Analysis of Pedestrian and Bicycle Crashes with Contributing Factor of "Other" or with No Contributing Factor

TxDOT Project 2024-TTI-G-1YG-0049

Authors:
Neal A. Johnson
Mahin Ramezani
Ehsan Jalifar
Minh Le
Joan G. Hudson

Prepared for the Behavioral Traffic Safety Section
Texas Department of Transportation

June 2024

TEXAS A&M TRANSPORTATION INSTITUTE
College Station, Texas 77843-3135

Texas A&M Transportation Institute

Table of Contents

List of Figures

List of Tables

Introduction

As part of the Walk. Bike. Safe. Texas project, the Texas A&M Transportation Institute (TTI) conducted a focused crash analysis looking at pedestrian and bicycle crash reports where the contributing factor of "other" was used or when no contributing factor was listed. The use of "other" or no contributing factor says nothing about the circumstances of the crash, the possible laws violated, and/or the behaviors of the road users involved.

The project team looked at the most recent five years of Texas Department of Transportation Crash Records Information System (CRIS) crash data (2018–2022) for fatal and suspected serious injury (KA) crashes involving a pedestrian or bicyclist where "other" was used as a contributing factor or where no contributing factor was listed. The analysis included 2,832 crash records (the number of crashes) involving 2,896 people (the number of people killed or injured).

In addition to the crash data details, the TTI team also analyzed weather data related to each of the crashes, specifically information on potential sun glare and other weather events such as rain, snow, sleet/hail, and fog to determine if weather may have been a factor.

This report is divided into four main sections: demographic characteristics, crash data findings, weather data findings, and conclusions. The aim of this report is to identify the types of behaviors, actions, and other factors that led to these crashes and to consider whether additional areas of concern should be included in messaging on pedestrian and bicycle crashes.

Demographic Characteristics

As part of this analysis of pedestrian and bicycle crashes with the contributing factor of "other" or without a contributing factor, it is important to understand who is involved in these crashes. The majority of crashes involve men, accounting for **69.2 percent** of pedestrian crashes and **88.0 percent** of bicyclist crashes, overall. Figure 1 and Figure 2 compare the numbers from this set of crashes from 2018-2022 with the numbers from the 2022 report looking at 2016-2020 KAB data, also completed under the Walk. Bike. Safe. Texas project, in which males were **65.9 percent** of pedestrian crash victims and **82.3 percent** of bicycle crash victims. These figures show that these crashes with the contributing factor of "other" or without a contributing factor skew even more toward males than females.

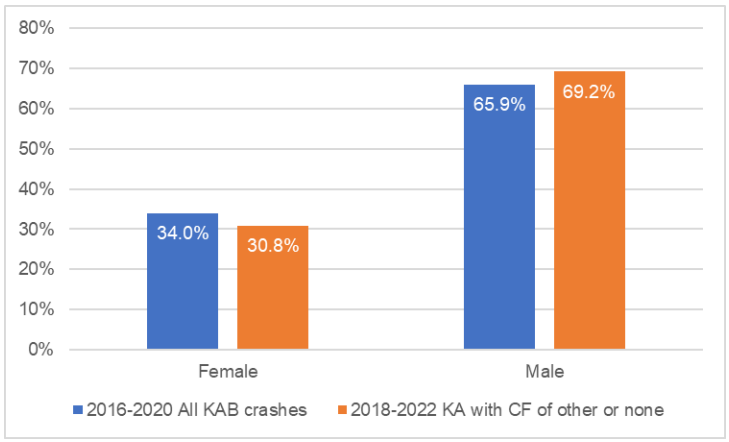

Figure 1. Gender of Pedestrians Comparison.

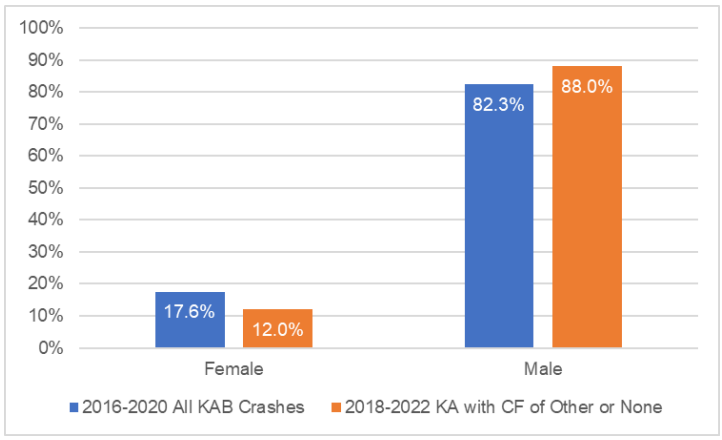

Figure 2. Gender of Bicyclists Comparison.

Table 1 and Table 2 show the age and gender breakdowns for the pedestrians and bicyclists, respectively, involved in these crashes. In terms of age, the highest percentages of pedestrians were in the 21–45 age range for both genders. For bicyclists, there is more variation with certain age groups showing higher percentages, such as females ages 16–20, 26–30, and 36–45 and males ages 26–35 and 51–55.

Table 1. Pedestrian Age and Gender.

Pedestrians		
Age Group	**Female**	**Male**
0-5	3.0%	1.6%
6-10	2.4%	1.5%
11-15	4.5%	3.1%
16-20	8.7%	6.7%
21-25	9.3%	10.2%
26-30	11.4%	11.7%
31-35	7.9%	9.9%
36-40	10.3%	9.3%
41-45	9.0%	7.2%
46-50	8.0%	8.9%
51-55	5.2%	6.3%
56-60	5.2%	7.5%
61-65	6.3%	6.4%
66-70	4.2%	4.3%
71-75	2.3%	2.7%
76+	2.4%	2.8%
Total	**30.8%**	**69.2%**

Table 2. Bicyclist Age and Gender.

Bicyclists		
Age Group	**Female**	**Male**
0-5	0.0%	0.6%
6-10	1.6%	3.5%
11-15	7.9%	7.8%
16-20	11.1%	7.3%
21-25	7.9%	5.0%
26-30	12.7%	8.2%
31-35	4.8%	9.5%
36-40	14.3%	7.8%
41-45	12.7%	6.7%
46-50	7.9%	7.6%
51-55	9.5%	11.2%
56-60	6.3%	9.7%
61-65	1.6%	8.6%
66-70	1.6%	3.2%
71-75	0.0%	1.5%
76+	0.0%	1.7%
Total	**12.0%**	**88.0%**

Crash Data Findings

This next task aimed to enhance the analysis of crash reports by identifying and extracting potential contributing factors not currently captured in the CRIS data. The goal was to use natural language processing (NLP) techniques to identify key phrases from crash report narratives and to provide additional context and details about these phrases.

Text data often contain variations due to different writing styles, typos, or inconsistencies. Preprocessing helps to normalize the text, such as converting all text to lowercase, removing punctuation, and handling contractions.

To normalize the data and improve model performance, a couple of data preprocessing techniques were used, including:

- Converting the text to all lowercase sentences.
- Removing stop words (e.g., “a,” “the,” “are,” and “many”).
- Lemmatizing words (i.e., reducing words to their base or root form, known as the lemma).
- Unifying the format for critical words.

After manually reviewing about 25 of the crash narratives, the team found that the narratives contained typos. Additionally, for the word

“unit,” a couple of formats or variants (e.g., unit 1, unit #1, u1, and u#1) were observed in the narratives. Since the unit number identifies whether the subject is the driver or pedestrian/bicyclist and plays a critical role in the modeling, these unit numbers were all normalized to the “u<number>” format, such as “u1” and “u2.” The results of the NLP models, which are explained in the following sections, also showed that this format unification improves the model performance.

The defined preprocessing techniques were implemented using the Natural Language Toolkit (NLTK) library in Python, a robust library commonly used in NLP. The resulting preprocessed narratives were fed to the subsequent tasks:

1. Word cloud.
2. Topic modeling.
3. Question and answer modeling.

Word Cloud

A word cloud is a visual representation of text data where the size of each word indicates its frequency or importance within the text. The word cloud first checks the frequency of the words within the crash narrative and then evaluates the importance of each word within the context. Some words like “fatal crash” might not be as common as the word “officer” in these data. However, the term “fatal crash” conveys much more important information than the word “officer.” The underlying word cloud model considers this matter. Figure 3 shows the word cloud generated using the NLTK library.

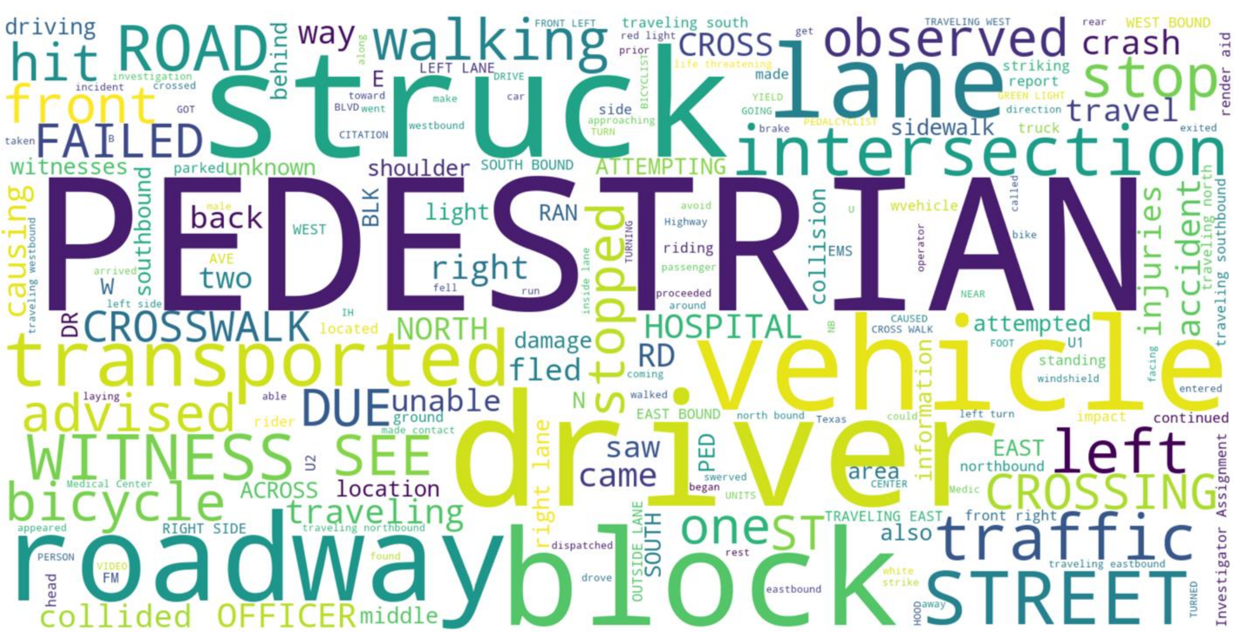

Figure 3. Crash Narrative Word Cloud.

Topic Modeling

Topic modeling is a technique used in NLP and machine learning to identify the main topics or themes present in a collection of text documents. The purpose of topic modeling in this task was to automatically identify hidden underlying contributing factors (topics) in a crash, based on the crash narratives. The BERTopic model was selected for its status as one of the cutting-edge open-source models available. Several rounds of model hyperparameter tuning, topic generation, and review by subject matter experts (SMEs) generated the potentially relevant terms (topics) related to crash contributing factors. The resulting potential terms were used in the question and answer modeling to evaluate their contribution in the crash.

For each topic generated by the BERTopic model, a group of narratives relevant to the specified topic was identified. Once again, the word cloud was used to get a better understanding of the main contributing factor in each group. In other words, a cluster/group of documents/narratives that implied the same potential contributing factor was obtained. The team found that using 3-grams for the word cloud representation provides the most tangible representation for the SMEs to interpret the results. The n-gram defines how many adjacent words are grouped together to form the tokens for the word clouds. For instance, “pedestrian crossed street” generates three 1-grams (i.e., “pedestrian,” “crossed,” and “streets”), two 2-grams (“pedestrian_crossed” and “crossed_street”), or one 3-gram (“pedestrian_crossed_street”).

Figure 4 illustrates the 3-gram word cloud for topics/potential contributing factors. In this example, one interpretation is that a vehicle **struck** a pedestrian walking. Thus, the action “struck” was selected to be analyzed in the question and answer modeling in the next section.

Figure 4. Example of the Generated 3-Gram Word Cloud Using the BERTopic Model.

Question and Answer Modeling

Four experts on the project reviewed the phrases generated from the topic modeling step, carefully selecting those they considered valuable as potential contributing factors to crashes. They identified phrases that required further investigation within the crash narratives to determine their presence and relevance. After data preprocessing, a refined list of phrases was created to be used in the next step of the analysis. This step involved applying Bidirectional Encoder Representations from Transformers (BERT) question and answer models to gain a deeper understanding of the contributing factors. Leveraging the expertise of these reviewers ensured that the analysis focused on the most insightful phrases, enhancing the overall utility of the findings.

BERT QA Model

BERT is a cutting-edge NLP model developed by Google. BERT understands the context of words in a sentence by considering the surrounding words. BERT is highly effective for various NLP tasks, including question answering (QA). In QA tasks, a model is given a text

and a question about that text. The model's job is to find the answer within the text. In this task, two BERT QA models were fine-tuned:

- **Model 1: Who Did the Action/Who Caused the Condition:** This model answers questions about who performed the action or caused the situation described by the identified phrase. For example, if the phrase is "left the scene," the model identifies who left the scene. This model only provides output if it can find a clear answer.
- **Model 2: Additional Details:** This model provides further context and details about the phrase. For example, the model describes the circumstances under which someone "left the scene." This model provides output for all reports that contain the identified phrase, even if Model 1 did not find an answer.

After application of these models, valuable information about various contributing factors to crashes not currently captured in the CRIS data was extracted (Table 3 and Table 4).

Table 3. Output Format from BERT QA Models for Sample Phrases.

Phrase	Model 1 Output	Model 2 Output
left scene/fled	Identified the individual who left the scene	Provided additional context about the circumstances of fleeing
drugs/alcohol	Identified the person involved with drugs or alcohol	Described the influence or effect of drugs/alcohol on the crash
struck/striking	Identified who or what was struck	Provided details about the incident

Table 4. Sample Outputs from BERT QA Models.

Phrase	Model 1 Output	Model 2 Output
fell	u2	u1 state drive southbound 700 block n foster road, u2 suddenly jump barrier separate road nearby dirt, fall lane directly
left scene/fled	u3	u3 flee scene fail stop render aid
cross	pedestrian	pedestrian cross street, travel westbound, mcdonald

Results and Discussions

Table 5 presents the results of BERT QA Model 1, which was designed to identify who performed the action described by specific phrases in crash narratives. The accuracy percentage of the model for each phrase indicates the proportion of correct answers among those manually labeled for evaluation purposes. The last column displays the total number of narratives containing each phrase. The accuracy values reflect the model's performance in correctly identifying the actor or subject of the action for the labeled instances. For example, the model achieved an accuracy of 92.8 percent for the phrase "fell," based on 97 labeled instances out of 215 narratives containing this

phrase. This high accuracy indicates the model's effectiveness in correctly identifying who fell in the crash narratives. Similar evaluations were conducted for other phrases, such as "fault," "cross," and "left scene/fled," demonstrating varying levels of accuracy and labeled data availability.

Table 5. BERT Q&A Model 1 Accuracy on the Labeled Crash Narratives.

Phrase	Accuracy (%)	Number of Manually Labeled Data	Number of Narratives Containing the Phrase
fell	92.8	97	215
fault	84.2	38	97
cross	81.6	76	911
left scene/fled	73.8	65	499
turned left	73.8	42	72
turned right	69.7	33	37
failed yield	93.1	72	212
struck	46.4	56	1,936
swerved	88.2	51	198
exited vehicle	91.3	23	68
jump	75	20	80
ran	86.1	36	556

Not all narratives have ground truth labels, and further labeling is required to obtain more accurate measurements. Additionally, some phrases did not have enough labeled data, so they were not evaluated in this table. Evaluating these phrases can be considered for future work. These phrases include:

- Bystander.
- Body found.
- Disregarding oncoming traffic.
- Roll.
- Skateboard.
- Soliciting.
- Standing.
- Stepped.
- Walking.
- Worker.
- Suddenly appeared.
- Drug/alcohol.

Moreover, some phrases only describe conditions and do not involve specific units or actors, resulting in outputs only from Model 2. Phrases like "visibility/dark" and "sun" describe environmental conditions rather than actions performed by specific entities. However, these phrases could be additional contributing factors to crashes involving pedestrians and bicyclists that are not currently in the contributing factor list.

The following are some challenges that were encountered during the analysis process using BERT QA models:

- For the action "struck," the model often found it difficult to discern the context, such as differentiating between "u1 struck u2" and "u1 struck the windshield," leading to a 50 percent accuracy rate. Additionally, subjects and objects were frequently confused, particularly when pedestrians were reported as striking something instead of being struck.
- The model also struggled with the phrase "toxicology/drugs/alcohol," often confusing the verb "drug" with the noun. The model did not clarify who had toxicology results, and often the narrative included orders for toxicology without actual results.
- For the phrase "cross," misinterpretations occurred when the phrase described crossing within a bike lane or lane changes, rather than crossing a road, or unrelated descriptions like crossing a bridge.
- Similarly, for "turn left/right," phrases like "u1 turned and left the scene" or mentions of "left turn lane" were incorrectly interpreted because they did not relate directly to the crash event.
- General issues included typographical errors, inconsistencies in narrative styles, lemmatization errors (e.g., "play+ing" matching "dis+play"), handling negations (e.g., "not + verb"), and phrases like "running out of gas" being misinterpreted. These challenges highlight the complexities of accurately analyzing and interpreting crash narratives using NLP models.

Weather Data Findings

Sun glare and detailed weather data were not present in CRIS and needed to be captured from external sources before appending these data to the crash data. TTI used datasets and tools developed by the National Oceanic and Atmospheric Administration (NOAA) Southern Regional Climate Center (SRCC). The sun glare and weather data analysis were conducted using SRCC data archives, which include a number of NOAA climate datasets.

Researchers used the same methodology described in *Applying Advanced Techniques to Datamine Pedestrian Crash Data.*[1] Generally, the weather information was estimated at the location and time of the crash. Automated surface observing system (ASOS) stations, which serve as the United States' primary surface weather observing network, are located within Texas. These ASOS sites report weather elements such as temperature, precipitation, relative

[1] Le, M., Pratt, M., Oliaee, A., Das, S., Ramezani, M., Wu, J., Guo, S. *Applying Advanced Techniques to Datamine Pedestrian Crash Data*. Texas A&M Transportation Institute, Center for Transportation Safety, November 2023.

humidity/dewpoint, wind speed, cloud cover, and visibility on an hourly basis.

These data elements were captured using the Applied Climate Information Systems application programming interface. Since the ASOS data occur at set points that may not be near the crash locations, the weather data must be interpolated across a 0.1° × 0.1° interpolated grid (about 10,000 × 10,000 meters). The climate element's value was captured and appended to the dataset. Given that ASOS sites are typically at airports/airfields, the distribution tends to be denser in urban areas (i.e., higher accuracy). Thus, SRCC assessed the quality of interpolated data and subsequently removed 19 crashes. Additionally, 28 crashes were missing coordinates that were critical for the analysis. Researchers were only able to geolocate 14 of them, yielding a total dataset of 2,799 crashes. Because cloud cover values cannot be easily interpolated, the cloud cover reading reported at the nearest ASOS station was used if available.

Solar glare information was determined at each crash location and about 10 minutes prior to the reported crash time. This temporal adjustment was made because the reported crash times are typically a little later than the time that the crash occurred according to research.[2] The solar glare was determined by the solar declination, solar zenith angle, solar elevation, hour angle, solar azimuth, and unit 1's direction of travel from the CRIS data in 45° increments (north, northeast, etc.). The solar azimuth and direction of travel were compared to determine whether "horizontal glare" was possible at the time of the crash.

There were 244 crashes (8.7 percent) flagged as solar glare possibly being a contributing factor (when no factor was provided in the analysis dataset). There were 82 crashes (2.9 percent) where the sun glare flag could not be determined. Figure 5 shows the distribution of crashes by hour and the glare flag. The distribution is consistent with the expected sun position and angle at the various hours of the day. The hours with the highest probability of glare-flagged crashes occurred near sunrise and sunset, which is when the sun sits lowest in the sky. The relative dips in the non-glare crashes between 6 and 8 a.m. and again between 4 and 6 p.m. coincide with the increase in glare crashes compared to other nearby hours. This reinforces that solar glare is possibly a factor during those hours.

[2] Kidd, B., M. Le, C. Poe, S. Joshua, and J. Short. *Evaluating Recurring and Nonrecurring Congestion Impacts within Phoenix Metropolitan Region, Arizona*. Transportation Research Board, 2012.

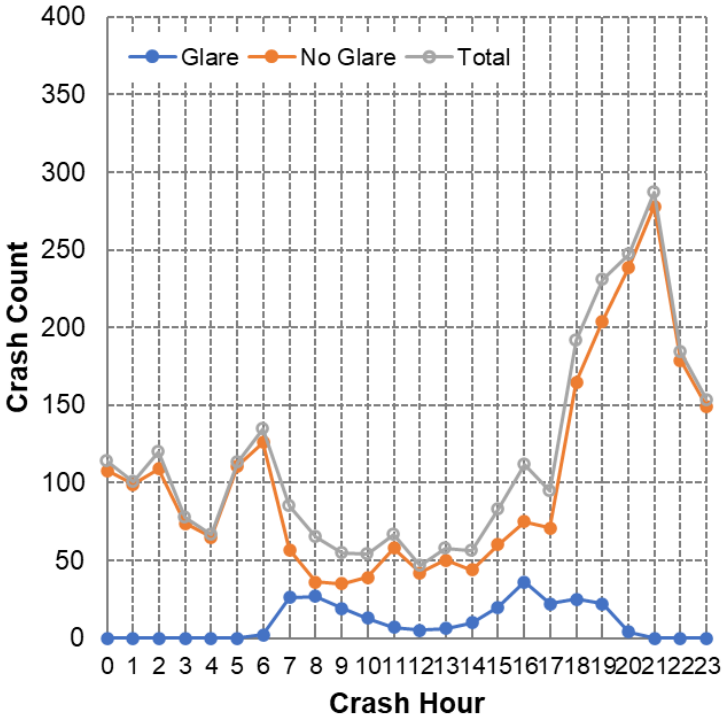

Figure 5. Distribution of Crashes by Hour and Glare Flag.

Figure 6 shows the distribution of crashes by the hourly precipitation variable from the SRCC hourly interpolated data. About 96 percent of the crashes occurred in hours with no precipitation, and the rest of the crashes occurred in hours with as much as 0.62 inches of precipitation, which could have been a contributor to those crashes. This seems reasonable given the average annual rainfall for Texas is only 27.25 inches. But rainfall totals vary across climatic regions, ranging from less than 14 inches in West Texas (e.g., El Paso and Odessa) to more than 54 inches in East Texas (e.g., Beaumont and Center).[3] The data are plotted on a logarithmic *y*-axis to improve the visibility of the small percentages for the non-zero data points.

[3] WeatherSTEM. Annual Precipitation. https://learn.weatherstem.com/modules/learn/lessons/182/19.html#:~:text=The%20average%20annual%20rainfall%20for,climatic%20regions%20of%20the%20state. Accessed May 2024.

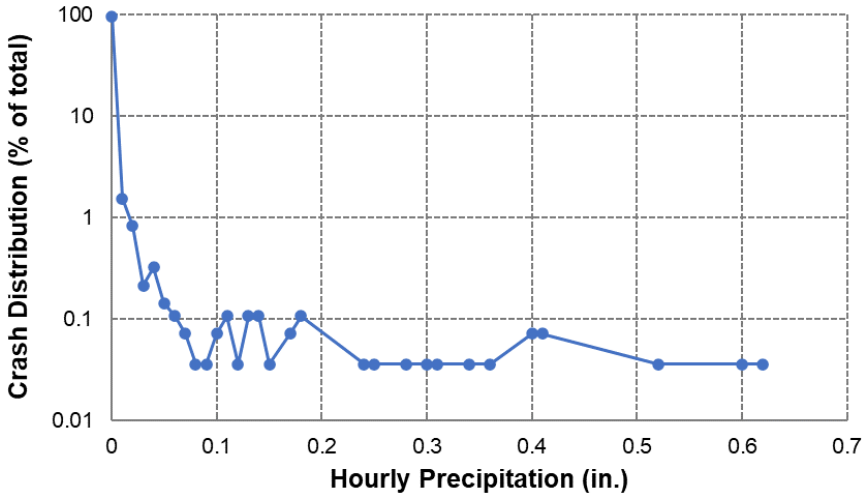

Figure 6. Distribution of Crashes by Precipitation (Logarithmic).

Researchers also examined the correlation between surface conditions and weather conditions based on available CRIS data. Table 6 shows that 88.5 percent of the crashes in the dataset were noted as dry (surface condition) and clear or cloudy (weather condition), followed by 5.8 percent noted as wet or standing water and rain conditions, which seems intuitive. The only counter-intuitive finding was the 3 percent wet or standing water and clear or cloudy conditions. This may be due to the "wet recovery" period when the pavement is still wet even after it stops raining.

Table 6. CRIS Surface Conditions versus Weather Conditions.

Surface Condition	Clear or Cloudy	Rain	Fog	Sleet, Hail, or Snow	Other
Dry	88.5%	0.0%	0.5%	0.0%	0.5%
Wet or standing water	3.0%	5.8%	0.3%	0.2%	0.1%
Slush, ice, or snow	0.1%	0.0%	0.0%	0.2%	0.0%
Other	0.5%	0.0%	0.0%	0.0%	0.2%

Conclusions

The findings from the crash analysis task highlight the potential of using NLP techniques to extract detailed and previously uncaptured information from crash reports. Text mining revealed some key words that have high accuracy, pointing to circumstances of the crash that are not revealed in the contributing factors or other places in the crash report. For example, "standing" was a key word used in a high percentage of crash narratives for Model 2, which pointed to crashes involving an unintended pedestrian, someone who exits the vehicle after a crash or breakdown and is struck. This finding points to the

potential for text mining to be a more efficient method of crash typing than reviewing police crash report narratives. Analyzing pedestrian and bicycle crashes without having to read the narrative and study the crash diagram would save many hours of work. If a practitioner or researcher was trying to understand a particular crash type or circumstance, they would ideally be able to use text mining to quickly access the desired crash dataset.

By identifying and analyzing specific phrases related to crash actions and conditions, the project team can propose adding new contributing factors to the CRIS data. This enhancement will enable more comprehensive crash data analysis and contribute to improving road safety. However, for a more accurate evaluation, more data need to be labeled correctly. Additionally, having checks on spelling and more uniform standards for writing narratives would greatly improve the usefulness of this approach to analyzing crash narratives. This also points to the need for robust training on how to accurately complete crash reports for police officers.

The weather data showed that glare from the sun was a possible factor in almost 10 percent of these crashes without a contributing factor or with the contributing factor of "other," which points to a new issue to consider when addressing traffic fatalities—and pedestrian and bicyclist fatalities more specifically. The precipitation data showed that the vast majority of crashes (96 percent) occurred when there was no recorded precipitation. However, the remaining 4 percent could still have been affected by weather conditions. The CRIS data weather conditions also provided further insight into how a surface may still be wet or have standing water, even in clear or cloudy conditions, due to the "wet recovery" period. This change in surface conditions could be something a driver would not necessarily expect given the current weather conditions.

As shown in the demographic characteristics, males have higher instances of being involved in crashes where the factor of "other" was used or where there was no contributing factor. Males could display more risk-taking behavior, which is not easily categorized in the existing contributing factors list.

The findings of this report indicate that the addition of new contributing factors to the crash report form may be warranted. The project team will consider recommendations of additional contributing factors that should be added to the current list of options, based on this analysis.

In terms of implications for outreach messaging to the public on transportation safety, and for the Walk. Bike. Safe. Texas project in particular, there seem to be a few key areas of safety messaging to consider, some new and some that need to be reemphasized. The following messaging could be included:

- Drivers run the risk of not seeing pedestrians and bicyclists when there is sun glare, especially at certain times of day and times of year.

- Consider the weather conditions when you are driving, walking, and biking and how you might need to adjust to conditions.
- Be mindful of the surface conditions in addition to the weather where a road may still be slick due to a recent rain shower that has left the area.
- Consider the risks of exiting your vehicle within the roadway environment or standing in the roadway where there is potential for being hit by another motor vehicle.
- Observe general safety rules on the roadway (as evidenced by the words “fell” and “ran”).
- Males are a key demographic group to try to reach.
- Follow right-of-way rules, specifically as they involve people who walk and bike.