Text Mining to Decipher Free-Response Consumer Complaints: Insights From the NHTSA Vehicle Owner's Complaint Database

Mahtab Ghazizadeh, **Anthony D. McDonald**, and **John D. Lee**, University of Wisconsin–Madison, Madison, WI, USA

Objective: This study applies text mining to extract clusters of vehicle problems and associated trends from free-response data in the National Highway Traffic Safety Administration's vehicle owner's complaint database.

Background: As the automotive industry adopts new technologies, it is important to systematically assess the effect of these changes on traffic safety. Driving simulators, naturalistic driving data, and crash databases all contribute to a better understanding of how drivers respond to changing vehicle technology, but other approaches, such as automated analysis of incident reports, are needed.

Method: Free-response data from incidents representing two severity levels (fatal incidents and incidents involving injury) were analyzed using a text mining approach: latent semantic analysis (LSA). LSA and hierarchical clustering identified clusters of complaints for each severity level, which were compared and analyzed across time.

Results: Cluster analysis identified eight clusters of fatal incidents and six clusters of incidents involving injury. Comparisons showed that although the airbag clusters across the two severity levels have the same most frequent terms, the circumstances around the incidents differ. The time trends show clear increases in complaints surrounding the Ford/Firestone tire recall and the Toyota unintended acceleration recall. Increases in complaints may be partially driven by these recall announcements and the associated media attention.

Conclusion: Text mining can reveal useful information from free-response databases that would otherwise be prohibitively time-consuming and difficult to summarize manually.

Application: Text mining can extend human analysis capabilities for large free-response databases to support earlier detection of problems and more timely safety interventions.

Keywords: latent semantic analysis, LSA, hierarchical clustering, cluster visualization, free-response data analysis, vehicle failures, crash analysis, vehicle technology, Toyota unintended acceleration, Ford/Firestone tire recall

Address correspondence to Mahtab Ghazizadeh, University of Wisconsin–Madison, 3217 Mechanical Engineering, 1513 University Ave., Madison, WI 53706, USA; ghazizadeh@wisc.edu.

HUMAN FACTORS
201X, Vol. XX, No. X, Month 2014, pp. 1–15
DOI: 10.1177/0018720813519473

INTRODUCTION

Recent advances in vehicle technology, such as Internet connectivity, navigation systems, collision warning systems, and adaptive cruise control, promise to change the driving experience in the coming years. These changes will extend the impressive contributions that technology has made to automotive safety, but their full effect is uncertain. Driving simulators, naturalistic data, and crash statistics all contribute to a better understanding of how drivers respond to changing vehicle technology (Evans, 1993; Fisher, Caird, Rizzo, & Lee, 2011; Jermakian, 2012; McLaughlin, Hankey, & Dingus, 2008), but other sources are needed for a more complete understanding. Analysis of free-response incident descriptions, such as those contained in customer surveys and complaint databases, could provide a valuable resource to complement knowledge from other sources (Ghazizadeh & Lee, 2012; Lehto, Park, Park, & Lehto, 2007).

The National Highway Traffic Safety Administration's (NHTSA) vehicle owner's complaint database encompasses more than 612,262 incident reports (as of July 10, 2012), based on Vehicle Owner's Questionnaire (VOQ) complaint entries. The complaints have been filed through NHTSA's Internet Vehicle Owner's Questionnaire (IVOQ; NHTSA, Office of Defects Investigation [ODI], n.d.), hotline VOQ, consumer letters, and other channels since January 1, 1995. Few published reports document formal analysis of this database. The most prominent of these reports examined complaints related to the Toyota and Lexus unintended acceleration problem since 1999, concluding that not all of the causes of this problem had been identified and rectified (Kane, Liberman, DiViesti, & Click, 2010). The results from Kane et al. (2010) suggest that a more thorough analysis of the database might

reveal trends corresponding to previous recall-inducing problems and trends that signal future problems.

The vehicle owner's complaint database includes information regarding the vehicle (e.g., manufacturer's name, vehicle make and model, and model year), incident (e.g., involvement of fire or crash), number of injuries and/or deaths, fuel type, and other descriptors of the incident. In addition to this categorical information, the database contains a field, "Description of the Complaint," that describes consumers' account of the vehicle problem and its consequences. This study focuses on the free-response incident descriptions contained in this field, using an analytic framework that efficiently handles the complexity of narrative data and extracts clusters of complaints, in a way that is very difficult for a human analyst, particularly if clusters of similar complaints need to be extracted and tracked over time.

Previous work has explored the utility of computerized text analysis techniques in decoding narrative data related to incidents. Lehto and colleagues used machine learning methods to categorize motor vehicle crashes (maintained by an insurance company), occupational injuries (filed with a workers' compensation insurance provider), and general population injuries (from the U.S. National Health Interview Survey [NHIS]; Lehto, Marucci-Wellman, & Corns, 2009; Lehto & Sorock, 1996; Noorinaeini & Lehto, 2006; Wellman, Lehto, Sorock, & Smith, 2004). A variety of Bayesian and regression methods were used (i.e., fuzzy Bayesian, naïve Bayesian, singular value decomposition [SVD] Bayesian, and SVD regression) to assign narratives to a preexisting set of categories. The results showed that these models were highly sensitive and accurate, producing results that agreed with expert assignments in as many as 90% of cases. One review of textual analysis in health care research found that textual data were beneficial for identifying injury cases, extracting additional information about those cases, and assessing the accuracy of a numerical coding scheme for injury identification (McKenzie, Scott, Campbell, & McClure, 2010). Such semi-automated approaches that use a combination of predetermined categories and statistical methods (e.g., Bayesian models and clustering) have clear value.

A more automated approach would use the textual data to identify coherent and correlated subsets of incidents, without committing to a predefined set of categories. Such an approach can highlight trends in particular incident clusters that could go unnoticed if the narratives are indexed by predefined categories. This study follows such an approach, using latent semantic analysis (LSA) to identify similar incidents (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Dumais, 2004; Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988). LSA uses only word co-occurrence to assess the semantic similarity of incidents and does not consider grammatical structure or the sequential relationship between terms (Deerwester et al., 1990). This "bag of words" representation of documents makes it possible to treat text as numerical data.

LSA has been successfully used in classifying and retrieving text-based patient record data (Chute, Yang, & Evans, 1991) as well as extracting patterns and concepts from psychiatric narratives (Cohen, Blatter, & Patel, 2008). LSA has also been used to flag fall-related injury cases based on unstructured text-based medical records (Tremblay, Berndt, Luther, Foulis, & French, 2009). In general, LSA has the capacity to analyze collections of hundreds of thousands of documents, but collections of millions of documents might require sampling or other simplifying techniques (Dumais, 2004). To our knowledge, no paper has been published in *Human Factors, Ergonomics*, or *Applied Ergonomics* using LSA to interpret incident reports; however, LSA is a robust method, with a long history that suggests it is suitable for application to incident reports, such as those considered in this paper (Landauer & Dumais, 1997).

The goal of this study was to extract information about severe incidents (i.e., involving deaths and injuries) from the vehicle owner's complaint database. More broadly, this study aims to demonstrate the utility of text mining approaches in analyzing textual data contained in free-response consumer review or complaint databases. No a priori hypothesis was made regarding the outcomes—the analysis was conducted in a fully exploratory fashion. However, it was expected that the technological changes over the past decades might lead to changes in the types and frequencies of complaints.

TABLE 1: Number and Percentage of Cases in the Database Organized by the "Injury" and "Deaths" Fields and the Median Length of Complaints

Injury and Death Involvement	Number of Cases	Proportion of the Database (%)	Median Number of Words in a Complaint
At least one death reported	**1,858**	**0.30**	**267**
No deaths and at least one injury reported	**18,489**	**3.02**	**246**
No injuries or deaths reported	221,092	36.11	250
No injuries reported, missing fatality data	2,026	0.33	256
At least one injury reported, missing fatality data	13,953	2.28	225
No deaths reported, missing injury data	6,912	1.13	59
Missing both injury and fatality data	347,932	56.83	212
Overall	612,262	100.00	222

Note: Bolded rows correspond to cases evaluated in this analysis.

METHOD

Vehicle Owner's Complaint Database

The vehicle owner's complaint database used in this study contained 612,262 incidents reported since 1995 (NHTSA, ODI, 2011). This analysis focused exclusively on the free-response description provided by drivers in the Description of the Complaint field. Two additional fields (i.e., "Injured," reporting the number of persons injured, and "Deaths," reporting the number of fatalities) were used to identify the relevant cases. More specifically, the two severity categories investigated were (a) entries involving at least one death and (b) entries involving at least one injury but no deaths. The number of incidents in each category was unbalanced. Incidents involving injury accounted for approximately 3% of the database, whereas fatal incidents accounted for approximately 0.3% of the database. Table 1 shows the distribution of cases defined by the "Injured" and "Deaths" fields. The top two rows, bolded in the table, correspond to the cases evaluated in this analysis. The remaining rows were excluded because of our focus on severe incidents. To account for the disparity in the size of each category, a uniform random sample of 2,000 complaints involving injury was analyzed. This sample is approximately equivalent to the number of fatal incidents, that is, 1,858 entries. The sample was generated using the "some" function in R's car package (Fox & Weisberg, 2013). The representativeness of this sample was confirmed through a cross-validation approach, which achieved similar results for all analyses on repeated samples of 2,000 complaints involving injury.

Data Preprocessing

The raw data from the Description of the Complaint field were prepared for the text mining analysis with a number of preprocessing steps using the tm package (Feinerer, Hornik, & Meyer, 2008) and the Snowball package (Hornik, 2009) in R 2.15.2 (R Development Core Team, 2011). Table 2 describes these preprocessing steps.

The steps described in Table 2 are common in text mining. The most contentious step in this process is the removal of custom stop words. In general, there are no specific guidelines for the choice of custom stop words, and thus they should be selected through a thorough consideration of the goals of the analysis and the potential for information loss. Misspellings and other text entry errors were not addressed directly; however, a general review of the data suggested that misspellings were rare. The preprocessing resulted in a reduction of the median number of terms (stemmed words) in the injury and fatal complaints to 112 and 111 terms, respectively.

TABLE 2: Summary, Descriptions, and Justification of the Preprocessing Steps Used in This Analysis

Preprocessing Steps	Description/Justification
1. Raw text data were converted into a corpus data structure	The corpus data structure is the data structure used for text analysis in the tm package (Feinerer, 2013).
2. The case of each letter in the corpus was converted to lowercase and punctuation was removed	This step facilitates classification of words. Punctuation and capitalization do not provide a significant benefit in this type of analysis because word co-occurrence is based on the frequency of words occurring together, rather than their position in sentences or their position relative to punctuation.
3. Stop words were removed from the data	Two types of stop words were removed from the data: generic stop words such as "a" and "the" and custom stop words. Removal of generic stop words is a standard step in many text mining applications as these words occur exponentially more frequently than others and thus provide very little benefit for classification. The removal of custom stop words served the same goal in this analysis. These custom stop words consisted of 57 words drawn from three categories: vehicle-specific words (e.g., "vehicle," "drive," "road"), incident-specific words (e.g., "accident," "investigation," "incident"), and vehicle problem irrelevant words (e.g., "daughter," "husband," "travel").
4. The remaining words were reduced to stems	Stemming is a method of reducing variance in the data by converting words to their radical. For example, "accelerator," "accelerate," "accelerating," and "acceleration" were all reduced to their radical, "acceler." Like the stop word removal step, this step is common in many text mining applications and improves data mining outcomes.

After the corpus was preprocessed, a term–document matrix was created. This matrix provides a complete mapping of the terms found in the corpus to the individual documents (here, complaints). Each row of the term–document matrix corresponds to a term in the corpus and each column to a complaint. The values in the matrix cells describe the frequency of each term in each document (Feinerer, 2013). Sparse terms (i.e., terms that occurred in fewer than 5% of the complaints) were removed.

Analysis

LSA provides an efficient way of identifying similar documents by reducing the dimensions of the term–document matrix. Similar to principal components analysis, dimensionality reduction is achieved through a reduced-rank SVD performed on the term–document matrix, such that only the k largest singular values are retained. The reduced-dimension SVD representation is the best k-dimensional approximation to the original matrix, in terms of least squares. Each term and document is represented as a k-dimensional vector in the new space derived by the SVD. The distances (dissimilarities) between documents are then calculated in this reduced-dimension space using cosine distance.

Based on the distance between documents, clusters of similar complaints can be identified using divisive hierarchical clustering that starts with all complaints grouped in a single cluster. The complaints are then split into smaller clusters based on the cosine distance between them,

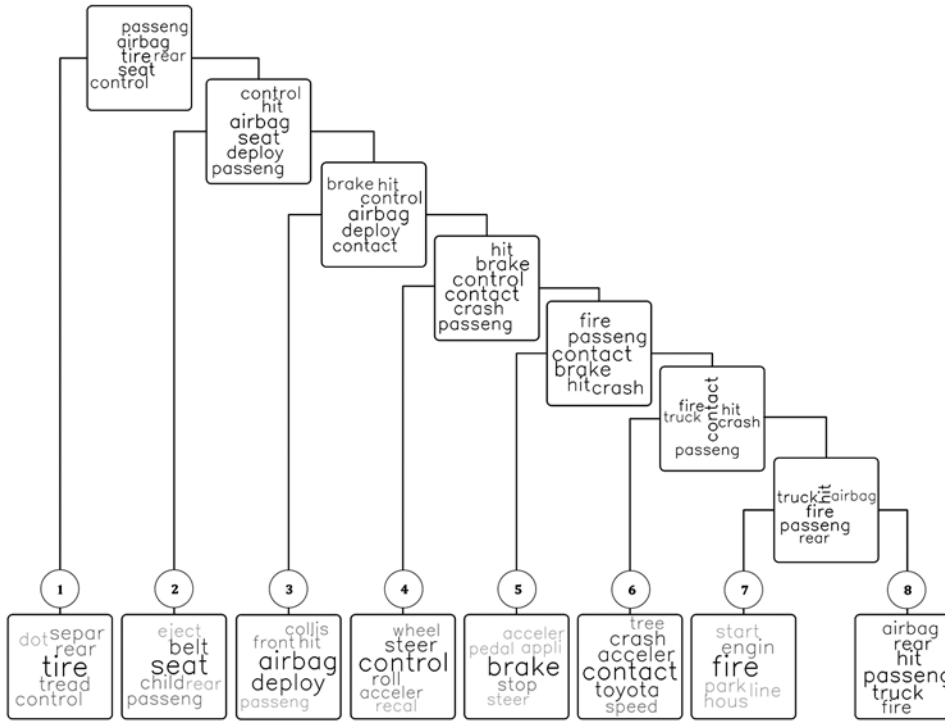


Figure 1. Hierarchical clustering results for fatal incidents.

that is, complaints with similar content, as defined by cosine distance, fall into one cluster (Zhao & Karypis, 2002). The divisions continue until a stopping criterion is reached. The sequential division of complaints into clusters is depicted by a dendrogram—a tree representation of clusters resulting from the successive division of the clusters.

RESULTS

The analysis was conducted using the lsa package in R (Wild, 2011). Separate analyses were conducted for complaints related to each severity level, that is, fatal incidents and incidents involving injury. Figures 1 and 2 show the resulting hierarchies of clusters, which Tables 3 and 4 describe in more detail. The choice of stopping criterion for clustering depends on the goal and scope of the analysis. Here, clustering was stopped one step before a split occurred that resulted in two clusters with the same most frequent term. For the incidents involving injury, the number of final clusters was chosen as six because further segmentation into seven clusters resulted in two clusters with "airbag" as the most common term. With this criterion, hierarchical clustering was stopped at eight clusters for fatal incidents.

The wordcloud package (Fellows, 2012) was used to visualize the contents of each cluster. The final clusters in Figures 1 and 2 are each labeled with a number, which is used to identify them in the corresponding tables. For simplicity, the results presented after these figures and tables refer to clusters by their most frequent stem in title case; thus, Cluster 1 in Figure 1 is referred to as the Tire cluster. In the case where the most frequent stem is incomplete, that is, "passeng" in Cluster 8 in Figure 1, the cluster is referred to by the simplest completion: the Passenger cluster.

Fatal Incidents

The 1,858 reports representing fatal incidents were reduced to a set of eight clusters. Half of the clusters' most common terms correspond to vehicle components. These include "tire," "seat," "airbag," and "brake." The remaining

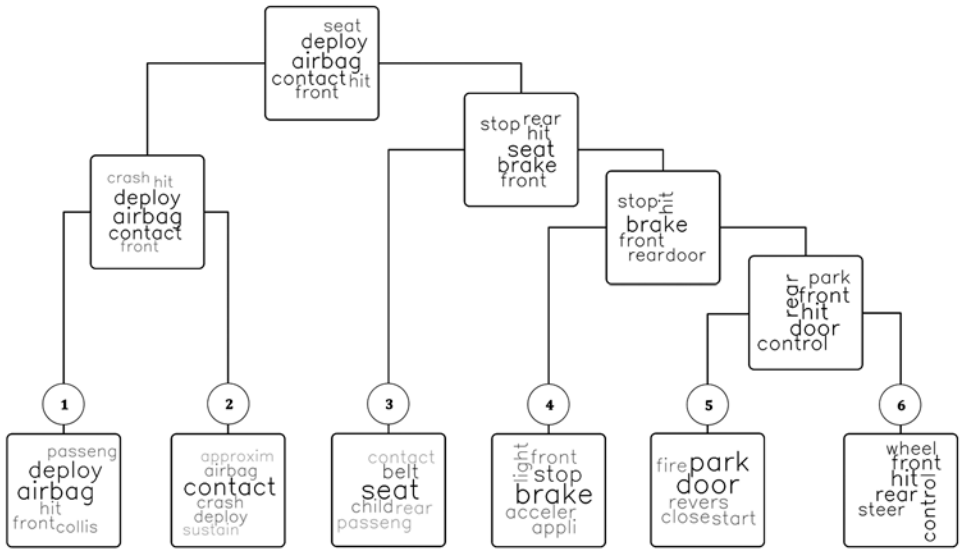


Figure 2. Hierarchical clustering results for incidents involving injury.

most common terms are "control," "contact," "fire," and "passeng." Figure 1 suggests that incident descriptions involving tires (the Tire cluster) are the most different from the rest of the clusters; the Tire cluster has the largest cosine distance from the complaints within other clusters and is the first cluster that separates, followed by Seat, Airbag, Control, Brake, Contact, Fire, and Passenger (in this order). Word clouds at each node visualize the corresponding cluster by its most frequent terms. These terms and their frequencies are listed in Table 3.

Table 3 provides more detail regarding each cluster and the most frequent terms suggest themes for each. The total number of cases in Table 3 (1,641) is less than the number of input cases (1,858) because the removal of stop words and sparse terms eliminates all terms from 217 incidents. Note that these frequencies are in terms of the total number of complaints in the cluster in which a particular term is mentioned. Based on this information, the Tire cluster seems to represent incidents involving a rear tire tread separation and a subsequent loss of control. The theme is typified in the following complaint: "While driving 65 mph rear passenger's tire separated, causing vehicle to skid out of control, and eventually rollover, resulting in minor injuries and a fatality." Although other clusters have less straightforward interpretations, the common terms show a distinct difference in the type of incident represented by each cluster. One term that is common across several clusters, but particularly prevalent in the Contact cluster, is the stem "acceler," referring to "acceleration." The theme characterized by terms such as "contact," "acceler," and "toyota" is unsurprising due to the well-documented Toyota unintended acceleration phenomenon, which resulted in a series of investigations and recalls (U.S. Department of Transportation, 2011). Analysis of the vehicle make (using the field "Vehicle/equipment model") for complaints containing the "acceler" stem revealed that among the 108 complaints containing "acceler," Toyota was the most frequent make, reported in 48 complaints. The next most frequent make was Ford, which was reported in 12 complaints. Of the complaints that reported Toyota as the make, 30 also specifically mentioned the term "toyota" in the incident description.

Incidents Involving Injury

The 2000-complaint sample of incidents involving injuries was separated into six distinct clusters (see Figure 2). Four of the clusters share the most common term with fatal incident clusters: "airbag," "contact," "seat," and "brake."

TABLE 3: The Most Frequent Terms in Each Cluster Identified in Fatal Incidents

Cluster	# of Complaints	Most Frequent Terms	Frequency of the Most Frequent Terms[a]
1	221	tire, rear, tread, separ, control, dot	354, 107, 94, 91, 77, 49
2	155	seat, belt, child, passeng, eject, rear	226, 108, 63, 38, 34, 30
3	182	airbag, deploy, hit, collis, front, passeng	208, 164, 57, 47, 44, 31
4	147	control, steer, roll, acceler, wheel, recal	115, 60, 39, 21, 21, 17
5	99	brake, stop, appli, acceler, pedal, steer	129, 31, 18, 14, 13, 11
6	162	contact, acceler, toyota, crash, speed, tree	126, 77, 69, 67, 34, 32
7	36	fire, engin, park, hous, line, start	49, 12, 7, 6, 6, 6
8	639	passeng, hit, truck, fire, airbag, rear	135, 117, 114, 94, 93, 91

[a]The frequency of a term can be greater than the number of complaints within the corresponding cluster because a term might have been repeated more than once in a single complaint.

TABLE 4: The Most Frequent Terms in Each Cluster Identified in Incidents Involving Injury

Cluster	# of Complaints	Most Frequent Terms	Frequency of the Most Frequent Terms[a]
1	743	airbag, deploy, hit, front, passeng, collis	656, 540, 192, 174, 129, 115
2	248	contact, airbag, crash, deploy, approxim, sustain	486, 118, 118, 106, 52, 52
3	194	seat, belt, child, rear, passeng, contact	293, 99, 50, 45, 44, 34
4	230	brake, stop, acceler, front, appli, light	237, 120, 54, 49, 45, 44
5	68	door, park, revers, close, fire, start	60, 57, 15, 12, 11, 11
6	359	hit, rear, front, control, wheel, steer	133, 110, 108, 96, 77, 76

[a]The frequency of a term can be greater than the number of complaints within the corresponding cluster because a term might have been repeated more than once in a single complaint.

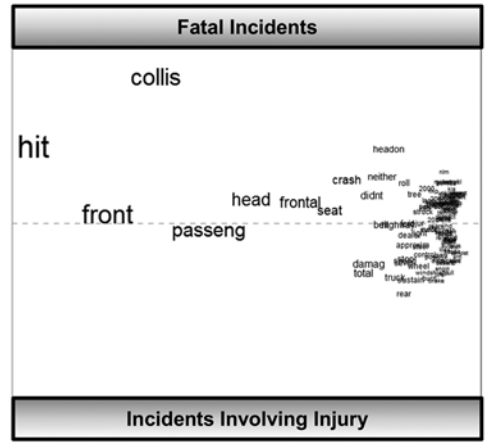
The two remaining clusters have "door" and "hit" as the most common terms, although in the door cluster, the term "park" follows "door" closely (60 and 57 occurrences for "door" and "park," respectively) and in the Hit cluster, the terms "rear" and "front" are both close to "hit" in frequency (133, 110, and 108 occurrences for "hit," "rear," and "front," respectively). The representativeness of the sample was verified using a tenfold cross-validation, which involved repeating the analysis on 10 folds (i.e., samples) of 2,000 cases drawn randomly from the injury data (without replacement). The process identified six clusters for 9 of the 10 folds and five clusters for the last fold. In all 10 folds the "airbag," "contact," "seat," "brake," and "door" clusters were present. The last cluster was "hit" in eight of the 10 folds, whereas one fold repeated the "airbag" cluster and the other had "rear" instead of "hit." Therefore the random sample was deemed to be reliable and representative.

Table 4 summarizes the themes of the incidents included in each cluster in Figure 2. For example, the general theme of the Airbag cluster is frontal collisions involving airbag deployment or malfunction, with passengers in the vehicle. One such accident is described in the following complaint: "2005 Hyundai Santa Fe involved in a severe rollover and all air bags never deployed. The vehicle flipped over 6x with 3 passengers in it. One was ejected from the back and the front two sustained injuries that could have been prevented."

Comparing Common Clusters Across Severity Levels

The hierarchical clustering analysis of incident descriptions resulted in several clusters with the same most common terms for fatal incidents and incidents involving injury, that is, "seat," "airbag," "brake," and "contact." Comparing these common clusters across the two severity levels can highlight differences between incidents at each level. However, it is difficult to quantitatively compare the clusters because they are based on different corpora, with different prominent terms and LSA dimensions. As such, it would be nearly impossible to ensure that the assumptions underlying any statistical test are met. Even so, the clusters can be compared qualitatively by graphing the relative frequencies of the terms extracted from complaints within them. After normalizing for the number of complaints in each cluster, terms can be plotted by overall frequency and frequency relative to each severity level.

Figure 3 compares complaints in the Airbag clusters identified in fatal incidents (182 complaints) and incidents involving injury (743 complaints). The vertical axis shows the frequency of each term relative to each level of severity—the proximity of a term to each pole reflects the relative frequency of that term in the corresponding incident level; terms near the top of the plot are strongly associated with fatal incidents, terms near the bottom are strongly associated with incidents involving injury, and terms near the dotted gray line bisecting the graph are either common across both categories or generally infrequent. The size and horizontal position of the terms reflect their average frequency across all complaints assigned to the Airbag clusters. The plot shows that the terms "collis[ion]," "hit," "headon," and "crash" are more strongly associated with fatalities, whereas the terms "damag[e]," "rear," and "total" are more strongly associated with injuries. The terms "front" and "passeng[er]" are common in both levels of severity. The Airbag clusters in the two severity levels had the same six most frequent terms ("airbag," "deploy," "hit," "collis[ion]," "front," "passeng[er]"), which is not surprising given the circumstances that necessitate airbag deployment and mark failures thereof. The comparison



[too loose]

Figure 3. Comparison of the Airbag clusters identified in fatal incidents and incidents involving injury. The vertical axis shows the frequency of each term relative to each level of severity and the size and horizontal position of the terms reflect their average frequency. In plotting this graph, the term "airbag" was removed from both clusters, as it had a much higher frequency than the other terms and would make it difficult to see any other terms. In addition, those terms that occurred in fewer than 10% of the reports were removed to reduce clutter.

in Figure 3 highlights differences between the two clusters that were otherwise concealed.

Figure 4 shows how this graphical comparison can be extended to incorporate other levels of incident severity. This graph shows Contact cluster terms for fatal incidents, incidents involving injury, and a 2000-complaint sample of incidents involving no fatalities or injuries, but marked in the data as having involved a crash (hereafter, "minor incidents"). The graph shows several terms that are strongly associated with all three levels. "airbag" and "deploy" are most common in the injury incidents, and "toyota" and "acceler[ation]" are most common in fatal incidents. Minor incidents are represented by a diverse set of terms mostly involving repairs, for example, "dealer," "repair," and "replace." These themes reflect the multiple meanings of the term "contact." Contact can be a noun that defines the individual who called in the complaint (as in "the contact stated that her transmission failed"), a verb that implies some type of collision (as in "the car contacted the tree"), or a verb implying

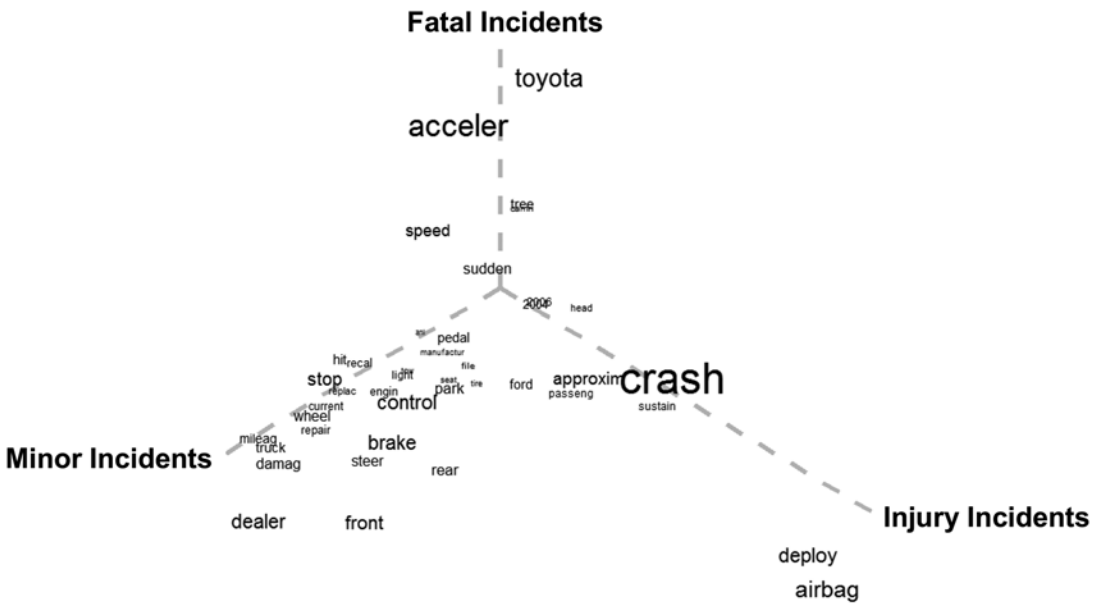


Figure 4. Comparison of the Contact clusters identified in fatal incidents, incidents involving injury, and minor incidents. The size of the terms reflects their average frequency. In plotting this graph, the term "contact" was removed from all three clusters, as it had a much higher frequency than the other terms and would make it difficult to see any other terms. In addition, those terms that occurred in fewer than 10% of the reports were removed to reduce clutter.

some type of communication (as in "he noticed the leak and contacted the dealer"), among other definitions. It is important to consider the variety of possible definitions in this type of analysis when interpreting clusters and avoid assuming a single definition for any term.

Together these analyses indicate how seemingly similar clusters differ according to incident severity. In general, clusters characterize a type of incident and are often centered on a vehicle component, such as tire, or a consequence, such as fire or contact. Similar comparisons could be useful for clusters across other dimensions beyond incident severity. For example, focusing on only fatal incidents, one might create a graph that compares different manufacturers based on their association with the prominent terms in the Contact cluster to better understand the relationship between problems and vehicle makes.

Time Trend Analysis of Clusters

One benefit of this analysis approach is that cluster labels are assigned to each complaint, facilitating the analysis of clusters with respect to other fields of data associated with the original complaint. Incident date and report date are particularly interesting in this case because they can describe time trends in incident frequency and indicate emerging issues that might require intervention. Figure 5 shows the number of complaints in each cluster of fatal incidents reported each year (from 1995 to 2011). The largest disturbance occurs in the Tire cluster around 2000, which corresponds to the Ford/Firestone SUV tire tread separation issue. NHTSA launched a formal investigation into the incidents in 2000 (NHTSA, 2000); however, the time trend clearly shows indications of an increase in tire-related complaints as early as 1998. An analysis of the complaints, similar to what has been done here, may have led to detection as early as 1998 or 1999. The fatal incident data do not show any currently emerging trends, although the Passenger and Contact trends have been highly variable over the past 5 years. Both Passenger and Airbag clusters (with "airbag" as a prominent term) have their peaks in 2003, followed by a generally decreasing trend, although

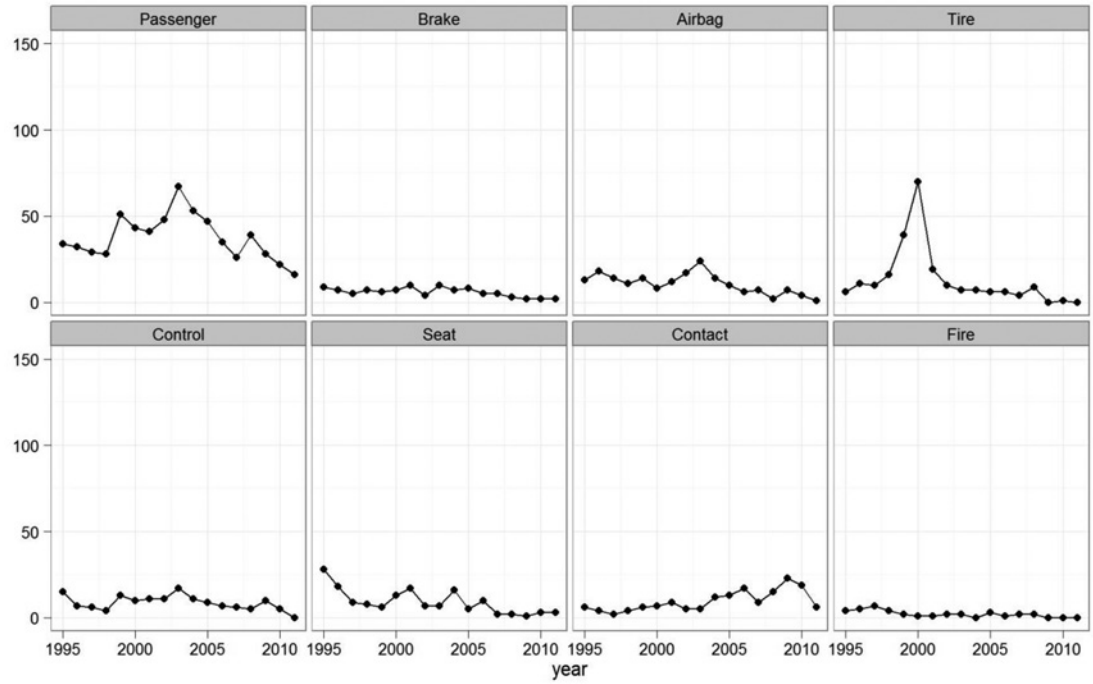


Figure 5. The number of complaints in each cluster of fatal incidents (based on the incident year).

it is difficult to explain the Passenger cluster trend because it represents a variety of themes. The Brake and Fire incidents generally occur infrequently and have very low variability. Incidents in the Control and Seat clusters have small peaks across the 17-year period; however, no consistent increasing or decreasing trend is evident for them.

Figure 6 shows the time trends for the clusters in incidents involving injury. Similar to fatal incidents, the Brake cluster has a low level of incidence and variance. The Airbag cluster has a generally downward trend (although with a few spikes). The overall improving trend may be the result of technological advancements. The Park cluster has a generally low rate of incidents, with some minor peaks over the years. The Contact cluster is concerning because it seems to be generally increasing since 2006. The initial increase seems to reflect the Toyota unintended acceleration issue, although the incident rate has remained at a high level since 2009. This assertion is supported by the fact that the frequency of the terms "toyota," "sudden," and "acceleration" in the complaints peaked in 2006 and 2009. The Hit cluster has fluctuated over the years, with a steady decline during the past 5 years. It is important to note that the Hit cluster is the most general of the injury clusters because it includes several terms with similar frequencies as shown in Table 4. This generality may make this cluster a "catchall" for the incidents not grouped in the other, more coherent clusters.

The analysis so far has focused on the reported date of incident rather than the date of report filing. The report filing date is an important factor in understanding the predictive capabilities of this process. Because reporting is voluntary, report dates may be driven by external events such as announcement of recalls and news stories rather than the incidents themselves. One way to evaluate the effect of such events is to plot the complaint frequency by date of report filing relative to these external events. Figures 7 and 8 show such plots for fatal incidents in the Tire and Contact clusters, respectively. These clusters were selected because they are strongly connected with two widely publicized problems: the Toyota unintended acceleration problem and the Ford/Firestone tire tread separation problem. The

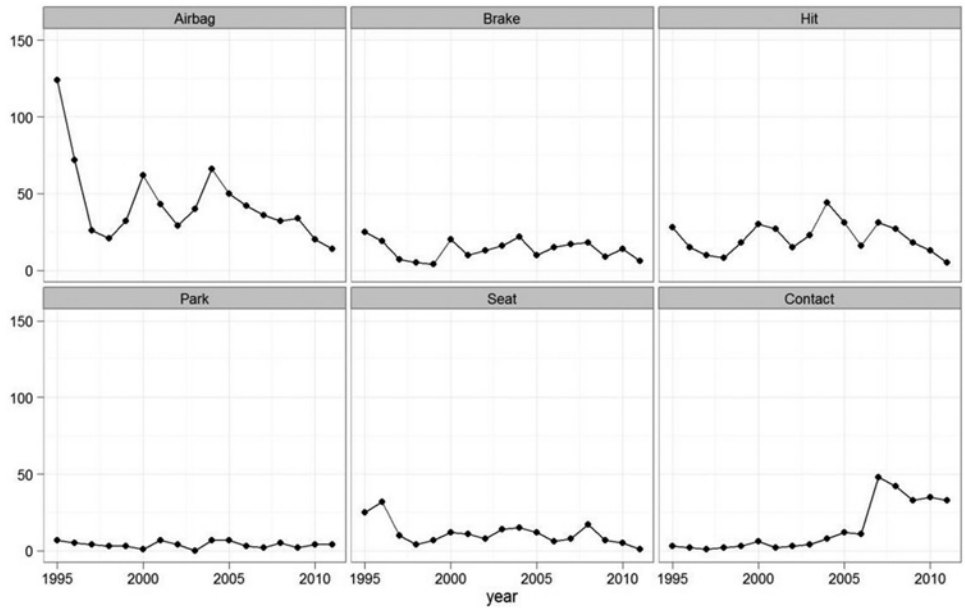


Figure 6. The number of complaints in each cluster of incidents involving injury (based on the incident year).

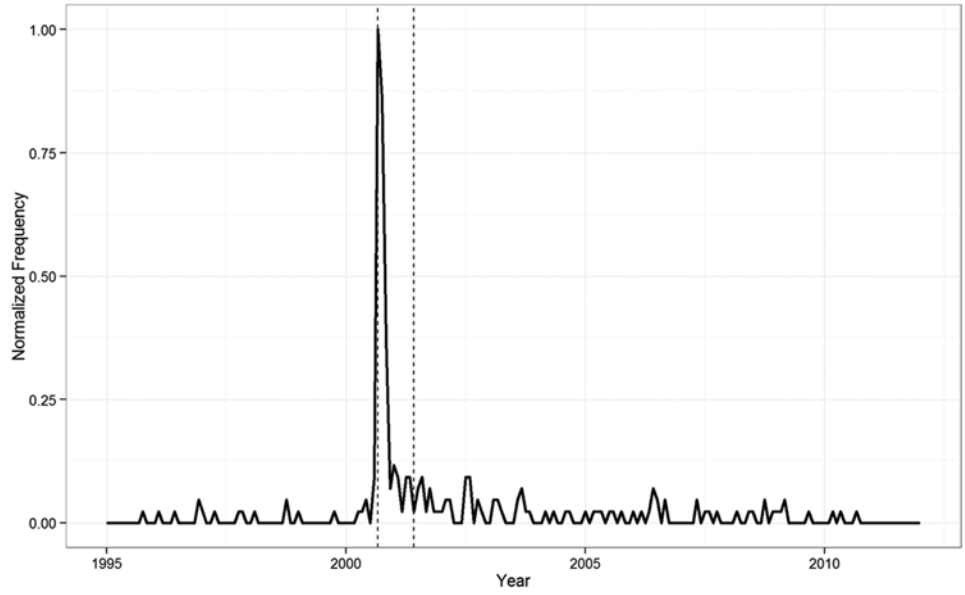


Figure 7. Normalized frequency (on the 0–1 scale) of complaints in the fatal incidents' Tire cluster by report date. The Tire cluster includes many reports related to tread separation. Dashed lines denote the dates of recall or replacement announcements.

Toyota unintended acceleration problem resulted in two recalls on October 5, 2009, and January 21, 2010 (NASA Engineering and Safety Center, 2011). The Ford/Firestone tread separation problem resulted in a recall from Firestone on August 9, 2000 (NHTSA, 2000), and a major replacement announcement by Ford on May 22, 2001 (NHTSA, 2001). The dashed lines on each figure

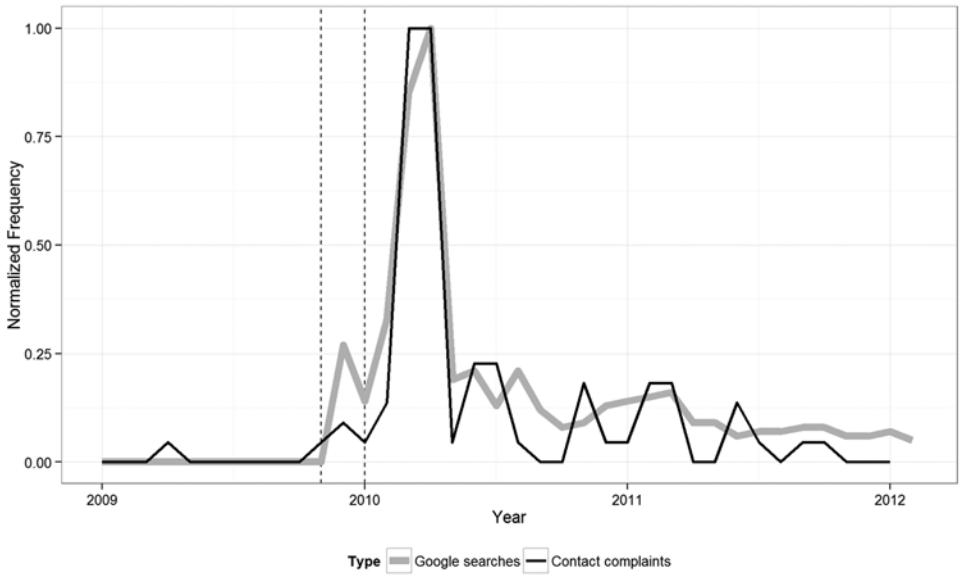


Figure 8. Normalized frequency (on the 0–1 scale) of complaints in the fatal incidents' Contact cluster by report date and Google searches for "unintended acceleration." Dashed lines denote the dates of recall announcements.

show these dates. In both plots, major spikes in reports occurred after the recalls. In the Tire cluster, the highest count of incident reporting occurred immediately after the initial recall. In contrast, the largest spike in Contact cluster incident reporting occurred several months after the initial recall. This increase in the public awareness is reflected in the trend of "unintended acceleration" searches on Google in the United States (see Figure 8)—a steep increase started in November 2009, and the peak search frequency was observed in February and March 2010 (Google, 2013). A 16% drop in Toyota sales in January 2010 (coinciding with Toyota's recalling of millions of vehicles and halting the sale of several models) provides further evidence of the public reaction ("Toyota Sales Drop," 2010). The replacement announcement in the Tire cluster seems to have had little effect on the report frequency. Together the plots in Figures 7 and 8 suggest that public awareness has a significant effect on reporting behavior. This influence might limit the predictive capacity of the method described here as related to the complaint database—a limitation that is an artifact of complaint submission behavior and public awareness rather than the analytic method.

DISCUSSION

The results demonstrate the utility of text mining approaches in analyzing free-response complaints. LSA and hierarchical clustering produced generally coherent clusters of complaints, and the time trends of these clusters captured several high-profile vehicle safety problems: tread separation and unintended acceleration. Clusters of fatal incidents and incidents involving injury showed common themes, as well as important differences. Terms such as "seat," "airbag," "brake," and "contact" occur frequently in clusters of both severity levels. In contrast, fatal incidents frequently include terms such as "tire," "control," "fire," and "passeng[er]." Similar clusters identified in both severity levels can be compared to highlight aspects of incidents shared between severity levels and those that were highly associated with a particular level of severity. Such comparison was illustrated using the Airbag cluster as an example, where "collis[ion]," "hit," "headon," and "crash" were more strongly associated with fatalities, whereas "damag[e]," "rear," and "total" were more strongly associated with injuries. These results generally confirm the utility of LSA as a

semiautomated approach of incident coding in identifying emerging themes of incidents.

As with any form of automation, careful attention to the limits of text mining is needed to avoid overreliance and misinterpretation of its output. In particular, ascribing causal relationships between component failures and incident outcomes should be done with care because LSA treats documents as bags of words, ignoring word dependencies, sequences, and nuances of word meaning. A human analyst can easily discern cases where faulty components are responsible for the incident from cases where an incident triggers a component failure, but LSA cannot differentiate these scenarios. An example of the former is a brake failure that causes the vehicle to run into the leading vehicle, and an example of the latter is when an abrupt braking occurs and leads to a crash where the airbag fails to deploy. Even though LSA does not consider word dependencies, it is still powerful in identifying themes in the narrative data and guiding analysts to more in-depth investigations of the identified documents.

The time trends presented here are based on raw incident frequencies—changes in the number of licensed drivers or registered vehicles are not taken into account. Although the lack of such adjustment might mask some trends and inflate others, when the trends in clusters are considered relative to each other, the patterns that stand out are those that indeed merit attention. Also, when the focus is on spikes from a year to the next, the effect of denominator changes would be minimal. Follow-up investigations on the identified problems can incorporate adjustments.

Analysis of the date of recall announcements/news reports found a noteworthy relationship between announcements and an upswing in the number of complaints—media coverage might be responsible for an increase in incident reports. However, in the absence of statistics describing the audience of the complaints database and the demographic information of complainants, no formal analysis of the effect of recalls and associated media coverage on complaints can be performed. Furthermore, because the complaint database is compiled based on a voluntary reporting system, it is not an exhaustive or unbiased record of incidents. Individuals reporting complaints may have varying degrees of technical expertise and might have been influenced by media reports, as reflected by the large upswing in reports following the news coverage of the unintended acceleration incidents. Also, in many cases the individual filing the report (e.g., a lawyer) has been absent from the scene of the incident, making inferences about those filing the reports even more difficult. This analysis did not directly address this issue; however, preprocessing, selection criteria for severity, and sparsity cutoffs focused the analysis on records containing relatively complete descriptions. The issues of sample representativeness and exposure are relevant to other methods of evaluation (e.g., simulator and naturalistic studies, crash data analyses) as well, but are less salient in those domains.

Many of the decisions throughout the analysis were based on heuristics. For example, the choice of stop words was partially driven by finding the most frequent noninformative terms (e.g., "vehicle"). The choice of stop words can influence the outcomes, both in the preprocessing phase and in clustering. For example, the term "passeng" (the stemmed form of "passenger") occurs in several of the clusters. One could argue that "passenger" is a stop word in the context of vehicle complaints and should be removed, as would be the case with term frequency inverse document frequency weighting, which removes very frequent and very infrequent words (Oza, Castle, & Stutz, 2009). Such a decision might change the composition of some of the clusters. Parameters such as the stopping criterion in clustering and the sparsity cutoff introduce additional degrees of freedom. As such, even though the modeling framework used here is quantitative, major qualitative considerations enter the analysis. An explicit consideration of the philosophy and techniques of qualitative data analysis and mixed-methods approaches could inform similar analyses, compared to using either a pure quantitative or qualitative approach (Johnson & Onwuegbuzie, 2004).

CONCLUSION

Text mining, specifically LSA and hierarchical cluster analysis, can reveal useful information

from free-response databases that is too time-consuming and difficult to summarize manually. The value of such techniques is likely to grow in the coming years as the automotive industry adopts new technologies that might have unanticipated consequences (e.g., keyless ignition and car hacking). Computerization of vehicles introduces failure modes that are difficult to detect during design, particularly for those systems that change the role of the driver (Merat & Lee, 2012). Although incident reports have important limitations, they still offer an important complement to the traditional methods of simulator-based testing and crash statistics.

The procedures outlined here can be useful in monitoring the vehicle owner's complaint database for emerging problems. More generally, text analysis approaches can be applied to a variety of consumer complaint databases or even Twitter feeds to capture the temporal and spatial trends in opinions, habits, or events (Golder & Macy, 2011). Whereas assigning incidents to predefined categories can let emerging issues go unnoticed, an exploratory analysis like the one described here can indicate emerging themes months or years before they are revealed in crash or naturalistic driving data. This method can be easily systematized (in a number of steps including subset selection, preprocessing, LSA, and clustering) and run periodically to automate the detection of emerging problems and identifying and addressing trends before more lives are endangered. For example, running the analysis at monthly intervals can identify new themes of vehicle problems that merit detailed investigation of the corresponding complaints. This can be achieved through a human–technology partnership in which the identified clusters are reintegrated into the database as an index that directs analysts to the noteworthy subsets of the database, rather than depending on analysts to read large volumes of narrative data and interpret themes.

ACKNOWLEDGMENTS

The authors would like to thank the members of the Cognitive Systems Laboratory (CSL) at the University of Wisconsin–Madison for their valuable input on an earlier version of this article. We would also like to thank William Horrey and the three anonymous reviewers of the article for their insightful comments.

KEY POINTS

- This work demonstrated the utility of text mining and hierarchical clustering in analyzing a free-response database, the National Highway Traffic Safety Administration's vehicle owner's complaint database.
- Clusters of complaints were identified and compared across subsets of data pertaining to fatal incidents and incidents involving injury. Clusters with seat, airbag, brake, and contact as the most frequent terms were identified at both severity levels.
- The time trends showed increased complaints surrounding the Ford/Firestone tire recall and the Toyota unintended acceleration recall.
- Further analyses suggested that increases in complaints might be influenced by recall announcements.

REFERENCES

Chute, C. G., Yang, Y., & Evans, D. A. (1991). Latent semantic indexing of medical diagnoses using UMLS semantic structures. In *Proceedings of the 15th Annual Symposium on Computer Application in Medical Care*. Washington, DC: IEEE Computer Society Press, 185–189.

Cohen, T., Blatter, B., & Patel, V. (2008). Simulating expert clinical comprehension: Adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative. *Journal of Biomedical Informatics*, *41*, 1070–1087.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, *38*, 188–230.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988, June). *Using latent semantic analysis to improve access to textual information*. Paper presented at the CHI'88 Conference on Human Factors in Computing Systems, Washington, DC.

Evans, L. (1993). Restraint effectiveness, occupant ejection from cars, and fatality reductions. *Accident Analysis & Prevention*, *22*, 167–175.

Feinerer, I. (2013). *Introduction to the tm package text mining in R*. Retrieved from http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54.

Fellows, I. (2012). *Wordcloud: Word clouds. R package version 2.0*. Retrieved from http://CRAN.R-project.org/package=wordcloud

Fisher, D. L., Caird, J. K., Rizzo, M., & Lee, J. D. (2011). Handbook of driving simulation for engineering, medicine, and psychology: An overview. In D. L. Fisher, M. Rizzo, J. K. Caird, & J. D. Lee (Eds.), *Handbook of driving simulation for engineering, medicine, and psychology* (pp. 1.1–1.16). Boca Raton, FL: CRC Press.

Fox, J., & Weisberg, S. (2013). *Car: Companion to applied regression version 2.0-18*. Retrieved from http://cran.r-project.org/web/packages/car/index.html

Ghazizadeh, M., & Lee, J. D. (2012, October). *Consumer complaints and traffic fatalities: Insights from the NHTSA vehicle owner's*

complaint database. Paper presented at the Human Factors and Ergonomics Society 56th annual meeting, Boston, MA.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, *333*(6051), 1878–1881.

Google. (2013). *Google trends*. Retrieved from http://www.google.com/trends/

Hornik, K. (2009). *Snowball: Snowball stemmers. R package version 0.0-7*. Retrieved from http://CRAN.R-project.org/package=Snowball

Jermakian, J. S. (2012). Crash avoidance potential of four large truck technologies. *Accident Analysis & Prevention*, *49*, 338–346.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*(7), 14–26.

Kane, S., Liberman, E., DiViesti, T., & Click, F. (2010). *Toyota sudden unintended acceleration*. Rehoboth, MA: Safety Research & Strategies.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lehto, M. R., Marucci-Wellman, H., & Corns, H. (2009). Bayesian methods: A useful tool for classifying injury narratives into cause groups. *Injury Prevention*, *15*, 259–265.

Lehto, M. R., & Sorock, G. S. (1996). Machine learning of motor vehicle accident categories from narrative data. *Methods of Information in Medicine*, *35*, 309–316.

Lehto, X., Park, J. K., Park, O., & Lehto, M. R. (2007). Text analysis of consumer reviews: The case of virtual travel firms. In M. J. Smith & G. Salvendy (Eds.), *Human interface and the management of information: Methods, techniques and tools in information design* (pp. 490–499). Berlin, Germany: Springer.

McKenzie, K., Scott, D. A., Campbell, M. A., & McClure, R. J. (2010). The use of narrative text for injury surveillance research: A systematic review. *Accident Analysis & Prevention*, *42*, 354–363.

McLaughlin, S. B., Hankey, J. M., & Dingus, T. A. (2008). A method for evaluating collision avoidance systems using naturalistic driving data. *Accident Analysis & Prevention*, *40*, 8–16.

Merat, N., & Lee, J. D. (2012). Preface to the special section on human factors and automation in vehicles designing highly automated vehicles with the driver in mind. *Human Factors*, *54*, 681–686.

NASA Engineering and Safety Center. (2011). *Technical support to the National Highway Traffic Safety Administration (NHTSA) on the reported Toyota Motor Corporation (TMC) unintended acceleration (UA) investigation*. Hampton, VA: Author.

National Highway Traffic Safety Administration. (2000). *Firestone tire recall*. Retrieved from http://www.nhtsa.gov/PR/FirestoneRecall

National Highway Traffic Safety Administration. (2001). *Statement of Michael P. Jackson, Deputy Secretary of Transportation, before the Subcommittees on Telecommunications, Trade and Consumer Protection and Oversight and Investigation of the Committee on Energy and Commerce, U.S. House of Representatives*. Retrieved from http://www.nhtsa.gov/nhtsa/announce/press/Firestone/DOTState.html

National Highway Traffic Safety Administration, Office of Defects Investigation. (2011). *Vehicle owner's complaint database*. Retrieved from http://www-odi.nhtsa.dot.gov/downloads/

National Highway Traffic Safety Administration, Office of Defects Investigation. (n.d.). *Internet Vehicle Owner's Questionnaire*. Retrieved from https://www-odi.nhtsa.dot.gov/ivoq/

Noorinaeini, A., & Lehto, M. R. (2006). Hybrid singular value decomposition: A model of human text classification. *International Journal of Human Factors Modelling and Simulation*, *1*, 95–118.

Oza, N., Castle, J. P., & Stutz, J. (2009). Classification of aeronautics system health and safety documents. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *39*, 670–680.

R Development Core Team. (2011). *R: A language and environment for statistical computing* (Unpublished manuscript). Vienna, Austria: Author.

Toyota sales drop 16 percent in January. (2010, February 3). *Washington Times*. Retrieved from http://www.washingtontimes.com/news/2010/feb/03/toyota-sales-drop-16-percent-in-january/

Tremblay, M. C., Berndt, D. J., Luther, S. L., Foulis, P. R., & French, D. D. (2009). Identifying fall-related injuries: Text mining the electronic medical record. *Information Technology and Management*, *10*, 253–265.

U.S. Department of Transportation. (2011). *Technical assessment of Toyota electronic throttle control (ETC) systems*. Washington, DC: National Highway Traffic Safety Administration.

Wellman, H. M., Lehto, M. R., Sorock, G. S., & Smith, G. S. (2004). Computerized coding of injury narrative data from the National Health Interview Survey. *Accident Analysis & Prevention*, *36*, 165–171.

Wild, F. (2011). *lsa: Latent semantic analysis. R package version 0.63-3*. Retrieved from http://CRAN.R-project.org/package=lsa

Zhao, Y., & Karypis, G. (2002, November). *Evaluation of hierarchical clustering algorithms for document datasets*. Paper presented at the Eleventh International Conference on Information and Knowledge Management, McLean, VA.

Mahtab Ghazizadeh is a PhD candidate and graduate research assistant in the Industrial & Systems Engineering Department at the University of Wisconsin–Madison. She earned her MS in industrial engineering from the University of Iowa in 2009.

Anthony D. McDonald is a PhD candidate and graduate research assistant in the Industrial & Systems Engineering Department at the University of Wisconsin–Madison, where he earned his MS in 2012.

John D. Lee is a professor in the Industrial & Systems Engineering Department and director of the Cognitive Systems Laboratory (CSL) at the University of Wisconsin–Madison. He earned his PhD in mechanical engineering from the University of Illinois at Urbana-Champaign in 1992.

Date received: April 1, 2013
Date accepted: November 30, 2013