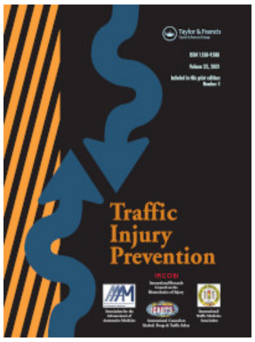

Traffic Injury Prevention

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gcpi20

Crash narrative classification: Identifying agricultural crashes using machine learning with curated keywords

Jisung Kim, Amber Brooke Trueblood, Hye-Chung Kum & Eva M. Shipp

To cite this article: Jisung Kim, Amber Brooke Trueblood, Hye-Chung Kum & Eva M. Shipp (2021) Crash narrative classification: Identifying agricultural crashes using machine learning with curated keywords, Traffic Injury Prevention, 22:1, 74-78, DOI: 10.1080/15389588.2020.1836365

To link to this article: https://doi.org/10.1080/15389588.2020.1836365

TRAFFIC INJURY PREVENTION
2021, VOL. 22, NO. 1, 74–78
https://doi.org/10.1080/15389588.2020.1836365

Crash narrative classification: Identifying agricultural crashes using machine learning with curated keywords

Jisung Kim[a], Amber Brooke Trueblood[b], Hye-Chung Kum[c], and Eva M. Shipp[b]

[a]Mobility Division, Transportation Planning, Texas A&M Transportation Institute, College Station, Texas; [b]Center for Transportation Safety, Crash Analytics Team, Texas A&M Transportation Institute, College Station, Texas; [c]Population Informatics Lab, Department of Health Policy and Management, School of Public Health, Texas A&M University, College Station, Texas

ABSTRACT

Objective: Traditionally, structured or coded data fields from a crash report are the basis for identifying crashes involving different types of vehicles, such as farm equipment. However, using only the structured data can lead to misclassification of vehicle or crash type. The objective of the current article is to examine the use of machine learning methods for identifying agricultural crashes based on the crash narrative and to transfer the application of models to different settings (e.g., future years of data, other states).

Methods: Different data representations (e.g., bag-of-words [BoW], bag-of-keywords [BoK]) and document classification algorithms (e.g., support vector machine [SVM], multinomial naïve Bayes classifier [MNB]) were explored using Texas and Louisiana crash narratives across different time periods.

Results: The BoK-support vector classifier (SVC), BoK-MNB, and BoW-SVC models trained with Texas data were better predictive models than the baseline rule-based algorithm on the future year test data, with F1 scores of 0.88, 0.89, 0.85 vs. 0.84. The BoK-MNB trained with Louisiana data performed the closest to the baseline rule-based algorithm on the future year test data (F1 scores, 0.91 baseline rule-based algorithm vs. 0.89 BoK-MNB). The BoK-SVC and BoK-MNB models trained with Texas and Louisiana data were better productive models for Texas future year test data with F1 scores 0.89 and 0.90 vs. 0.84. The BoK-MNB model trained with both states' data was a better predictive model for the Louisiana future year test data, F1 score 0.94 vs. 0.91.

Conclusions: The findings of this study support that machine learning methodologies can potentially reduce the amount of human power required to develop key word lists and manually review narratives.

ARTICLE HISTORY
Received 7 April 2020
Accepted 8 October 2020

KEYWORDS
Machine learning; crash narratives; agricultural crashes; bag-of-words; document classification algorithms

Introduction

The agriculture, forestry, fishing, and hunting sector has grown 14.5% from 2009 to 2019 to 2,303,600 jobs (U.S. Bureau of Labor Statistics 2020). Along with this growth, fatal occupational injury rates are higher in the sector compared to the national average (U.S. Bureau of Labor Statistics 2019a). The fatal occupational injury rate was 6.7 times higher in the agriculture, forestry, fishing, and hunting sector compared to all other sectors and industries in 2018 (U.S. Bureau of Labor Statistics 2019a). In 2018, 39.6% of fatal occupational injuries were transportation incidents across all sectors and industries, whereas 47.7% of fatal occupational injuries among the agriculture, forestry, fishing, and hunting industry were transportation incidents (U.S. Bureau of Labor Statistics 2019b). Little research has examined the frequency, severity, or factors that contribute to crashes in the sector. The lack of research hinders the identification of the most effective countermeasures for preventing crashes involving farm equipment or reducing their severity.

Studies have analyzed state crash data to identify factors affecting farm equipment crashes (Gerberich et al. 1996; Hughes and Rodgman 2000; Lacy et al. 2003; Peek-Asa et al. 2007; Gkritza et al. 2010; Harland et al. 2014, 2018; Ramirez et al. 2016; Ranapurwala et al. 2016). In most of these studies, farm equipment–related crashes are identified solely based on structured data fields in the crash data, such as farm equipment vehicle body style. However, there are at least 2 scenarios where farm equipment–related crash identification methods based on the vehicle type field in the structured records can potentially be misleading. First, equipment such as grass cutters and construction equipment can be used for agricultural purposes but can be used for a number of nonagricultural purposes (e.g., grass cutting by municipal government). Second, identifying farm crashes based on the structured data could result in missing crashes that involve another vehicle being utilized for agricultural purposes (e.g., farm equipment not classified as a farm vehicle or equipment in the structured data). Both cases lead to potential misclassification of farm equipment crashes when only structured crash data fields are relied upon to classify or identify crashes.

CONTACT Amber Brooke Trueblood a-trueblood@tti.tamu.edu 3135 TAMU, College Station, TX, 77843-3135.
Associate Editor Brian D. Stemper oversaw the review of this article.
Supplemental data for this article is available online at https://doi.org/10.1080/15389588.2020.1836365

There are 2 prior studies that go beyond using structured data to identify farm equipment crashes through the addition of crash narratives (Gkritza et al. 2010; Trueblood et al. 2019). Crash narratives are open-text fields completed by the reporting police officer. Crash narratives provide crash details that may not be available in the structured data. Gkritza et al. (2010) used a basic method to identify crashes for which narratives included either a "tractor" or "combine" keyword for their analysis. Trueblood et al. (2019) developed a full keyword list of 130 inclusion keywords and 38 exclusion keywords that was applied to crash narratives from Louisiana from 2010 to 2015. The authors presented a semi-automated method to identify agricultural crashes, as well as a gold standard data set used to evaluate the developed methodology. The semi-automated methodology reduced the search space for narratives by up to 60% depending on the research question. In addition, using information from the narratives, the study found that 33.5% of agricultural crashes identified based on vehicle body style were not farm equipment.

Though there are limited papers that have used narratives to explore farm equipment crashes, there are several articles that have used crash narratives for crash analysis (Nayak et al. 2010; Gao and Wu 2013; Rakotonirainy et al. 2015; Pour-Rouholamin et al. 2016; Das et al. 2019; Boggs et al. 2020; Zhang et al. 2020). Most recently, Boggs and colleagues (2020) utilized text analysis to extract information from crash narratives to explore automated vehicle crashes.

The objective of the current article is to investigate which machine learning methods perform best for analyzing crash narratives to identify agricultural crashes and to evaluate how well the models transfer to different settings (e.g., future years of data, different states). We compared different data representations (e.g., bag-of-words [BoW], bag-of-keywords [BoK]) and document classification algorithms (e.g., support vector machine [SVM], multinomial naïve Bayes classifier [MNB]) using Texas and Louisiana crash narratives across different time periods.

Methods

Data collection

Crash narratives were obtained from the Louisiana Department of Transportation and Development (LaDOTD) for 2005 to 2017 and the Texas Department of Transportation's (TxDOT) Crash Records Information System for 2010 to 2016. Two manually labeled data sets were created using the crash narratives for agricultural crashes, which were defined as crashes that involved at least one farm equipment vehicle based on each state's vehicle type definition. This research was approved by the Texas A&M Institutional Review Board (IRB 2016-0592D).

Data preparation

Crash researchers manually labeled each crash narrative with one of 3 target class labels: Narrative includes evidence of "agricultural" (farm equipment related), "nonagricultural" (not related), or "ambiguous" (unable to classify). Crash narratives labeled as agricultural and nonagricultural were used as examples for training the machine learning algorithm and validation because the ambiguous labels meant that there was not sufficient information in the narrative, even for a human expert to determine whether farm equipment was involved in the crash or not. For example, a narrative with no vehicle details could not be confirmed to be an agricultural crash or not. The data were then split into training and test data sets for each state. The test data set was the last year of available data, 2017 for Louisiana and 2016 for Texas. Table 1 summarizes the training and test data sets for each state by target class.

Analysis

Different data representations (e.g., BoW, BoK) and document classification algorithms (e.g., SVM, MNB) were explored using Texas and Louisiana crash narratives across different time periods. The Louisiana prediction models were trained using the Louisiana agricultural training data set from 2005 to 2016 and validated by prediction accuracy on 2017 data. Similarly, the Texas agricultural training data set from 2010 to 2015 was used to train the Texas prediction models with validation based on prediction accuracy on 2016 data. Next, the Louisiana and Texas models were tested to determine how well they transfer to the other state. The following section summarizes the steps taken to preprocess and tune the models.

Bag-of-keywords development

Previously, a simple rule-based classification method involved a manually curated 2-step keyword list of inclusion terms and exclusion terms (Trueblood et al. 2019). The keyword lists were designed to be generalizable across multiple states and time periods. The inclusion list includes terms (words or phrases) that indicate that a crash narrative is likely to be agricultural, such as "farmer," "cotton," "wheat," "planter," "seeder," and "John Deere." The exclusion list has terms suggesting the context surrounding a crash, indicating that although an inclusion keyword (e.g., "tractor") was used, it is not an agricultural crash, such as "construction," "grass cutter," "lawn mower," "parish tractor," and "work zone." Because the keywords in those lists are a subset of the words appearing in the collection of crash narratives, a crash narrative can be represented as BoK only using those keywords and their frequencies.

Using the previously developed keyword list, a crash narrative can be represented by a 277-dimensional vector (calculated before reducing to root forms); see Table S1 (see online supplement). The keyword list treated some forms of morphologically variant words as separate keywords that were reduced to their root forms or semantic grouping. For example, "farm worker" and "farmworker" are different entries in the keyword lists. There is no difference between "farm worker" and "farmworker" for document classification; thus, semantically related keywords were merged from the original keyword list into one single keyword feature.

Table 1. Number of class examples in different datasets.

Data set	Number of examples Agricultural	Nonagricultural	Ambiguous	Total
Louisiana training (2005–2016)	148	199	320	793
Louisiana test (2017)	21	26	40	87
Louisiana total	169	225	361	880
Texas training (2010–2015)	305	153	827	1285
Texas test (2016)	52	42	159	253
Texas total	357	195	986	1,538

After this semantic grouping, 154 keywords remained in the inclusion and exclusion terms. In addition, a keyword distribution analysis found that "horse" shows up only in one narrative in the training data (LA agricultural), as "horse power," which is a wrong example of the word in the keyword list meant to reference the animal. Thus, in the rest of the analysis, "horse" was excluded. Thus, the final BOK feature vector is composed of 153 keywords.

Text normalization

A word has many morphological variations. In principle, the meaning is of primary interest and not their forms; therefore, such variation should be reduced before creating any feature vector representation of a document using the vocabulary. First, words were changed to lowercase, and then they were reduced to their root forms using the Snowball English stemmer from the NLTK Python library (Rossum and Drake 2011). Finally, stop words or common English words that are not meaningful (e.g., the, a, etc.) were removed.

Term frequency–inverse document frequency weighting

Bag-of-words representation treats every word as equally important. For example, "vehicle" and "driver" frequently appear in crash narratives. Almost all narratives have these words. These frequent words do not provide many clues for machine learning algorithms to learn models that distinguish different classes of crash narratives. However, rarer words such as "harvester" and "cultivator" can be more informative. Tf-idf weighting is based on this intuition. Tf-idf reweights the direct count (term frequency) of a term into floating point values using the inverse document frequency (idf). With idf weighting, a frequent word gets less weight. Scikit-learn implementation of tf-idf weighting (*sklearn.feature_extraction.text.TfidfVectorizer*) was utilized.

Hyperparameter selection through cross-validation

To select the best performing hyperparameters for support vector classifier (SVC) and MNB, 5-fold cross-validation was performed with exhaustive grid search on predefined hyperparameter values.

Training imbalanced classes

One of the main issues in training a classification model is the imbalance in the classes. For example, in the Texas data set, 357 crash narratives out of 552 narratives are agricultural. Thus, if an algorithm labeled all to be the agricultural class, it would be 65% accurate and 15% better than a coin toss. Often models built using the raw training data with imbalanced classes do not perform well because the smaller classes are not properly learned. For an MNB model training, the smaller class examples were up-sampled to match the number of the larger class examples.

Performance metrics

Accuracy is the most common performance metric for classifiers, which is defined as the number of correctly predicted samples divided by the number of total samples. However, with an imbalanced data set where examples in one class outnumber other classes by a large proportion, it can be misleading, because by simply assigning most of test examples to the large class, a classifier can achieve high accuracy, as discussed in the previous section. To address this limitation of accuracy as a measure, F_1 macro scores are often used as a better performance metric in classification. F_1 is a harmonic average of the precision and recall. For multiple class classification, the average of metrics is taken to measure overall performance. Macro averaging was adopted, which simply computes independently computed metrics for each class and then averages those metrics, thus treating each class with equal importance.

Model development process

The scikit-learn implementation of the SVM classifier (*sklearn.svm.SVC* class) and MNB classifier (*sklearn. naive_bayes.MultinomialNB*) were utilized. For SVM, the kernel was set to "linear" and a brute force search was conducted over specified values (using *sklearn.model_selection.GridSearchCV*) for the regularization parameter C of SVC. For MNB, the smoothing parameter (α) was set using scikit-learn *GridSearchCV* function. Additional references used in model development can be found in the Appendix (see online supplement).

Results

Texas model (trained on Texas 2010–2015)

The first 6 years of Texas data ($N = 458$), 2010–2015, were used to train the model. Predictive model performance was tested with the last year (2016), the "same state test," and how well the model transfers when tested with the Louisiana data, the "cross-state test." The training data had twice as many agricultural examples as nonagricultural examples (67% agricultural vs. 33% nonagricultural). A comparison of the different models, a combination of document representation (BoW vs BoK) and machine learning algorithms (SVC vs. MNB), is presented in Table S2 (see online supplement).

As a baseline comparison, the first row presents results from the simple rule-based classification using the keyword list. BoK-SVC had the best results for both the 2016 Texas data and Louisiana data at about 88% performance on F_1. BoK-MNB also had good performance for both tests, at 89% and 87% on F_1. In comparison, both best-performing BoW representation models on both tests did not perform as well: 85% (BoW-SVC) with the same state test and 70% (BoW-MNB) with the cross-state test on F_1. Compared to the simple keyword-based classification, building an SVC with the keywords improved the F_1 score by 4% (from 84% to 88%) with the same state test.

Louisiana model (trained on Louisiana 2005–2016)

More years of data were available for Louisiana than for Texas, so the model was built with the first 12 years of data, from 2005 to 2016 ($N = 347$), and then tested on 2017 data as well as Texas data (see Table S3, online supplement). In the Louisiana training data, there were fewer agricultural examples (43% agricultural vs. 57% nonagricultural). As with the Texas model, BoK representation performed better than BoW for both SVC and MNB algorithms. But BoK-MNB performed best overall based on F_1 (89% with the same state test vs. 85% for BoK-SVC). However, overall the rule-based keyword performed the best when validated on Texas data ($F_1 = 89\%$).

Combined model (trained on Texas 2010–2015 and Louisiana 2005–2016)

Machine learning algorithms tend to perform better with larger and more balanced training data. Thus, Texas and Louisiana training data were combined, and predictive model performance was tested on the last year of combined data (see Table S4, online supplement). The combined training data included 56% agricultural and 44% nonagricultural examples, with a total of 807 examples. Again, BoK clearly outperformed BoW representation, with the MNB algorithm performing better than SVC. The model built using the combined data improved the performance of Texas test data to F_1 equal to 92%, but it had no impact on the Louisiana test data.

Discussion

Overall, the BoK models trained with the Texas data or the combined data had better predictive performance on the Texas 2016 data and had better or comparable performance on the Louisiana 2017 test data set compared to the baseline rule-based algorithm. The result suggests that MNB and SVC algorithms can learn the decision boundary as well as the baseline rule-based algorithm. Even the BoW-SVC model trained with the Texas data performed better than the baseline algorithm. Unlike the baseline rule-based algorithm, MNB and SVC models do not distinguish keywords in the "exclusion" lists and those in the other 6 categories, which could reduce the keyword curation efforts. This suggests that at least with the BoK representation of crash narratives, keywords do not have to be divided into inclusion and exclusion keyword lists and a set of rules does not need to be manually developed to achieve comparable performance to the rule-based algorithm. The benefit is a reduction in manpower often required to review narratives to develop these keywords and rules. Another benefit of using MNB or SVC with BoK instead of a set of manual rules is clear in the example of the "horse" keyword. There was a single crash narrative in the Louisiana data set that included "horse" as part of "horse power." The rule-based algorithm always classifies any crash narrative having horse in it as agricultural related even though it could mean something else, such as horse power. In contrast, MNB and SVC may learn that horse is not useful when distinguishing between agricultural and nonagricultural crashes.

The performance of the models with BoW representation was not far behind when tested on the same state test compared with the models with BoK representation. The differences in performance between the best-performing BoK model and the best-performing BoW model tested on the same state are 3% (Texas) and 4% (Louisiana). The BoW models performed significantly worse if they were trained with a single state's data and tested on the other state's data. This might imply that there are differences between words/phrases used in different states' data. As reported in Table S4, when the BoW models were trained with the combined data, the BoW model performance improved.

When resources are limited for annotating crash narratives with document class labels, BoW-SVC or BoW-MNB can be used to provide a baseline document classification system before investing more resources into developing a keyword list. To adopt this approach, the number of labeled documents (examples) should be large enough to capture the variance in terms of use of words in the target corpus, which will be greater than that with BoK representation due to the higher dimensionality.

Overall, crash narratives were found to be relatively short in terms of the number of words and sentences. On average, a crash narrative in the Louisiana data set had 279 words and 8 sentences, and the Texas data set had 116.8 words and 2.3 sentences. A crash narrative is not as short as a product review or a tweet. However, some short narratives do not have enough information to capture the context of words in BoW representation. This implies that there would be a limitation in improving classification performance even by adding more training examples due to this inherent short document length. This is also true with BoK representation. About 52% of Louisiana narratives and 50% of Texas crash narratives contained no or only one keyword, including both inclusion and exclusion keywords. This aspect of crash narratives may explain why MNB or SVC models could not outperform the rule-based algorithm because there is not sufficient information to form more complex decision boundaries utilizing relationships between words or keywords. Because each feature is most likely simply independent in short narratives, the MNB algorithm performs comparable to or better than SVC algorithm. It should be noted that the naive independence assumptions of MNB models between keyword features are unrealistic for the data, which may be a side effect of not having many training examples with multiple keywords. Based on these limitations future research should expand the current data set to include additional years of narratives and states, which would result in a more generalizable keyword list.

The BoW model development did not include any complicated dimensionality reduction techniques such as word embedding. However, it is not far behind the keyword-based models in terms of performance, especially with the Texas data set, which we believe still to be a viable alternative to the manually curated keyword list. More research is needed on dimensionality reduction with BoW models to improve BoW model performance.

In summary, the findings of the article indicate that machine learning methods can be applied to identifying agricultural crashes and can be transferred to different settings (e.g., future years of data, other states). Overall the current article demonstrates that the machine learning methodologies (e.g., BoK-MNB trained with Louisiana 2005–2016 data and the BoK-SVC trained with Texas 2010–2015 data) performed better than the baseline rule-based algorithm (e.g., semi-automated method; Trueblood et al. 2019) and can be effective in identifying agricultural crashes. Ultimately, the use of machine learning methodologies reduces the amount of human power traditionally required to develop keyword lists and manually review narratives. The results from this article can be applied to multiple settings and built on through the expansion of a more general keyword list. Furthermore, future work should build on these results to capture agricultural domain knowledge into computational forms (e.g., taxonomy) beyond.

Acknowledgment

The authors thank the LaDOTD for assistance with providing the data and providing clarification as needed.

Funding

This research was supported by CDC/NIOSH under Cooperative Agreement No. U50 OH07541 to the Southwest Center for Agricultural Health, Injury Prevention, and Education at the University of Texas Health Science Center at Tyler. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of CDC/NIOSH.

ORCID

Amber Brooke Trueblood http://orcid.org/0000-0002-1404-347X
Hye-Chung Kum http://orcid.org/0000-0002-6882-8053
Eva M. Shipp http://orcid.org/0000-0002-4034-8031

Data availability statement

The data sets generated during the current study are not publicly available due to data use agreements. Data can be requested through TxDOT and LaDOTD.

References

Boggs A, Wali B, Khattak A. 2020. Exploratory analysis of automated vehicle crashes in California: a text analytics & hierarchical Bayesian heterogeneity-based approach. Accid Anal Prev. 135:105354. doi:10.1016/j.aap.2019.105354

Das S, Kong X, Tsapakis I. 2019. Hit and run crash analysis using association rules mining. J Transport Safety Security. 1–20. [CrossRef][doi:10.1080/19439962.2019.1611682

Gao L, Wu H. 2013. Verb-based text mining of road crash report. Paper presented at: 92nd Annual Meeting of the Transportation Research Board; Washington, DC; January, 13–7.

Gerberich S, Robertson L, Gibson R, Renier C. 1996. An epidemiological study of roadway fatalities related to farm vehicles: United States, 1988 to 1993. J Occup Environ Med. 38(11):1135–1140. doi:10.1097/00043764-199611000-00013

Gkritza K, Kinzenbaw C, Hallmark S, Hawkins N. 2010. An empirical analysis of farm vehicle crash injury severities on Iowa's public road system. Accid Anal Prev. 42(4):1392–1397. doi:10.1016/j.aap.2010.03.003

Harland K, Bedford R, Wu H, Ramirez M. 2018. Prevalence of alcohol impairment and odds of a driver injury or fatality in on-road farm equipment crashes. Traffic Inj Prev. 19(3):230–234. doi:10.1080/15389588.2017.1407924

Harland K, Greenan M, Ramirez M. 2014. Not just a rural occurrence: differences in agricultural equipment crash characteristics by rural-urban crash site and proximity to town. Accid Anal Prev. 70:8–13.

Hughes R, Rodgman E. 2000. Crashes involving farm tractors and other farm vehicles/equipment in North Carolina 1995-1999. April 1, 2000. [accessed 2020 Mar 23]. https://rosap.ntl.bts.gov/view/dot/4790.

Lacy J, Hunter W, Huang H. 2003. Study of farm vehicle crashes in North Carolina. Transp Res Rec. 1840(1):178–185.

Nayak R, Piyatrapoomi N, Weligamage J. 2010. Application of text mining in analysing road crashes for road asset management. In Proceedings of the 4th World Congress on Engineering Asset Management (WCEAM 2009). September, 28–30. Athens, Greece.

Peek-Asa C, Sprince NL, Whitten PS, Falb SR, Madsen MD, Zwerling C. 2007. Characteristics of crashes with farm equipment that increase potential for injury. J Rural Health. 23(4):339–347. doi:10.1111/j.1748-0361.2007.00112.x

Pour-Rouholamin M, Zhou H, Zhang B, Turochy R. 2016. Comprehensive analysis of wrong-way driving crashes on Alabama interstates. Transp Res Rec. 2601(1):50–58.

Rakotonirainy A, Chen S, Scott-Parker B, Loke SW, Krishnaswamy S. 2015. A novel approach to assessing road-curve crash severity. J Transport Safety Security. 7(4):358–375.

Ramirez M, Bedford R, Wu H, Harland K, Cavanaugh J, Peek-Asa C. 2016. Lighting and marking policies are associated with reduced farm equipment-related crash rates: a policy analysis of nine Midwestern US states . Occup Environ Med. 73(9):621–626. doi:10.1136/oemed-2016-103672

Ranapurwala S, Mello E, Ramirez M. 2016. A GIS-based matched case-control study of road characteristics in farm vehicle crashes. Epidemiology. 27(6):827–834.

Rossum GV, Drake FL. 2011. The Python language reference manual: for Python version 3.2. Bristol (UK): Network Theory Ltd.

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

U.S. Bureau of Labor Statistics. 2019a. Table 4. Fatal occupational injuries counts and rates by selected industries. Economic News Release. 2017–2018. December 17 [accessed 2020 23 Mar]. https://www.bls.gov/news.release/cfoi.t04.htm.

U.S. Bureau of Labor Statistics. 2019b. Table A-2. Fatal occupational injuries resulting from transportation incidents and homicides, all United States, 2018. Injuries, Illnesses, and Fatalities.December 17 [accessed 2020 23 Mar]. https://www.bls.gov/iif/oshwc/cfoi/cftb0323.htm.

U.S. Bureau of Labor Statistics. 2020. Employment by major industry sector. Employment Projections September 1 [accessed 2020 Sept 16]. https://www.bls.gov/emp/tables/employment-by-major-industry-sector.htm.

Zhang X, Green E, Chen M, Souleyrette R. 2020. Identifying secondary crashes using text mining techniques. J Transport Safety Security. 12(10):1338–1321.