Accident Analysis and Prevention 150 (2021) 105899

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Discovering latent themes in traffic fatal crash narratives using text mining analytics and network topology

Keneth Morgan Kwayu [a,*], Valerian Kwigizile [a], Kevin Lee [b], Jun-Seok Oh [a]

[a] Dept. of Civil and Construction Engineering, Western Michigan Univ., 4601 Campus Dr., G-238, Kalamazoo, MI, 49008-5316, United States
[b] Dept. of Statistics, Western Michigan Univ., 1903 W Michigan Ave, Kalamazoo, MI, 49008-5152, United States

ARTICLE INFO

Keywords:
Structural topic modeling
Network topology
Network centrality measures
Traffic crashes

ABSTRACT

The proliferation of digital textual archives in the transportation safety domain makes it imperative for the inventions of efficient ways of extracting information from the textual data sources. The present study aims at utilizing crash narratives complemented by crash metadata to discern the prevalence and co-occurrence of themes that contribute to crash incidents. Ten years (2009–2018) of Michigan traffic fatal crash narratives were used as a case study. The structural topic modeling (STM) and network topology analysis were used to generate and examine the prevalence and interaction of themes from the crash narratives that were mainly categorized into pre-crash events, crash locations and involved parties in the traffic crashes. The main advantage of the STM over the other topic modeling approaches is that it allows the researchers to discover themes from documents and estimate how the topic relates to the document metadata. Topics with the highest prevalence for the angle, head-on, rear-end, sideswipe and single motor vehicle crashes were crash at stop-sign, crossing the centerline, unable to stop, lane change maneuver and run-off-road crash, respectively. Eigenvector centrality measure in network topology showed that event-related topics were consistently central in articulating the crash occurrence. The centrality and association between topics varied across crash types. The efficacy of generated topics in classifying crashes by type was tested using a machine learning algorithm, Random Forest. The classification accuracy in the held-out sample ranged between 89.3 % for sideswipe crashes to 99.2 % for single motor vehicle crashes. High classification accuracy suggests that automation of crash typing and consistency checks can be accomplished effectively by using extracted latent themes from the crash narratives.

1. Background

Globally, vehicle crash fatalities have resulted in 1.25 million deaths with an average of about 3.827 deaths per day (Association for Safe International Road Travel, 2019). More than half of these deaths involved young adults. Traffic crashes as a cause of death is ranked ninth leading cause of mortality globally and expected to climb the ladder into the fifth position by 2030 if no necessary actions are taken. In the U.S.A, recent fatal motor vehicle crashes statistics have shown that 36,560 people died in 2018 (Highway Traffic Safety Administration, 2019). The report stipulates various factors that are associated with crash occurrences. The fatalities were mainly composed of passenger car occupants (35 %) followed by light truck occupants (27 %), pedestrians and bicyclists (20 %), motorcyclists (14 %) and large trucks and buses (4%). Drivers' attributes played a significant role in crash occurrence. Alcohol was a significant contributor in about 29 % of these crashes while 41 % of the occupants that died in a crash were unrestrained. The above and similar information on fatality by vehicle type, drivers' attributes, vehicle and environmental attributes and series of events leading to a crash are important in understanding crash causes. Innovative data analyses such as text and data mining analytics which uses unconventional methods and data sources to understand the effects and interactions of crash contributing factors, are needed.

The computational text analyses have gained considerable attention in various research domains due to the proliferation of digitized text available for organizational science research. Text mining, complemented with data mining techniques, have been used to analyze publication activities and trends in scientific research journals including transportation-related journals (Abuhay et al., 2018; Das et al., 2016), analyzing open-ended surveys (Roberts et al., 2014; Trappey et al.,

* Corresponding author.
E-mail addresses: kenethmorgan.kwayu@wmich.edu (K.M. Kwayu), valerian.kwigizile@wmich.edu (V. Kwigizile), k.lee@wmich.edu (K. Lee), jun.oh@wmich.edu (J.-S. Oh).

https://doi.org/10.1016/j.aap.2020.105899
Received 10 August 2020; Received in revised form 25 October 2020; Accepted 10 November 2020
Available online 4 December 2020

2013) opinion mining (Das et al., 2019), urban activities pattern classification (Hasan and Ukkusuri, 2014), and detecting medical descriptions frauds and abuses (Zafari and Ekin, 2019), among many others.

Transportation is one of the domains that have an enormous amount of textual data. It is the system that is run by and for people, and therefore it is almost automatic to have textual content in different forms such as user-generated content and internal information asset (Kinra et al., 2019). Huge textual databases are available in the area of transportation safety such as crash data narratives which contain a police officer's written description of the crash incident and consumer complaints having records of vehicle defects. The information from these textual sources can be utilized to gain deeper insights into factors contributing to crash occurrence. For example, Kwayu et al. (2020) utilized the crash narratives to conduct the semantic analysis and classification of drivers' hazardous action at signalized intersections. The proposed algorithm was able to identify "disregard traffic control" and "fail to yield" hazardous actions from crash narratives with a decent level of accuracy. Further, the developed textual-based algorithm proved to be promising in detecting possible errors that were made by the police officers while coding hazardous actions in the crash reports. In a similar token, Das et al. (2018) examined the consumer complaints available in the National Highway Traffic Safety Administration (NHTSA) database to discern major vehicle defects. The NHTSA database of consumers contains a list of complaints by each vehicle type with the level of injury that has occurred. The database is important because about 6.35 percent of the fatalities in the US occurred as a result of vehicle manufacturing defects (Das et al., 2018). Exploratory text mining complemented by the Empirical Bayes data mining approach assisted in discerning major vehicle defects that were mainly associated with airbags, braking systems, seatbelts, and speed control.

Different text mining applications can be applied to extract new insights from textual data such as thematic analysis, content analysis, supervised modeling, unsupervised modeling and natural language processing (Banks et al., 2018). Amongst many text mining applications, topic modeling has been applied in the transportation safety domain to identify latent structure within documents. It is part of the unsupervised modeling category that has been used to detect patterns and prevalence of different crash contributing factors in transportation safety data. Roque et al. (2019)utilized a latent Dirichlet allocation to identify and examine the co-occurrence patterns of attributes related to run-off-road crashes in road safety reports. The directed graphs were used to explore relationships between words. Kuhn (2018) utilized the incident reports from the aviation safety reporting system to identify latent topics and trends. The structural topic modeling was used as it offers ways to connect the available incident metadata and the generated latent themes. The analysis was able to identify additional unreported connections of different contributing factors to incidents in the aviation safety industry. Different issues were identified, including fuel pump tank, and landing gear issues. Robinson et al. (2019) utilized the same aviation safety database and applied temporal topic modeling, combined with subject matter expert review. The generated topics were presented to subject matter experts (SMEs) for evaluation and interpretation. There was a consensus among SMEs on the topic interpretation, and trend based on most probable terms and most likely narratives for a given topic. In another study, Brown (2016) utilized topic modeling among other text mining techniques to better understand the contributor to rail accidents. The generated topics and other textual features from the crash narratives were found to significantly improve the prediction of extreme accident costs.

Network topology theories and analysis has been used widely in the field of social and behavioral science to analyze social relations, computer science, telecommunication, and transportation, to mention a few. Quantitative narrative network analysis has been used to explore the large text corpora in social sciences using network properties such as structure robustness of the network and node centrality measures such as node degree, normalized betweenness, and eigenvector centrality (Golbeck, 2013). For example, Zhong et al. (2020) developed an LDA-based network analysis to visualize factors contributing to accidents in the construction industry. The network analysis measured the relationship between keywords in the accident narratives using degree centrality and eigenvector centrality. Link network analysis explores graphically the association between objects and has been used to detect frauds in financial institutions and insurance companies. It is a means of reducing the high-dimensional association between objects and identify the most important associations within a network (Miner et al., 2012). A similar analogy can be followed in examining the association between pre-crash events, locations, and involved parties in a crash for different crash types. In computational science, Abuhay et al. (2018) analyzed the publication activities in the journal of computational science using graph theory. Static and dynamic collaboration networks were used to investigate multidisciplinary collaboration among scientific communities. The results revealed that conferences are a suitable platform for encouraging collaboration development in researches among authors. Static and dynamic network theories that were applied in Abuhay et al. study can be leveraged in traffic safety studies to capture how different crash contributing factors have been evolving over time. In transportation, network theory has been a fundamental tool in analyzing the operation and management of transportation facilities and services. Topics in transportation that incorporate network theory include the optimization and design of transportation networks, the geography of the transportation network and the science of transportation networks (Monteiro et al., 2012). Such mentioned applications in various fields can be effectively leveraged in the transportation safety domain to understand the interaction and relative importance of different crash contributing factors.

In the USA, millions of crash narratives are available from various crash databases. The crash narratives, like any other unstructured data, are less utilized or incorporated in decision-making systems compared to structured crash metadata, as they are not easily manageable. Unstructured data usually exists in an unorganized format that offers no or little insight unless indexed and stored in an organized fashion (Berman, 2013). The inherent format of unstructured data, which for our case are the crash narratives, exacerbates difficulties in data preprocessing and information extraction. This study aims at utilizing crash narratives complemented by crash metadata to discern persistent crash factors and co-occurrence of factors that contribute to crash incidents using structural topic modeling (STM) approach and network topology analysis. The fatal crash narratives obtained from the online crash database, Michigan Traffic Crash Facts (MTCF), are used as a case study, but the proposed framework can be applied to any domain with versatile textual data sources. The specific objectives of this study are threefold. First, to understand the prevalence and trend of different topics emanated from the crash narratives. Secondly, to estimate the relationships between the generated topics and structured crash metadata. Thirdly, to understand both direct and indirect associations of the generated topics using the network topology. The last objective advances our understanding of the crash scenarios by providing a holistic overview of the existing association between crash events, locations, and involved parties in a crash - the information that would otherwise be difficult to obtain using only structured crash metadata.

2. Methodology

Topic models are unsupervised probabilistic models that enable users to search and explore the documents based on the underlying themes that form a document (Blei, 2012). Over time, different topic modeling approaches have been developed to cover the static and dynamic aspects such as temporal topic evolution and the topic hierarchy. Topic models allow for a mixture of topics to represent a single document. The simplest probabilistic topic model is the Latent Dirichlet Allocation (LDA), whereby the topic proportion for a given document is

estimated using Dirichlet distribution (Blei et al., 2003). LDA assumes independence between topics. But in most cases, this assumption is violated as it is more likely for the topics to be correlated with one another. The Correlated Topic Model (CTM) allows the topic proportions to exhibit correlations following the logistic-normal distribution. In most cases, the CTM has shown to provide a better fit than LDA (Blei and Lafferty, 2007). Another branch of topic modeling is the dynamic topic modeling which is designed to capture the evolution of topics in large corpora of document organized in sequential order (Blei and Lafferty, 2006). Supervised topic modeling approaches can also be used in cases where a modeler is interested in generating topics conditional on the response variable (Rafla and Gauba, 2011). This allows for the estimation of underlying latent themes or topics that best predict a given response variable in the test dataset. The aforementioned topic models contributed to the development of structural topic modeling (STM)-a topic model that was used in this study. The general background, formulation and the main advantages of structure topic modeling are provided hereafter.

2.1. Structural topic modeling (STM)

In this study, we utilized one of the most recent topics modeling approach called structural topic modeling (STM) proposed by Roberts et al. (2016). It combines and advances the aforementioned model frameworks of the correlated topic model (CTM), the Dirichlet-Multinomial Regression (DMR) topic model, and the Sparse Additive Generative (SAGE) topic model (Roberts et al., 2019). Like any other topic modeling approach, it is a generative model of word counts which begins by specifying the generating process of document-topic and topic-word distributions of the documents. An overview of the modeling approach is succinctly described by the pioneers of the STM algorithms (Roberts et al., 2016). The main advantage of the STM over the other topic modeling approaches is that it allows the researchers to discover topics and estimate how the topic relates to the document metadata.

2.1.1. Model estimation process

In Latent Dirichlet modeling (LDA), the topic proportion, $\overrightarrow{\theta}_d$ which is drawn from Dirichlet distribution is assumed to be a random variable, common across all the documents. For the case of STM, this parameter is assumed to be drawn from logistic-normal distribution conditioned on document metadata. A major limitation of LDA is the inability to model correlation or variability among topics. However, in most cases we expect the topics to be correlated (Blei and Lafferty, 2007). The logistic-normal distribution is a more flexible option than Dirichlet distribution and it allows for the estimation of correlations between topics. (Zafari and Ekin, 2019). Given a document d with vocabulary size V, topic prevalence X with topical prevalence covariates p, and K topics, the $\overrightarrow{\theta}_d$ can be estimated as shown in Eq. 1.

$$\overrightarrow{\theta}_d | X_d\gamma,\ \Sigma\ \sim\ LogisticNormal(\mu =\ X_d\gamma,\ \Sigma) \tag{1}$$

where, X_d is a *1 x p* vector, γ is a *p x K-1* matrix of coefficients for topic prevalence and variance , Σ is *K-1 x K-1* covarince matrix. Also, the word proportions or probabilities β within each topic is a random variable drawn from the Multinomial Logit model conditional on baseline word distribution (m), topic (k), document-level covariates, y_d and the interaction of topics and covariates as shown in Eq. 2.

$$\beta_{d,k,v} \propto exp(m_v + \kappa_{k,v} + \kappa_{y_d,v} + \kappa_{y_d,k,v}) \tag{2}$$

where the $\kappa_{k,v}$, $\kappa_{y_d,v}$ and $\kappa_{y_d,k,v}$ are the topic-specific deviation, the covariate group deviation and the interaction between the topic deviation and covariate deviation, respectively. The κ terms provide the adjustment based on a given topic and covariate data.

The topic assignment z_i is based on the $\overrightarrow{\theta}_d$. For a given selected topic, the word assignment is based on the estimated β_d as shown in Eq. 3 and Eq. 4;

$$z_{d,n} | \overrightarrow{\theta}_d\ \sim\ Multinomial(\overrightarrow{\theta}_d) \tag{3}$$

$$w_{d,n} | z_{d,n}, \beta_{d,k=z_{d,n}}\ \sim\ Multinomial(\beta_{d,k=z_{d,n}}) \tag{4}$$

The overall description of the topic distribution and word distribution can be nicely presented using graphical plate notation in Fig. 1, designed by Roberts et al. (2016).

2.1.2. Selecting the number of topics

Essentially, there is no clear unified and correct approach for selecting the optimal number of topics. In most cases, it will involve the use of statistical tools and experts' judgment based on the research question at hand (Roberts et al., 2016; Zafari and Ekin, 2019). A simple way to evaluate topic models that are used to identify themes and assist in interpretation, rather than to predict a knowable state is to look at the qualities of each topic and discern whether they are reasonable. However, there are various statistical data-driven tools that can assist a researcher to obtain a tentative number of topics to be estimated. STM R package, which was used in this study for generating the topics, offers data-driven statistical diagnostic tools such as held-out likelihood estimation, residual check, semantic coherence, and exclusivity measure. The residual check tests the overdispersion of the variance of a multinomial function (Taddy, 2012). Higher residuals indicate that more topics are needed to capture extra variance present in the data. A model with the lowest residuals is therefore desirable. Semantic coherence is a measure of how words in a topic co-occur together. Higher levels are obtained when most probable words in a topic co-occur together. Semantic coherence can be complemented by other metrics that combine frequency and exclusivity of a word or term in a topic such as the FREX metric (Bischof and Airoldi, 2012).

2.1.3. Topic-word assignment

Different measures can be used to explore words that represent a topic. Having multiple criteria for selecting words associated with a given topic helps to grasp the comprehensive meaning of a topic. A simpler metric populates words based on the probabilities. However, this metric will be biased towards most common words in a document that are likely to spread across multiple topics. To counteract this setback, other metrics have been developed such as FREX, Lift, and Score. FREX select words that are common for a given topic but rare in other topics. This measure provides a balance between the frequency and exclusivity of words in a topic. Mathematically, it is a weighted harmonic mean of a word based on exclusivity and frequency (Bischof and Airoldi, 2012). Lift divides the frequency of a word in each topic with its frequency in other topics. The score works similarly as Lift but uses the natural logarithm of frequency instead of frequency (Roberts et al., 2019).

2.1.4. Topic interpretation and validation

Once the topic has been generated, it is a task of the researcher to define the context of each topic. This is one of the most critical stages as it affects the subsequent analyses and ultimately inferences and conclusions. The representative words in each topic based on different metrics such as the highest probability, FREX, Score and Lift, are used to get the first impression of the context encircling a given topic. The next step involves the examination of documents that are highly associated with a given topic to gain a deeper understanding of the topic. In this study, a single document was represented by a single police fatal crash narrative. Each topic was assessed by using word assignment metrics and crash narratives that were highly associated with a topic in question. The STM package in R programing language has a function, *findThought* that was used to pull out documents that are highly associated with topics using topics posterior probabilities. The review of words and

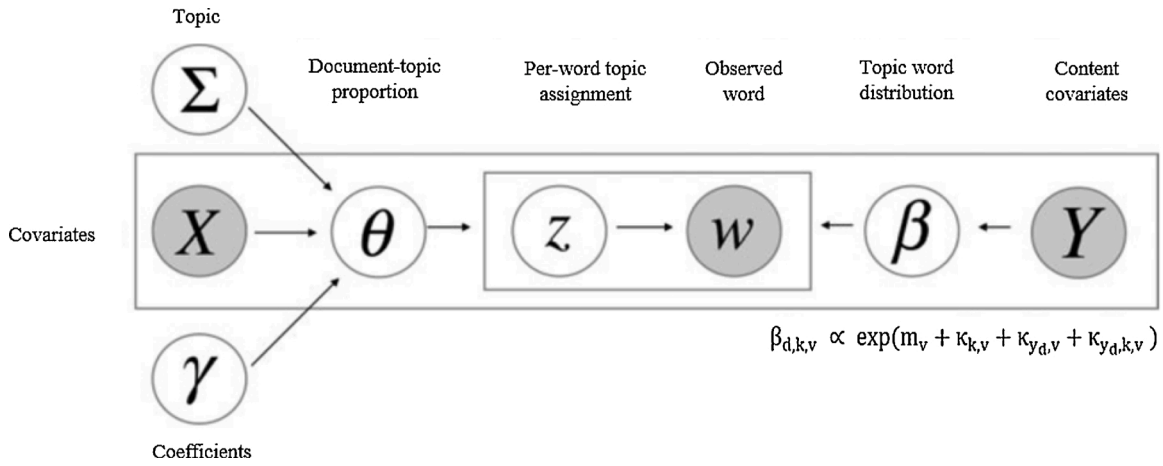

Fig. 1. Graphical plate notation of structural equation modeling (Roberts et al., 2016).

representative narratives for each topic were used to interpret and assign a label to each topic.

2.2. Assessment of topic association using network topology

Network theory is used to visualize the pairwise relationship between discrete objects (Metcalf et al., 2016). It encompasses a collection of techniques that allow researchers to find relationships among subjects using relational data that are organized in matrix form (Chiesi, 2015). For our case, the objects were the inferred topics categorized into events, locations and involved parties in a fatal crash. A network is made up of nodes and links. Each node represents a given topic while the link represents a relationship between inferred topics. In this study, the strength of the relationships between inferred topics was obtained using the Pearson correlation coefficient matrix. The representation of topics in a network offers a convenient way of describing the relationships between pre-crash events, locations, and involved parties in a crash. It may also reveal how multiple factors can contribute to crash incidence through investigation of direct and indirect links between factors. Various network centrality metrics can be used to measure the influence or importance of a topic in network topology (Marsden, 2015). In this study, we used Eigenvector Centrality to understand the influence of the topic in a network that describes a crash incidence. Eigenvector centrality assigns a relative score to a topic in a network while considering the importance of its neighbors. A given node in question will have a higher score if it is connected to high-scoring nodes compared to if it was connected to lower-scoring nodes. Mathematically, the scores so-called principal Eigenvectors are determined using the adjacency matrix (Golbeck, 2013). Intuitively, a topic with a high Eigen score in the crash network diagram indicates the centrality of a topic in the make-up of a chain of factors that led to a crash incident.

3. Data source and preprocessing step

Police officer written narratives of fatal crashes that occurred in Michigan between 2009 through 2018 were used as input in the STM framework. A police crash report consists of a police officer's written narrative articulating the series of events, involved party types, crash locations, environmental conditions, and first aid responses, among others. The investigation of a traffic crash involves a thorough examination of all elements contributing to a crash that could help to explain series of events based on factual data (North Carolina DOT, 2006). The sanitized Michigan police crash reports are available online in the Michigan Traffic Crash Facts (MTCF) database. Bulk access to UD-10 s can be enquired from the Michigan Criminal Justice Information Center (CJIC). Overall, the acquired dataset had a total of 9202 traffic fatal crashes that occurred in Michigan during the ten years of analysis. The crash reports were available in portable document format (see Fig. 2) and it was impractical to manually extract the crash narratives from each report. A script was created in Java programming language to efficiently loop through the sanitized UD-10 reports, extracting the crash ID and the crash narratives and map the array of results in a more manageable file format. The accuracy of the extracted narratives and crash ID was conducted through manual examination of randomized sample UD-10 reports to make sure that there was no information mismatch. On average, each narrative contained about 80 words. The crash ID was used to merge the extracted crash narratives with other indexed crash metadata. The crash metadata that was used in this study includes crash type, time of the day, age of involved parties in the fatal crash and the year of crash occurrence.

4. Results and discussions

This section presents the results and discussions of various analyses that were performed which include the topic selection process, analysis of the prevalence of the topic on crash metadata and analysis of topics co-occurrences using network topology. The discussion of the result is mainly pivoted on how the results and methodology used in this study can be integrated with previous and future traffic safety studies that utilized structured data sources to gain deeper insights on traffic crash causes.

4.1. Topic selection

The first task of this study was to select the number of topics that are to be estimated by the STM model using experts' judgment and statistical data-driven diagnostic tools. Fig. 3 presents model diagnostic results of residuals while Fig. 4 shows the plot of semantic coherence against exclusivity. The diagnostics plots present topic model performance for the various number of topics ranging from twenty to one hundred. The preferred model is the one that offers low residuals, high semantic coherence and exclusivity. The STM with sixty topics yielded the lowest residuals. From Fig. 4, high semantic coherence and exclusivity found at the upper right corner of the graph were mostly populated by the STM model with sixty topics. The STM model with sixty topics was tentatively selected for further evaluation using experts' opinion which was mainly based on the interpretability of the inferred topics. Consequently, the final STM model with sixty topics was used in the subsequent analysis. It should be noted that the final model to be used doesn't have to concur with the results of the diagnostics plots. The diagnostic results only provide a guide for the researcher to investigate the suitability of the inferred topics for further analysis.

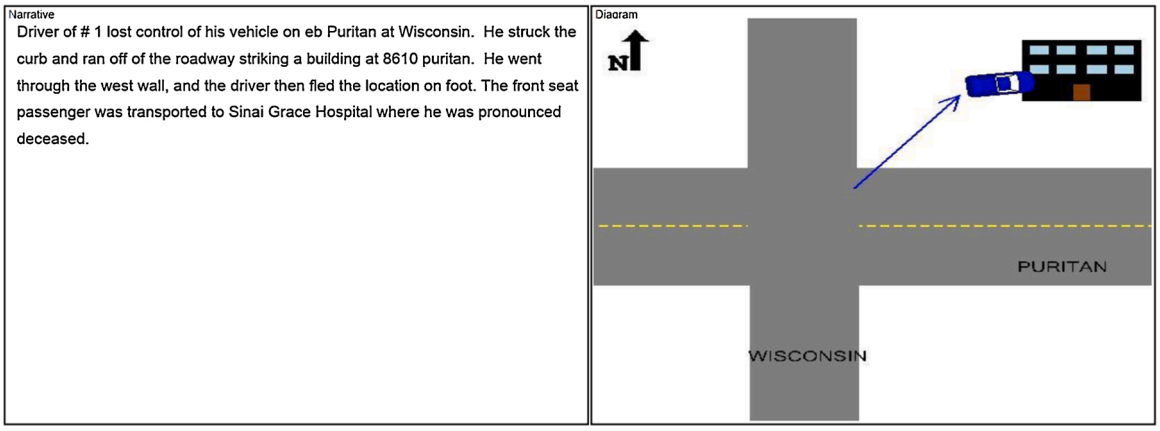

Fig. 2. Excerpt of the crash report in portable document format consisting of a crash narrative.

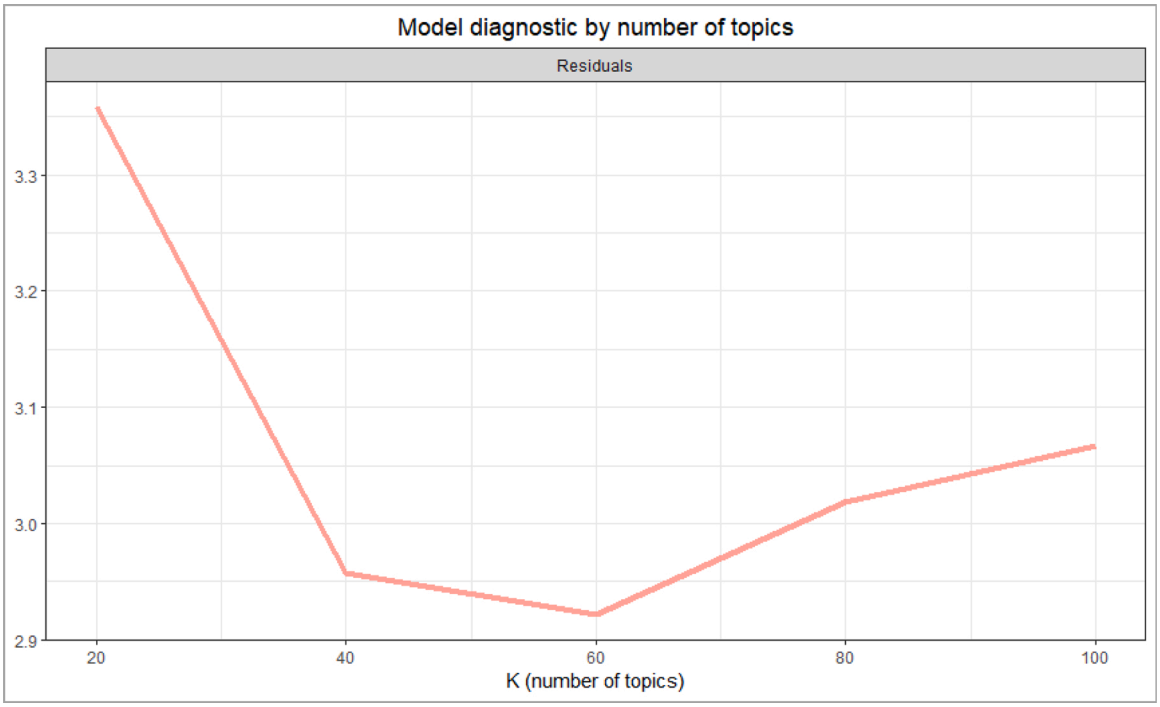

Fig. 3. Residual check at different number of topics.

4.2. Topic description and interpretation

Word assignment based on the highest probability, FREX, Score and Lift measures were used to infer the context of each topic. Further, the most representative crash narratives for each topic were carefully examined to aid the interpretation of each topic. Each topic was labeled successfully using word assignment and crash narratives. The topics of interest in this study are those that were categorized as either describing a pre-crash event, crash location or the involved party in a crash. Twenty-five out of sixty topics matched the above-mentioned categories and were retained for further analysis. The remainder of the topics that didn't fit in any of the categories were excluded from the analysis. Such topics described other issues about the crash incident such as first-aid response, the direction of travel, driver's unit number, road names, reference number to the additional investigation report, among others. Table 1 provides the most representative words for each of the labeled topics using the metrics already described in the methodology section. Different word assignment metrics are provided to avoid giving a bias interpretation of a topic. Table 2 provides examples of sample crash narratives that were highly associated with each topic.

The crash narratives give a practical demonstration of how each of the inferred topics fits in the description of a crash. Most of the inferred topics described a pre-crash event such as loss control, avoiding collision, passing maneuver, speeding, failure to yield, turning left and run-off-road. The topics related to involved parties in a crash articulated the involvement of a passenger, child, motorcyclist and commercial vehicle in a crash. The location-related topics described fatal crashes that occurred at stop-controlled intersections, signal-controlled intersections, driveways or parkways and on freeways, particularly freeway exit ramps.

The inferred topics present unique cases that have been extensively researched in previous studies such as run-off-road crashes, alcohol, seatbelt usage, fail to yield, turning left, secondary crashes, pedestrians, and bicyclists (Rolison et al., 2018). The contribution of this study is to examine how each topic connects to a larger network of pre-crash events, locations or involved parties in a crash. Such interactions between pre-crash events, location and involved party in traffic crashes that can provide an in-depth insight to crash contributing factors has not

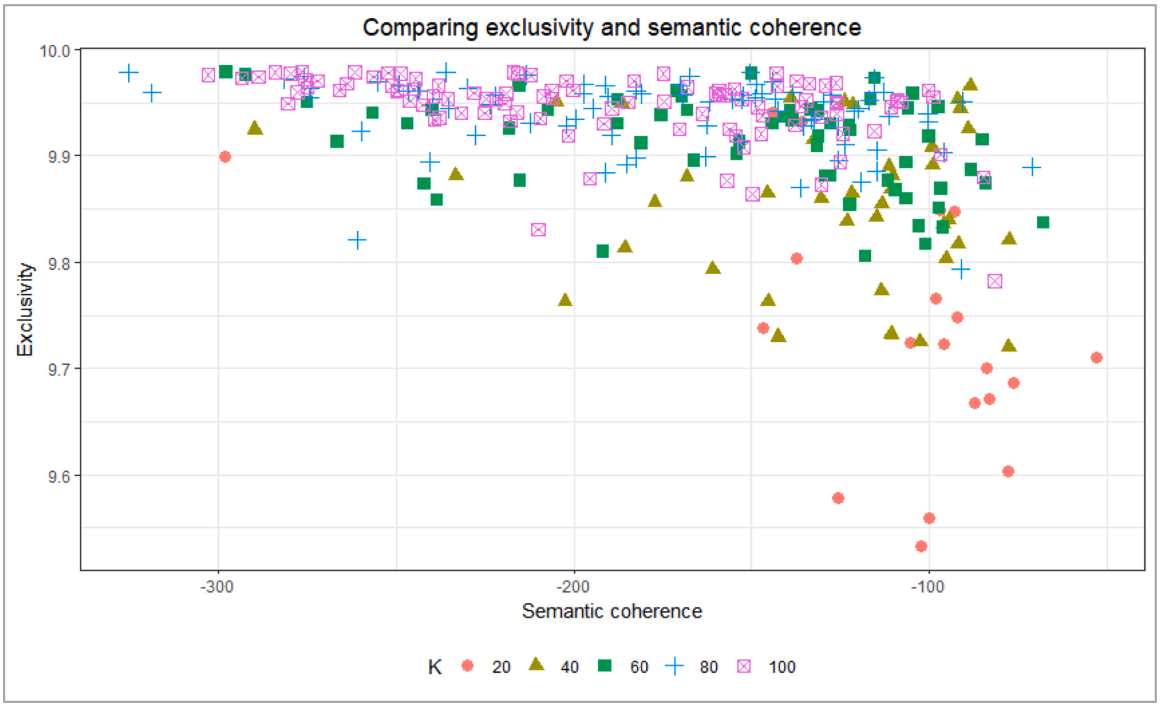

Fig. 4. Semantic coherence and exclusivity for different topic models.

Table 1
Word assignment of each topic.

Topic label	Group	Highest Prob	FREX	Lift	Score
AvoidCollision	Event	avoid, collis, unabl	avoid, collis, swerv	anim, avoid, swerv	avoid, collis, swerv
CrossCenterline	Event	center, cross, line	center, line, head-	mackinac, line, center	center, line, cross
Driveway|Parkway	Event	park, drive, tow	park, lot, main	broadway, rusko, scooter	broadway, park, lot
FailYield	Event	fail, pull, way	yield, way, fail	baselin, mertz, yeild	fail, yield, way
FrontDamage	Event	front, damag, caus	damag, straight, heavi	collison, extens, damag	damag, collison, front
GoingIntoDitch	Event	went, ditch, continu	airborn, went, embank	guid, wire, fenc,	ditch, pole, went,
Intoxication	Event	blood, result, alcohol	test, alcohol, blood	forens, test, corey	blood, alcohol, forens
LaneChange	Event	lane, right, left	lane, shoulder, right	-wheeler, merg, guard	lane, right, shoulder
LostControl	Event	control, lost, slid	lost, lose, slid	welch, ici, slipperi	control, lost, slid
PassingManuever	Event	pass, away, attempt	pass, away, anoth	ludington, pass, away	pass, ludington, away
RanoffRoad	Event	roadway, left, ran	overturn, roadway, ran	enbank, re-ent, -correct	roadway, ran, overturn
Seatbelt	Event	insid, found, trap	seatbelt, insid, trap	rbuxton, seatbelt, belt	trap, seatbelt, rbuxton
SecondaryCrash	Event	crash, prior, complet	crash, prior, occur	averi, amend, crash	crash, averi, complet
Speeding	Event	speed, high, rate	high, rate, speed	kevin, rate, high	speed, high, rate
TurnLeft	Event	turn, left, onto	turn, make, onto	capit, highland, turn	turn, left, make
UnableStop	Event	rear, stop, end	end, push, slow	assur, brooklyn, closur	rear, end, stop
RampExit	Location	exit, ramp, run	ramp, exit, bus	approch, ramp, bus	exit, ramp, bus
RedLight	Location	light, red, traffic	green, light, flash	terrac, solid, green	light, red, green
StopSign	Location	stop, sign, intersect	sign, stop, intersect	linco, sign, rds	stop, sign, intersect
Bicyclist	PartyType	median, bicyclist, travel	bicyclist, median, bicycl	bicyclist, sullivan, barricad	median, bicyclist, bicycl
ChildInvolved	PartyType	ford, fire, depart	child, health, ford	allegi, attack, elli	ford, elli, henri
CommericalVehicle	PartyType	trailer, truck, semi	trailer, tractor, semi	boat, dane, haul	trailer, truck, semi-truck
Motorcycle	PartyType	motorcycl, rider, deer	motorcycl, rider, deer	larri, motorcycl, motorcyl	motorcycl, rider, larri
PassengerInvolved	PartyType	passeng, side, driver	passeng, seat, driver	golf, seat, passeng	passeng, side, seat
Pedestrian	PartyType	pedestrian, cross, walk	walk, pedestrian, street	lewi, crosswalk, auburn	pedestrian, walk, lewi

been examined by using text mining and network topology approaches.

4.3. Prevalence of topics on crash metadata

One of the main advantages of structural topic modeling is that it allows the researcher to easily associate the topics with the covariates or metadata. The incorporation of metadata in the STM framework enriches our understanding of the possible relationships between the inferred topics from unstructured crash narratives and structured crash metadata.

Fig. 5 shows the overall topic prevalence across all the fatal crash narratives. The prevalence of topics highlights important underlying themes in traffic crashes that may require extensive analyses from multiple data sources.

Topics that had higher proportions were the ones that described the pre-cash events such as run-off-road, lane change and going to the ditch. A topic describing fatal crashes at the stop-sign intersection was the most prevalent location-based topic followed by a topic describing crashes at the signal-controlled intersection. Further, the most prevalent topic describing the involved parties in a crash was passenger-related crashes followed by pedestrian-related crashes and motorcycle-related crashes. The ran-off-road topic generally covered two main scenarios. First, the

Table 2
Examples of most likely narrative for some topics.

Topic	Group	Sample narrative
TurnLeft	Event	*vehicle #2 was w/b on pontiac trail. driver of vehicle #1 was attempting to turn left into a driveway and turned infornt of vehicle #2.*
Speeding	Event	*..shows motorcycle coming out of alley in a high rate of speed -lost control and struck garbage dumpster. motorcycle reported stolen from wixom p.d.*
FailYield	Event	*vehicle #2, failed to yeild right of way to vehicle #1 and drove into path of vehicle #2 causing vehicle #2 to strike vehicle #1.*
RanoffRoad	Event	*#1 was traveling south on angevine rd and ran off the roadway right, re-entered the roadway and ran off the roadway left and struck a tree.*
LaneChange	Event	*unit 2 was traveling north on m5 in the left lane. the driver of unit 2 stated that unit 1 changed lanes from the far-right lane, to the middle lane. unit 2 driver stated it appeared unit 1 was going directly into his lane so he swerved slightly to the left on the shoulder.*
StopSign	Location	*veh #1 n/b on neff rd failed to stop for a stop sign at mcbride rd. as #1 entered the intersection it struck #2, who was e/b on mcbride rd.*
RedLight	Location	*unit 1 n/b on livernois ave approaching a red signal at wattles rd. unit 2 was proceeding w/b wattles rd to s/b livernois ave on a green signal. unit 1 disobeyed the red signal, entered the intersection.*
Driveway| Parkway	Location	*vehicle #1 was reversing from driveway and struck a motorized wheelchair that was traveling south on the private drive. an elderly woman fell from the motorized wheelchair.*
Passenger Involved	PartyType	*according to the driver and rear passenger, the front seat passenger sat on the window ledge with his upper torso outside of the vehicle while in motion.*
Child Involved	PartyType	*#2 was w/b on ellis and was distracted by children in the backseat. she ran the stopsign at s. moorland rd and was struck by a n/b #1. both vehicles ran off the road and rolled over.*
Pedestrian	PartyType	*pedestrian crossing the street from longfellow traveling w/b to e/b entered the roadway with no crosswalk was then struck by vehicle #2. writer obs none of the streetlamps to be functioning.*
Bicyclist	PartyType	*both unit 1 and 2 were traveling e/b on stoll rd. unit 1 struck unit 2 from behind on the paved portion of the roadway, throwing the bicyclist from the bike. unit 1 driver stated she did not see the bicyclist at all on the roadway…*
Motorcycle	PartyType	*unit 1 and the rider slid across the pavement striking the curb. both the motorcycle and the rider vaulted into the air. the rider came to rest in the grass. the motorcycle came to rest in the roadway and began to burn due to a ruptured fuel tank. wofford was not wearing a helmet.*

topic mainly gathered instances when the single vehicle ran off the roadway and hit a fixed object. Secondly, it captured aftermath scenario of the multivehicle crashes whereby the impact of collision between two or more vehicle caused one or more vehicles to lose control and veer or pushed away from the roadway. It is also the case when the ran off road vehicle struck another vehicle while trying to re-enter the roadway resulted into different types of multivehicle crashes.

The topic proportions across each crash type were examined to discern the most prevalent topics for each crash type. This can generally assist in understanding the most critical events, party types, or locations that are associated with a certain crash type. Moreover, the analysis is expanded by interacting the topics with other crash metadata namely the time of the day and drivers' age with the crash type as shown in Figs. 7 and 8 respectively.

The definitions of each of the crash types are well described with illustrative diagrams in the Michigan traffic crash report instruction manual under the crash type section (Michigan Department of State Police, 2018). Angle crash occurs when the direction of travel is perpendicular for both drivers and there is a side impact of approximately 90 degrees. A head-on collision occurs when the direction both vehicles involved in a collision is towards each other. When two vehicles are approaching head-on and at least one vehicle is attempting a left turn is called a head-on left turn. Rear-end crash occurs when the vehicles are traveling in the same direction and the leading vehicle is struck by the following vehicle. The leading vehicle could either be going straight(rear end), turning left (rear end-left turn), or turning right (rear end-right turn). Sideswipe crashes involve vehicles that were either traveling in the same direction or opposite direction and make a side contact. A single motor vehicle crash involves only one motor vehicle as defined in the Michigan crash reporting manual. It includes those cases in which a motor vehicle was the only traffic unit and the only motor vehicle involved that collided with a bicyclist, pedestrian, engineer (railroad train), animal, or any other non-motorized object.

4.3.1. The topic prevalence across angle fatal crashes

The proportion of topics that described angle crashes were mostly found at the stop-sign intersection (*StopSign*) compared to signal-controlled intersections (*RedLight*). This was intuitively correct as, at signal-controlled intersections, vehicle movements are more regulated than vehicle movements at the stop-controlled intersections. The difference in the topic proportion of angle crashes at stop-sign was moderated by the time of the day as shown in Fig. 8. A topic describing fatal crashes at the signal-controlled intersections which involved running a red light and other violations as described by sample narrative in Table 2 occurred mostly at nighttime. Further, the topic proportion was higher in young driver crashes compared to older driver crashes as shown in Fig. 7. Conversely, the topic describing crashes at stop-controlled intersection occurred mostly during the daytime with no significant difference in topic proportion between drivers' age. Red-light running (RLR) has long been known as one of the leading causes of urban crashes (Retting et al., 1999; Schattler and Datta, 2004). Exploratory analysis such as the study by Fu et al. (2013) can be combined with text-based analysis as outlined in this section to explore further the influence of different contributing factors on driving violations at intersections.

The event-related topics that had higher proportions in angle crashes include failure to yield and turning left. The topic proportion of failure to yield was higher in older-related crashes and during the daytime. Turning left topic was also skewed towards the older-related crashes and mostly during the daytime. The findings align with previous studies that used surveys and structured crash data. Older drivers have been commonly reported to make errors at intersections, such as failing to comply with signals and signs and making improper turns (Langford and Koppel, 2006; McGwin and Brown, 1999).

The passenger-related topic was more prevalent among young drivers with no significant difference in topic proportion by the time of the day. By using naturalistic driving study, accidents records and survey, studies have shown that the presence of a passenger affects the driver's behavior (Cooper et al., 2005; Lam et al., 2003; Orsi et al., 2013). Generally, the elevated crash risk among young drivers carrying a peer passenger has been reported. Conversely, the presence of passengers has shown to have a positive or protective effect among older drivers.

4.3.2. The topic prevalence across head-on fatal crashes

The head-on collisions were mostly populated by event-related topics such as crossing the centerline, turning left, or lane change maneuver. Crossing the centerline describes an event when a driver crosses the centerline of the road as he could not stay in lane or preparing to perform a passing maneuver. If such maneuver is performed recklessly, it could result in direct head to head collision with the vehicle on the opposite lane. Turning left topics in the head-on collision was more prevalent among older drivers while crossing the centerline was more prevalent among young drivers. Older drivers' involvement in a head-on

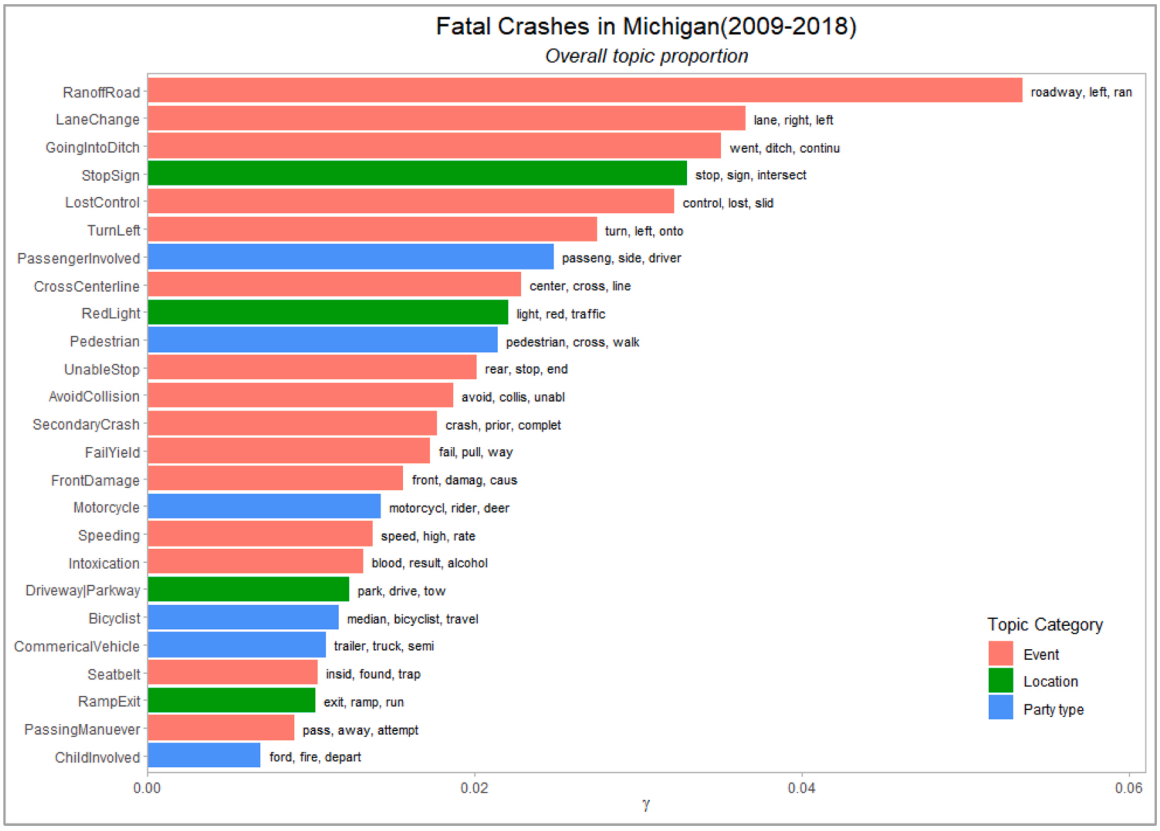

Fig. 5. Overall topic distribution.

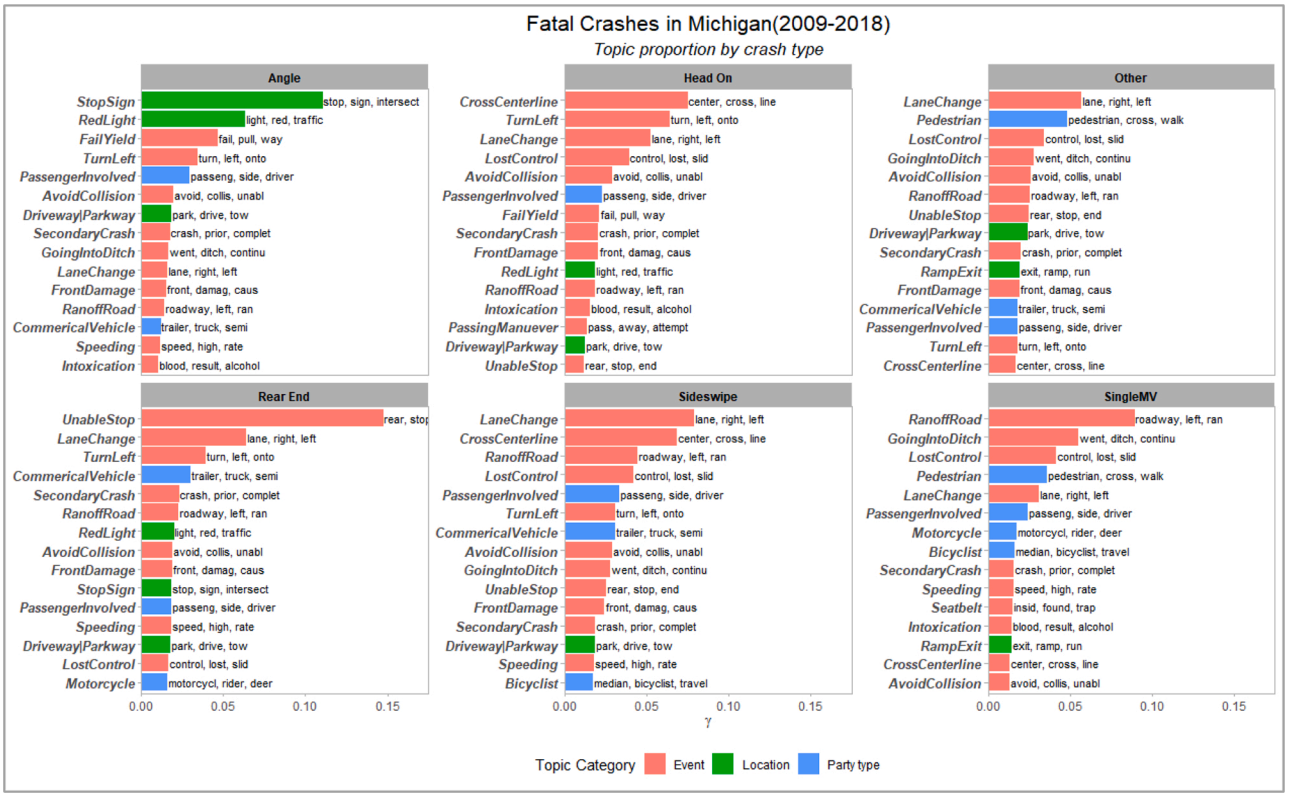

Fig. 6. Topics proportion by crash type.

collision while making a left turn may partly be contributed by their age-related decline in sensory, perceptual, cognitive and motor driving abilities (Lombardi et al., 2017). Overrepresentation of crossing the centerline topic, lost control and speeding topics among young drivers in head-on crashes suggest higher risk-taking and reckless driving behavior for this age group. Previous studies have reported risk-taking behaviors and increased road violations among young drivers compared to older drivers (Rolison et al., 2018).

4.3.3. The topics prevalence across rear-end fatal crashes

Rear-end fatal crashes were mostly populated by event-related topics such as the driver being unable to stop, performing a lane-changing maneuver and turning left as shown in Fig. 6. Unable to stop topic was more prevalent among older drivers which may partially be attributed to slower reaction time. Crashes involving a commercial vehicle also emerges as the most prevalent involved-party related topic for this crash type. A naturalistic study conducted by Bianchi Piccinini et al. (2017) on commercial vehicles shows that rear-end crashes were mostly triggered by drivers' adoption of shorter headways while following. Other factors that have been reported include drivers' fatigue and driving-related visual scanning mismatches (Victor et al., 2013; Woodrooffe et al., 2014). In-vehicle safety measures and automation can significantly deter the risk of commercial vehicle involvement in rear-end crashes.

4.3.4. The topics prevalence across sideswipe crashes

The sideswipe crashes were mostly associated with event-related topics describing lane changing or driver crossing the centerline. Both events were more prevalent in older driver crashes and during the daylight. Lane changing maneuver place a driver in a great risk due to the presence of a blind spot area. It requires the driver to make a quick judgment about the available safe gap and it is an act that disrupts the traffic flow (Munro et al., 2010). Older drivers have been reported to make more errors during lane changing maneuvers compared to young drivers which is consistent with this study findings. Most of the predictors of lane change errors in older drivers were reported to be associated with their visuospatial skills (Sivak et al., 2007) and lack of attention (Munro et al., 2010).

4.3.5. The topics prevalence across single-vehicle fatal crashes

Single-vehicle crashes represented about half of the vehicle fatalities in the case study area. The most prevalent topics for single motor vehicle crashes were run-off-road instances. Contributing factors to the run-off-road crash as reported by McLaughlin et al. (2009) are distractions (i.e. most contributing factor), low visibility, low-friction conditions, changes in roadway boundaries, short following distance, fatigue, and late route selection, among others. The proportion of run-off-road topic was higher among young drivers compared to older drivers. Factors such as excessive speed, loss of control, distraction, or inattention have been reported as primary causes of crash among young drivers (Braitman et al., 2008). This could explain the observed higher prevalence of run-off-road topic among young drivers in our study. With respect to time of the day, run-off-road topic appeared to have a higher proportion during daytime compared to nighttime. Involved party topics in single-vehicle fatal crashes were pedestrians, passengers, and motorcyclists. A pedestrian related topic was prevalent in older drivers than young drivers crashes and mostly at night-time compared to daytime. The passenger-related topic was also more prevalent among young drivers compared to older drivers with no much difference between daytime and nighttime. As for motorcycle-related crashes, there was no much difference in topic proportion by age group, but the topic was more pronounced at daytime than nighttime.

4.3.6. Prevalence of topic over the years by crash type

Tracking the prevalence of inferred topics over time can assist traffic safety and practitioners to assess the effectiveness of various countermeasures that have been designed to combat specific pre-crash events at a given location or driving population segment. The trends of the five

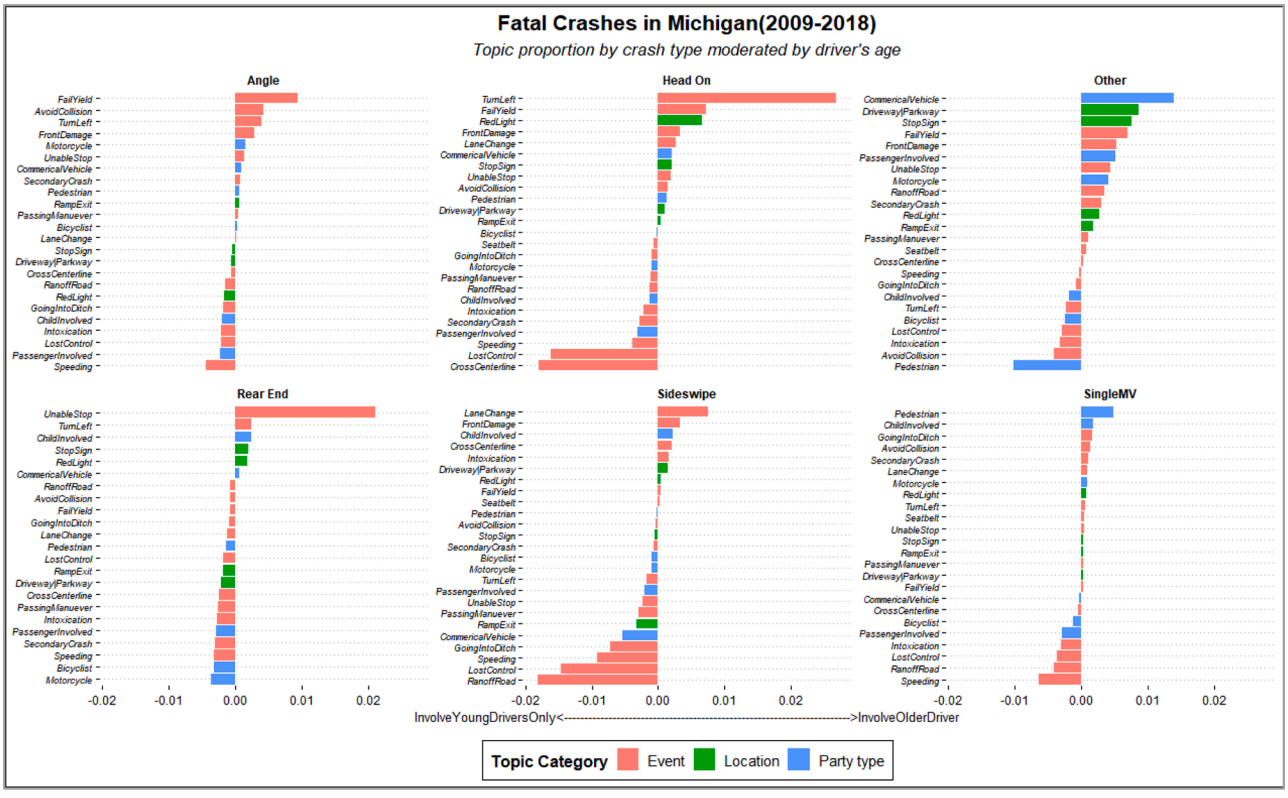

Fig. 7. Topic proportion by crash type moderated by driver's age.

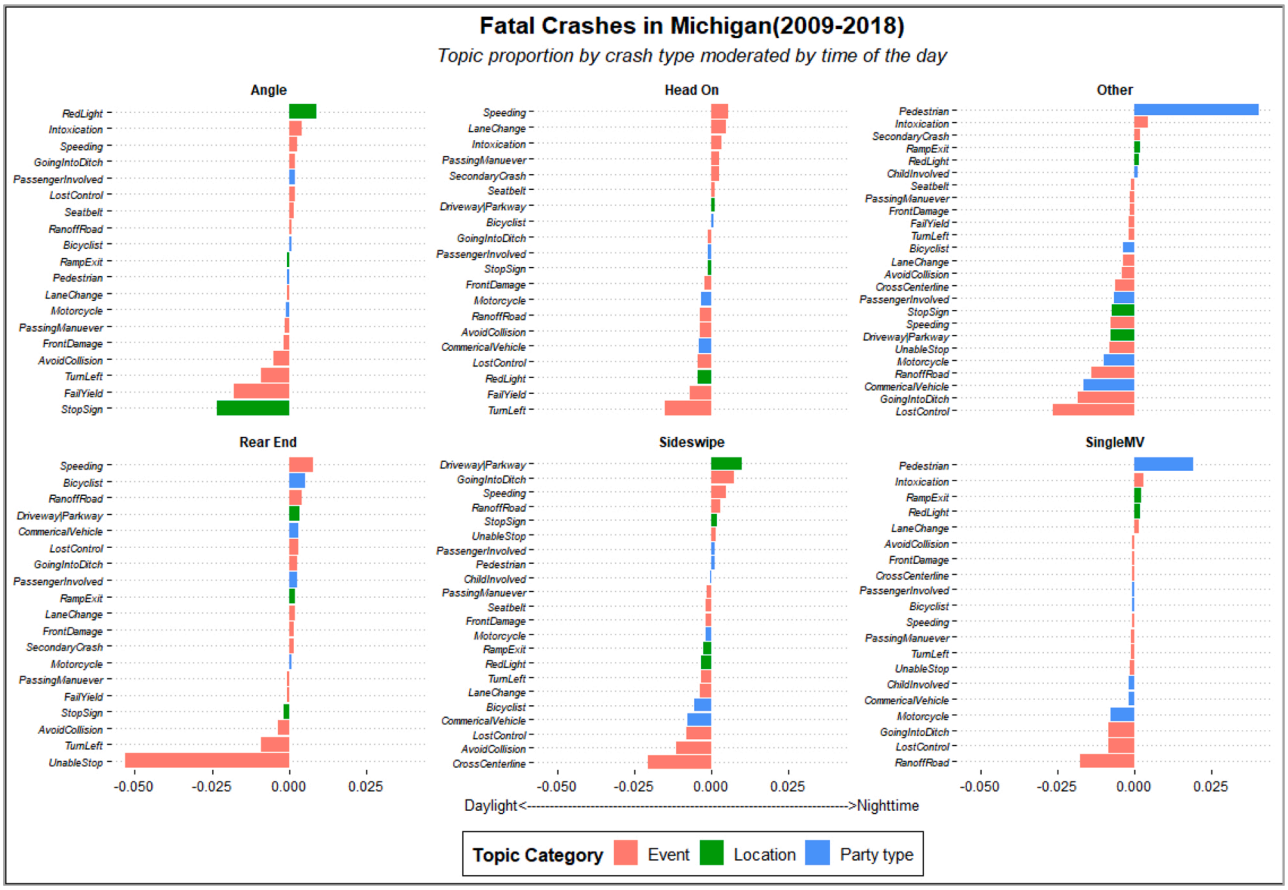

Fig. 8. Topics proportion by crash type moderated by the time of the day.

most prevalent topics for each crash type that led to a fatality were assessed over ten years from 2009 to 2018 as shown in Fig. 9.

The prevalence of top topics in angle fatal crashes such as crashes at stop-sign and signal-controlled were found to decrease over the years. There was a slight increase in the passenger-related topics and failed to yield topics over time. In the previous section, it was shown that passenger-related crashes and failed to yield topics were skewed mostly to young drivers and older drivers respectively. The observed upward trend of the passenger-related topic suggests that there is a need for further research on this area. For example, distraction and social influence have been reported as the main factors that are associated with increased fatal crash risk among young drivers carrying passengers (Ehsani et al., 2015).

The prevalence of the topic describing turning-left movement on head-on crashes, which mostly occurred at intersections was decreasing over time while crossing the centerline topic was increasing over time. In the previous section, crossing the centerline topic was more prevalent among young drivers who were involved in head-on crashes. The results further suggest the need for more extensive studies covering various aspects of the driving behaviors among young drivers.

Unable to stop at assured distance and lane changing which led to rear-end crashes seemed to be increasingly prevalent over time while topic such as turning left was decreasing over time. Also, lane changing maneuver which was the most prevalent topic in sideswipe crashes exhibited an upward trajectory while other topics showed no change or slight decrease over the years.

The lane changing topic exhibited an upward trend in sideswipe crashes. Recall, this topic was more prevalent among older drivers. An upward trend of lane changing topic calls for further investigations on how older drivers decreased sensory, perceptual, cognitive and motor driving abilities can be carefully compensated in the design of in-vehicle safety features and other engineering countermeasures.

Most of the factors in single-vehicle crashes topics which described pre-crash events were decreasing over time. The decline in topic prevalence was more noticeable on run-off-road and lost control cases.

4.4. Assessing the co-occurrence of topics using network topology

The network topology was used to analyze the association between inferred topics articulating crash events, locations and involved parties in a crash. The network topology is made up of nodes and links. The nodes represent the inferred topics while the links between nodes depict the relationship between nodes using the Pearson correlation matrix as a weighting factor. Eigenvector was used to measure the centrality of each node within a network. The larger the eigenvalue the more central a given event, location or party type in explaining the occurrence.

Figs. 10–15, shows the network topology that was created for each crash type. Further, Figs. 16 and 17 show the network topology for older driver-related crashes and young drivers-related crashes respectively. Variations in strength between topics were observed across the topics by crash types and drivers' age cohorts. Most of the positive associations were between events such as lost control & ran off the road, lane change & cross the centerline, secondary crash & intoxication, and turn left & fail to yield. The positive associations between location and events include turning left & driveways or parkways, fail to yield & stop sign and turn left & red light. Party type and event-related topics positive relationships include motorcycle & speeding, lane change & bicyclist, passenger involved & seatbelt usage, and child involved & seatbelt usage.

Such a pair of associations between topics articulating events, locations and involved parties in a crash could help in understanding possible causes of fatal crashes. For example, the positive association

Fig. 9. Prevalent of the high-proportion topics over time by crash type.

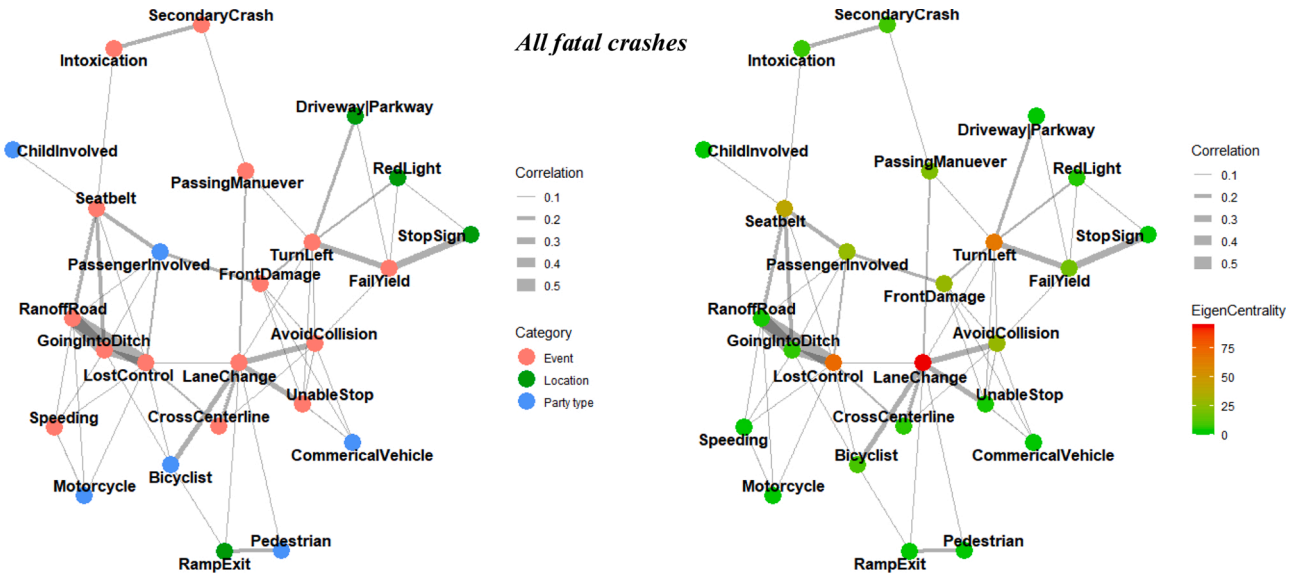

Fig. 10. Network topology and centrality measures between topics for all fatal crashes.

between secondary crash topic and intoxication topic which was in most cases was prevalent among young drivers (Fig. 7) could indicate higher susceptibility of intoxicated young drivers to be involved in a fatal crash. Such kind of fatal crashes has a higher likelihood of initiating a chain of secondary crashes. By using a network diagram of fatal crashes in Fig. 10, it can also be revealed that such association of intoxication and secondary crashes have other direct links to seatbelts and passing maneuver topics. Also, slight variations of other topics having a direct link to intoxication & secondary crashes topics can be observed across crash types. For example, failed to yield topic has a direct link to secondary

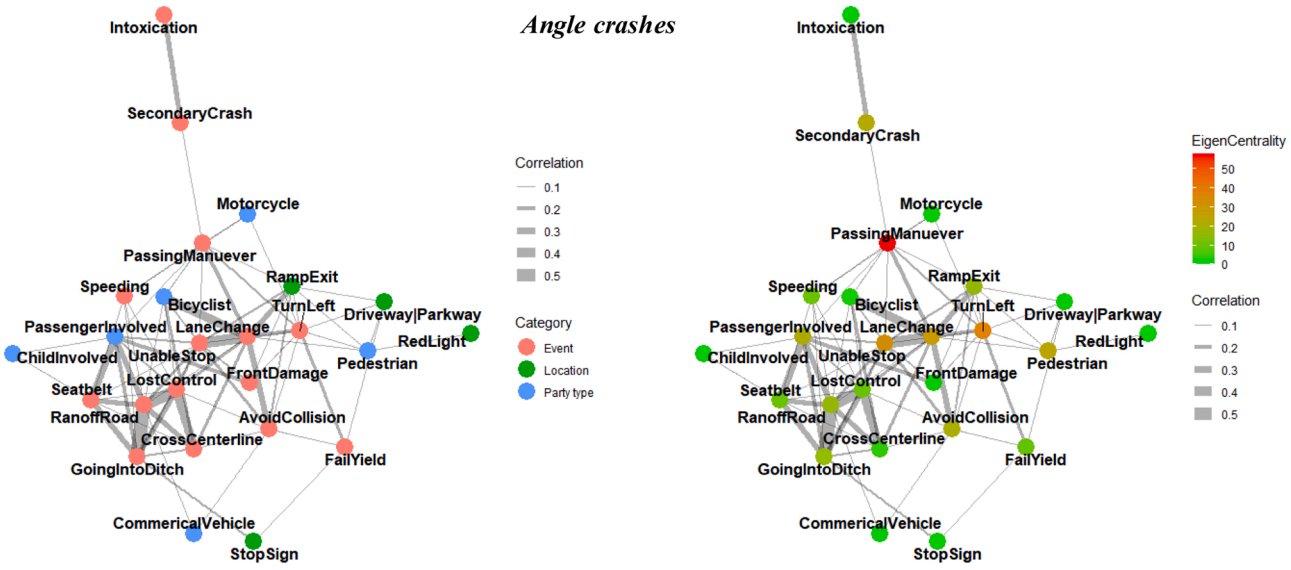

Fig. 11. Network topology and centrality measures between topics for angle fatal crashes.

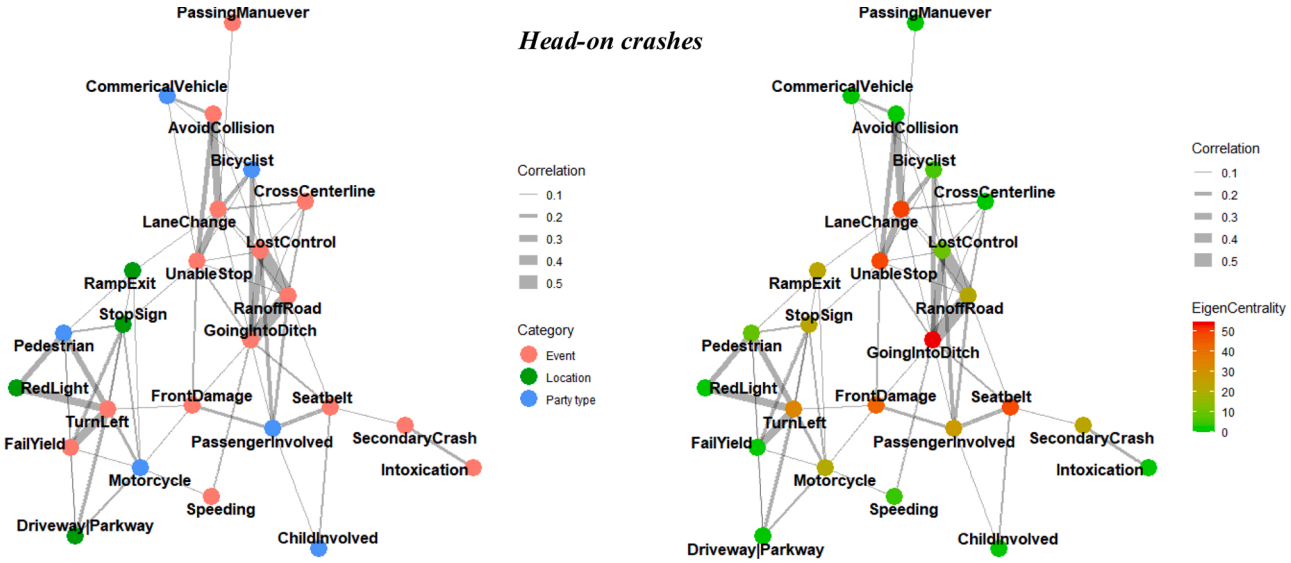

Fig. 12. Network topology and centrality measures between topics for head-on fatal crashes.

crashes in single motor vehicle crashes and ran red light topic has a direct link to the intoxication topic in rear-end crashes. A similar analysis can be used for any topic or pair of topics in a crash network topology of interest to the researcher. Consequently, a chain of crash contributing factors can be discerned using crash network topology which is vital in the appraisal of holistic crash countermeasures.

Overall, event-related topics were more central in explaining crash scenarios than location-based or involved party topics as indicated by Eigen centrality value. Lane change and lost control topics were central for all the crash types as shown in Fig. 10. A slight movement in Eigen centrality between nodes was observed across crash types (Figs. 11–15). For example, topics that had higher centrality value in angle fatal crashes were run-off-road, lost control, unable to stop and lane change topics. Similar results of Eigen centrality were obtained for head-on crashes. Turn left and avoiding collision topics had the highest Eigen centrality values for single motor vehicle crashes while fail to yield and turn left were most central topics in sideswipe crashes. Run-off-road and lost control were central in rear-end crashes. Topics that had higher Eigen centrality value were the most important topics that connected other topics in explaining a series of events, locations and involved parties in a crash.

4.5. *Efficacy of inferred topics on the classification of crash types*

The efficacy of the generated topics on the classification of crash types was tested using a machine-based classification method. The Random Forest model was deployed utilizing the inferred topics as features for classifying the crash types. The dataset was divided into training (70 percent) and testing dataset (30 percent). The repeated cross-validation process was used to estimate the model parameters. The calibrated model was then tested in a held-out sample. Table 3 tabulates the accuracy for each crash type using the test dataset. Accuracy is defined as the percentage of crashes that were correctly classified into their respective crash type using inferred topics from the crash narratives. The highest accuracy of 99.2 % was obtained in single motor vehicle crashes. The lowest performance was in sideswipe crashes, having 89.3 % accuracy. The importance of a topic in the classification of the fatal crashes by crash types are shown in Fig. 18. For instance, the

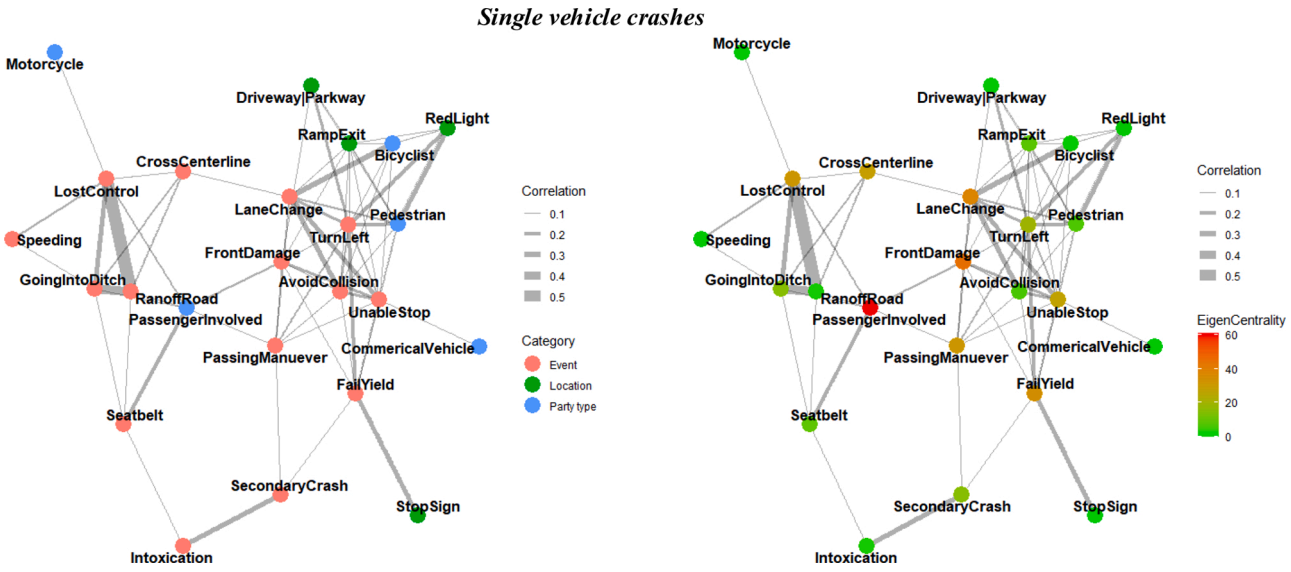

Fig. 13. Network topology and centrality measures between topics for single vehicle fatal crashes.

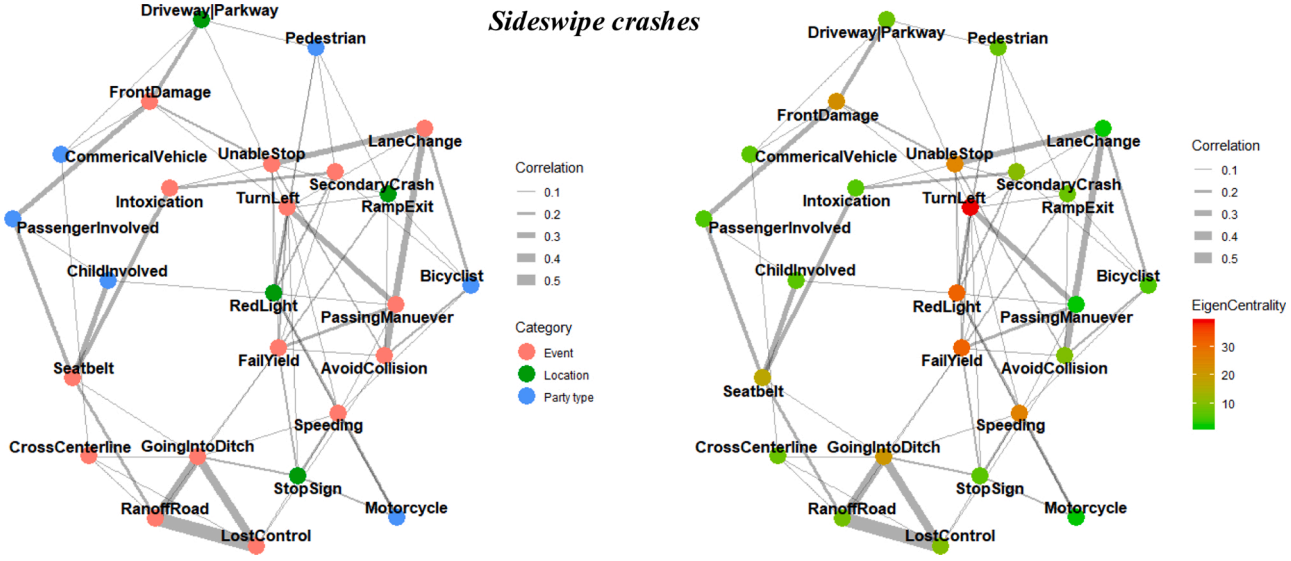

Fig. 14. Network topology and centrality measures between topics for sideswipe vehicle fatal crashes.

location-based topic *RampExit* was the most important feature used to classify single motor vehicle crashes followed by *UnableStop*. Recall, *RampExit* was not the most prevalent topic in single motor vehicle crashes. Therefore, the most prevalent topic for a given crash type was not necessarily the most important classifier. The relative importance of the classifier was, therefore, a function of its prevalence and exclusivity for a given crash type. The classification results suggest that the pre-crash events, locations and involved party's information extracted from the crash narratives can be used to automate the classification of a crash by type. The same procedure can be useful for spotting possible crash typing inconsistencies in the crash reports.

5. Conclusions

The present study utilized text and data mining techniques to understand the prevalence and interaction of various factors that are attributable to crash events. The incorporation of text data in the analysis which for our case was the police crash narratives enabled us to gain valuable insights into the prevalence and interactions between pre-crash events, crash location and involved party in a crash.

Among the factors that hinder the integration of textual data in various domains despite the proliferation of digital textual archives is difficulty in automating the information extraction process. A significant effort is needed in the text data processing which may involve data extraction, data cleaning, and information extraction. By using the most recent text and data mining techniques, this study showed that most of the text processing tasks can now be accomplished automatically. In this study, a java-based algorithm was created to extract crash ID and crash narratives automatically from the police crash reports which were in portable document format. The crash narratives were later joined with structured crash metadata. As for the information extraction, the structural topic modeling (STM) was used to generate topics from crash narratives. The STM offers a smooth integration between inferred topics and structured data as discussed in the methodology section. An automated data processing and information extraction process which was demonstrated in this study can greatly enhance a data-driven decision-making process whereby informed decisions are made using diverse structured and unstructured data sources.

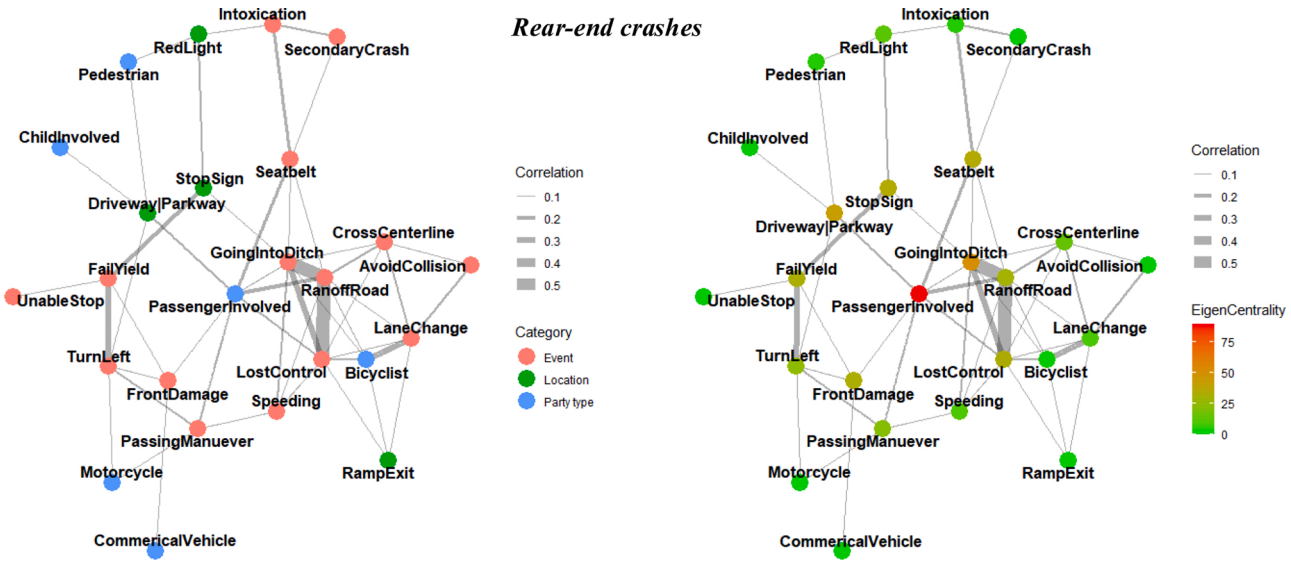

Fig. 15. Network topology and centrality measures between topics for rear-end vehicle fatal crashes.

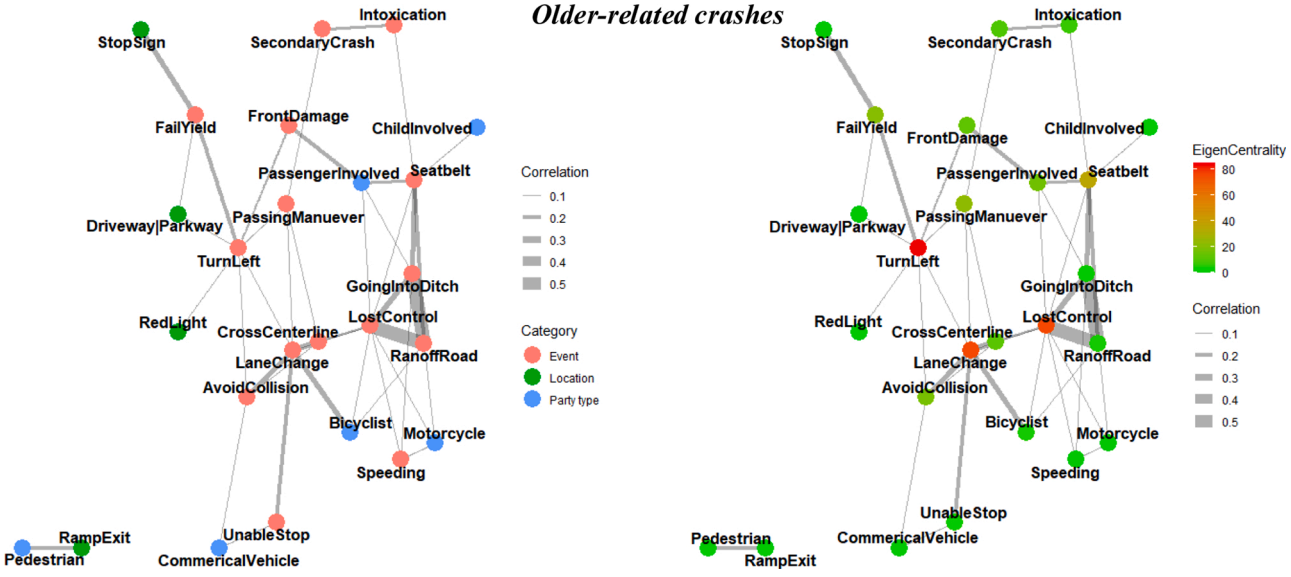

Fig. 16. Network topology and centrality measures between topics for older-related crashes.

The topics discovery from unstructured textual data and estimation of topics' relationships to the indexed crash metadata has the potential to augment new insights in the transportation safety domain. For example, the STM generated topics from fatal crash narratives indicated that topics such as passenger involvement in a crash, crossing the centerline, intoxication, and speeding were more prevalent to young drivers while topics such as turning left, failed to yield, lane changing were prevalent among older drivers. The observed prevalence of topic over time in conjunction with findings from previous studies provides a basis for further research on topics such as the effect of a peer passenger, reckless driving, distraction, and social influence on young drivers. Further, the findings emphasize the need for innovative countermeasures that can compensate for older drivers' decline in visual, cognitive, and sensory abilities related to driving.

The determination of direct and indirect links between crash contributing is among the key elements in the modeling of traffic crashes. This study offers an alternative and efficient way of exploring such complex interaction of crash contributing factors using a crash network topology. The study illustrated how the direct and indirect links between topics can be used to find the most causal chain of events leading up to a crash. Ultimately, the crash network topology can assist in the filtering of topics that have the highest predictive or explanatory value in crash causation models.

The effectiveness of crash prevention programs and other subsequent use of crash reports are dependent upon proper and complete crash investigation and report writing (Alaska Department of Public Safety, 2016). The indexed crash attributes such as crash type, most harmful events, weather conditions, time of the day, among others are manually coded by the police officer at the crash scene after a thorough examination of all elements that contributed to a crash. Such a procedure is prone to human errors. For example, the police officers might mistakenly select an incorrect crash type code from a list of predefined codes available in the crash report instruction manual. The text mining and machine learning become a handy tool under such circumstances whereby extracted information from the crash narrative can assist in quality control of the crash reports. The potential efficacy of the proposed quality control procedure has been demonstrated in this study. The generated themes from crash reports were used to assign a crash

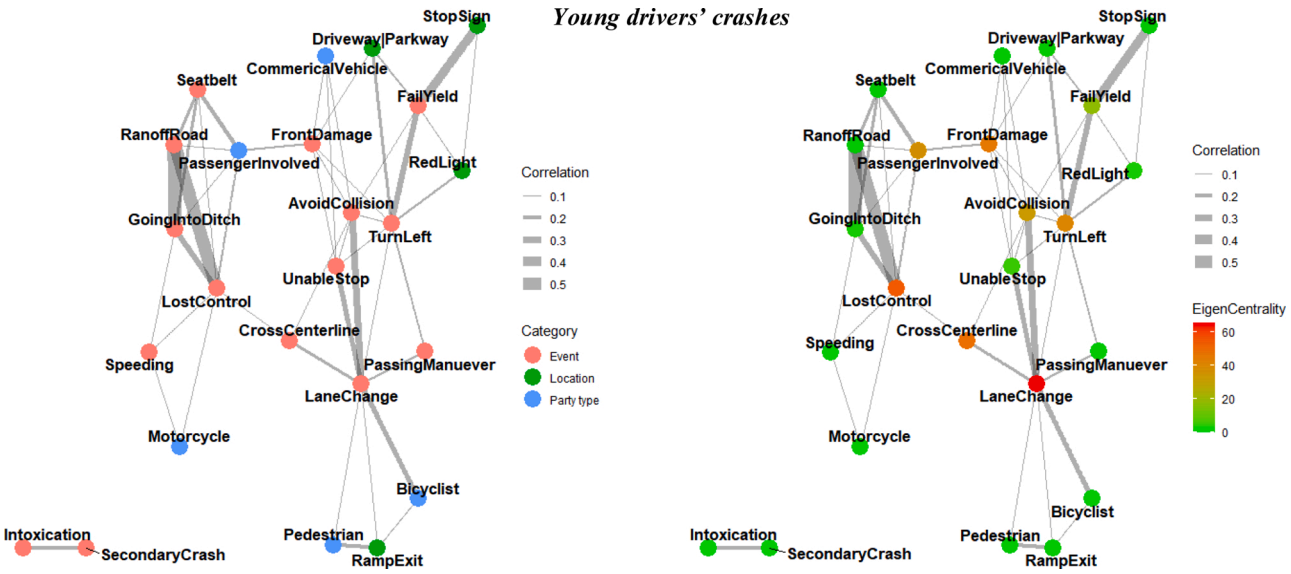

Fig. 17. Network topology and centrality measures between topics among young drivers' crashes.

Table 3
Held-out accuracy for different crash types.

Crash Type	Angle	Head-On	Other	Rear-End	Sideswipe	Single-MV
Accuracy	99.1 %	98.1 %	92.5 %	98.2 %	89.3 %	99.2 %

type label to each crash incident with decent accuracy ranging from 89 % to 99 %. Therefore, the proposed framework can be part of the advanced and rigorous quality control of crash reports.

It should be noted that the results obtained from this study apply mostly to the State of Michigan, but the text-based analytical framework developed in this study can be applied in any safety study with versatile textual data archives. Future endeavors could involve expanding the analysis of the crash narratives and the crash metadata to cover multiple US states or regions. The underlying topics and interactions between topics are expected to change by states or regions as there is a wide

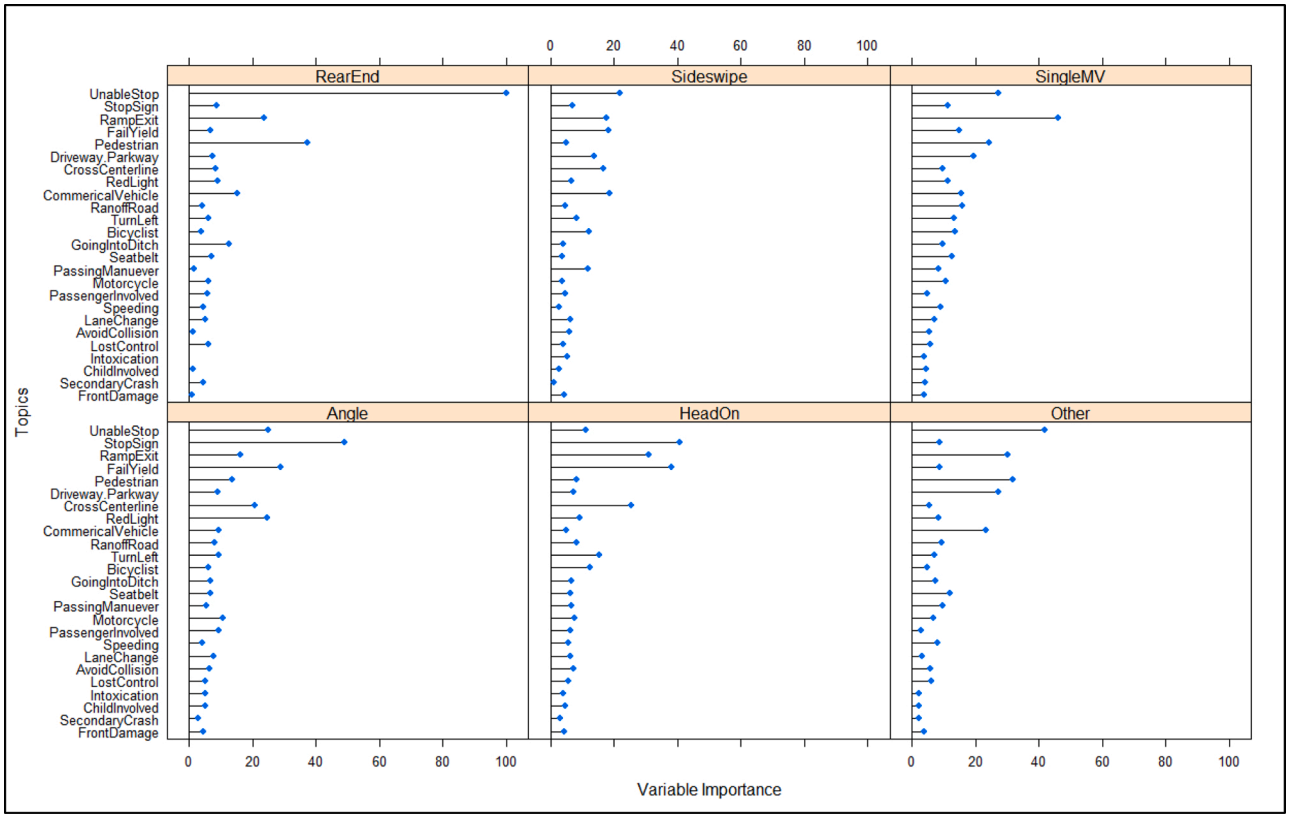

Fig. 18. Importance of a variable in classifying the crash types.

variation of driving behaviors and the state's leniency to traffic violations across states or regions.

CRediT authorship contribution statement

Keneth Morgan Kwayu: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft. **Valerian Kwigizile:** Formal analysis, Supervision, Resources, Writing - review & editing. **Kevin Lee:** Methodology, Formal analysis, Writing - review & editing. **Jun-Seok Oh:** Formal analysis, Supervision.

Declaration of Competing Interest

The authors report no declarations of interest.

References

Abuhay, T.M., Kovalchuk, S.V., Bochenina, K., Mbogo, G.K., Visheratin, A.A., Kampis, G., Krzhizhanovskaya, V.V., Lees, M.H., 2018. Analysis of publication activity of computational science society in 2001–2017 using topic modelling and graph theory. J. Comput. Sci. 26, 193–204. https://doi.org/10.1016/j.jocs.2018.04.004.

Alaska Department of Public Safety, 2016. Alaska Motor Vehicle Collision Report(12-200) Instruction Manual.

Association for Safe International Road Travel, 2019. Road Safety Facts — Association for Safe International Road Travel [WWW Document]. URL https://www.asirt.org/safe-travel/road-safety-facts/ (Accessed 2.2.20).

Banks, G.C., Woznyj, H.M., Wesslen, R.S., Ross, R.L., 2018. A review of best practice recommendations for text analysis in r (and a user-friendly app). J. Bus. Psychol. 33 (4), 445–459. https://doi.org/10.1007/s10869-017-9528-3.

Berman, J.J., 2013. In: Berman, J.J.B.T.-P. (Ed.), Chapter 1 - Providing Structure to Unstructured Data. Morgan Kaufmann, Boston, pp. 1–14. https://doi.org/10.1016/B978-0-12-404576-7.00001-0 of B.D. (Ed.).

Bianchi Piccinini, G., Engström, J., Bärgman, J., Wang, X., 2017. Factors contributing to commercial vehicle rear-end conflicts in China: a study using on-board event data recorders. J. Safety Res. https://doi.org/10.1016/j.jsr.2017.06.004.

Bischof, J.M., Airoldi, E.M., 2012. Summarizing topical content with word frequency and exclusivity. In: Proceedings of the 29th International Conference on Machine Learning. ICML, 2012.

Blei, D.M., 2012. Surveying a suite of algorithms that offer a solution to managing large document archives. Cs.Princeton.Edu 77–84. https://doi.org/10.1145/2133806.2133826.

Blei, D.M., Lafferty, J.D., 2006. Dynamic topic models. Proc. 23rd Int. Conf. Mach. Learn. - ICML' 06 113–120. https://doi.org/10.1145/1143844.1143859.

Blei, D.M., Lafferty, J.D., 2007. A correlated topic model. Ann. Appl. Stat. 1 (1), 17–35. https://doi.org/10.1214/07-AOAS114.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. https://doi.org/10.1016/b978-0-12-411519-4.00006-9.

Braitman, K.A., Kirley, B.B., McCartt, A.T., Chaudhary, N.K., 2008. Crashes of novice teenage drivers: characteristics and contributing factors. J. Safety Res. https://doi.org/10.1016/j.jsr.2007.12.002.

Brown, D.E., 2016. Text mining the contributors to rail accidents. IEEE trans. Intell. Transp. Syst. 17 (2), 346–355. https://doi.org/10.1109/TITS.2015.2472580.

Chiesi, A.M., 2015. Network analysis. International Encyclopedia of the Social & Behavioral Sciences, second edition. https://doi.org/10.1016/B978-0-08-097086-8.73055-8.

Cooper, D., Atkins, F., Gillen, D., 2005. Measuring the impact of passenger restrictions on new teenage drivers. Accid. Anal. Prev. https://doi.org/10.1016/j.aap.2004.02.003.

Das, S., Sun, X., Dutta, A., 2016. Text mining and topic modeling of compendiums of papers from transportation research board annual meetings. Transp. Res. Rec. J. Transp. Res. Board 2552 (1), 48–56. https://doi.org/10.3141/2552-07.

Das, S., Mudgal, A., Dutta, A., Geedipally, S.R., 2018. Vehicle consumer complaint reports involving severe incidents: mining large contingency tables. Transp. Res. Rec. 2672 (32), 72–82. https://doi.org/10.1177/0361198118788464.

Das, S., Dutta, A., Lindheimer, T., Jalayer, M., Elgart, Z., 2019. YouTube as a source of information in understanding autonomous vehicle consumers: natural language processing study. Transp. Res. Rec. https://doi.org/10.1177/0361198119842110.

Ehsani, J.P., Haynie, D.L., Luthers, C., Perlus, J., Gerber, E., Ouimet, M.C., Klauer, S.G., Simons-Morton, B., 2015. Teen drivers' perceptions of their peer passengers qualitative study. Transp. Res. Rec. https://doi.org/10.3141/2516-04.

Fu, C., Pei, Y., Wu, Y., Qi, W., 2013. The Influence of Contributory Factors on Driving Violations at Intersections: An Exploratory Analysis. Adv. Mech. Eng. 5, 905075 https://doi.org/10.1155/2013/905075.

Golbeck, J., 2013. Network structure and measures. Analyzing the Social Web. https://doi.org/10.1016/b978-0-12-405531-5.00003-1.

Hasan, S., Ukkusuri, S.V., 2014. Urban activity pattern classification using topic models from online geo-location data. Transp. Res. Part C Emerg. Technol. https://doi.org/10.1016/j.trc.2014.04.003.

Highway Traffic Safety Administration, 2019. 2018 Fatal Motor Vehicle Crashes: Overview.

Kinra, A., Kashi, S.B., Pereira, F.C., Combes, F., Rothengatter, W., 2019. Chapter 8 - textual data in transportation research: techniques and opportunities. In: Antoniou, C., Dimitriou, L., Pereira Big Data and Transport Analytics, F.B.T.-M.P (Eds.), Mobility Patterns, Big Data and Transport Analytics. Elsevier, pp. 173–197. https://doi.org/10.1016/B978-0-12-812970-8.00008-7.

Kuhn, K.D., 2018. Using structural topic modeling to identify latent topics and trends in aviation incident reports. Transp. Res. Part C Emerg. Technol. 87, 105–122. https://doi.org/10.1016/j.trc.2017.12.018.

Kwayu, K.M., Kwigizile, V., Zhang, J., Oh, J.-S., 2020. Semantic N-Gram feature analysis and machine learning–Based classification of drivers' hazardous actions at signal-controlled intersections. J. Comput. Civ. Eng. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000895.

Lam, L.T., Norton, R., Woodward, M., Connor, J., Ameratunga, S., 2003. Passenger carriage and car crash injury: a comparison between younger and older drivers. Accid. Anal. Prev. https://doi.org/10.1016/S0001-4575(02)00091-X.

Langford, J., Koppel, S., 2006. Epidemiology of older driver crashes - identifying older driver risk factors and exposure patterns. Transp. Res. Part F Traffic Psychol. Behav. https://doi.org/10.1016/j.trf.2006.03.005.

Lombardi, D.A., Horrey, W.J., Courtney, T.K., 2017. Age-related differences in fatal intersection crashes in the United States. Accid. Anal. Prev. 99, 20–29. https://doi.org/10.1016/j.aap.2016.10.030.

Marsden, P.V., 2015. Network centrality, measures of. International Encyclopedia of the Social & Behavioral Sciences, second edition. Elsevier Inc., pp. 532–539. https://doi.org/10.1016/B978-0-08-097086-8.43115-6

McGwin, G., Brown, D.B., 1999. Characteristics of traffic crashes among young, middle-aged, and older drivers. Accid. Anal. Prev. https://doi.org/10.1016/s0001-4575(98)00061-x.

McLaughlin, S.B., Hankey, J.M., Klauer, S.G., Dingus, T.A., 2009. Contributing Factors to Run-Off-Road Crashes and Near-Crashes Final Report.

Metcalf, L., Casey, W., Metcalf, L., Casey, W., 2016. Chapter 5 – graph theory. Cybersecurity and Applied Mathematics, pp. 67–94. https://doi.org/10.1016/B978-0-12-804452-0.00005-1.

Michigan Department of State Police, 2018. UD-10 Traffic Crash Report Instruction Manual.

Miner, G.D., Elder, J., Nisbet, R.A., 2012. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. https://doi.org/10.1016/C2010-0-66188-8.

Monteiro, J., Robertson, G., Atkinson, B., 2012. Networks in Transportation–Theory. Proc. Ctrf 47th Annual Conference 1–21.

Munro, C.A., Jefferys, J., Gower, E.W., Muñoz, B.E., Lyketsos, C.G., Keay, L., Turano, K. A., Bandeen-Roche, K., West, S.K., 2010. Predictors of lane-change errors in older drivers. J. Am. Geriatr. Soc. 58 (3), 457–464. https://doi.org/10.1111/j.1532-5415.2010.02729.x.

North Carolina DOT, 2006. North Carolina Crash Report Instruction Manual.

Orsi, C., Marchetti, P., Montomoli, C., Morandi, A., 2013. Car crashes: the effect of passenger presence and other factors on driver outcome. Saf. Sci. https://doi.org/10.1016/j.ssci.2013.01.017.

Rafla, N.I., Gauba, D., 2011. Supervised topic models David. Conf. Proc. Midwest Symp. Circuits Syst. (Midwest Symp Circuits Syst) 1–8.

Retting, R.A., Ulmer, R.G., Williams, A.F., 1999. Prevalence and characteristics of red light running crashes in the United States. Accid. Anal. Prev. https://doi.org/10.1016/S0001-4575(99)00029-9.

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G., 2014. Structural topic models for open-ended survey responses. Am. J. Pol. Sci. 58 (4), 1064–1082. https://doi.org/10.1111/ajps.12103.

Roberts, M.E., Stewart, B.M., Airoldi, E.M., 2016. A model of text for experimentation in the social sciences. J. Am. Stat. Assoc. 111 (515), 988–1003. https://doi.org/10.1080/01621459.2016.1141684.

Roberts, M.E., Stewart, B.M., Tingley, D., 2019. Stm: r package for structural topic models. J. Stat. Softw. 91, 2. https://doi.org/10.18637/jss.v000.i00.

Robinson, S.D., 2019. Temporal topic modeling applied to aviation safety reports: a subject matter expert review. Saf. Sci. 116 (March), 275–286. https://doi.org/10.1016/j.ssci.2019.03.014.

Rolison, J.J., Regev, S., Moutari, S., Feeney, A., 2018. What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records. Accid. Anal. Prev. 115 (March), 11–24. https://doi.org/10.1016/j.aap.2018.02.025.

Roque, C., Lourenço Cardoso, J., Connell, T., Schermers, G., Weber, R., 2019. Topic analysis of Road safety inspections using latent dirichlet allocation: a case study of roadside safety in Irish main roads. Accid. Anal. Prev. https://doi.org/10.1016/j.aap.2019.07.021.

Schattler, K.L., Datta, J.K., 2004. Driver behavior characteristics at Urban signalized intersections. Transp. Res. Rec. https://doi.org/10.3141/1862-03.

Sivak, M., Schoettle, B., Reed, M.P., Flannagan, M.J., 2007. Body-pillar vision obstructions and lane-change crashes. J. Safety Res. 38 (5), 557–561. https://doi.org/10.1016/j.jsr.2007.06.003.

Taddy, M.A., 2012. On estimation and selection for topic models. J. Mach. Learn. Res.

Trappey, C., Wu, H.-Y., Liu, K.-L., Lin, F.-T., 2013. Knowledge discovery of service satisfaction based on text analysis of critical incident dialogues and clustering methods. Proc. - 2013 IEEE 10th Int. Conf. E-Bus. Eng. https://doi.org/10.1109/ICEBE.2013.40. ICEBE 2013 265–270.

Victor, T., Bärgman, J., Dozza, M., Rootzén, H., Lee, J.D., Ahlström, C., Bagdadi, O., Engström, J., Zholud, D., Ljung-Aust, M., 2013. Safer glances, driver inattention, and crash risk: an investigation using the SHRP 2 naturalistic driving study. Initial Analyses From the SHRP 2 Naturalistic Driving Study: Addressing Driver Performance and Behavior in Traffic Safety. https://doi.org/10.17226/22621.

Woodrooffe, J., Blower, D., Bao, S., Bogard, S., Flannagan, C., Green, P.E., LeBlanc, D., 2014. Performance Characterization and Safety Effectiveness Estimates of Forward

Collision Avoidance and Mitigation Systems for medium/heavy Commercial Vehicles. Ann Arbor, MI.
Zafari, B., Ekin, T., 2019. Topic modelling for medical prescription fraud and abuse detection. J. R. Stat. Soc. Ser. C Appl. Stat. 68 (3), 751–769. https://doi.org/10.1111/rssc.12332.
Zhong, B., Pan, X., Love, P.E.D., Ding, L., Fang, W., 2020. Deep learning and network analysis: classifying and visualizing accident narratives in construction. Autom. Constr. https://doi.org/10.1016/j.autcon.2020.103089.