was not turning, then the pedestrian was more likely to be at fault (81%). If the pedestrian dashed or darted out into the road, the pedestrians were at fault in all cases analysed.

3.3 Framework for crash-narrative analysis

The current study developed a framework for solving the classification problem using crash-narrative data. The steps are as follows:

3.3.1 Step 1: data collection. The first stage is to retrieve the crash narratives' digital information. The crash reports are handwritten in many cases and are not recorded electronically. However, the job of digitizing the crash reports has started in many states. Louisiana, for instance, holds an electronic database of crash-report narratives.

Table 2: At-fault determination based on crash group description (Austin data)

Pedestrian location	Party at fault (%)			
	Motorist	Pedestrian	Both	Undefined
Crossing roadway (vehicle turning)	78.6	11.2	3.1	7.1
Crossing roadway (vehicle not turning)	14.3	81.4	2.9	1.4
Unusual circumstances	58.2	23.6	18.2	0.0
Dash/dart-out	0.0	100.0	0.0	0.0
Walking along roadway	50.0	9.1	40.9	0.0
Crossing expressway	0.0	93.8	6.3	0.0
Pedestrian in roadway (circumstances unknown)	10.0	50.0	0.0	40.0
Multiple threat/trapped	0.0	87.5	12.5	0.0
Crossing driveway or alley	85.7	0.0	0.0	14.3
Other/unknown (insufficient details)	14.3	14.3	14.3	57.1
Backing vehicle	100.0	0.0	0.0	0.0
Unique midblock	60.0	20.0	20.0	0.0
Off roadway	66.7	0.0	33.3	0.0
Bus-related	50.0	50.0	0.0	0.0
Working or playing in roadway	100.0	0.0	0.0	0.0
Waiting to cross	100.0	0.0	0.0	0.0
Total	**44.6**	**41.9**	**8.5**	**5.0**

*3.3.2 **Step 2: data cleaning.*** Data cleaning can be done with text-mining algorithms. It is possible to use available lexicons to extract redundant phrases. However, domain-specific lexicons are needed. For instance, when researching crash reports, numerical values are sometimes vital. Vehicle 1 and Vehicle 2 usually indicate at-fault and not-at-fault vehicles, respectively, in a two-vehicle crash. Removing all numerical data from the textual data would not be a good strategy for crash-narrative analysis. This process helps in performing feature extraction in the form of n-grams (i.e. sequences of *n* number of words) or features (as shown in Fig. 4).

*3.3.3 **Step 3: predictive modelling application.*** Many studies have developed innovative machine-learning tools to solve the classification problem. Several machine-learning models can be examined to select the best model with the lowest rate of misclassification. This study investigated three machine-learning models to determine the most suitable model. The framework for using training and test data is shown in Fig. 4.

Before applying the machine-learning algorithms, the text-mining tools were applied to reduce noise in the data set. The existence of excess words and redundant characteristics in narratives is one of the most widespread problems in crash-narrative analysis. Additionally, to render the classification more robust, phrases or sections of phrases with similar meanings were compressed into the same word. Redundant-word removal was performed to prepare the final data set. Future studies could perform more robust data cleaning to improve the precision of the model based on domain-specific lexicons.

4. Results and discussion

4.1 Machine-learning models

The research team used three machine-learning algorithms to determine the classification types of pedestrian crashes from the two data sets. Textual data from both data sets was split into a training set (around 70%) and a testing set (around 30%) through random sampling:

(i) Dallas (training data: 60 crashes, testing data: 30 crashes), and
(ii) Austin (training data: 205 crashes, testing data: 90 crashes)

The classes for these two databases were developed manually by an expert group. For the Dallas data set, manual reading of the crash reports did not provide enough evidence to determine the intention of the pedestrian in each crash. These crashes were identified as ‘unknown’. The Austin data set had similar issues. When the performance of each model was evaluated, crash reports classed as ‘unknown’ were excluded in the model development. Assigning data points to the training and testing data sets was done using stratified resampling. The models built using the training data were then used on the testing data to evaluate their performance. Table 3 presents the results for the three models when classifying intended

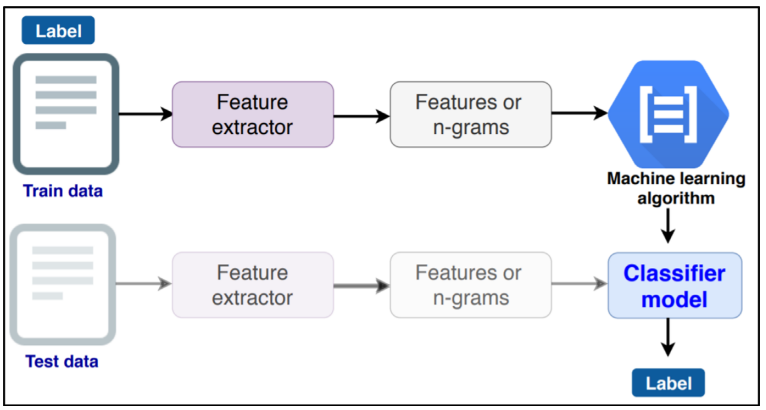


Fig. 4: Machine-learning framework for pedestrian crash-type prediction

Table 3: Confusion matrix for the predicted classes (Dallas data)

Model	Intention class (observed)	Training (60 crashes)		Testing (30 crashes)	
		Intended (predicted)	Unintended (predicted)	Intended (predicted)	Unintended (predicted)
SVM	Intended	26	14	10	6
	Unintended	8	12	6	8
RF	Intended	25	15	10	6
	Unintended	8	12	6	8
XGBoost	Intended	29	11	12	4
	Unintended	6	14	5	9

and unintended crashes from the Dallas data in the form of a confusion matrix.

Table 4 presents the results for the models when classifying at-fault motorist and pedestrian crashes from the Austin data.

The measures true positive (TP) and false positive (FP) are instances of correct and incorrect classification per actual class, respectively. True negative (TN) and false negative (FN) are instances of correct and incorrect rejection per actual class, respectively [40]. Some of the common performance measures are:

(i) Recall or sensitivity $= \frac{TP}{TP+FN}$ = effectiveness of positive-level identifications,
(ii) Specificity $= \frac{TN}{TN+FP}$ = effectiveness of negative-level identifications,
(iii) Precision $= \frac{TP}{TP+FP}$ = class agreement of data labels with positive labels,
(iv) Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$ = overall accuracy,
(v) Balanced accuracy $= \frac{TP}{TP+FN} \times 0.5 + \frac{TN}{TN+FP} \times 0.5$ = balanced accuracy, and
(vi) F-score $= \frac{2 \times Precision \times Recall}{Precision+Recall}$ = weighted average of recall and precision

Table 5 lists the performance measures of the three machine-learning models for both the training and the testing data from the Dallas data set. Based on the performance-measure values, XGBoost showed better performance than the other two models.

Table 6 lists the performance measures of the three models for both the training and the testing data from the Austin data set. Based on the performance-measure values, XGBoost again showed better performances than the other two models. The performance of XGBoost exceeded the performance of the algorithms for several reasons, including: (i) it was able to perform feature selection automatically and capture high-order associations without breaking down, and (ii) it included an additional randomization parameter to decrease the correlation of each tree.

4.2 Log odds ratio from bigrams of the crash narratives

The *odds* of the usage of word z in group *i*e is: $O_{kz}^{(i)} = f_{kz}^{(i)} / (1 - f_{kz}^{(i)})$ (where $f_{kz}^{(i)} = \frac{y_{kz}^{(i)}}{n_k^{(i)}}$; $y_{kz}^{(i)}$ denotes the Z-vector of word frequencies from documents of class *i* in topic *k*). The odds ratio between the two document groups is $\theta_{kz}^{(g1-g2)} = O_{kz}^{(g1)} / O_{kz}^{(g2)}$. This ratio is typically given for single words in isolation or used as a measurement to rank words. As machine-learning models are black-box in nature,

Table 4: Confusion matrix for the predicted classes (Austin data)

Model	At-fault class (observed)	Training (205 crashes)		Testing (90 crashes)	
		Motorist (predicted)	Pedestrian (predicted)	Motorist (predicted)	Pedestrian (predicted)
SVM	Motorist	70	35	31	16
	Pedestrian	44	56	22	21
RF	Motorist	68	37	30	17
	Pedestrian	46	54	22	21
XGBoost	Motorist	75	30	75	30
	Pedestrian	39	61	39	61

Table 5: Performance measures of the machine-learning models (Dallas data)

Model	Data set	Sensitivity	Specificity	Accuracy	Balanced accuracy	Precision	F-score
SVM	Training	0.6500	0.6000	0.6333	0.6250	0.7647	0.7027
	Testing	0.6250	0.5714	0.6000	0.5982	0.6250	0.6250
RF	Training	0.6250	0.6000	0.6167	0.6125	0.7576	0.6849
	Testing	0.6250	0.5714	0.6000	0.5982	0.6250	0.6250
XGBoost	Training	0.7250	0.7000	0.7167	0.7125	0.8286	0.7733
	Testing	0.7500	0.6429	0.7000	0.6964	0.7059	0.7273

Table 6: Performance measures of the machine-learning models (Austin data)

Model	Data set	Sensitivity	Specificity	Accuracy	Balanced accuracy	Precision	F-score
SVM	Training	0.6667	0.5600	0.6146	0.6133	0.6140	0.6393
	Testing	0.6596	0.4884	0.5778	0.5740	0.5849	0.6200
RF	Training	0.6476	0.5400	0.5951	0.5938	0.5965	0.6210
	Testing	0.6383	0.4884	0.5667	0.5633	0.5769	0.6061
XGBoost	Training	0.7143	0.6100	0.6634	0.6621	0.6579	0.6849
	Testing	0.7021	0.6047	0.6556	0.6534	0.6600	0.6804

the research team performed odds ratio analysis to provide some inference of the modelling algorithms. This approach helped in understanding which word or word pair was used as the identifier for the classification of crash types. Fig. 5 displays the log odds ratios for the most common bigrams (i.e. pairs of consecutive words) from the Dallas reports of fatal pedestrian crashes. Based on this data, the odds of a crash report associated with a motorist at fault including a variant of the phrase 'left turn' were 1.6 times those of a crash report associated with a pedestrian at fault including the same; this is shown in Fig. 6. Bigrams such as 'stop sign', 'left turn', 'red light' and 'the crosswalk' had odds ratios greater than 1. This indicates that motorists were at fault in intersection-related crashes due to poor judgement. On the other hand, running-related crashes can be defined as 'pedestrian-at-fault' crashes, and the higher odds ratios provided evidence that crash-narrative analysis using machine learning was capable of classifying these crashes.

5. Conclusions

Conventional police crash reports contain inadequate information about the types of pedestrian crash. The use of these conventional reports can hinder the development of effective countermeasures to prevent pedestrian crashes. Pedestrian crash typing is helpful in describing the pre-crash scenarios to better define the sequence of events and key contributing factors leading to pedestrian crashes. This study used pedestrian crash-typing data and manual classification extraction using an expert group. The high prediction power of the XGBoost classifiers indicates that this machine-learning technique was able to classify pedestrian crash types (intended vs. untended and pedestrian at fault vs. motorist at fault) with the highest accuracy rate (up to 77% for the training data and 72% for the testing data). This suggests that unknown patterns and trends can be uncovered and examined using powerful machine-learning models. This provides grounds for applying quantitative modelling techniques in addressing

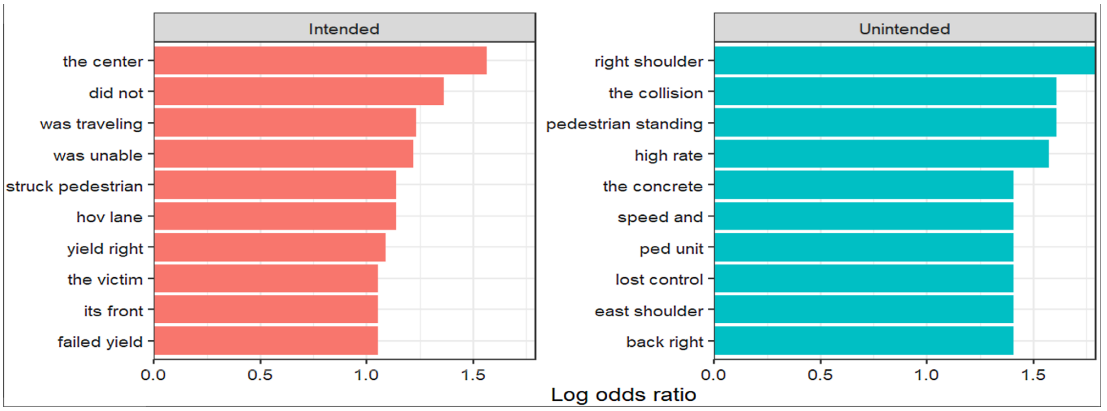


Fig. 5: Log odds ratios for the bigrams from the crash narratives (Dallas data: what is the intention?)

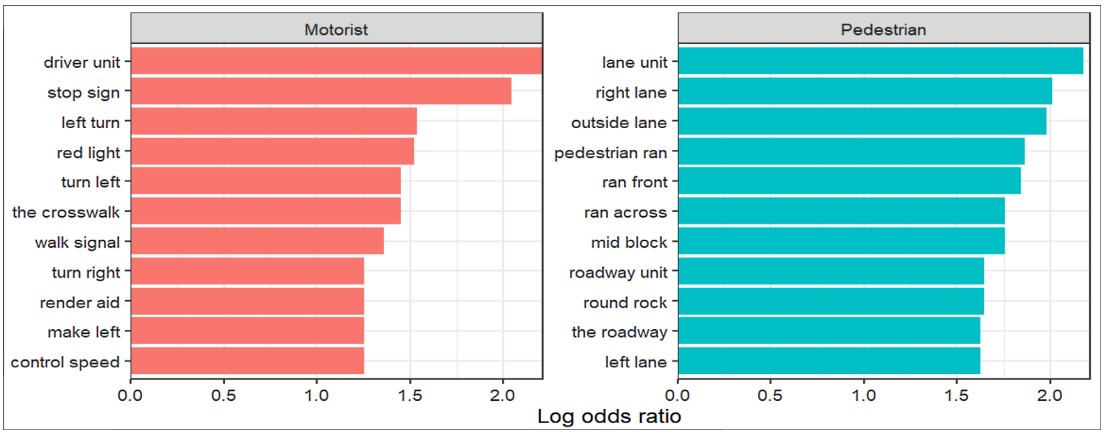


Fig. 6: Log odds ratios for the bigrams from the crash narratives (Austin data: who is at fault?)

crash-typing issues for non-motorist crashes. This study has demonstrated that machine-learning tools identify crash types from crash narrative-free text. Since crash-narrative reports are unused in many cases, as the work of reviewing them systematically is too labour-intensive for many agencies and practitioners, the framework developed in this study has the potential to be used in tackling other crash-related classification tasks (for example, collision type) using crash narratives.

The current study is not without limitations. First, the classification accuracies are not very high. Second, the sample size of the current study is small. Third, the results of the study are conditional on the information provided in the textual content of the police reports. Future studies, with the inclusion of a larger sample, will be able to develop a robust crash-narrative lexicon of stop words and trigger words (words with high association values with crash injury). This will help in reducing misclassification of the crash type.

Conflict of interest statement. None declared.

References

1. National Center for Statistics and Analysis. *Traffic Safety Facts, 2016 Data: Pedestrians*. 2018. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812493 (18 May 2020, date last accessed).
2. National Center for Statistics and Analysis. *Traffic Safety Facts, 2016 Data: Bicyclists and Other Cyclists*. 2018. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812507 (18 May 2020, date last accessed).
3. Pedestrian & Bicycle Information Center. *About PBCAT*. http://www.pedbikeinfo.org/pbcat_us/about.cfm (18 May 2020, date last accessed).
4. Terranova S. 2017 FARS/CRSS pedestrian bicyclist crash typing manual. *Technical report*. National Highway Traffic Safety Administration 2017.
5. Amsden M, Huber T. Bicycle crash analysis for Wisconsin using a crash typing tool (PBCAT) and geographic information system (GIS). *Technical report*. Wisconsin Department of Transportation 2006.

6. Das S, Jha K, Fitzpatrick K, *et al*. Pattern identification from older bicyclist fatal crashes. *Transp Res Rec* 2019; **2673**: 638–49.
7. Das S, Bibeka A, Sun X, *et al*. Elderly pedestrian fatal crash-related contributing factors: applying empirical Bayes geometric mean method. *Transp Res Rec* 2019; **2673**:254–63.
8. Ernst M. Mean streets 2004: how far have we come? Pedestrian safety, 1994–2003. *Technical report*. Surface Transportation Policy Project 2004.
9. Schneider RJ, Stefanich J. Application of the location–movement classification method for pedestrian and bicycle crash typing. *Transp Res Rec* 2016; **2601**:72–83.
10. Berkow M, van Hengel D, Blanc B, *et al*. Improvements to statewide collision reporting to understand sidewalk-related bicycle collisions. *Technical report*. Transportation Research Board 2017.
11. Gopalakrishnan K, Khaitan SK. Text mining transportation research grant Big Data: knowledge extraction and predictive modeling using fast neural nets. *Int J Traffic Transp Eng* 2017; **7**:354–67.
12. Jin W, Srihari RK, Hay Ho H, *et al*. Improving knowledge discovery in document collections through combining text retrieval and link analysis techniques. In: *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, Omaha, NE, USA, 2007, pp. 193–202.
13. Abdat F, Leclercq S, Cuny X, *et al*. Extracting recurrent scenarios from narrative texts using a Bayesian network: application to serious occupational accidents with movement disturbance. *Accid Anal Prev* 2014; **70**:155–66.
14. Marucci-Wellman H, Lehto M, Corns H. A combined Fuzzy and Naïve Bayesian strategy can be used to assign event codes to injury narratives. *Inj Prev* 2011; **17**:407–14.
15. Smith GS, Timmons RA, Lombardi DA, *et al*. Work-related ladder fall fractures: identification and diagnosis validation using narrative text. *Accid Anal Prev* 2006; **38**:973–80.
16. Bertke SJ, Meyers AR, Wurzelbacher SJ, *et al*. Comparison of methods for auto-coding causation of injury narratives. *Accid Anal Prev* 2016; **88**:117–23.
17. Vallmuur K, Marucci-Wellman HR, Taylor JA, *et al*. Harnessing information from injury narratives in the 'Big Data' era: understanding and applying machine learning for injury surveillance. *Inj Prev* 2016; **22**:34–42.
18. Bondy J, Lipscomb H, Guarini K, *et al*. Methods for using narrative text from injury reports to identify factors contributing to construction injury. *Am J Ind Med* 2005; **48**:373–80.
19. Bunn TL, Slavova S, Hall L. Narrative text analysis of Kentucky tractor fatality reports. *Accid Anal Prev* 2008; **40**:419–25.
20. Williamson A, Feyer AM, Stout N, *et al*. Use of narrative analysis for comparisons of the causes of fatal accidents in three countries: New Zealand, Australia, and the United States. *Inj Prev* 2001; **7**:15–20.
21. Chen W, Wheeler KK, Lin S, *et al*. Computerized 'learn-as-you-go' classification of traumatic brain injuries using NEISS narrative data. *Accid Anal Prev* 2016; **89**:111–17.
22. Marucci-Wellman HR, Lehto MR, Corns HL. 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* 2015; **84**:165–76.
23. Wang Z, Shah AD, Tate AR, *et al*. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. *PLoS One* 2012; **7**:e30412.
24. Chatterjee S. A connectionist approach for classifying accident narratives. *Ph.D. Thesis*. Purdue University 1998.
25. Das S, Mudgal A, Dutta A, *et al*. Vehicle consumer complaint reports involving severe incidents: mining large contingency tables. *Transp Res Rec* 2018; **2672**:72–82.
26. Beanland V, Fitzharris M, Young KL, *et al*. Driver inattention and driver distraction in serious casualty crashes: data from the Australian National Crash In-Depth Study. *Accid Anal Prev* 2013; **54**:99–107.
27. Brown DE. Text mining the contributors to rail accidents. *IEEE Trans Intell Transp Syst* 2016; **17**:346–55.
28. Fitzpatrick CD, Rakasi S, Knodler MA. An investigation of the speeding-related crash designation through crash narrative reviews sampled via logistic regression. *Accid Anal Prev* 2017; **98**:57–63.
29. Graves JM, Whitehill JM, Hagel BE, *et al*. Making the most of injury surveillance data: using narrative text to identify exposure information in case-control studies. *Injury* 2015; **46**:891–7.
30. Nayak R, Piyatrapoomi N, Weligamage J. Application of text mining in analysing road crashes for road asset management. In: *Engineering Asset Lifecycle Management: Proceedings of the 4th World Congress on Engineering Asset Management (WCEAM 2009)*, Athens, Greece, 2010, pp. 49–58.
31. Pollack KM, Yee N, Canham-Chervak M, *et al*. Narrative text analysis to identify technologies to prevent motor vehicle crashes: examples from military vehicles. *J Safety Res* 2013; **44**:45–49.
32. Sorock GS, Ranney TA, Lehto MR. Motor vehicle crashes in roadway construction workzones: an analysis using narrative text from insurance claims. *Accid Anal Prev* 1996; **28**:131–8.
33. Williams TP, Betak JF. Identifying themes in railroad equipment accidents using text mining and text visualization. In: *International Conference on Transportation and Development*, Houston, TX, USA, 2016; pp. 351–57.
34. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001; **29**:1189–232.
35. Bishop CM. *Pattern Recognition and Machine Learning*. New York: Springer, 2011.
36. Breiman L. Random forests. *Mach Learn* 2001; **45**:5–32.
37. Ho TK The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell* 1998; **20**:832–44.
38. SVM Learning. *History of Support Vector Machines*. https://www.svms.org/history.html (18 May 2020, date last accessed).
39. Smola AJ, Schölkopf B. A tutorial on support vector regression. *Stat Comput* 2004; **14**:199–222.
40. Labatut V, Cherifi H. Accuracy measures for the comparison of classifiers. In: *Proceedings of the 5th International Conference on Information Technology*, Amman, Jordan, 2011; pp. 1–5.
41. Le M, Srinivas G, Neal J, *et al*. Understanding Dallas district pedestrian safety issues. *Technical report*. Texas A&M Transportation Institute 2019.
42. Fitzpatrick K, Iragavarupu V, Brewer M, *et al*. Characteristics of Texas pedestrian crashes and evaluation of driver yielding at pedestrian treatments. *Technical report*. Texas A&M Transportation Institute 2014.
43. Hudson J, Boya D. Pedestrian and bicycle crash analysis work. *Technical report*. Texas A&M Transportation Institute 2019.
44. Texas Legislature. *Texas Transportation Code*. 2019. https://statutes.capitol.texas.gov/Docs/SDocs/Transportationcode.pdf (18 May 2020, date last accessed).

45. Austin Pedestrian Advisory Council. *Pedestrian Safety: A Guide to Applicable Laws in Austin*. https://www.austintexas.gov/edims/document.cfm?id=271172 (18 May 2020, date last accessed).

46. City of Austin. *Pedestrian Safety and the Law*. https://austintexas.gov/sites/default/files/files/Transportation/Ped_Safety_and_the_Law_Austin.pdf (18 May 2020, date last accessed).