Accident Analysis and Prevention 159 (2021) 106211

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Identification and analysis of misclassified work-zone crashes using text mining techniques

Md Abu Sayed [a,c], Xiao Qin [a,c,*], Rohit J. Kate [b,c], D.M. Anisuzzaman [b], Zeyun Yu [b,c]

[a] Department of Civil and Environmental Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI, 53201, United States
[b] Department of Computer Science, University of Wisconsin-Milwaukee, Milwaukee, WI, 53201, United States
[c] Institute for Physical Infrastructure and Transportation (IPIT), University of Wisconsin-Milwaukee, Milwaukee, WI, 53201, United States

ARTICLE INFO

Keywords:
Work zone
Crash data
Missclassified
Crash narrative
Text mining
Noisy-OR

ABSTRACT

Work zone safety management and research relies heavily on the quality of work zone crash data. However, it is possible that a police officer may misclassify a crash in structured data due to: restrictive options in the crash report; a lack of understanding about their importance; lack of time due to police officers' work load; and ignorance of work zone as one of the crash contributing factors. Consequently, work zone crashes are under representative in crash statistics. Crash narratives contain valuable information that is not included in the structured data. The objective of this study is to develop a classifier that applies text mining techniques to quickly find missed work zone (WZ) crashes through the unstructured text saved in the crash narratives.

The study used three-year crash data from 2017 to 2019. The data from 2017 to 2018 was used as training data, and the 2019 data was used as testing data. A unigram + bigram noisy-OR classifier was developed and proven to be an efficient and effective means of classifying work zone crashes based on key information in the crash narrative. The ad-hoc analysis of misclassified work zone crashes sheds light on when, where and the plausible reasons as to why work zone crashes are more likely to be missed.

1. Introduction

Work zone activities are essential for maintaining good roadways, supporting economic development and competition, and improving safety. While road work is temporary, the poor decisions and mistakes made by motorists that lead to work zone crashes can have lasting impacts. According to the Federal Highway Administration (FHWA), 27,037 people, or 773 per year, died in work zone crashes in the U.S. from 1982 through 2017 (CDC, 2020). In Wisconsin, more than 2600 work zone crashes took place every year over the past five years, resulting in 5200 injuries and 50 deaths (WisDOT, 2020). Work zone safety for both motorists and workers is an urgent issue that must be addressed through better design, operations and management.

Observational safety analysis has been instrumental in identifying potential deficiencies in work zone design and traffic operations. Examples of safety analyses based on crash data include: crash rate estimation across different work zone configurations (Cheng et al., 2012; Daniel et al., 2000; Elias and Herbsman, 2000; Khattak et al., 2002); crash pattern identification and categorization(Garber and Zhao, 2002; Graham et al., 1978; Weng et al., 2016); work zone crash prediction (Li and Bai, 2009b; Meng et al., 2010); and evaluating the safety of innovative work zone designs and management strategies (Li and Bai, 2009a; Maze et al., 2005; Rahman et al., 2017; Ullman et al., 2008). All of the aforementioned examples are dependent on the completeness and accuracy of work zone crash data. Although narrative contains valuable information about a specific crash, (such as a work zone crash), the crash in the structured data of the narrative may not have been coded or recorded as that specific crash type.

Work zones near traffic, whether they involve major road construction, utility work, or emergency vehicles at the side of the road, always present some risk to both drivers and workers. Identifying and analyzing historical work zone crashes can save lives; however, work zone crashes are missed in the structured data for a variety of reasons: restrictive reporting options in tabular forms (Blackman et al., 2020; Ullman and Scriba, 2004; Wang et al., 1996); lack of understanding about the importance of work zone crashes, overloaded by work during crash reporting time (Graham and Migletz, 1983); and misclassification of work zone areas and/or work zone activities (Wang et al., 1996, Farmer,

* Corresponding author.
E-mail addresses: sayed@uwm.edu (M.A. Sayed), qinx@uwm.edu (X. Qin), katerj@uwm.edu (R.J. Kate), anisuzz2@uwm.edu (D.M. Anisuzzaman), yuz@uwm.edu (Z. Yu).

https://doi.org/10.1016/j.aap.2021.106211
Received 10 August 2020; Received in revised form 25 February 2021; Accepted 24 May 2021
Available online 11 June 2021
0001-4575/© 2021 Elsevier Ltd. All rights reserved.

2003).

Generally, a police officer makes certain judgments about a crash based on the severity of the crash and the driver. A fatal crash is usually given the highest reporting priority, compared with property damage crashes which usually receive a lower priority (Ye and Lord, 2011). Work zone crashes that happened at or near work zones with less severe or no injuries are not reported in structured data (Wang et al., 1996). In addition, the probability of reporting an injury crash increases with the age of the driver (i.e., for young children, it is 20–30 %; and for persons over 60, it is 70 %); and the number of vehicles involved. (Hauer and Hakkert, 1988). A crash involving a younger or female driver has a lower probability of being reported (Amoros et al., 2006). Estimates based solely on structured data reported by the police greatly underestimate the results of the safety analysis (Abay, 2015). Safety analysts often resort to a manual review of the crash narrative in order to recover the missed work zone crashes. The narrative is the detailed description of a crash by law enforcement officers. While crash reports contain valuable information, manually reviewing them is time-consuming and labor intensive.

Crash narratives include a significant amount of information that is presented in an unstructured text format. For example, a work zone narrative may include any combination of the words "orange barrels", "orange signs", "flags", "flagging operations", "workers", or "flashing lights". Text mining techniques use automatic information extraction and text classification, providing a predictable, consistent and efficient way of reviewing thousands of crash reports in a matter of minutes. The goal of this study is to develop a text mining-based work zone classifier that uses crash narratives in order to quickly recover missed work zone crashes and to develop a better understanding of the circumstances under which work zone crashes are more likely to be missed. The information gained from this study will be helpful in providing recommendations on how to better collect and analyze work zone data.

2. Literature review

Text mining was introduced as a way to enable machine-supported analysis of text (Feldman and Dagan, 1995). Information retrieval, natural language processing, information extraction, text summarization, opinion mining and sentiment analysis are some of important areas of text mining research (Allahyari et al., 2017). Text mining has become both popular and necessary in many fields, including financial services, health care, transportation, communication and media, information technology and internet, political analysis, public administration and legal services (Gupta and Lehal, 2009; Inzalkar and Sharma, 2015; Maheswari and Sathiaseelan, 2017).

Most text mining algorithms require some text preprocessing, such as tokenization, filtering, lemmatization, stemming, etc. Once preprocessing has been completed, algorithms for classification, clustering, or information extraction are applied to the text. Some commonly used clustering algorithms are hierarchical clustering, k-means clustering, and probabilistic clustering and topic models (e.g., probabilistic latent semantic analysis, latent Dirichlet allocation) (Allahyari et al., 2017). Examples of popular classification algorithms include naive Bayes, nearest neighbor, decision tree, decision rule, support vector machine, logistic regression, Rocchio's algorithm, neural network, associative classifier, and centroid based classifier (Allahyari et al., 2017; Brindha et al., 2016; Korde and Mahender, 2012).

In highway safety analysis, most of the text mining-based studies are conducted using social media and medical data, while a few studies are conducted using crash narratives. Text mining techniques used to identify a specific type of crashes are primarily based on keywords, or words that are direct or indirect indicators of certain unique and specific crash characteristics. Sorock et al. applied Haddon's injury epidemiology model of crash phases to identify pre-crash vehicle activities and various work zone crashes from automobile insurance claim narratives. In a pilot study, the authors manually selected a set of work zone-related words and showed that the keyword "construction" had maximum frequency in the dataset (Sorock et al., 1996). Williamson et al. extracted patterns of events of fatal injuries from crash narratives based on a pre-established text search mechanism (Williamson et al., 2001). Zheng et al. identified secondary crashes by using the keywords' distance, which was calculated by the absolute difference of indexes between two types of keywords: relationships keywords (RKWs) and events keywords (EKWs) (Zheng et al., 2015).

Rakotonirainy et al. used a keyword selection approach that automatically selects keywords in the narratives. The authors used text mining to identify curve-related crash factors and their associated severity from insurance claim reports. The words mentioned only in curve-related crashes were selected as keywords, and the keywords with high frequencies were used as the main factors contributing to curve-related crashes (Rakotonirainy et al., 2015). Gao and Wu developed a verb-based text mining method by applying various Natural Language Processing (NLP) techniques that automatically identify the sequence of crash events from crash narratives (Gao and Wu, 2013). Their method utilized syntactic and semantic information from the text to overcome the limitations of previous methods that used predefined keywords. However, the process was not completely automatic, as the words with similar meaning had to be grouped together manually. Trueblood et al. developed a classifier tool in Excel to identify agricultural crash from crash narratives. The authors prepared two lists of keywords (agricultural and nonagricultural) manually and used the lists to search keywords in the narratives for identifying the agricultural crashes (Trueblood et al., 2019). However, their classifier assigns equal weight to the narratives that are related to agricultural crash, so it may not be effective for large data sets in which narratives are more relevant to agricultural crash.

Existing research also uses structured data reported by the police and data from other sources to recover missed crashes. For example, Watson, Watson, & Vallmuur used police reported structured data with Hospital Admitted Patients (HAP), Emergency Department Information System (EDIS) and Injury Surveillance Unit(ISU) data using 'separation principle' approach to recover missed crashes (Watson et al., 2015). Salifu & Ackaah conducted survey at medical and among drievers and link those data to police report to find missed crashes (Salifu and Ackaah, 2012). Cheung & Braver used vehicle identification number(VIN) to find missed fatalities in single-unit truck crashes from Trucks Involved in Fatal Accidents (TIFA) and Fatality Analysis Reporting System (FARS) data (Cheung and Braver, 2016). The TIFA data is prepared from police reports and the interview of truck owners. Thomas, Thygerson, Merrill, & Cook used hospital and survey data to detect missed crashes that were not found in the structured data of the police report (Thomas et al., 2012).

While past research has focused on analyzing various aspects of traffic crashes from crash narratives, none of the studies emphasized missed work zone crashes. Their methods are either complicated, time-consuming, external data dependent or require substantial manual intervention, which does not meet our research goals. Moreover, the data from other sources are not easily accessible to the public and can be costly. This study develops a work zone crash classifier that is simple, for example, there is no need to manually prepare any keyword lists, and no hyperparameters to fine tune. It is also computationally efficient and easy to implement.

3. Data collection

The dataset comprised 377,479 crash reports, including crash narratives, that occurred between January 1, 2017 and October 31, 2019 that were acquired from the Wisconsin Department of Transportation (WisDOT) through the WisTransPortal data hub. A construction zone flag (CONSZONE) within the crash data indicates whether "*a crash occurred in a construction, maintenance, or utility work zone or is related to activity within a work zone*". The reported work-zone (WZ) crashes make

up 2.27 %, 2.49 %, and 1.93 % of total crashes for years 2017, 2018 and 2019, respectively. Narratives were included in 94.21 % of the reported WZ crashes and 77 % of the non-work zone (NWZ) crashes. The ratio of WZ to NWZ crashes is 1:36, which is a highly imbalanced dataset. The two following sample crash narratives were randomly chosen from the dataset to illustrate the structure of crash narratives.

WZ crash narrative example: "Entering construction zone with right lane closure. Unit 1 driver stated unit 2 and a semi were straddling center line. Unit 1 driver stated thought unit two was merging to right lane toward hwy c exit and tried to pass unit 2. Unit 1 driver stated himself and semi were straddling traffic lane to stop other drivers from passing on right as right lane was closed ahead. Unit 2 stated unit 1 attempted to pass on left shoulder but ran out of room due to portable warning sign. Unit 2 driver stated unit 1 driver side swiped driver side."

NWZ crash narrative example: "Unit #2 was stopped in the inside straight lane of eastbound university ave., at a red light at the intersection with n. Midvale blvd. Unit #1 was traveling in the same lane directly behind unit #2, and was unable to stop in time to avoid a rear end collision with unit #2. The roadway was wet, and the weather conditions were rainy."

The numeric values within the narratives usually represent date, time, driver and road information. The narratives have a certain formality but can still be flexible in the sequence of events. In the WZ narrative, some sentences contain words that indicate WZ (e.g., "construction zone", "right lane closure", "portable warning sign"), while others do not contain any WZ indicators. In fact, the latter cannot be distinguished from sentences that could have been in a NWZ narrative. This observation is true of other WZ narratives as well; only a few words are indicative of a WZ while the rest of the narrative is not, suggesting that presence of just a few words can be used to identify a WZ narrative without having a deep understanding of the entire narrative. Additionally, there are no such words in the narrative that specifically indicate NWZ.

4. Methodology

This section describes the principles and procedures used in the method for identifying missed WZ crashes. The study uses a probabilistic approach in which word probabilities were combined using the noisy-OR method.

4.1. The nature of noisy data

The 2017 and 2018 work zone crash data were used to train a classifier (described later) to categorize a narrative as either WZ or NWZ and the NWZ narratives of 2019 (Data was available till October 31, 2019) were used as testing data to recover missed WZ crashes. The narratives corresponding to reported WZ crashes (i.e., marked under CONSZONE flag) were used as examples of WZ narratives to train the classifier. Similarly, the narratives corresponding to reported NWZ crashes (i.e., not marked under CONSZONE flag) were used as examples of NWZ narratives. The method did not require the manual annotation of training examples, a task that usually requires the huge effort of training a classifier.

However, the training dataset created does include a high level of noise. On one hand, many narratives of reported WZ crashes may not contain any relevant information about the WZ. For example, the officer may have already indicated a crash as WZ by using the CONSZONE flag, hence not feeling the need to mention it in the narrative. However, WZ crashes are known to be missed, and there are narratives corresponding to reported NWZ crashes that are actually WZ. The classifier may have difficulty learning from such noisy training data.

4.2. Data cleaning and pre-processing

Several text mining techniques for data cleaning and pre-processing were applied to prepare the data. The key terminologies from the text mining domain are introduced here:

- Corpus is the collection of all of the narratives.
- Tokenization is the process of breaking up the sentence into a token. A token can be words, numbers, unigram, or bigram. The terms unigram and bigram are used interchangeably as the token in this study.
- Collection frequency (cf) is the number of times a token occurred in the corpus.
- Term frequency (tf) is the number of times a token occurred in a narrative.
- Document frequency (df) is the number of documents/narratives that contain a token. Only the tokens with high df values in WZ narratives will have a high impact on the model.

In the training dataset, the narratives were first lower-cased to merge the occurrences of the same word in different cases. Then, all punctuations and special characters (e.g., ! " # $ % & ' () * +, -. / : ; < = >? @ [\] ^ _ ` {|} ~) were removed from the narratives. Next, the narratives were converted into tokens to build a vocabulary list from the training set. The narratives may include spelling errors and/or words in multiple forms, such as "zone" and "zones" or "construction" and "construct", which are common issues when mining unstructured text data. While some text mining techniques can handle these issues, there is no guarantee the problem will be solved completely. Furthermore, improper processing of these words may lead to new problems. Thus, the words in the vocabulary list were kept as-is.

4.3. Computing unigram/bigram probabilities

Unigrams, or single words (e.g., "flagman", "barrel"), as well as *bigrams*, or consecutive words (e.g., "orange barrel", "construction zone"), can be highly indicative of WZ. Hence, the method used in this study uses both unigrams and bigrams. The probability that a unigram or bigram in the corpus indicates WZ is computed using simple frequency counts (see Eq. 1).

$$\textit{Probability Score}\ (w) = \frac{\text{Positive Count(w)} + 1}{\text{Positive Count (w)} + \text{Negative Count(w)} + 2} \quad (1)$$

where ***w*** is a unigram or a bigram, a positive count indicates the number of occurrences of a unigram or bigram in the WZ narratives. A negative count indicates the number of occurrences of a unigram or bigram in the NWZ narratives. A simple version of Laplace smoothing adds one in the numerator and two in the denominator of the equation, which assumes each unigram or bigram occurred once in the WZ narrative and once in the NWZ narrative. Laplace smoothing ensures that among the unigrams and bigrams that have zero negative counts, those with higher positive counts receive higher probability scores. Without smoothing, the unigrams and bigrams would receive an unrealistic probability score of 1, for instance, just because they occurred in a few WZ narratives and no NWZ narrative.

Probability scores for the unigrams and bigrams in the training data were estimated. The unigrams and bigrams that have a less than 0.25 probability score were discarded because they are unlikely to impact the classification decision. This truncation threshold was set to low due to imbalanced data (i.e., the number of NWZ narratives are much higher than the number of WZ narratives). The unigrams and bigrams that appeared fewer than four times in the training narratives also were discarded. The remaining are called positive unigrams and positive bigrams in this study. Consequently, there are two lists of positive tokens: Positive Unigrams contain all of the positive unigrams, and Positive Bigrams contain all of the positive bigrams.

4.4. Combining unigram/bigram probabilities

A narrative's probability of being WZ is computed by combining the probability scores of the unigrams and bigrams in the narrative. The noisy-OR method was applied to combine probabilities ((Zagorecki and Druzdzel, 2004), a method commonly used in Bayesian networks (Oniśko et al., 2001; Vomlel, 2006). Noisy-OR is a probabilistic extension of logical "or". In logical "or", if any one of the inputs is equal to true, the output is equal to true. The output is equal to false only when all inputs are equal to false. In noisy-OR, the inputs and outputs are probabilities instead of true/false values, hence the term "noisy".

Analogous to logical "or", if any one of the input probabilities in noisy-OR is high (i.e. close to 1), then the combined probability is high. But unlike logical "or", the combined probability in noisy-OR is even higher if more input probabilities are high. The combined probability is low (i.e. close to 0) only when all input probaibilities are low. The noisy-OR combined probability is mathematically computed as shown in Eq. 2. In the equation, the probability score of a narrative is computed by combining the probability scores of the unigrams and bigrams occurring in it.

$$Probability\ Score\ (N) = 1 - \prod_{i,j=1}^{n} (1 - P_i)^j \quad (2)$$

where N is a given narrative, P_i indicates the probability of i^{th} unigram or bigram as computed from the training data, and j equals the number of occurrences of that i^{th} unigram or bigram in the crash narrative N.

It should be clear from Eq. 2 that if neither a unigram nor a bigram in a narrative has a high probability score, the probability score of the narrative will be close to zero. On the other hand, a single unigram or bigram with a high probability score will result in a high probability score of the entire narrative. Furthermore, presence of more unigrams and bigrams with high probability scores will only make the combined probability score higher. Another advantage of using the noisy-OR method is that it is resistant to the type of noise in the training data that was mentioned earlier; this is because the unigrams and bigrams with high probability scores will not necessarily be the ones that occur in high percentages of WZ narratives (which we know are noisy), but they will be the ones that occur more often in WZ narratives than in NWZ narratives.

4.5. Crash verification and model performance evaluation

The main objective of this study is to find missed WZ crashes from the crash narratives. The performance of the proposed method on the test dataset was manually reviewed. Since the test data are unlabeled, it is not possible to manually check all possible WZ crashes from the huge test data (over 80,000 cases). Initially, it was anticipated that a threshold could be set up for the classification score to separate the cases into WZ and NWZ. However, the classification scores show that many cases have very small differences. Hence, the performance of the model was evaluated by sorting the results in descending order so that the most probable scenarios are at the top. The authors reviewed and manually classified (NWZ or WZ) the top 100 narratives with the highest probability scores using only unigrams, as well as the top 450 narratives with the highest probability scores using both unigrams and bigrams. Each reviewer was assigned an equal number of samples to eliminate any reviewer bias. This study avoids using any external data (such as work zone inventory data) for evaluating the model performance.

5. Results and analysis

The results of unigram and unigram + bigram are compared and discussed in this section. The characteristics of missed crashes are analyzed from spatial and temporal perspectives, along with other features. The additional analysis is expected to provide insight on the circumstances in which crashes are not reported as WZ related so that recommendations can be made for improving future data collection.

5.1. The analysis of positive unigram and positive bigram

The 2017–2018 crash data were cleaned and preprocessed, showing 10,875 unigram and 96,550 bigram words (tokens) in the corpus. Table 1 presents the top ten positive unigrams and bigrams and their corresponding probability scores. As shown in Table 1, the bigram approach extracted more WZ-related information than the unigram approach. However, despite high probability scores, some positive unigrams did not carry meaningful information such as "Kampo", "Kucej", or "Werych". While "Kampo", "Kucej", and "Werych" may appear only in WZ cases, at a very low frequency, meaning including them in the Positive Unigram list may degrade the model's performance. For example, if a narrative has many such unigrams, the noisy-OR may tend to classify it as a WZ crash even if it's not.

The document frequency (df) and collection frequency (cf) of the training set were calculated to examine how the positive unigrams and bigrams with high probability scores influence the proposed method. The classifier performance did not degrade much due to lower document frequency(df) of the less meaningful positive unigrams and bigrams. Thus, an important positive token should have both high df and cf values and with high probability score.

Table 2 populates a list of the top 15 important positive unigrams and bigrams ranked by df, cf and probability score (p_r) in a decreasing order. In the positive unigram list, the token "construction" is the most important because it has the highest df and cf values. The most important token in the positive bigram list is "construction zone". Approximately 35.15 % of the WZ crash narratives contain the token "construction", whereas 16.16 % of the WZ crash narratives contain "construction zone". Table 2 shows that the Positive Bigram list offers more specific WZ crash information and higher probability scores than the unigram list.

The unigram method will classify a narrative as a WZ crash if it contains the token "construction" ($p_r = 0.89$) from the positive unigram list only because the threshold value is greater than or equal to 0.89. The df of "construction" is much higher compared to other unigrams in the list, so the misclassification rate by the unigram method will be higher. Compared with the positive unigram "construction", the positive bigram "construction zone" ($p_r = 0.90$) is more contextual and has a higher df than the remaining bigrams in the list. A narrative with the presence of "construction zone" instead of "construction" is more likely to be correctly classified as a WZ crash. The manual review result shows that all of the NWZ narratives that contain "construction zone" are true WZ crashes. However, 22 NWZ narratives that contain "construction" are not WZ crashes.

Positive tokens such as "fst", "kampo" and "kicmol" in the Positive Unigram list do not carry any meaningful information. These unigrams have a small df with high probability scores, meaning they should be discarded to reduce the misclassification rate. The positive token lists

Table 1
Top Ten Positive Unigrams and Bigrams by Probability Score.

Rank	Positive Unigram		Positive Bigram	
	Positive Words	Probability	Positive Words	Probability
1	flagman	0.960	active construction	0.990
2	taper	0.947	in construction	0.988
3	barreled	0.937	temporary cement	0.983
4	dividers	0.929	zone where	0.972
5	roadworks	0.923	construction crew	0.971
6	kampo	0.917	zone lane	0.964
7	unfinished	0.917	interstate is	0.960
8	flaggers	0.917	no workers	0.960
9	kucej	0.909	flag person	0.957
10	werych	0.900	workers present	0.956

Table 2
Top 15 Positive Unigrams and Positive Bigrams By df, cf, and Probability Score*.

Rank	Positive Unigram				Positive Bigram			
	Token	cf	df	Pr	Token	cf	df	Pr
1	construction	2960	2088	0.89	construction zone	966	826	0.9
2	zone	1181	972	0.45	the construction	763	625	0.82
3	closed	743	588	0.44	a construction	484	437	0.77
4	barrels	407	314	0.69	to construction	320	312	0.73
5	closure	265	191	0.61	was closed	242	228	0.51
6	orange	192	152	0.34	construction barrels	195	167	0.77
7	barrel	228	147	0.56	lane closed	212	161	0.67
8	temporary	170	126	0.37	construction unit	158	156	0.78
9	zoo	219	123	0.56	under construction	151	149	0.79
10	cones	166	122	0.45	construction area	158	136	0.82
11	workers	120	110	0.52	road construction	145	135	0.68
12	barriers	119	97	0.49	work zone	161	132	0.92
13	barricades	107	78	0.42	the zoo	206	120	0.67
14	attenuator	145	74	0.47	construction and	123	120	0.7
15	worker	95	67	0.51	zoo interchange	181	114	0.67

* cf = collection frequency in WZ narratives, df = document frequency in WZ narratives, p_r = probability.

also contain names of locations such as "zoo" in unigram and "the zoo[1] " in bigram. The presence of those tokens can cause the noisy-OR method to misclassify NWZ crashes as WZ crashes.

5.2. Comparing unigram and unigram + bigram methods

The preceding section explains that in using the noisy-OR method, the unigram method may not be effective as expected. Positive unigrams with high cf values may have low probability values because the same unigrams also appear in the NWZ crash narratives. The problem can be mitigated by adding some context to the noisy-OR approach, such as in the form of bigrams. The ordered positive bigram list provides more contextual information related to WZ. This section provides empirical evidence of using the noisy-OR method as a text classifier to identify missed WZ crashes from narratives. The section also explores the classification outcomes of unigram and unigram + bigram when compared with gold label, or manual reviewing.

The 100 narratives with the highest probability scores in each classifier were manually reviewed. The top 100 narratives of the unigram noisy-OR classifier included 65 actual WZ crashes, while the top 100 narratives of the unigram + bigram noisy-OR classifier included 78 actual WZ crashes. The unigram + bigram noisy-OR narratives that were correctly classified contained more contextual positive bigrams such as "construction zone", "under construction", "construction worker" and "lane closed" with high df values in the WZ training set.

A close review of 35 unigram noisy-OR cases that were misclassified shows that they contain WZ-related positive unigrams such as "construction", "barrels", "attenuator", "barricades", "orange" and some noisy words such as "carrao", "kampo", "melloch". These noisy unigrams have high df values in the WZ training set, indicating their popularity in the WZ crash narratives. On the contrary, the unigram + bigram noisy-OR misclassified 22 cases from its top 100 narratives. A close review of these 22 cases reveals that the unigram portion of unigram + bigram noisy-OR contains few positive unigrams but with high probability scores; the bigram portion contains a longer list of positive bigrams with moderate probability values. Thus, the comparison reaffirms that unigram + bigram noisy-OR tackled the noisy tokens more successfully than unigram noisy-OR.

[1] Zoo interchange construction is the most complex and expensive highway project in Wisconsin's history, which began in 2014 with an expected completion date of 2022.

5.3. Classification accuracy rate of unigram + bigram results

Further analysis was performed to quantify the classification accuracy rate against the case rank of the unigram + bigram method. Starting from the highest-ranked cases, the number of correctly-identified WZ crashes is counted over the 50-case intervals, as shown in Fig. 1.

Based on the 450 cases reviewed, two observations can be made from Fig. 1: a) more than 50 % of cases correctly classified till the fifth interval (201–250), and b) the model performance degrades rapidly from 80 % in the first interval [0–50] to 12 % in the last interval [401–450]. The fitted quadratic equation has a R^2 value of 0.9668, suggesting a strong and consistent trend for the descending accuracy rate. The findings are good news for an agency who wants to estimate the effort of a manual review for missed WZ crashes, as the manual effort seems manageable and quantifiable.

The probabilistic distribution of narrative length was plotted for WZ and NWZ crashes, respectively, in Fig. 2. The distribution was inspired by a study that shows that narratives not designated by officers as speed-related crashes have a longer length on average than non-speed related crashes (Fitzpatrick et al., 2017). Fig. 2 shows that the narrative length of actual NWZ crashes is approximately normally distributed, while missed WZ crashes are slightly skewed toward the left. The two distributions are statistically different at a 5% level of significance (two sample t-test, p=<0.0001).

Moreover, the average narrative length of reported WZ crashes is 104, and Std. is 68 (sample size:1989), which is a statistically significant difference between NWZ (two sample t-test, p=<0.001) and missed WZ (two sample t-test, p=<0.001). Though it is expected that long narratives would have more positive tokens than short narratives, no correlations are observed between the length of narratives and the number of positive tokens for reported WZ and NWZ and missed WZ. In other words, there is not enough evidence to claim that long narratives tend to classify crashes more accurately than short narratives.

5.4. Analysis of missed WZ crashes

Further analysis was conducted on the crash time and location for a better understanding of the circumstances under which a WZ crash is missed. Fig. 3 illustrates the distribution of police reported WZ crashes and missed WZ confirmed in this study by time of day, day of week, and month of year.

In 2017–2019, 70.96 % of all reported WZ crashes and in 2019, 73.13 % of the missed WZ crashes identified in this study occurred during daylight hours from 8 a.m. to 6 p.m., as shown in Fig. 3(a). Among daytime WZ crashes, a high percentage of missed cases occurred

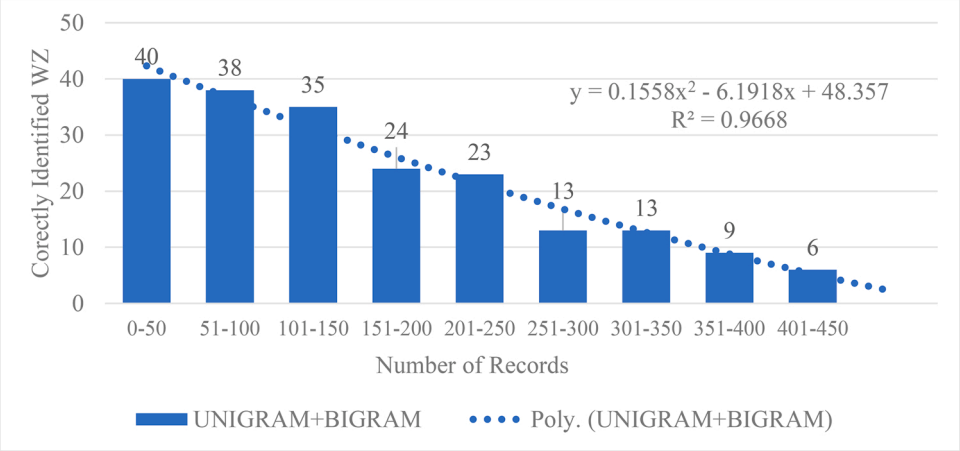

Fig. 1. Accuracy of (Unigram + Bigram) Noisy-OR.

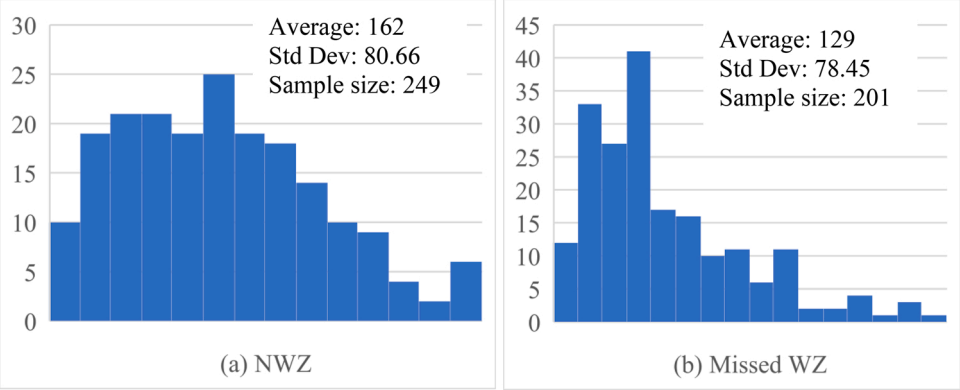

Fig. 2. Histogram of narrative length for a) NWZ and b) Missed WZ.

in the midmorning (10am), early afternoon (1pm) and late afternoon when traffic is busiest, from 4 p.m. to 5 p.m. It is plausible that crashes are missed when traffic is high or when construction activities are intense. The day of week distribution suggests that the WZ crashes are probably missed throughout the week, especially on Monday, Wednesday and Saturday, as shown in Fig. 3(b). Fig. 3(c) also displays the monthly distribution of reported WZ crashes versus missed WZ crashes, showing that a high percentage of missed cases are observed in the summertime, especially in July and August when construction activities are extensive and intensive.

Fig. 4 shows the distribution of missed WZ crashes compared to reported WZ crashes by highway class. The evidence shows that most missed WZ crashes occurred in urban areas, including urban city streets (43.11 %), urban state highways (16.89 %) and urban interstate highways (15.33 %). The rural interstate highway system has the best performance in terms of a low ratio of missed crashes to reported crashes. The next best performance is from state highways, where the ratio is close to 1. Urban city streets have the highest ratio of missed crashes to reported crashes, particularly urban city streets which have only 20 % of the total reported WZ crashes but make up 43 % of missed WZ crashes identified in this study. Cheng et al. stated that construction work zones are usually assumed to be long term works, but maintenance or utility works are usually short term and temporarily, which may not be known to driver in advance (Cheng et al., 2012). Since many crashes on urban streets involve utility work zones, it is plausible that police may not consider those as construction zone related.

Comparisons were conducted for other structured data fields, including weather conditions, pavement conditions, light conditions, and injury severity. The results show similar distributions between all reported WZ crashes and missed WZ crashes, mainly due to the lack of variety since most WZ crashes, reported or missed, occur during clear or cloudy weather, on dry pavement, in the daytime, and involve less severe injuries.

In summary, the unigram + bigram noisy-OR method is an effective and efficient method for classifying and recovering missed WZ crashes from narratives. According to Fig. 1, a review of the top 450 cases of the unigram + bigram noisy-OR identified 201 WZ crashes as missed, which is more than 8% of reported WZ crashes from 01/01/2019 to 10/31/2019. Moreover, the decreasing trend of finding missed WZ crashes suggests the chance may be 12 % or lower after the first 450. Additionally, 450 crashes is a tiny fraction of the pool of potentially missed WZ crashes (i.e., 125,509 NWZ crashes in 2019), which is very helpful to an agency that wants to prioritize and estimate the level of effort of a manual review.

An analysis of missed cases suggests the 73.13 % of the missed WZ crashes identified in the study occurred from 8 a.m. to 6 p.m. with a high percentage in the afternoon from 4 p.m. to 5 p.m. A high percentage of WZ crashes that are misclassified are observed in July and August when the construction activities are extensive and intensive. 43 % of the missed WZ crashes identified in this study occurred on urban city streets.

6. Conclusion and lessons learned

In this study, a keyword-based text classifier was developed using the noisy-OR combined probability to identify misclassified WZ crashes from the crash narratives of police reports. Specifically, the unigram + bigram noisy-OR classifier was created and proven to be an effective means to recover WZ crashes from those police officers did not flag as

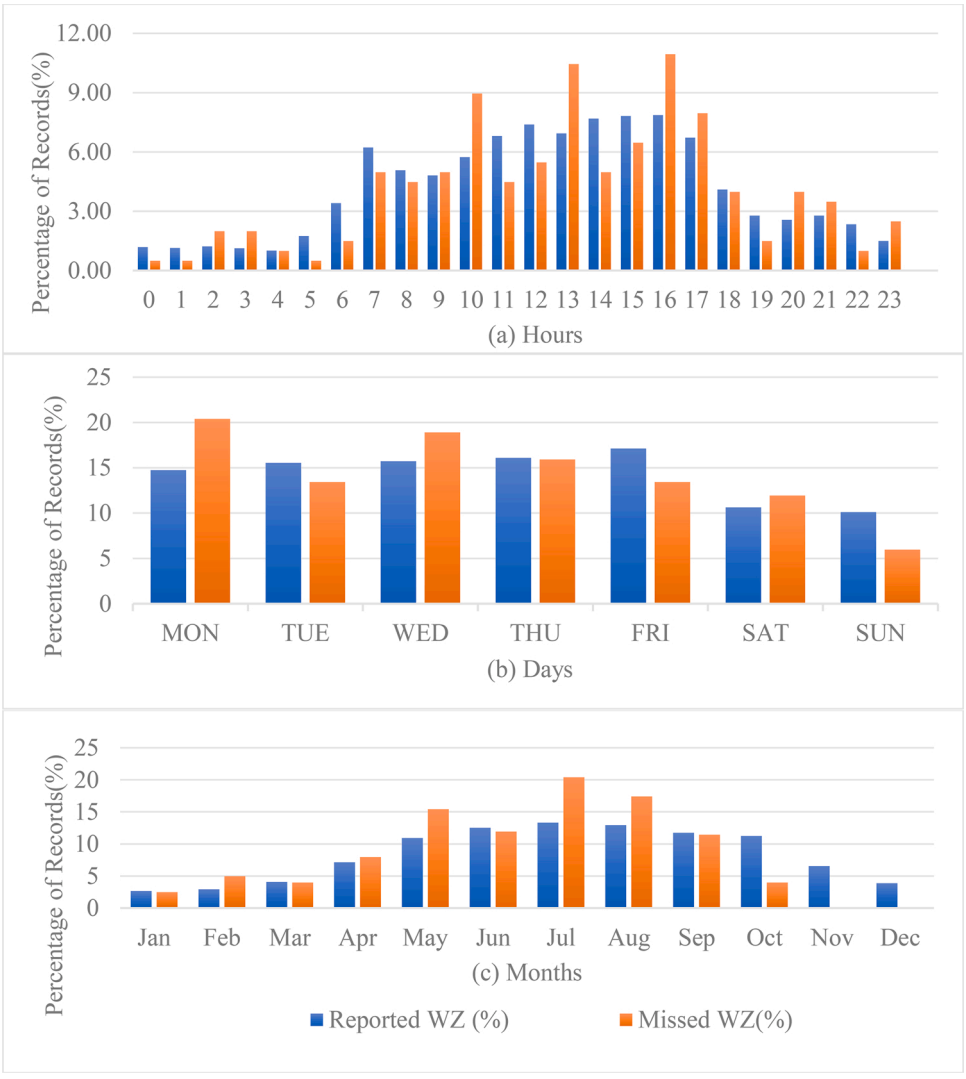

Fig. 3. WZ crash analysis by a) hour, b) day and c) month.

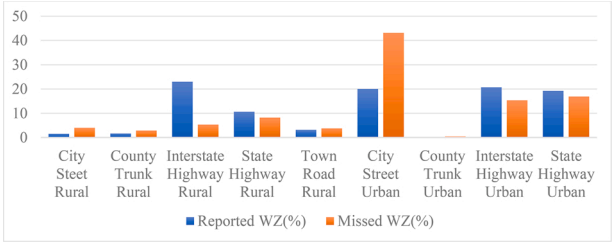

Fig. 4. WZ crash analysis by highway class.

construction zone crashes. The narrative in a flagged WZ crash may not contain any relevant information linking it to a work zone, and narratives from unflagged WZ crashes may contain information related to a work zone. The noisy-OR method was used because of its ability to work effectively despite the high level of noise in the unstructured text or crash narratives. Moreover, noisy-OR does not require much training time, is computationally efficient and is easier to implement.

The authors used 377,479 crash reports from January 1, 2017 through October 31, 2019. The 2017 and 2018 WZ and NWZ crashes were treated as training data, and the 2019 NWZ crashes were used as the testing data. A manual review of the top 450 cases classified as WZ crashes in the testing data recovered 201 missed WZ crashes, which was 0.24 % of the testing data. The review also indicated that beyond 450 cases, the chance of having missed WZ crashes may be very low. A follow-up analysis revealed that 73.13 % of the missed crashes occurred from 8 a.m. to 6 p.m., with a high percentage happening from 4 p.m. to 5 p.m. A large percentage of those crashes occurred in the summer (July and August) and 43 % occurred on urban city streets.

The narratives of the cases that have high noisy-OR scores but are not WZ crashes were carefully reviewed and categorized into the five following groups:

1) Cases with positive words for location or address such as “the Zoo”, “Zoo interchange”: This issue is caused primarily by major roadway construction projects that span over multiple years, multiple stages and phases and multiple areas.
2) Cases with positive words for (temporal) traffic control devices such as “concreate barrier”, “median cement”, “attenuator” and “barriers”: Many of these devices, such as median concrete barriers, are permanently deployed to channelize traffic or to protect overpass and underpass structures such as an attenuator near a bridge or at a gore area.
3) Cases with weak positive words for traffic situations such as “congestion” or “backup” which are caused by non-WZ events (i.e., regular congestion or secondary crashes).

4) Cases with strong positive words such as “orange construction” or even “construction zone” whose situations are actually not related to a work zone location or work zone activities.
5) Undecided cases, even after a manual review: The authors were conservative and categorized undecided crashes from this study as NWZ crashes.

A location and/or time that a work zone crash occurred can certainly improve WZ classification in types 1 and 5. Such information, however, has to be linked to and retrieved from a different data source or system such as a lane closure system or a work zone management system. Application of advance text mining techniques may help improve classification accuracy for cases in types 2 and 3. Unfortunately, no good solutions are available for cases in type 4, but such cases rarely occur. Nevertheless, the discussion underscores the importance of properly documenting the presence of a work zone or work zone activities in the crash narrative.

Author statement

Md Abu Sayed: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Roles/Writing – original draft **Xiao Qin:** Conceptualization, Data collection, Investigation, Writing – review & editing, Supervision **Rohit J Kate:** Methodology, Validation, Writing – review & editing **D M Anisuzzaman:** Methodology **Zeyun Yu** Conceptualization.

All the authors of the manuscript titled “Identification and Analysis of Misclassified Work-Zone Crashes Using Text Mining Techniques”, AAP_2020_1466, declare that there is no conflict of interest.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

This study was supported by a grant from the Wisconsin Department of Transportation (WisDOT) Bureau of Transportation Safety (FG-2020-UW-MILWA-05069). Thanks to the University of Wisconsin-Madison Traffic Operations and Safety (TOPS) Laboratory for making crash narrative data available for this research.

References

Abay, K.A., 2015. Investigating the nature and impact of reporting bias in road crash data. Transp. Res. Part A Policy Pract. 71, 31–45. https://doi.org/10.1016/j.tra.2014.11.002.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K., 2017. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. ArXiv Preprint ArXiv:1707.02919.

Amoros, E., Martin, J.L., Laumon, B., 2006. Under-reporting of road crash casualties in France. Accid. Anal. Prev. 38 (4), 627–635. https://doi.org/10.1016/j.aap.2005.11.006.

Blackman, R., Debnath, A.K., Haworth, N., 2020. Understanding vehicle crashes in work zones: analysis of workplace health and safety data as an alternative to police-reported crash data in Queensland, Australia. Traffic Inj. Prev. 21 (3), 222–227. https://doi.org/10.1080/15389588.2020.1734190.

Brindha, S., Prabha, K., Sukumaran, S., 2016. A survey on classification techniques for text mining. ICACCS 2016 - 3rd International Conference on Advanced Computing and Communication Systems: Bringing to the Table, Futuristic Technologies from Arround the Globe, *01*(I) 1–5. https://doi.org/10.1109/ICACCS.2016.7586371.

Cheng, Y., Parker, S., Ran, B., Noyce, D., 2012. Enhanced analysis of work zone safety through integration of statewide crash and lane closure system data. Transp. Res. Rec. 2291, 17–25. https://doi.org/10.3141/2291-03.

Cheung, I., Braver, E.R., 2016. Undercounting of large trucks in federal and state crash databases: extent of problem and how to improve accuracy of truck classifications. Traffic Inj. Prev. 17 (2), 202–208. https://doi.org/10.1080/15389588.2015.1034273.

Daniel, J., Dixon, K., Jared, D., 2000. Analysis of fatal crashes in Georgia work zone. Transp. Res. Rec. 1715, 18–23. https://doi.org/10.3141/1715-03.

Elias, A.M., Herbsman, Z.J., 2000. Risk analysis techniques for safety evaluation of highway work zones. Transp. Res. Rec. 1715, 10–17. https://doi.org/10.3141/1715-02.

Farmer, C.M., 2003. Reliability of police-reported information for determining crash and injury severity. Traffic Inj. Prev. 4 (1), 38–44. https://doi.org/10.1080/15389580309855.

Feldman, R., Dagan, I., 1995. Knowledge Discovery in textual databases (KDT). International Conference on Knowledge Discovery and Data Mining (KDD) 112–117 https://doi.org/10.1.1.47.7462.

Fitzpatrick, C.D., Rakasi, S., Knodler, M.A., 2017. An investigation of the speeding-related crash designation through crash narrative reviews sampled via logistic regression. Accid. Anal. Prev. 98, 57–63. https://doi.org/10.1016/j.aap.2016.09.017.

Gao, L., Wu, H., 2013. Verb-based text mining of road crash report. Transportation Research Board, 92nd Annual Meeting 5–16. Retrieved from. http://trid.trb.org/view/2013/C/1241434.

Garber, N.J., Zhao, M., 2002. Distribution and characteristics of crashes at different work zone locations in Virginia. Transp. Res. Rec. 1794, 19–28. https://doi.org/10.3141/1794-03.

Graham, J.L., Migletz, J., 1983. Collection of work-zone accident data. Transp. Res. Rec. 15–18.

Graham, J.L., Paulsen, R.J., Glennon, J.C., 1978. Accident analysis of highway construction zones. Transp. Res. Rec. 693, 25–32.

Gupta, V., Lehal, G.S., et al., 2009. A survey of text mining techniques and applications. J. Emerg. Technol. Web Intell. 1 (1), 60–76.

Hauer, E., Hakkert, A.S., 1988. Extent and some implications of incomplete accident reporting. Transp. Res. Rec. 1185 (January), 1–10.

Inzalkar, S., Sharma, J., 2015. A survey on text mining-techniques and application. Int. J. Res. Sci. Eng. 24, 1–14.

Khattak, A.J., Khattak, A.J., Council, F.M., 2002. Effects of work zone presence on injury and non-injury crashes. Accid. Anal. Prev. 34 (1), 19–29. https://doi.org/10.1016/S0001-4575(00)00099-3.

Korde, V., Mahender, C.N., 2012. Text classification and classifiers: a survey. Int. J. Artif. Intell. Appl. 3 (2), 85.

Li, Y., Bai, Y., 2009a. Effectiveness of temporary traffic control measures in highway work zones. Saf. Sci. 47 (3), 453–458. https://doi.org/10.1016/j.ssci.2008.06.006.

Li, Y., Bai, Y., 2009b. Highway work zone risk factors and their impact on crash severity. J. Transp. Eng. 135 (10), 694–701. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000055.

Maheswari, M.U., Sathiaseelan, D.J.G.R., 2017. Text mining: survey on techniques and applications. Int. J. Sci. Res. 6 (6), 45–56.

Maze, T., Burchett, G., Hochstein, J., 2005. Synthesis of Procedures to Forecast and Monitor Work Zone Safety and Mobility Impacts. Report. Retrieved from. http://www.intrans.iastate.edu/publications/_documents/t2summaries/wz_road_close.pdf.

Meng, Q., Weng, J., Qu, X., 2010. A probabilistic quantitative risk assessment model for the long-term work zone crashes. Accid. Anal. Prev. 42 (6), 1866–1877. https://doi.org/10.1016/j.aap.2010.05.007.

Oniśko, A., Druzdzel, M.J., Wasyluk, H., 2001. Learning Bayesian network parameters from small data sets: application of Noisy-OR gates. Int. J. Approx. Reason. 27 (2), 165–182. https://doi.org/10.1016/S0888-613X(01)00039-1.

Rahman, M.M., Strawderman, L., Garrison, T., Eakin, D., Williams, C.C., 2017. Work zone sign design for increased driver compliance and worker safety. Accid. Anal. Prev. 106 (May), 67–75. https://doi.org/10.1016/j.aap.2017.05.023.

Rakotonirainy, A., Chen, S., Scott-Parker, B., Loke, S.W., Krishnaswamy, S., 2015. A novel approach to assessing road-curve crash severity. J. Transp. Saf. Secur. 7 (4), 358–375. https://doi.org/10.1080/19439962.2014.959585.

Salifu, M., Ackaah, W., 2012. Under-reporting of road traffic crash data in Ghana. Int. J. Inj. Contr. Saf. Promot. 19 (4), 331–339. https://doi.org/10.1080/17457300.2011.628752.

Sorock, G.S., Ranney, T.A., Lehto, M.R., 1996. Motor vehicle crashes in roadway construction workzones: an analysis using narrative text from insurance claims. Accid. Anal. Prev. 28 (1), 131–138. https://doi.org/10.1016/0001-4575(95)00055-0.

Thomas, A.M., Thygerson, S.M., Merrill, R.M., Cook, L.J., 2012. Identifying work-related motor vehicle crashes in multiple databases. Traffic Inj. Prev. 13 (4), 348–354. https://doi.org/10.1080/15389588.2012.658480.

Trueblood, A.B., Pant, A., Kim, J., Kum, H.C., Perez, M., Das, S., Shipp, E.M., 2019. A semi-automated tool for identifying agricultural roadway crashes in crash narratives. Traffic Inj. Prev. 20 (4), 413–418. https://doi.org/10.1080/15389588.2019.1599873.

Ullman, G.L., Scriba, T.A., 2004. Revisiting the influence of crash report forms on work zone crash data. Transp. Res. Rec. 1897, 180–182. https://doi.org/10.3141/1897-23.

Ullman, G.L., Finley, M.D., Bryden, J.E., Srinivasan, R., Council, F.M., 2008. Traffic safety evaluation of nighttime and daytime work zones. Transportation Research Board, (NCHRP Report 627). Retrieved from. http://www.trb.org/Publications/Blurbs/160500.aspx.

Vomlel, J., 2006. Noisy-or classifier. Int. J. Intell. Syst. 21 (3), 381–398. https://doi.org/10.1002/int.20141.

Wang, J., Hughes, W.E., Council, F.M., Paniati, J.F., 1996. Investigation of highway work zone crashes: what we know and what we don’t know. Transp. Res. Rec. 1529, 54–62. https://doi.org/10.3141/1529-07.

Watson, A., Watson, B., Vallmuur, K., 2015. Estimating under-reporting of road crash injuries to police using multiple linked data collections. Accid. Anal. Prev. 83, 18–25. https://doi.org/10.1016/j.aap.2015.06.011.

Weng, J., Zhu, J.Z., Yan, X., Liu, Z., 2016. Investigation of work zone crash casualty patterns using association rules. Accid. Anal. Prev. 92, 43–52. https://doi.org/10.1016/j.aap.2016.03.017.

Williamson, A., Feyer, A.M., Stout, N., Driscoll, T., Usher, H., 2001. Use of narrative analysis for comparisons of the causes of fatal accidents in three countries: New Zealand, Australia, and the United States. Inj. Prev. 7 (SUPPL. 1) https://doi.org/10.1136/ip.7.suppl_1.i15.

Ye, F., Lord, D., 2011. Investigation of effects of underreporting crash data on three commonly used traffic crash severity models. Transp. Res. Rec. 2241, 51–58. https://doi.org/10.3141/2241-06.

Zagorecki, A., Druzdzel, M., 2004. An empirical study of probability elicitation under noisy-OR assumption. In: Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004, 2, pp. 880–885.

Zheng, D., Chitturi, M.V., Bill, A.R., Noyce, D.A., 2015. Analyses of multiyear statewide secondary crash data and automatic crash report reviewing. Transp. Res. Rec. 2514 (2514), 117–128. https://doi.org/10.3141/2514-13.