Research Article

TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Transportation Research Record
1–14
© National Academy of Sciences:
Transportation Research Board 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/03611981211002523
journals.sagepub.com/home/trr
SAGE

Topic Models from Crash Narrative Reports of Motorcycle Crash Causation Study

Subasish Das[1], Anandi Dutta[2], and Ioannis Tsapakis[1]

Abstract

The Motorcycle Crash Causation Study (MCCS) is a matched case-control study that contains a very wide list of crash contributing factors associated with motorcycle crash occurrences. It contains information such as motorcycle information, rider information, and associated trip information. This study also provides crash narrative information that presents an in-depth narrative discussion of the crash causation. Because of the plethora of information, it is critical to investigate MCCS-related data. Some studies examined the structured information in MCCS datasets. There is no in-depth study that has examined the unstructured textual contents in the MCCS data. This study aims to mitigate this research gap by applying different natural language processing tools (e.g., text mining, topic modeling). Fatal and non-fatal crash narratives are clustered separately to gain insights pertaining to the injury level. The findings of this study will contribute to the ongoing studies on MCCS to better understand the crash causation mechanism associated with motorcycle crashes.

Motorcycle crashes and fatalities remain a significant public health problem as fatality rates have increased substantially in comparison with other vehicle types in the United States. Motorcycles account for only 3% of the total number of vehicles but amount to around 14% of the total crashes. In 2017, 5,172 motorcyclists (around 17% of all traffic fatalities) were killed in traffic crashes. Human errors account for 94% of the total crashes, and the rest are because of environmental and vehicle-related factors (*1*).

The Motorcycle Crash Causation Study (MCCS) is one of the large-scale studies sponsored by the National Highway Traffic Safety Administration (NHTSA). This study collected an enormous amount of data on motorcycle crashes through rigorous post-crash inspections and control motorcycle observations and interviews performed in Orange County, CA. The database delivers data from 351 injury crashes and 702 paired control observations. This study collected detailed information on rider/motorcycle, passenger, and other vehicles. Some of the critical variables are driver demographics, environmental contributors, crash contributing factors, relevant variables associated with motorcycles and other vehicles, information on injury levels, and information on clothing/helmet. The study also collected observations of the crash, control riders, passengers, and motorcycles and other vehicles involved in the crashes. Additionally, this dataset provides detailed crash narrative reports (*2*). Analysis of key causal factors for motorcycle crashes is often difficult, given the lack of detailed information in the police-reported crash information (*3–5*). An on-scene in-depth investigation represents an ideal setup for acquiring a large set of variables. This dataset provides the researchers with a grand opportunity to understand the causal issues associated with traffic crashes. As MCCS data is not very old, a limited number of studies have been conducted.

The current study aims to answer three key research questions (RQ): RQ1: Is MCCS data representative when comparing with national statistics? RQ2: What are the key insights from the detailed crash narratives? RQ3: How can the new insights help in improving motorcycle safety? The current study conducted a comparison between the key variables that are included in both MCCS and General Estimates System (GES) databases to answer the first research question. Answering this question is important to show the representativeness of

[1] Texas A&M Transportation Institute, San Antonio, TX
[2] Department of Computer Science, University of Texas at San Antonio, San Antonio, TX

Corresponding Author:
Subasish Das, s-das@tti.tamu.edu

MCCS data. This study collected all MCCS crash narrative reports and documented them in a structured dataset to utilize different natural language processing (NLP) tools to explore the knowledge behind the large, unstructured set of textual data. As crash narrative reports are usually less examined, an in-depth analysis of MCCS crash narrative analysis can produce a framework of such analysis to explore hidden insights in these reports. Analysis of the MCCS crash narrative helps in answering the last two research questions.

Literature Review

The literature review is divided into three major sections: crash causation studies, motorcycle-related safety studies, and crash narrative studies.

Crash Causation Studies

To examine motorcycle crash risk factors, Chawla et al. employed the data recently made available by the Federal Highway Administration MCCS (*3*). Logistic regression models are estimated to identify the vehicle and rider characteristics associated with motorcycle crashes. The results suggest that motorcycle crash risks are related to the physical status, age, and education level of the rider. In addition to these factors, multiple risk factors that the rider has the ability to adjust were found to be significantly associated with motorcycle crash risk. These adjustable risk factors include speed, type of motorcycle, helmet coverage, trip destination, motorcycle ownership, and traffic violation history. These factors may be associated with the riders' inclination to take risks.

Wali et al. analyzed 321 motorcycle injury crashes from the MCCS data (*6*). These were all considered nonfatal injury crashes that represented the vast majority (82%) of motorcycle crashes. An anatomical injury severity scoring system, termed as Injury Severity Score (ISS), was analyzed and provided an overall score by accounting for the potential of multiple injuries to various body parts on a rider.

Note that there were two other major crash causation studies, which are worth mentioning here. The National Motor Vehicle Crash Causation Study (NMVCCS) was an extensive data collection to determine the underlying reasons for crashes, such as roadways, vehicles, drivers, and environmental factors, in addition to critical precrash events (*7*). The NMVCCS data collected between 2005 and 2007 at crash scenes were analyzed by Choi to identify intersection-related crashes (crashes that had critical pre-crash events coded as crossing over, turning right, or turning left at an intersection) factors (*8*). With the Large Truck Crash Causation Study (LTCCS) database, Hallmark et al. investigated large truck lane departure crash contributing factors (*4*).

Motorcycle-Related Safety Studies

Kostyniuk et al. examined the trends and patterns of motorcycle crashes in 2005 with the objective to augment motorcycle-related fatalities in Michigan (*9*). The analysis of the crash patterns suggested aging to be the leading factor of crash fatalities. Similarly, Ryb et al. concentrated on the impact of advanced age on the consequence of injured individuals in a Maryland dataset linking hospital discharge documents and police reports (*5*). This study examined whether this mortality rise reflected alterations in accident or case fatality levels in any specific age category. Medina et al. conducted a 39-segment highway review method using correlation, analysis of variance, and multiple regression analyses, along with a motorcycle drivers' study (*10*). Results of the research suggested that the primary road components connected with motorcycle collision rates were the type and width of cross-section, junction density, published speed limit, presence of on-street parking, pavement faults, and housing development.

Eustace et al. used Ohio crash information from 2003–2007 to explore the likelihood of a motorcyclist being seriously wounded in a collision and the concerning risk variables (*11*). The findings identified several key factors such as female rider, speeding behavior, drug or alcohol impairment, not using a helmet, segment-related single motorcycle crashes, and curved segments. Chen and Fan developed a multinomial logit (MNL) model to investigate and identify significant contributing factors to estimate the pedestrian-vehicle crash severity in North Carolina, United States (*12*). The results portray the factors that significantly increase the probability of fatalities and disabling injuries include the driver's physical condition, pedestrian age, weekend, light condition, and speed limit.

Since motorcycles do not have an attached occupant room, lowering the number of accidents is key to decreasing accidents and deaths. Safety training is one countermeasure to reduce motorcycle accidents. For drivers of all ages, three states currently demand driver instruction: Florida, Maine, and Rhode Island. Furthermore, there are also 16 states that require permits for applicants under a specified age to be able to drive. The primary purpose was to determine whether driver-education-requiring states have a lower risk of motorcycle collision for those subject to the obligation than the states without a necessity. In addition, Shaheed and Gkritza used a latent class strategy to explore variables influencing collision seriousness resulting in one-vehicle motorcycle accidents using motorcycle crash data from 2001 to 2008 in

Iowa (*13*). The findings of the assessment indicated an important connection between the serious outcomes of collision injury and crash-specific variables such as speed, run-off highway, helmet-free riding, impaired riding, and many more. The findings of the model fit and analysis highlighted the need for collision segmentation and proved the latent class strategy to be a successful instrument for modeling the severity outcomes of a motorcycle crash.

Using a full Bayesian formulation, Cheng et al. developed five models representing different correlations of weather conditions on motorcycle crash injuries commonly experienced in crash data, and compared fitness and performance at four different levels of severity (*14*). Results exhibited that the designs with parameter differences in series and severity had a superior fit and precise forecast of crashes. Farid et al. conducted a crash severity analysis of motorcycle crashes on low-volume roadways (*15*). If all other conditions are the same, speed and impairment are associated with a high likelihood of motorcycle crashes with severe injuries. Das et al. applied deep learning on 5 years (2010–2014) of Louisiana at-fault motorcycle rider-involved crashes (*16*). The prediction accuracy of the test data was 94%, which is very high precision when compared with statistical model precision outcomes.

Crash Narrative Studies

Crash narratives provide an unstructured form of crash information. With many narrative crash reports, manual interpretation is not feasible. Text mining methods, a very effective tool in exploring unstructured textual data, can identify trends, insights, and anomalies in complex textual data with limited efforts. NLP tools such as text mining have been widely used in transportation research (*17–29*). Many transportation studies applied different text mining tools to explore insights from unstructured crash narrative data (*30–42*). These studies generally targeted two research directions: (1) identify hidden trends from unstructured textual contents; and (2) classify crash type or severity type from the crash narrative texts. For example, Das et al. used pedestrian crash narrative reports to classify the collision types by applying several machine learning algorithms (*39*).

The literature review reveals that many crash narrative studies provide important insights, which are difficult to attain by analyzing conventional crash databases. It is also found that text mining on MCCS crash narrative reports has not been conducted yet. This gap infers that there is a need for further research and an in-depth investigation of MCCS data. This study aims to identify the trends from crash narratives by applying several NLP tools to mitigate the current research gap.

Methodology

Data Collection and Analysis

The first research question of this study examines whether MCCS data is representative enough to represent motorcycle crash data. This study collected MCCS data from the Highway Safety Information System (HSIS) platform (https://www.hsisinfo.org/index.cfm). MCCS data contains a limited number of crashes with an extra-ordinary range of information on several key features. It is important to compare the distributions of the key variables of MCCS with an established dataset. As MCCS contains all injury levels, a comparison between MCCS and the Fatality Analysis Reporting System (FARS) will not be appropriate. Another comparative data is National Automotive Sampling System (NASS) dataset GES. GES data is generated from a nationally representative sample of police-reported traffic crashes of all injury levels. Table 1 shows the distribution of the key attributes by MCCS and GES (2001–2015). Alcohol impairment crashes are overrepresented in MCCS. It is interesting that cloudy weather crashes are disproportionately high in MCCS crashes. The percentages of different roadway types in MCCS are not represented when compared with GES. Younger riders are disproportionately high in representation in the MCCS dataset. The proportion of segment level crashes is 0.5 when compared with GES data. The descriptive statistics show that MCCS data is somewhat representative of GES data.

This study collected the crash narrative data in pdf formats. The data contains information based on the following key information:

- Crash number: identification of the motorcycle (351 cases)
- CrashType; type of crash
- MC1: details of the motorcycle/rider
- OV1: details of other vehicle 1
- OV2: details of other vehicle 2
- RiderInj: details of rider injury
- ODriverInj: details of other vehicle driver/occupant injury
- Human: details of human-related errors and other issues
- Vehicle: details of vehicle(s)
- Environmental: details on crash environment
- Crash_Narr: detailed crash narrative.

This study used manual effort to transfer the information in the pdf into a spreadsheet. To understand the amount of textual information for each crash, an example fatal crash report is shown in Table 2 (all person-level information was removed).

Table 1. Comparison between Motorcycle Crash Causation Study (MCCS) and General Estimates System (GES) Motorcycle Crashes

Variable category	MCCS (2015)	GES motorcycle crashes (2001–2015)	Variable category	MCCS (2015)	GES motorcycle crashes (2001–2015)
Day of week			71–75	0.00	0.80
Fri/Sat/Sun	51.20	50.70	> 75 mph	2.00	11.7
Mon/Tue/Wed/Thu	48.80	49.30	*Number of lanes*		
Number of vehicles involved			One	3.70	2.90
None	24.20	44.30	Two	40.20	49.70
One	68.40	51.80	Three	25.60	10.00
Two	6.80	3.30	Four	16.80	12.20
Three	0.00	0.40	Five	7.70	5.20
Four	0.60	0.10	Six	3.10	1.20
Five or more	0.00	0.00	Seven, eight	2.30	0.50
Lighting			Others	0.60	18.40
Daylight	68.10	72.20	*Roadway type*		
Dusk	5.40	3.20	One-way	2.00	2.10
Dark, lighted	24.50	14.90	Two-way, divided, no median barrier	53.60	11.40
Dark, not lighted	1.40	8.3	Two-way, divided, with median barrier	2.90	14.10
Dawn	0.60	0.90	Two-way, undivided	31.60	48.20
Weather condition			Two-way, with a continuous left-turn lane	9.70	6.40
Clear	50.40	83.10	Other, specify	0.30	17.90
Cloudy	36.80	12.60	*Rider age*		
Drizzle	0.90	3.20	20 or under	8.30	7.00
Overcast	9.70	0.30	21–25	22.20	13.60
Not Reported	2.30	0.40	26–30	16.50	11.50
Other and unknown	0.00	0.40	31–35	8.80	8.80
Intersection type			36–40	6.80	8.90
Not at intersection	30.20	61.20	41–45	6.80	8.80
Four-leg intersection	29.06	20.20	46–50	9.70	9.90
T intersection	19.94	11.80	51–55	5.70	9.00
Y intersection	1.14	0.40	56–60	7.70	7.90
Roundabout; traffic circle	0.00	0.40	61–65	3.70	5.00
Alley/driveway	19.64	0.20	66–70	2.50	2.80
Multi-leg intersection	0.00	5.80	71–75	0.30	1.20
Posted speed limit			76–96	0.00	0.50
1–25 mph	9.40	11.00	Unknown	1.00	5.10
26–30 mph	2.90	7.60	*Rider gender*		
31–35 mph	11.10	20.50	Male	95.40	91.40
36–40 mph	21.40	9.20	Female	4.60	7.40
41–45 mph	36.20	16.20	Not reported	0.00	1.20
46–50 mph	9.10	2.60	*Alcohol involvement*		
51–55 mph	4.80	11.20	No	41.90	85.70
56–60 mph	1.10	2.00	Alcohol use; combined alcohol and drug	16.00	6.20
61–65 mph	2.00	5.10	Drug, medication	8.80	0.00
66–70 mph	0.00	1.50	Other	1.10	3.50

Topic Model

Latent Dirichlet allocation (LDA), a popular topic modeling method, can discover underlying topics from large text datasets (*43*). The modeling framework is designed in the form of probability distributions over topics and related words. A brief overview of the fundamental concepts of LDA is explained here.

Consider $D = \{d_1, d_2, ...d_M\}$ is a collection of M documents, with consideration of each document d_i having N words. The framework of LDA considers each document d_i as a topic mixture θ^i over T topics, which are characterized by vectors of word probabilities $\Phi^1, \Phi^2, ..., \Phi^T$. Consider topic word probability is as Φ and Dirichlet parameter is as α. LDA undertakes the following generative process for a document (*44*):

1. Sample a topic mixture vector $\theta \sim Dirichlet(\alpha)$, where $Dirichlet(.)$ is a function that generates values for random variables θ following the Dirichlet distribution.
2. Sample a topic index $t \sim Categorical(\theta)$, where t is a random integer from 1 to T, and then, sample a

Table 2. Example Crash Narrative Data

Variable	Details
CaseID	xxxx
CrashType	Motorcycle vs Fixed Object with Secondary Lay-Down Event. (on-scene investigation/on-scene photographs)
MV1	XXX Honda XXX
OV1	NA
OV2	NA
RiderInj	Both the rider and passenger sustained fatal injuries. The rider was pronounced dead at the scene. The passenger was taken to a trauma center where she died during surgery. Both were impaired from alcohol consumption according to the coroner's report.
ODriverInj	NA
Human	The XXX yr. old rider was driving while impaired and at a speed which was too fast for the roadway configuration.
Vehicle	The case vehicle involved was found to have extensive post-crash damage. It did appear to have been well maintained prior to the crash. The tires and brake pads looked to be near new. The post-crash investigation of the vehicle showed that the overall pre-crash mechanical condition was good.
Environmental	There were no environmental factors that were determined to be linked to the crash causation.
Crash_Narr	This is a single motorcycle crash with two fatalities occurring in XXX at XXX hrs. in XXX. It was dark, the weather was overcast, and the temperature was approximately 57 degrees F. The asphalt roadway was dry at the time. It was well worn and polished, but in good repair. The wind was light. The crash occurred at the entrance to a three-leg intersection that had no signal control devices. The surrounding area is comprised mostly of residential structures on the riders? right and a fenced in field to the left. The roadway in question at the crash scene runs in an east/west configuration and consists of four through lanes, two in each direction. In the opposite direction there is a left turn lane leading to the intersection which has a raised concrete divider both on the lanes left side and at its terminus, protecting the lane from on-coming traffic. The roadway has a right curve prior to the first harmful event, and then curves left before going straight. The roadway has concrete curbs on both sides, and a sidewalk on the Case Vehicles? right. There are overhead streetlights that were functioning properly. The posted speed limit is 45 mph. The side street is made of asphalt and in good condition. It has raised concrete curbs and no sidewalks. Case Vehicle was a XXX Honda XXX, predominately yellow in color. It was being operated by XXX yr. XXX and had a XXX yr. old XXX passenger. Both were wearing full-face helmets that appear to not have been strapped. Case vehicle was proceeding in a XXX direction at a very high rate of speed, (80–90 mph) when it failed to negotiate a right curve and sideswiped the center divider. Case vehicle then moved laterally and sharply to the right, striking the right curb. Case vehicle then continued XXX while making prolong contact with the curb, and fully ejected both the rider and passenger. The Case vehicle made a glancing contact with a 24 diameter wooden utility pole, and then continued in a WB direction for another 300 before coming to its FRP. When both occupants of the Case vehicle were ejected, they tumbled/slid on the sidewalk and then impacted into the utility pole. After which, they were deflected a few feet to the right and into a block wall and utility vent pipe where they came to rest.

Note: XXX = redacted information.

word $v \sim Categorical(\Phi_t)$, where v is an integer from 1 to V corresponding to the index of word w_v in the vocabulary, and $Categorical(.)$ generates values of nominal variables t and v. The process repeats N times for each word in document d_i.

Results and Discussions

Text Mining

Before performing text mining and topic modeling, several basic steps of data cleaning were performed. Stop words, redundant words, numbers, and punctuations are removed initially. Besides, this study used additional steps influenced by Zipf's law, including removing words that only occur once (*45*, *46*). Lemmatization uses vocabularies and morphological assessments to eradicate the inflectional endings of a word and convert it into its dictionary form. This study performed both stop word removal and lemmatization to clean the data for analysis.

Figure 1 illustrates the top 20 most frequent keywords based on properties of the words such as adjectives, verbs, noun phrases, and Rapid Automatic Keyword Extraction algorithm (RAKE) measure. The top key adjectives are "left," "right," "old," "other," "clear," "east," "rear," "southern," "front," and "unknown." The top verbs are "occurred," "has," "left," "wearing," "stated," "was," "are," "is," "stopped," and "were." Top noun phrases are "left turn," "roadway surface," "speed limit," "southern California," "right side," "crash site," "1 lane," "#1 lane," "turn lane," and "old male." A domain-independent, unsupervised, and language-

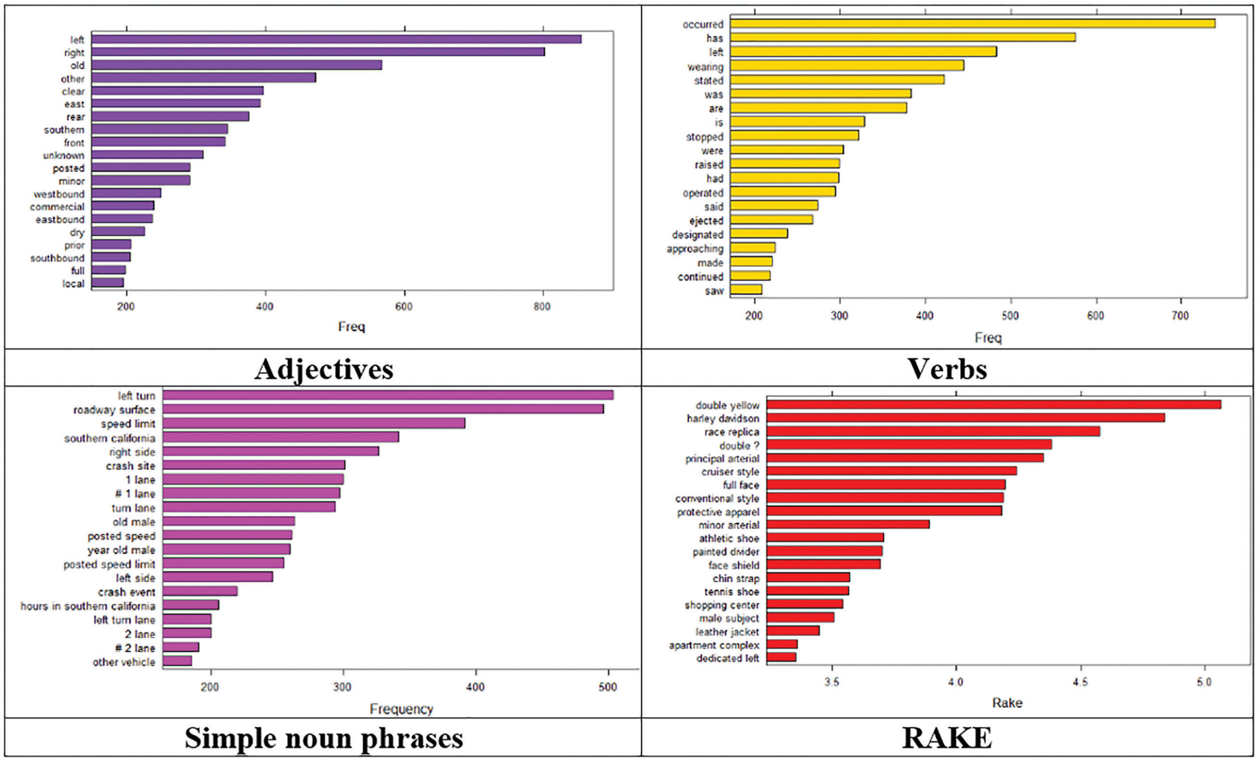


Figure 1. Bar plots of word frequency measures.
Note: RAKE = rapid automatic keyword extraction.

independent method for obtaining keywords from individual documents is known as RAKE (*47*). By defining its text into a set of candidate keywords, RAKE begins the keyword extraction on a document. First, the document text is categorized into an array of words by the specified word delimiters. This array is then separated into sequences of contiguous words at stop word positions and phrase delimiters. Words placed within a sequence are assigned the same position in the text and then considered a candidate keyword. Extracted keywords do not contain interior stop words because RAKE splits candidate keywords by stop words. Because of its ability to extract highly specific terminology, a strong interest in identifying keywords that contain interior stop words, such as the axis of evil, was expressed. RAKE searches for pairs of keywords that attach one another in the same order and at least twice in the same document to find the keywords. A combination of those keywords and their interior stop words creates a new candidate keyword. The sum of its member keyword scores is the score for the new keyword. Top RAKE keywords are “double yellow,” “Harley Davidson,” “race replica,” “double,” “principal arterial,” “cruiser style,” “full face,” “conventional style,” “protective apparel,” and “minor arterial.”

Co-occurrence of Terms

The co-occurrence of terms is another measure to understand the patterns from unstructured news report narratives. This study used the pipeline developed in “udpipe” to perform the co-occurrence analysis from the whole corpus (i.e., collection of texts or documents) (*48*). Some of the key co-occurrences of terms are described below (see Figure 2):

- At least seven clusters are visible with more than 300 co-occurrences (referred to as ‘cooc’ in Figure 2).
- A link is visible (top left) around the words “injury,” “straight,” “major,” “asphalt,” “surface,” “non-freeway,” “arterial,” and “principal.” The co-occurrence of “non-freeway,” “arterial,” and “surface” is 400.
- A dark line is present (top left) around the words “male,” “female,” “Caucasian driver,” and “year old.” These terms are also linked with the node associated with “mc1” and “rider.” MC1 indicates the rider of the first motorcycle.
- At the bottom, a large cluster is visible. The dark nodes (representing more than 500 co-occurrences) are seen linking words such as “lane,” “2,”

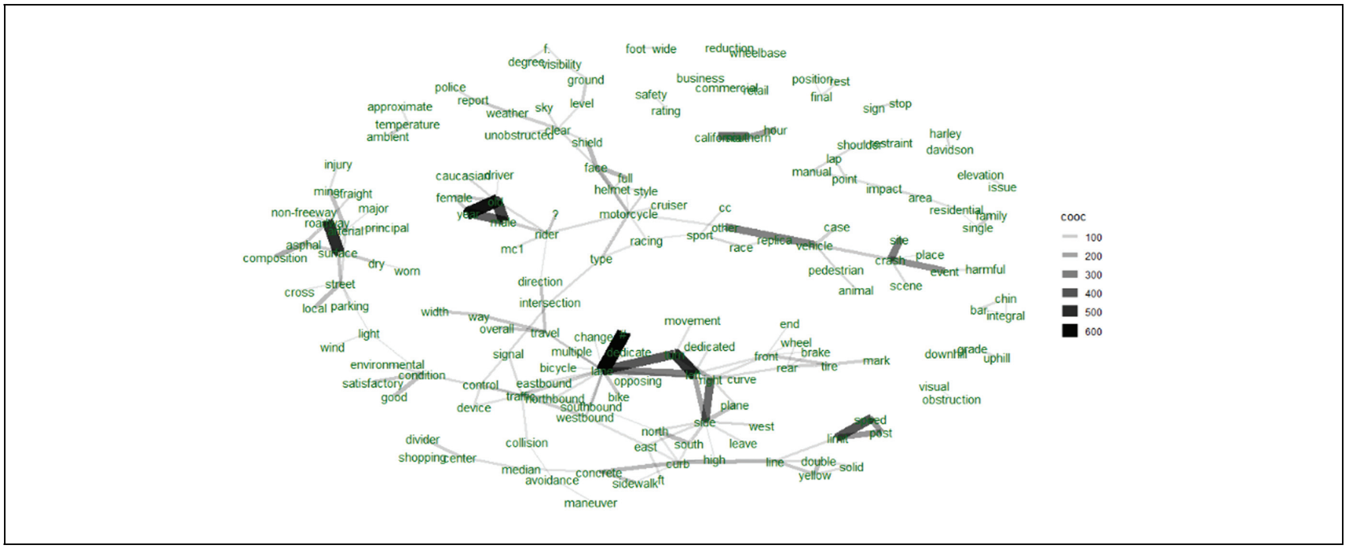


Figure 2. Co-occurrence plots of the top keywords.
Note: COOC = co-occurrences.

"dedicated," "left," "right curve," "side," and "plane." This cluster is also associated with motorcycle-condition-related keywords such as "wheel," "brake," and "tire." This particular cloud is closely associated with the causation factors that are detailed in the narratives.
- In the bottom, another cluster is visible that shows a link between "concrete," "sidewalk," "curb," "yellow," "solid," "line," and "posted speed limit."

The co-occurrence plots provide a broad overview of some of the key co-occurrences in MCCS crash narratives.

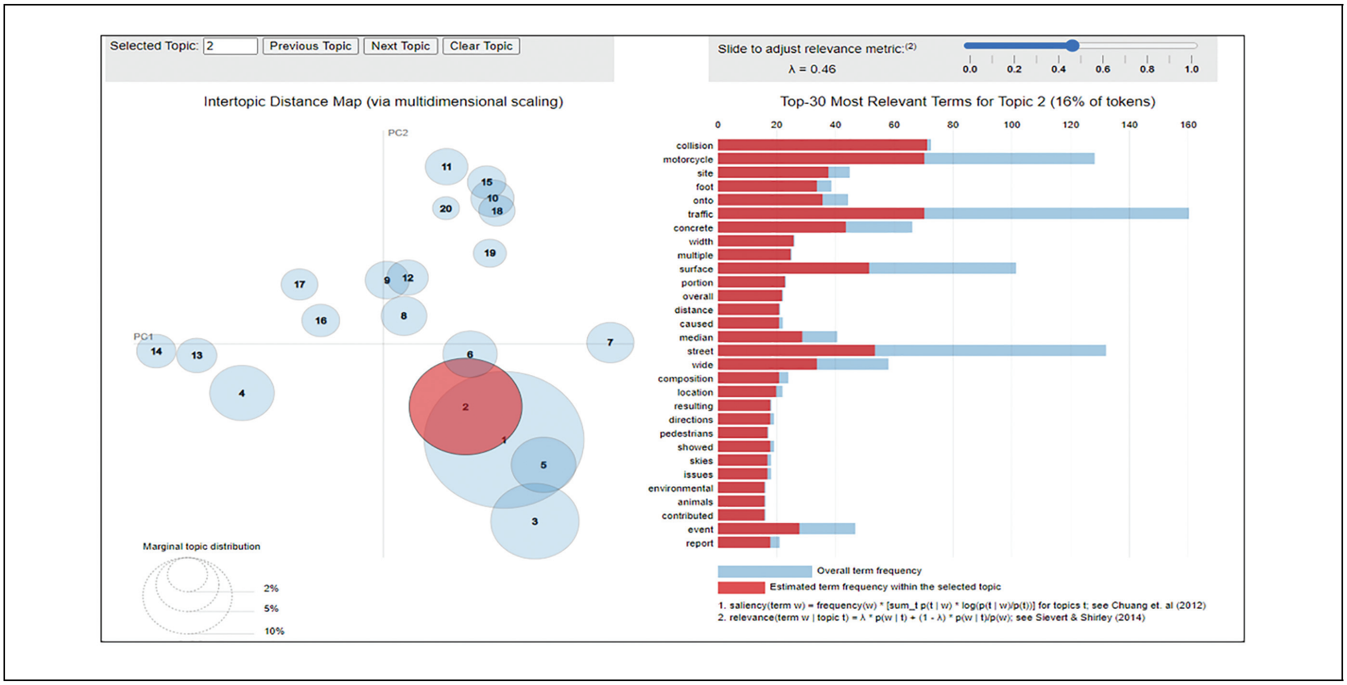
Visualizations of LDA Outcomes

Multidimensional Scaling Approach (Interactive Visualization). LDA is normally applied to thousands of documents, representing combinations of dozens to hundreds of topics which are modeled as distributions across thousands of terms. To mitigate these challenges, interactivity is the best technique to create LDA visualizations. Interactivity is a basic technique that is both compact and thorough. In this study, the LDAvis package was employed to develop interactive LDA models (*49*). Figures 3 and 4 illustrate interactive visualization screenshots of two topic models: 1) corpora based on fatal-crash-related crash narratives, 2) corpora based on non-fatal-crash-related crash narratives. The research team developed two web tools to demonstrate these interactive plots. The plots are comprised of two sections:

- The left section of the visualization represents a global perspective on the topic model. The topics are plotted as circles in the two-dimensional plane. By computing the distance between topics and projecting the inter-topic distances onto two dimensions using multidimensional scaling, the centers of these circles are placed in the visualization. Each topic's overall prevalence is then encoded using the areas of the circles to allow the research team to sort the topics in decreasing order of prevalence.
- The right section displays a horizontal bar chart. The bars represent the individual terms that are the most useful for interpreting the topics on the left, based on which topic is currently selected. This allows users to comprehend the meaning of each topic. A pair of overlaid bars represent both the corpus-wide frequency of a given term and the topic-specific frequency of the term.
- Both sections of this visualization are linked. When the user selects a topic (on the left), the visualization highlights the most useful terms (on the right) for interpreting the selected topic.

LDA Topic Models in Word Clouds. Open source R package "tidytext" was used to develop the topic models (*46*). MCCS comprises of 351 motorcycle crashes with extensive information. The crash reports contain narratives on four key features:

- Human-related information
- Vehicle-related information
- Rider injury information
- Details of crash reports.

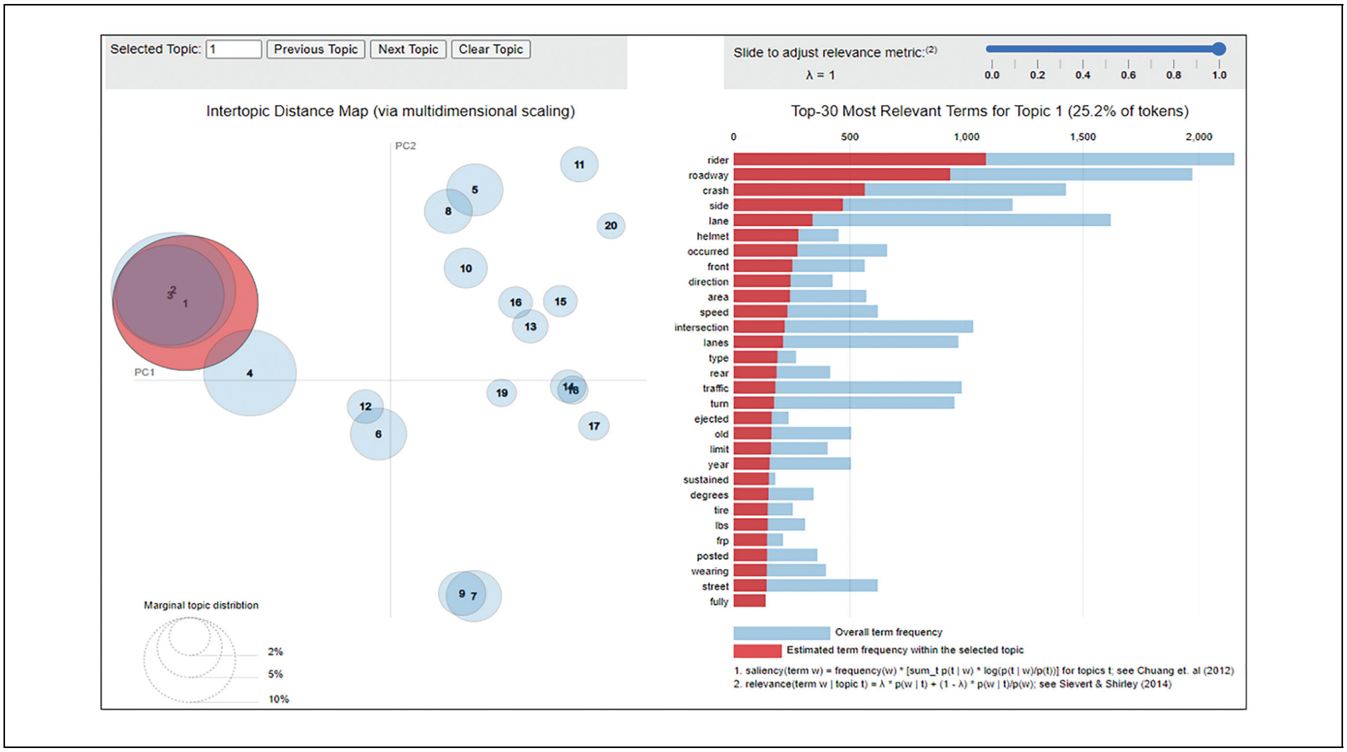


Selected Topic: 2

λ = 0.46

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3. Multidimensional scaling using fatal crash corpus.



Selected Topic: 1

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 4. Multidimensional scaling using non-fatal crash corpus.

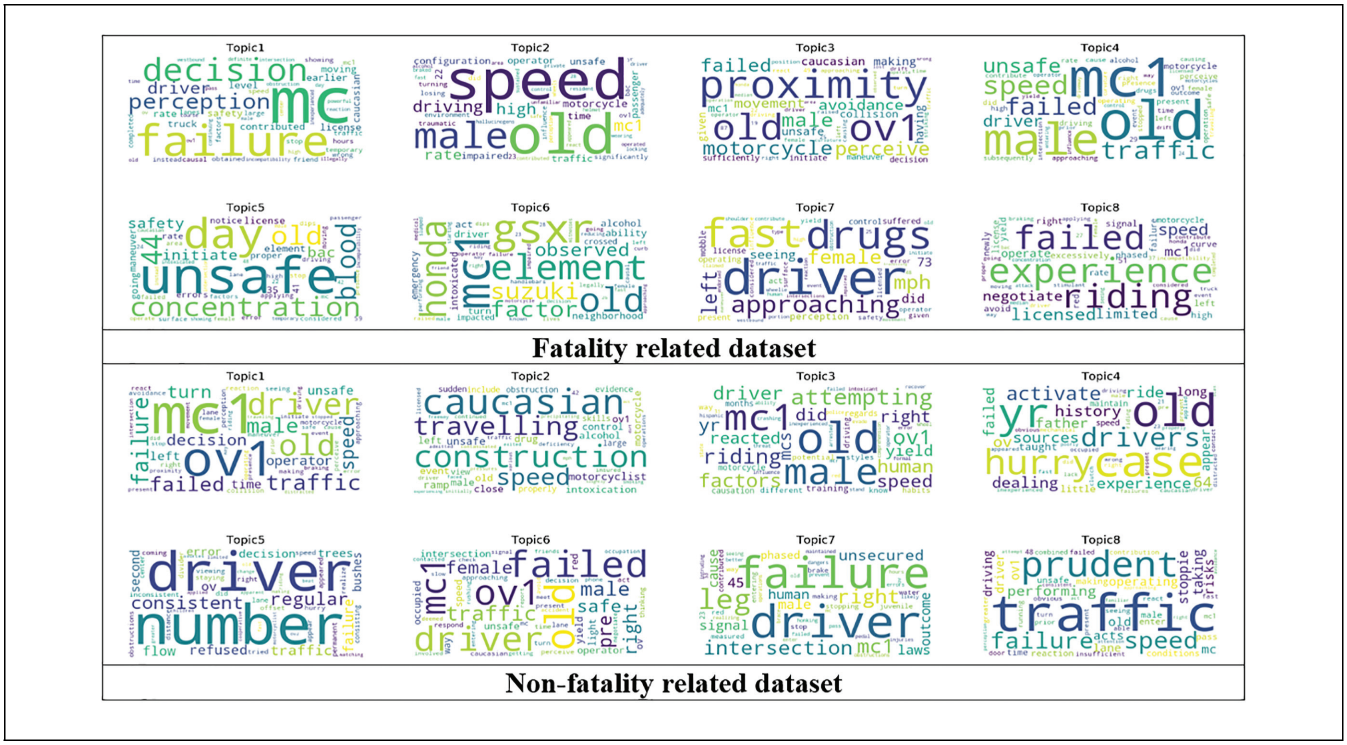


Figure 5. Topic models based on rider-related information.

Out of 351 crashes, 40 are fatal crashes, and 311 crashes are non-fatal. Based on the severity types and crash narrative categories (listed above), several topic models have been developed. The analysis has been conducted based on these eight corpora (i.e., group of texts or documents).

Figure 5 shows word clouds from topic models based on rider-related information. The figure shows eight topics for datasets with fatal crashes and eight topics for datasets with non-fatal crashes. For the fatal human cluster, “old” (Topic 2, Topic 3, Topic 4, Topic 5, and Topic 6) and “failed” (Topic 3, Topic 4, and Topic 8) are featured in many of the topics. Other prominent words include “male” (Topic 2, Topic 3, and Topic 4), “speed” (Topic 2, Topic 4, and Topic 8), “mc1” (Topic 4 and Topic 6), “unsafe” (Topic 5), “failure” (Topic 1), “experience” (Topic 8), and “proximity” (Topic 3). The cluster for non-fatal human featured many of the same words from the fatal human cluster. The words that were present in the greatest number of topics are “driver” (Topic 1, Topic 3, Topic 5, Topic 6, and Topic 7), “mc1” (Topic 1, Topic 3, Topic 6, and Topic 7), “old” (Topic 1, Topic 3, Topic 4, and Topic 6), and “failed” (Topic 1, Topic 4, and Topic 6). Other prominent words in the non-fatal crash cluster include “failure” (Topic 1, Topic 7, and Topic 8), “ov1” (Topic 1 and Topic 6), “number” (Topic 5), “traffic” (Topic 8), “Caucasian” (Topic 2), and “prudent” (Topic 8).

Figure 6 shows word clouds from topic models based on vehicle-related information. The figure shows four datasets with fatal crashes and eight topics for datasets with non-fatal crashes. For the fatal vehicle cluster, “damage” (Topic 1, Topic 3, Topic 4, and Topic 8), “condition” (Topic 1, Topic 3, Topic 4, and Topic 5), and “mc1” (Topic 5, Topic 6, and Topic 8) are featured in many of the topics. Words with similar meanings were also prominently present, such as “contribution” (Topic 2), “contributed” (Topic 5), and “contributory” (Topic 6). Other words like “pressure” (Topic 6 and Topic 8), “post” (Topic 1), “pads” (Topic 1), “California” (Topic 2), “recommended” (Topic 2), “causal” (Topic 7), “mc” (Topic 4), and “factors” (Topic 7) are prominent as well. The cluster for non-fatal vehicles had similar words like “condition” (Topic 1, Topic 2, and Topic 4), “broken” (Topic 3 and Topic 6), “sustained” (Topic 3 and Topic 4), and “damage” (Topic 1 and Topic 4). A few other prominent words in the non-fatal vehicle clouds include “major” (Topic 1 and Topic 3), “mc1” (Topic 4 and Topic 6), “steering” (Topic 1), “causation” (Topic 5), “factors” (Topic 5), “headlight” Topic 7), and “controlled” (Topic 7).

Figure 7 shows word clouds from topic models based on rider injury information. The figure shows eight topics for datasets with fatal crashes and eight topics for datasets with non-fatal crashes. For the fatal rider cluster, the

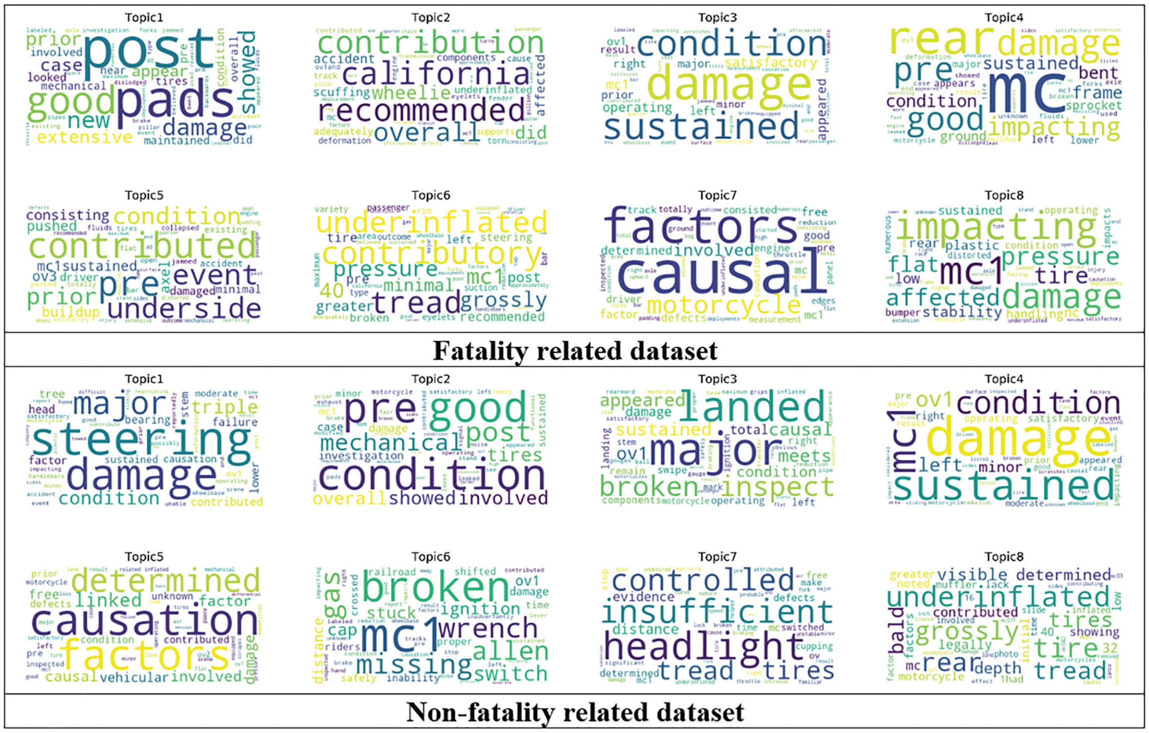


Figure 6. Topic models based on vehicle-related information.

words “injuries” (Topic 4, Topic 5, Topic 7, and Topic 8), “fractures” (Topic 4, Topic 6, and Topic 7), “sustained” (Topic 4, Topic 5, and Topic 8), “specified” (Topic 1 and Topic 2), and “right” (Topic 4 and Topic 6) were featured in many of the topics. Other visible words include “plate” (Topic 2), “fracture” (Topic 3), “operator” (Topic 5), “bilateral” (Topic 6), “contusion” (Topic 6), and “passed” (Topic 7). For the non-fatal rider cluster, “sustained” and “old” were visible in every topic. Other commons words include “male” (Topic 1, Topic 3, Topic 4, Topic 5, Topic 6, Topic 7, and Topic 8), “injuries” (Topic 1, Topic 3, Topic 4, Topic 5, Topic 7, and Topic 8), “left” (Topic 1, Topic 2, Topic 3, Topic 5, and Topic 7), “motorcycle” (Topic 2 and Topic 6), “score” (Topic 1), “treated” (Topic 2), and “test” (Topic 8).

Figure 8 shows word clouds from topic models based on crash reports. The figure shows eight topics for datasets with fatal crashes and eight topics for datasets with non-fatal crashes. For the fatal narrative, “lane” (Topic 1, Topic 2, Topic 3, and Topic 7) and “lanes” (Topic 2, Topic 5, Topic 6, Topic 7, and Topic 8) mean the same thing and are featured in several topics. Other prominent words include “mc1” (Topic 1, Topic 2, Topic 3, Topic 4, Topic 5, Topic 6, and Topic 7), “traffic” (Topic 1, Topic 2, Topic 3, Topic 7, and Topic 8), “left” (Topic 1, Topic 2, Topic 3, Topic 6, Topic 7, and Topic 8), “street” (Topic 1, Topic 3, and Topic 7), and “surface” (Topic 4, Topic 5, and Topic 8). The cluster for non-fatal narratives featured many of the same words such as “lanes” (Topics 1–8), “left” (Topic 1, Topic 2, Topic 3, Topic 5, Topic 6, Topic 7, and Topic 8), “mc” (Topic 2, Topic 3, Topic 5, Topic 6, and Topic 8), “motorcycle” (Topic 1, Topic 2, Topic 3, Topic 4, Topic 6, and Topic 8), and “intersection” (Topic 1, Topic 2, Topic 4, Topic 6, and Topic 7).

The key takeaways from Figures 5–8 are the following:

- *Rider-related information*: Failure, speed, unsafe, drug, and experience are key topics in fatal crash corpora. In non-fatal corpora, the keyword failure is present in four topics. Other latent topics are male, hurry, and Caucasian. The topic models clearly show that fatal crashes are associated with some of dominant factors. For non-fatal crashes, no dominant factor is visible.
- *Vehicle-related information*: The top topics of fatal crash corpora are post/pads, California, damage/sustained, rear, underside, tread (tire), causal factors, and tire damage. For non-fatal crash corpora, the key topics are steering, condition, inspect, damage/sustained, causation factors,

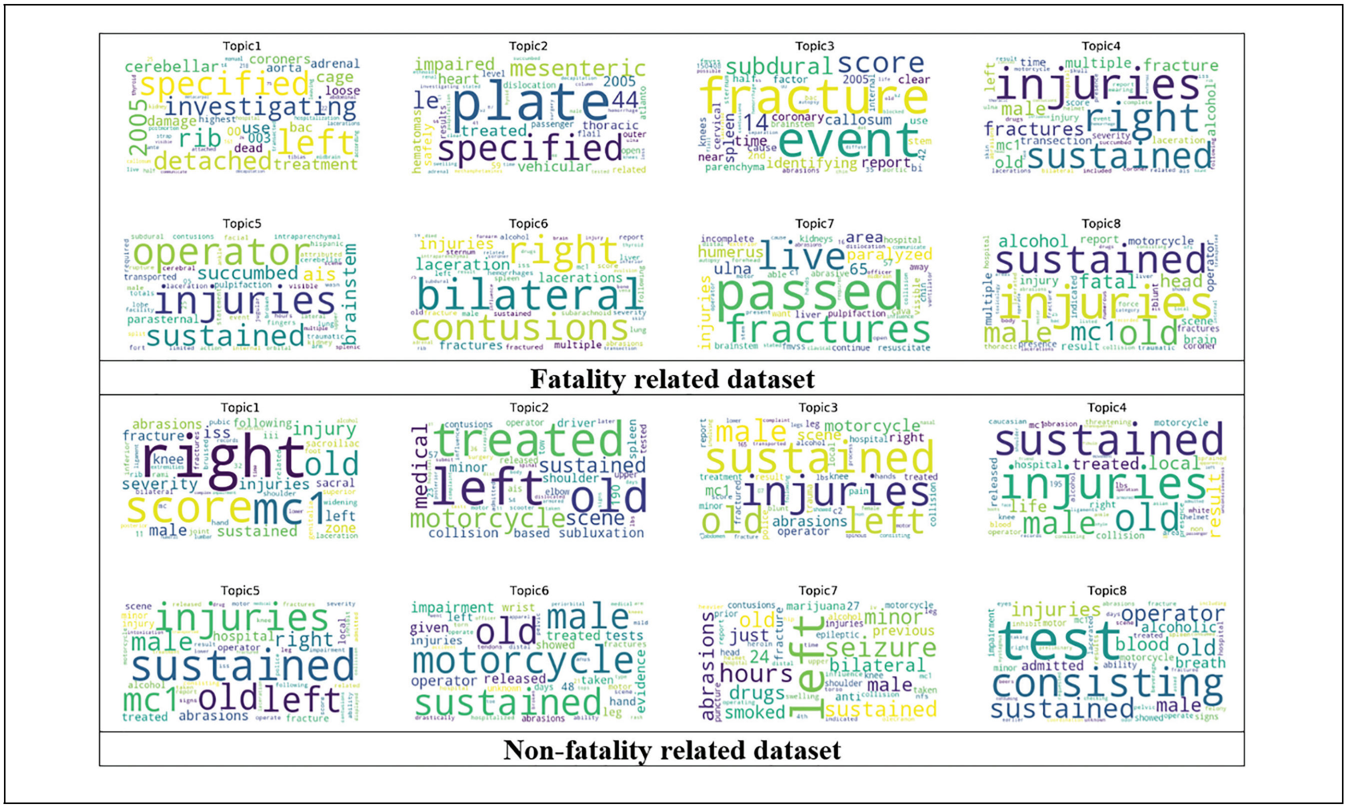


Figure 7. Topic models based on rider injury information.

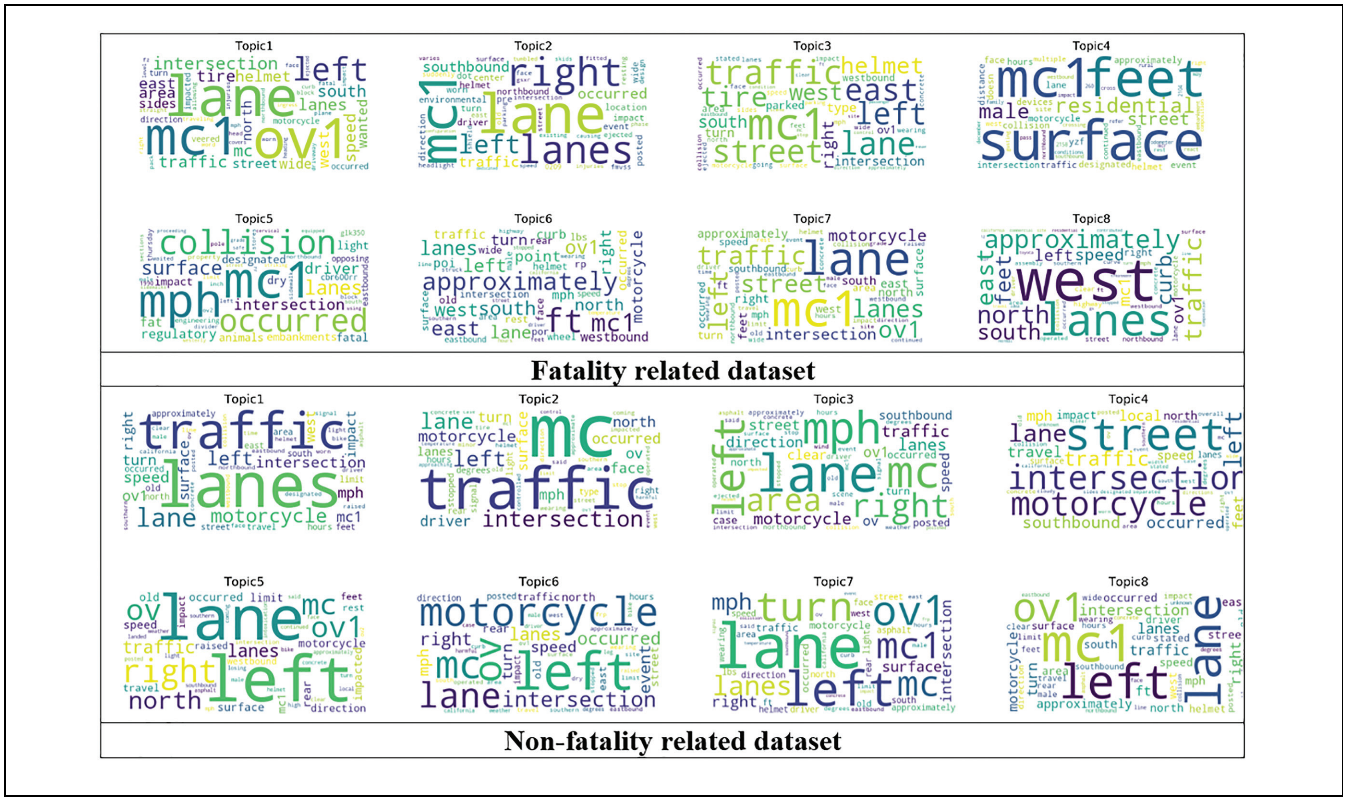


Figure 8. Topic models based on crash narrative reports.

broken, insufficient headlight, and inflated. Both fatal and non-fatal corpora provide different evidence of vehicle-related issues.
- *Rider injury information*: The top topics of fatal crash corpora are specified plate, fracture/event, injuries (three topics), bilateral contusions, and fractures. On the other hand, non-fatal contains topic such as injuries (three topics), male, seizure, and text. For both corpora, injury is a critical topic, which is obvious as the corpora are based on rider injury information.
- *Crash narrative reports*: For fatal crash corpora, lane is present in several topics. Other dominant topics are speed, residential, surface, and intersection. Lane keyword is also present in several topics in non-fatal corpora. Other latent keywords are intersection and turn. As crash narrative reports contain substantial information, there is a need for future research to comprehensively analyze each of the reports by clustering the reports into 20–30 subgroups based on the latent properties in the textual content.

Conclusions

MCCS data is comparatively new, and only a handful of studies examined the dataset to uncover new insights that can provide guidelines and suggest countermeasures to improve motorcycle safety. First, this study compared MCCS variable categories with GES variable categories to understand the representativeness of MCCS data to answer the first research question. The descriptive measures indicate that MCCS data is representative and it is worthy to explore the crash narrative reports to explore hidden insights in this dataset. To answer the second and third research questions, this study applied several NLP tools to present hidden insights from the unstructured text content. This effort has two major contributions:

- First, this study developed a structured database of MCCS crash narrative by converting the pdf-level information into a spreadsheet, which can be explored by other researchers in future.
- Second, this study developed a framework of NLP tools that can be replicated in other crash causation narrative investigations. As crash narratives provide a plethora of information, this study performed topic modeling into two broad text corpora based on two major injury levels (fatal and non-fatal). Investigations on both types of narratives identified some of the key traits for fatal and non-fatal motorcycle crashes.

This is one of a handful of research papers that addresses the narrative reports of the crash causation studies to reveal insights from the unstructured textual contents in crash narratives. Conventional crash data analytical methods have relied on raw statistics alone without considering the potential of crash narrative documents. Some of the key findings include:

- The co-occurrence analysis shows several risk clusters, such as crashes on curves on the right, inflated tire, crossing using dedicated left lane, brake failure, passing solid yellow line, and speeding over the posted speed limit.
- The highly representative keywords or risk factors in fatal crash reports are “unsafe speed,” “male,” “intersection,” “bilateral contusion,” and “fracture.”
- The highly representative keywords or risk factors in fatal crash reports are “make,” “broken,” “headlight,” “inflated,” “seizure,” and “lane.”

The findings of this study can help safety engineers in understanding the risk factors associated with these crashes. As the conventional crash countermeasures are mostly dedicated to vehicles and large vehicles, there is a need for significant efforts in designing motorcycle-specific countermeasures (i.e., barrier design for motorcyclists). As motorcyclists represent a small proportion of traffic exposure, suitable countermeasures can be applied to high-motorcycle-volume corridors.

This paper adds significant value to the ongoing studies of MCCS by providing additional contexts and new research scopes. The current study has limitations. First, the statistical significance test was not conducted while comparing MCCS and GES. Because of imbalanced data, the current study is limited to the comparison with percentage distribution levels. Future studies can aim to resolve the imbalanced data by exploring under-sampling, over-sampling, or cost-sensitive learning. Second, the current analysis is limited to only two broad groups: fatal and non-fatal crashes. In many cases, these two broad groups are not sufficient to identify intuitive information. As this analysis is limited to only MCCS data, future studies can collect data from different regions to examine the findings of the current study. As there are many cutting-edge machine learning and deep learning algorithms available, future researchers can explore the MCCS crash narrative data to solve several classification problems (i.e., crash severity types, helmet usage) using these innovative algorithms.

Author Contributions

The authors confirm the contribution to the paper as follows: study conception and design: S. Das; data collection: S. Das; analysis and interpretation of results: S. Das; draft manuscript

preparation: S. Das, A. Dutta, I. Tsapakis. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. National Highway Traffic Safety Administration (NHTSA). Traffic Safety Facts. Motorcycle: 2017. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812785MCCS. Accessed August 1, 2020.
2. Nazemetz, J. W., F. D. Bents, J. G. Perry, C. Thor, and Y. M. Mohamedshah. *Motorcycle Crash Causation Study: Final Report*. FHWA-HRT-18-064. FHWA, Washington D.C., 2019.
3. Chawla, H., I. Karaca, and P. T. Savolainen. Contrasting Crash- and Non-Crash-Involved Riders: Analysis of Data from the Motorcycle Crash Causation Study. *Transportation Research Record: Journal of the Transportation Research Board*, 2019. 2673: 122–131.
4. Hallmark, S., Y.-Y. Hsu, T. Maze, T. McDonald, and E. Fitzsimmons. *Investigating Factors Contributing to Large Truck Lane Departure Crashes Using the Federal Motor Carrier Safety Administration's Large Truck Crash Causation Study (LTCCS) Database*. Center for Transportation Research and Education, Iowa State University, 2009.
5. Ryb, G. E., P. C. Dischinger, J. A. Kufera, and T. J. Kerns; Association for the Advancement of Automotive Medicine (AAAM). Trends in Motorcycle Injuries, Deaths and Case Fatality Rates in the U.S.A. Population. *Annals of Advances in Automotive Medicine*, Vol. 53, 2009, 2 p.
6. Wali, B., A. J. Khattak, and N. Ahmad. Examining Correlations between Motorcyclist's Conspicuity, Apparel Related Factors and Injury Severity Score: Evidence from New Motorcycle Crash Causation Study. *Accident Analysis & Prevention*, Vol. 131, 2019: 45–62.
7. National Highway Traffic Safety Administration (NHTSA). *National Motor Vehicle Crash Causation Survey: Report to Congress*. Publication DOT HS 811 059. United States Department of Transportation, Washington, D.C., 2008, p. 47.
8. Choi, E.-H. *Crash Factors in Intersection-Related Crashes: An On-Scene Perspective: (621942011-001)*. Publication HS-811 366. American Psychological Association, 2010.
9. Kostyniuk, L. P., and A. D. Nation. *Motorcycle Crash Trends in Michigan: 2001-2005*. University of Michigan Transport Research Institute, Michigan, 2005, p. 50.
10. Medina, A. M. F., and J. C. T. Soto; Institute of Transportation Engineers (ITE). *Roadway Factors Associated with Motorcycle Crashes*. ITE, Washington, D.C., 2011.
11. Eustace, D., V. K. Indupuru, and P. Hovey. Identification of Risk Factors Associated with Motorcycle-Related Fatalities Ohio. *Journal of Transportation Engineering*, Vol. 137, No. 7, 2011, pp. 474–480.
12. Chen, Z., and W. (David) Fan. A Multinomial Logit Model of Pedestrian-Vehicle Crash Severity in North Carolina. *International Journal of Transportation Science and Technology*, Vol. 8, No. 1, 2019, pp. 43–52. https://doi.org/10.1016/j.ijtst.2018.10.001.
13. Shaheed, M. S., and K. Gkritza. A Latent Class Analysis of Single-Vehicle Motorcycle Crash Severity Outcomes. *Analytic Methods in Accident Research*, Vol. 2, 2014, pp. 30–38.
14. Cheng, W., G. S. Gill, T. Sakrani, M. Dasu, and J. Zhou. Predicting Motorcycle Crash Injury Severity Using Weather Data and Alternative Bayesian Multivariate Crash Frequency Models. *Accident Analysis & Prevention*, Vol. 108, 2017, pp. 172–180.
15. Farid, A., M. Rezapour, and K. Ksaibati. Modeling Severities of Motorcycle Crashes on Low-Volume Roadways. Presented at the 12th International Conference on Low-Volume Roads. Transportation Research Board, MT, 2019.
16. Das, S., A. Dutta, K. Dixon, L. Minjares-Kyle, and G. Gillette. Using Deep Learning in Severity Analysis of At-Fault Motorcycle Rider Crashes. *Transportation Research Record: Journal of the Transportation Research Board*, 2018. 2672: 122–134.
17. Das., S. Fatal Crash Reporting in Media: A Case Study on Bangladesh. Presented at 100th Annual Meeting of the-Transportation Research Board, 2021. (virtual).
18. Das, S. An Exploratory Analysis of Unmanned Aircraft Sightings using Text Mining. *Transportation Research Record: Journal of Transportation Research Board*, 2021. https://doi.org/10.1177/0361198120987230
19. Gu, Y., Z. S. Qian, and F. Chen. From Twitter to Detector: Real-Time Traffic Incident Detection Using Social Media Data. *Transportation Research Part C: Emerging Technologies*, Vol. 67, 2016, pp. 321–342.
20. Das, S., A. Dutta, and M. Brewer. Case Study of Trend Mining in Transportation Research Record Articles. *Transportation Research Record: Journal of Transportation Research Board*, 2020. 2674: 1–14.
21. Chen, F., and R. Krishnan. *Transportation Sentiment Analysis for Safety Enhancement. Technologies for Safe and Efficient Transportation*. US DOT University Transportation Center, 2013.
22. S., Das, and Dutta A. Twelve Years of Transportation Annual Meeting Hashtag: Implications for Networking and Research Trends. Presented at 100th Annual Meeting of theTransportation Research Board, 2021. (virtual). .
23. Das, S., X. Sun, and A. Dutta. Investigating User Ridership Sentiments for Bike Sharing Programs. *Journal of Transportation Technologies*, Vol. 5, 2015, pp. 69–74.
24. Sun, X., and S. Das. *User Sentiment Analysis with Louisiana Social Media Data for Better and Effective Crash Countermeasures*. Report No. 14-4TIRE. 2015.
25. Das, S., X. Sun, A. Dutta, and M. Zupancich. Twitter in Circulating Transportation Information: A Case Study on

Two Cities. Presented at 96th Annual Meeting of the Transportation Research Board, Washington D.C., 2017.

26. Zhang, Z., M. Ni, Q. He, J. Gao, J. Gou, and X. Li. Exploratory Study on Correlation between Twitter Concentration and Traffic Surges. *Transportation Research Record: Journal of Transportation Engineering*, 2016. 2553: 90–98.
27. Das, S., and A. Dutta. Characterizing Public Emotions and Sentiments in COVID-19 Environment: A Case Study of India. *Journal of Human Behavior in the Social Environment*, 2020. https://doi.org/10.1080/10911359.2020.1781015.
28. Das, S., Dutta A, Mudgal A., and Datta. S Non-fear-based Road Safety Campaign as a Community Service: Contexts from Social Media. In *Innovations for Community Services. I4CS 2020. Communications in Computer and Information Science* (S. Rautaray, G. Eichler, C. Erfurth, and G. Fahrnberger, eds.) Vol 1139. Springer, New Mexico, 2020.
29. Das, S., A. Dutta, G. Medina, L. Minjares-Kyle, and Z. Elgart. Extracting Patterns from Twitter to Promote Biking. *IATSS Research*, Vol. 43, 2019, pp. 51–59.
30. Bondy, J., H. Lipscomb, K. Guarini, and J. E. Glazner. Methods for Using Narrative Text from Injury Reports to Identify Factors Contributing to Construction Injury. *American Journal of Industrial Medicine*, 2005, 48: 373–380.
31. Bunn, T. L., S. Slavova, and L. Hall. Narrative Text Analysis of Kentucky Tractor Fatality Reports. *Accident Analysis & Prevention*, Vol. 40, 2008, pp. 419–425.
32. Williamson, A., A. M. Feyer, N. Stout, T. Driscoll, and H. Usher. Use of Narrative Analysis for Comparisons of the Causes of Fatal Accidents in Three Countries: New Zealand, Australia, and the United States. *Injury Prevention*, Vol. 7, 2001, pp. 15–20.
33. Chen, W., K. K. Wheeler, S. Lin, Y. Huang, and H. Xiang. Computerized "Learn-As-You-Go" Classification of Traumatic Brain Injuries Using NEISS Narrative Data. *Accident Analysis & Prevention*, Vol. 89, 2016, pp. 111–117.
34. Marucci-Wellman, H. R., M. R. Lehto, and H. L. Corns. A Practical Tool for Public Health Surveillance: Semi-Automated Coding of Short Injury Narratives from Large Administrative Databases Using Naïve Bayes Algorithms. *Accident Analysis & Prevention*, Vol. 84, 2015, pp. 165–176.
35. Wang, Z., A. D. Shah, A. R. Tate, S. Denaxas, J. Shawe-Taylor, and H. Hemingway. Extracting Diagnoses and Investigation Results from Unstructured Text in Electronic Health Records by Semi-Supervised Machine Learning. *PLoS One*, Vol. 7, 2012, pp. e30412.
36. Chatterjee, S. *A Connectionist Approach for Classifying Accident Narratives*. Theses and dissertations. Available from ProQuest, 1998, pp. 1–345.
37. Das, S., A. Mudgal, A. Dutta, and S. R. Geedipally. Vehicle Consumer Complaint Reports Involving Severe Incidents: Mining Large Contingency Tables. *Transportation Research Record: Journal of the Transportation Research Board*, 2018. 2672: 122–134.
38. Beanland, V., M. Fitzharris, K. L. Young, and M. G. Lenné. Driver Inattention and Driver Distraction in Serious Casualty Crashes: Data from the Australian National Crash In-Depth Study. *Accident Analysis & Prevention*, Vol. 54, 2013, pp. 99–107.
39. Das, S., M. Le, and B. Dai. Application of Machine Learning Tools in Classifying Pedestrian Crash Types: A Case Study. *Transportation Safety and Environment*, Vol. 2, No. 2, 2020, pp. 106–110.
40. Fitzpatrick, C. D., S. Rakasi, and M. A. Knodler. An Investigation of the Speeding-Related Crash Designation through Crash Narrative Reviews Sampled via Logistic Regression. *Accident Analysis & Prevention*, Vol. 98, 2017, pp. 57–63.
41. Graves, J. M., J. M. Whitehill, B. E. Hagel, and F. P. Rivara. Making the Most of Injury Surveillance Data: Using Narrative Text to Identify Exposure Information in Case-Control Studies. *Injury*, Vol. 46, 2015, pp. 891–897.
42. Nayak, R., N. Piyatrapoomi, and J. Weligamage. Application of Text Mining in Analysing Road Crashes for Road Asset Management. In *Engineering Asset Lifecycle Management* (D. Kiritsis, C. Emmanouilidis, A. Koronios, and J. Mathew eds.), Springer, London, 2010, pp. 49–58.
43. Foulds, J., L. Boyles, C. DuBois, P. Smyth, and M. Welling. Stochastic Collapsed Variational Bayesian Inference for Latent Dirichlet Allocation. *Proc., 19th International Conference on Knowledge Discovery and Data Mining ACM SIGKDD*, Chicago, IL, 2013, pp. 446–454.
44. Blei, D. M., A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 993–1022.
45. Hoffman, M. D., D. M. Blei, C. Wang, and P. John. Stochastic Variational Inference. *Journal of Machine Learning Research*, Vol. 14, 2013, pp. 1303–1347.
46. Silge, J., and D. Robinson. Tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *Journal of Open Source Software*, Vol. 1, 2016, p. 37.
47. Cramer, N., S. Rose, D. Engel, and W. Cowley. *Automatic Keyword Extraction from Individual Documents*. John Wiley & Sons, Ltd., Hoboken, NJ, 2010.
48. Wijffels, J. Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. R Package Version 0.*8*.2, 2019.
49. Sievert, C., and K. Shirley. LDAvis: Interactive Visualization of Topic Models. R Package Version 0.3.2. https://CRAN.R-project.org/package = LDAvis. Accessed August 1, 2020.