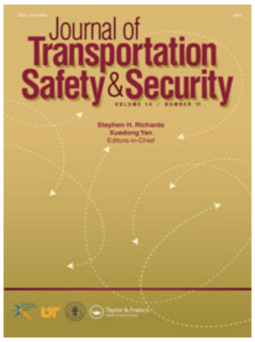


Journal of Transportation Safety & Security

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/utss20

Rules mining on hybrid electric vehicle consumer complaint database

Subasish Das, Zihang Wei & Anandi Dutta

To cite this article: Subasish Das, Zihang Wei & Anandi Dutta (2022): Rules mining on hybrid electric vehicle consumer complaint database, Journal of Transportation Safety & Security, DOI: 10.1080/19439962.2022.2147614

To link to this article: https://doi.org/10.1080/19439962.2022.2147614

Published online: 24 Nov 2022.

JOURNAL OF TRANSPORTATION SAFETY & SECURITY
https://doi.org/10.1080/19439962.2022.2147614

Rules mining on hybrid electric vehicle consumer complaint database

Subasish Das[a], Zihang Wei[b], and Anandi Dutta[c]

[a]Civil Engineering, Ingram School of Engineering, Texas State University, San Marcos, Texas, USA; [b]Zachry Department of Civil & Environmental Engineering, Texas A&M University, College Station, Texas, USA; [c]Department of Computer Science, The University of Texas at San Antonio, San Antonio, Texas, USA

ABSTRACT

The hybrid electric vehicle (HEV) is a critical transportation disruptive technology that is expected to be widely adopted in the current and future marketplace. Many nations are promoting the success of HEVs. As the technologies and designs of these vehicles are significantly different from conventional vehicles, it is also important to understand the technical and body-related issues associated with these vehicles. This study used the National Highway Traffic Safety Administration's vehicle owner's complaint database to explore the potential issues associated with HEVs. The acquired dataset was divided into two groups based on their involvement in traffic crashes. The study applied association rule mining and text mining methods to analyze vehicle consumer complaint data. The results of association rule mining showed a significant association between hybrid electric all-wheel-drive vehicles manufactured between 2010 and 2021 that do not have anti-lock brakes and cruise control in the crash-related vehicle complaints dataset. Non-HEV vehicles, manufactured between 1992 and 1999, with cruise control and anti-braking systems as well as 5-10 cylinders, appeared frequently in the crash-related complaint dataset. Mileage-related issues and comparatively older HEVs (2000-2009) are dominant in non-crash-related data. The results from the text mining method show that *brakes, mileage, failure, and crash* are key features for consumer complaints related to crashes and *brakes, battery, power, and recall* are the key features for consumer complaints not related to crashes. The sentiment analysis results show slightly higher negative sentiments in complaint reports associated with crashes. The findings of this study can provide some insights into this unexplored research area.

KEYWORDS
hybrid electric vehicle; vehicle consumer complaints; text mining; empirical Bayes data mining; vehicular defect

1. Introduction

In recent years, the advancement of hybrid electric vehicles (HEVs) has become faster and faster. According to the total sale of HEVs (with

CONTACT Subasish Das subasish@txstate.edu Civil Engineering, Ingram School of Engineering, Texas State University, 601 University Drive, San Marcos, TX 78666, USA

inclusion of electric vehicles or EVs) in the United States, which was 761,100 in 2020, there was an 8.7% growth compared with 2019 (United States Department of Transportation Bureau of Transportation Statistics, 2019). Since hybrid vehicles adopt many technologies different from those of traditional vehicles, drivers need to deal with different technology features than before. However, how drivers respond HEV technologies is still relatively unclear. Thus, there is a need to understand the drivers' response to the new technology features applied by HEVs. More importantly, despite the fast advancement of the HEV market share, no studies have investigated the consumers' perspective of HEVs regarding defects that could potentially lead to traffic crashes. By applying the findings of this study, HEV manufacturers can better improve the features that drivers are more concerned with and decrease potential accident risks.

HEVs may create many potential risks. Electric vehicle battery defects and electric system defects are two major problems that affect the consumers' experience with HEVs. The increasing risk of HEV and EV battery defects has made consumers become worried (Zhao et al., 2017). Some researchers have investigated the potential of HEV battery failures (Yang et al., 2020; Ren et al., 2018). According to one news report in Korea, because of the growth in HEV sales, the related consumers' complaints multiplied by 46 times in 4 years (EV Defect Claims Multiply 46 times in 4 years, n.d.). Hyundai recalled 82,000 HEVs, including its KONA model, due to battery flaws (Auto News: Hyundai Recalls Electric Cars including Kona, Ioniq on Battery Defect—Bloomberg, n.d.). Moreover, General Motors (GM) recalled its Chevrolet Bolts for a second time for a battery defect problem that could set the HEVs on fire (The Associated Press, 2021). Audi also recalled its very first HEV E-Tron due to the risk of battery blaze (Bloomberg, n.d.). Battery blazing incidents also happened on Tesla's Model S (EV) due to the vehicles' battery defects (A Tesla Model S Erupted 'like a Flamethrower.' It Renewed Old Safety Concerns about the Trailblazing Sedans, n.d.). With the number of HEVs and EVs being sold worldwide growing dramatically in recent years, the number of consumer complaints about HEVs and EVs has been skyrocketing. Currently, there are not enough studies that have investigated the patterns of consumer complaints on HEVs.

The data used in this study is from the National Highway Traffic Safety Administration (NHTSA) Office of Defect Investigation's (ODI) consumer complaint dataset. This dataset contains consumer complaint information received by NHTSA starting from January 1, 1995. The database contains consumer complaint data for all types of vehicles. This study selected the consumer complaint records related to HEV s. The current study aims to answer the following research questions:

RQ1: What are the patterns of associated factors in complaint dataset when an HEV was either involved with crash or non-crash events?

RQ2: How do complaining patterns differ in the complaint narratives when an HEV was either involved with crash or non-crash events?

This study can provide insights into consumer complaints about HEVs to provide a better understanding of the general public's concerns and worry about hybrid and electric vehicles. Based on the findings from this paper, HEV manufacturers can put more emphasis on solving the challenges in the future.

2. Earlier work and research context

There are several studies that have applied textual analysis to investigate transportation-related problems. This includes traffic incident prediction, incident duration detection, and consumer vehicle complaint data analysis.

Schulz et al. (2015) applied spatial-temporal-type clustering to predict small-scale traffic incidents based on information from Twitter posts. The proposed method can identify 32.14% of the real-world incidents that happened in Seattle. Salas et al. (2017) proposed a method to analyze public Twitter data in order to perform traffic incident detection. This method combines both Natural Language Processing (NLP) and Support Vector Machine (SVM). Kokkinos and Nathanail (2020) applied textual analysis methods to Twitter data to detect and classify traffic events by using several machine learning methods.

Other studies have used textual analysis methods to predict time loss due to incident delay and to find the factors that affect incident delay duration. For example, Pereira et al. (2013) introduced a topic modeling method that can extract real-time information from incident reports. They compared different prediction methods with and without textual features. The results indicate that the prediction models with textual features are consistently better than those without textual features. Based on Pereira et al.'s study, Li et al. (2015) further proposed a risk mixture hazard-based model with topics extracted from textual features using the topic modeling technique. Liu and Wang (2021) proposed a text-based analysis of the Accelerated Failure Time (AFT) model to explore the factors that affect subway incident delay duration and predict the time lost because of the incidents.

Several studies have applied text analysis methods to the NHTSA consumer complaint database. Ghazizadeh and Lee (2012) applied a text mining method that includes latent semantic analysis (LSA) to study the relationship between traffic fatalities and consumer complaints using data from the NHSTA consumer complaint database. Later, Ghazizadeh et al. (2014) applied LSA to investigate the patterns of consumer complaint data for both fatal incidents and incidents involving injury. They identified 8 clusters for fatal incidents and 6 clusters for incidents involving injury. Das et al. (2018) applied exploratory text mining and empirical Bayes (EB) data mining techniques on Fatality Analysis Reporting System (FARS) and NHSTA consumer complaint data to investigate the trends in consumer complaints about vehicle defects and groups of vehicle models with major defects. Later, Das et al. (2019) used the same databases to investigate the effectiveness of vehicle inspection regulations in different states of the United States. In terms of air transportation, Walton and Marion (2020) conducted a study on hazardous good incidents on aircrafts using a text mining method. They applied SAS Text Miner to identify 5 major topics of aircraft hazardous good incidents.

To the best knowledge of the authors, there are few related works that have investigated the consumer complaint data of HEVs. In this study, consumer complaint records related to HEVs were extracted from the NHTSA consumer complaints database.

3. Methodology

To accomplish the research goals, the current study is divided into three major tasks: 1) perform descriptive statistics and statistical significance tests in examining differences between crash and non-crash related HEV complaints, 2) conduct association rules mining to identify the patterns of association factors in crash and non-crash involved complaint data, and 3) perform text network analysis on crash and non-crash related complaint narratives.

3.1. Descriptive statistics

The data used in this study is based on consumer-provided complaints collected by NHTSA. There is a variable named “FUEL_TYPE,” which records vehicles’ fuel type, and “HE,” which stands for HEV. This study only selected records with the variable “FUEL_TYPE” equals “HE”. There are 1,835,710 consumer complaint records in total as of April 1st, 2021. Among these complaints, there are 1,827,648 non-HEV complaints with 1,706,316 no-crash and 121,332 crash-based complaints, and 8,062 HEV complaints with 7,200 no-crash and 862 crash-based complaints.

Table 1 lists the descriptive statistics of the final dataset by removing rows with missing (blank) entries. The number of reports associated with some level of injury or death is 387 (around 5.4% of total complaint entries), of which 353 reports are related to traffic crashes, which is approximately 91.2% of the total injury or death-related reports. Several similarities and dissimilarities can be observed between the HEV-related

Table 1. Descriptive statistics.

		HEV (7,198 records)		Non-HEV (191,029 records)	
Variable	Category	Count	Percentage	Count	Percentage
YEARTXT (Vehicle year)	1992–1999	21	0.30%	21,359	11.18%
	2000–2009	4,293	59.60%	125,811	65.86%
	2010–2021	2,873	39.90%	42,919	22.47%
	unk[1]	11	0.20%	940	0.49%
Crash (Whether relate to crash)	n	6,418	89.20%	171,226	89.63%
	y	780	10.80%	19,803	10.37%
MILES (Vehicle milage)	lt 10000	1,556	21.60%	19,356	10.13%
	10000–40000	1,382	19.20%	39,430	20.64%
	gt 40000	3,324	46.20%	124,242	65.04%
	unk	936	13.00%	8,001	4.19%
ANTI_BRAKES_YN (Whether have anti brake system)	n	5,805	80.60%	131,500	68.84%
	y	1,393	19.40%	59,529	31.16%
CRUISE_CONT_YN (Whether have cruise control)	n	5,957	82.80%	127,635	66.81%
	y	1,241	17.20%	63,394	33.19%
NUM_CYLS (Number of cylinders)	0–4 cy	5,459	75.80%	54,626	28.6%
	5–10 cy	677	9.40%	136,250	71.32%
	unk	1,062	14.80%	153	0.08%
DRIVE_TRAIN (Vehicle drivetrain)	4wd	140	1.90%	26,839	14.05%
	awd	350	4.90%	12,464	6.52%
	fwd	2,652	36.80%	54,579	28.57%
	rwd	227	3.20%	23,680	12.4%
	unk	3829	53.2%	73,467	38.46%
VEH_SPEED (Vehicle speed)	lt 35 mph	2,986	41.50%	74,420	38.96%
	35–60 mph	1,699	23.60%	66,890	35.02%
	gt 60 mph	744	10.30%	30,130	15.77%
	unk	1,769	24.60%	19,589	10.25%

Note: [1]*n = no, y = yes, gt = greater than, lt = less than, unk = unknown, cy = cycle, 4WD = four-wheel drive, AWD = all wheel drive, fwd = front wheel drive, rwd = rear wheel drive.*

complaints and non-HEV related complaint. Around 10% of the total complaints are related to crashes for both HEV and non-HEV-related complaints. In terms of vehicle mileage, there are 21.60% of HEV-related consumer complaints have vehicle milage less than 10000 while only 10.37% of the non-HEV related consumer complaints have vehicle mileage less than 10000. Moreover, 65.04% of non-HEV complaints are vehicles with over 40000 miles, and this value for HEV complaints is only 46.2%. 80.6% of HEV-related complaints do not have anti brake systems, while only 68.84% of non-HEV-related complaints do not have anti brake system. Similarly, 82.8% of HEV-related complaints do not have cruise control systems, while only 66.81% of non-HEV-related complaints do not have cruise control systems. Most vehicles (75.8%) have 0-4 cylinders in the HEV-related complaint. while most vehicles (71.32%) have 5-10 cylinders in the non-HEV-related complaint. As for vehicle drive trains, most vehicles in both HEV and non-HEV-related complaints are FWD.

3.2. Association rules mining

To answer the first research question (**RQ1:** What are the patterns of associated factors in the complaint dataset when an HEV was either involved with crash or non-crash events?), this study performed association rules mining. The association rules mining method has been effectively used to unearth unknown patterns or rules in various fields of science and engineering. 'Rules discovery' indicates the recognition of sets of items (i.e., vehicle complaint-related information and involvement in crashes in this study) that occur collectively in each transaction (i.e., each complaint report in this study). The use of association rules is motivated by the complex interactions among contributing factors associated with hybrid vehicle complaints. Interactions in the form of rules can provide valuable insights into the complaint data. The association rules mining method was first introduced by Agrawal et al. (1993). The rules are generated based on the relative count of the sets of items that occur individually and in combination in a dataset. It is worth mentioning that the generated rules do not indicate direct causation but rather a pattern or association among a large number of other association patterns. Using the *a priori* algorithm, introduced by Agrawal et al. (1993), association rules mining was used in this analysis. This algorithm utilizes simple and step-by-step ways to continually analyze candidate itemsets to discover frequent itemsets. Then, using the frequent itemsets as the new candidate itemsets, the algorithm explores to discover new frequent itemsets till there are no new frequent itemsets. The common practice is to determine a desirable level of support, higher confidence, and a lift value over at least one. The determination of the support threshold depends on the nature of the study.

An association rule can be defined as $A \to B$, where A *and* B are disjoint itemsets which can be the combinations of different variables. A is called the antecedent and B is called the consequent. Association rules can be measured by three indicator values: support, confidence, and lift. Support shows the proportion of all observations in the dataset that contain the itemsets in A and B. Confidence represents the proportion of observations that contain the itemsets in B among all observations containing itemsets in A. Lift is the proportion of observations that contain itemsets in both A and B among all observations containing itemsets in A over the proportion of observations that contain itemsets in both A and B among all observations in the dataset. The calculation process of support is demonstrated in equations 1 through 3.

$$S(A) = \frac{\sigma(A)}{N} \tag{1}$$

$$S(B) = \frac{\sigma(B)}{N} \tag{2}$$

$$S(A \to B) = \frac{\sigma(A \cap B)}{N} \tag{3}$$

where,
$\sigma(A)$= Number of incidents with A antecedent
$\sigma(B)$= Number of incidents with B consequent
$\sigma(A \cap B)$= Number of incidents with both A antecedent and B consequent
N = Total number of incidents
S(A)= Support of antecedent
S(B)= Support of consequent
S(A → B)= Support of the association rule (A → B)

The equations of confidence and lift are listed in equations 4 and 5, respectively. Confidence evaluates a generated rule's inference reliability. Higher confidence shows that the presence of B is highly visible in all observations that have A. The lift of the rule makes an association with the frequency of co-occurrence of the antecedent and the consequent to the expected frequency of co-occurrence.

$$C(A \rightarrow B) = \frac{S(A \rightarrow B)}{S(A)} \tag{4}$$

$$L(A \rightarrow B) = \frac{S(A \rightarrow B)}{S(A).S(B)} \tag{5}$$

where,
C(A → B)= Confidence of the association rule (A → B)
L(A → B)= Lift of the association rule (A → B)

The lift measure is used to determine the correlation between antecedent and consequent. A lift value above 1 indicates that the association between the antecedent and consequent is higher than the general proportion of the consequent in the dataset. This indicates that the antecedent and consequent are highly correlated. Similarly, if lift values are lower than 1, this means that the antecedent and consequent have less interdependence. A critical inference of the association rules is that the generated rules are not needed to be interpreted as causation rather than association.

4. Results

4.1. Association rules mining

4.1.1. Crash-related complaint data

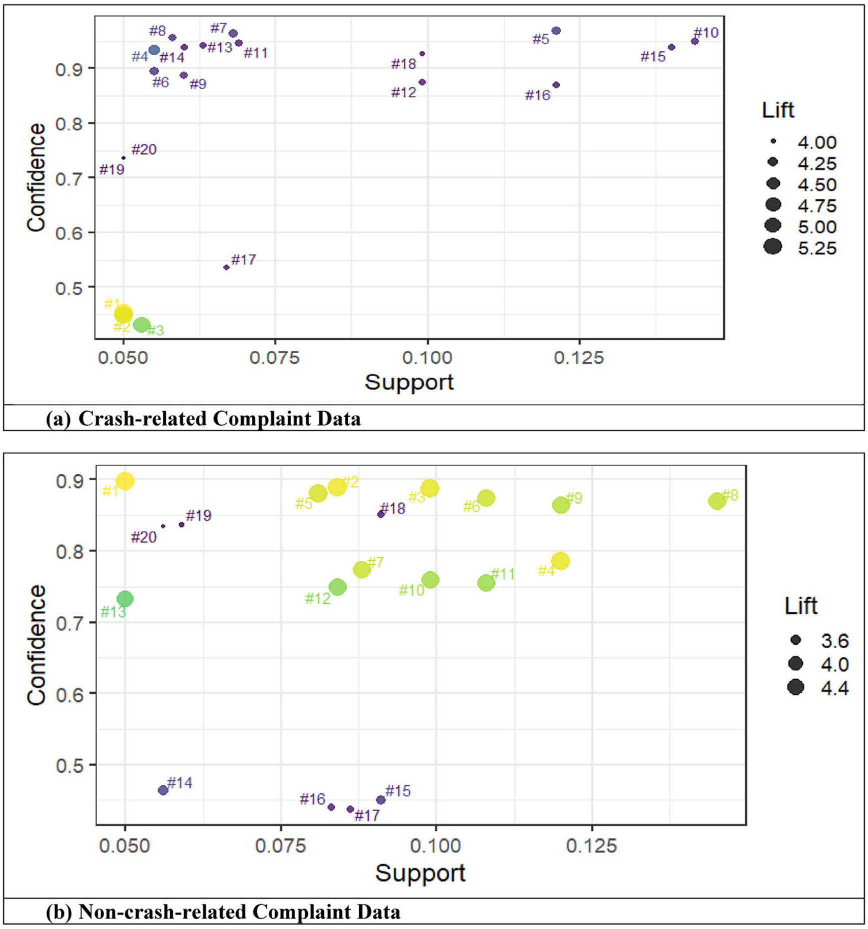
To understand the association patterns of patterns, association rules mining was applied to both HEV and non-HEV vehicle datasets. Table 2 presents the top 20 rules that contain highly associated characteristics in the crash-related HEV complaint data. Figure 1a shows the visual display of the measures of support, confidence, and lift values for the top 20 rules. It shows that the top rules of this group have low support and confidence

Table 2. Top 20 rules sorted by the lift values (crash related HEV complaint data).

No	Rules	S	C	L
#1	*yeartxt = 2010-2021, anti_brakes_yn = n, num_cyls = unk → drive_train = awd*	0.05	0.453	5.279
#2	*yeartxt = 2010-2021, cruise_cont_yn = n, num_cyls = unk → drive_train = awd*	0.05	0.448	5.219
#3	*yeartxt = 2010-2021, num_cyls = unk → drive_train = awd*	0.053	0.432	5.024
#4	*miles = gt 40000, anti_brakes_yn = y → cruise_cont_yn = y*	0.055	0.935	4.34
#5	*cruise_cont_yn = y, num_cyls = 0-4 cy, drive_train = fwd → anti_brakes_yn = y*	0.121	0.969	4.176
#6	*compdesc1 = vehicle speed control, anti_brakes_yn = y → cruise_cont_yn = y*	0.055	0.896	4.159
#7	*cruise_cont_yn = y, drive_train = fwd, veh_speed = lt 35 mph → anti_brakes_yn = y*	0.068	0.964	4.153
#8	*yeartxt = 2010-2021, cruise_cont_yn = y → anti_brakes_yn = y*	0.058	0.957	4.126
#9	*state = ca, anti_brakes_yn = y → cruise_cont_yn = y*	0.06	0.887	4.117
#10	*cruise_cont_yn = y, drive_train = fwd → anti_brakes_yn = y*	0.144	0.949	4.09
#11	*cruise_cont_yn = y, num_cyls = 0-4 cy, veh_speed = lt 35 mph → anti_brakes_yn = y*	0.069	0.947	4.083
#12	*yeartxt = 2000-2009, anti_brakes_yn = y, drive_train = fwd → cruise_cont_yn = y*	0.099	0.875	4.063
#13	*miles = 10000-40000, cruise_cont_yn = y → anti_brakes_yn = y*	0.063	0.942	4.061
#14	*miles = lt 10000, cruise_cont_yn = y → anti_brakes_yn = y*	0.06	0.94	4.051
#15	*cruise_cont_yn = y, num_cyls = 0-4 cy → anti_brakes_yn = y*	0.14	0.94	4.049
#16	*anti_brakes_yn = y, num_cyls = 0-4 cy, drive_train = fwd → cruise_cont_yn = y*	0.121	0.87	4.041
#17	*anti_brakes_yn = n, cruise_cont_yn = n, veh_speed = unk → miles = unk*	0.067	0.536	4.021
#18	*yeartxt = 2000-2009, cruise_cont_yn = y, drive_train = fwd → anti_brakes_yn = y*	0.099	0.928	3.998
#19	*cruise_cont_yn = n, drive_train = awd → num_cyls = unk*	0.05	0.736	3.986
#20	*anti_brakes_yn = n, cruise_cont_yn = n, drive_train = awd → num_cyls = unk*	0.05	0.736	3.986

Note: S: Support; C: Confidence; L: Lift; unk = Unknown.

measures. The rule that has the highest lift value is *yeartxt = 2010-2021, anti_brakes_yn = n, num_cyls = unk → drive_train = awd*.This rule can be explained as among all HEV crash complaints, 5% of the complaints are associated with HEVs manufactured between 2010 and 2021 that do not have anti-lock brakes and are all-wheel drive. Among all vehicles manufactured between 2010 and 2021 that do not have anti-lock brakes, 45.3% of them are all-wheel drive. The proportion of crash-related HEV complaints that are all-wheel drive is 5.279 times higher if the vehicles are manufactured between 2010 and 2021 and do not have anti-lock brakes. This finding is unique as anti-lock brakes are mandatory in vehicles manufactured after 2011. Rule #2 is like rule #1, except in this rule do not have anti-lock brakes is replaced with do not have cruise control. Rule #4 presents the characteristic that crash-related vehicle complaints with more than 40,000 miles and that have cruise control are associated with the presence of cruise control. Rule #5 shows that among all crash-related vehicle complaints, 12.1% are related to vehicles that have cruise control, have 0-4 cylinders, are front-wheel drive, and have anti-lock brakes. Generally, *yeartxt = 2010-2021*, *cruise_cont_yn = n* and *anti_brakes_yn = n*, and *drive_train = awd* show up in the top 2 rules with the highest lift values. This may indicate that all-wheel-drive vehicles manufactured between 2010 and 2019 without anti-lock brakes and cruise control frequently appear in the crash-related vehicle complaints dataset. Moreover, *cruise_cont_yn = y* and *anti_brakes_yn = y* appear together in many rules. Since *cruise_cont_yn = y* and *anti_brakes_yn = y* only exist in 17.2% and 19.4% of all vehicle complaints, these two features are highly associated in crash-related vehicle complaints.



(a) Crash-related Complaint Data

(b) Non-crash-related Complaint Data

Figure 1. Measures of support, confidence, and lift values for top 20 rules.

Table 3 presents the top 20 rules that contain highly associated characteristics in the crash-related non-HEV complaint data. The top rule that has the highest lift is *miles = lt 10000 + anti_brakes_yn = n + cruise_cont_yn = n* → *yeartxt = 2010-2021*. This can be explained as among all non-HEV crash-related complaints, 5% of the total complaints have vehicles with mileage less than 10000 miles and do not have anti brake systems nor cruise control. Among all vehicles with mileage less than 10000 miles and that do not have anti brake systems nor cruise control, 77.6% of them are manufactured between 2010 and 2021. The probability of crash-related non-HEV complaint data that are manufactured between 2010 and 2021 is 3.449 times higher if they have vehicle mileage less than 10000 and do not have anti brake systems nor cruise control systems. Rule #2 is *miles = gt 40000 + anti_brakes_yn = y + num_cyls = 5-10 cy* → *yeartxt = 1992-1999*. This can be explained as 5.1% of the total crash-related non-HEV complaints have vehicle milage over 40000 and 5-10

Table 3. Top 20 rules sorted by the lift values (crash related non-HEV complaint data).

No	Rules	S	C	L
#1	*miles = lt 10000 + anti_brakes_yn = n + cruise_cont_yn = n → yeartxt = 2010-2021*	*0.050*	*0.776*	*3.449*
#2	*miles = gt 40000 + anti_brakes_yn = y + num_cyls = 5-10 cy → yeartxt = 1992-1999*	*0.051*	*0.361*	*3.285*
#3	*miles = gt 40000 + cruise_cont_yn = y + num_cyls = 5-10 cy → yeartxt = 1992-1999*	*0.052*	*0.356*	*3.242*
#4	*miles = lt 10000 + cruise_cont_yn = n → yeartxt = 2010-2021*	*0.051*	*0.700*	*3.111*
#5	*miles = gt 40000 + cruise_cont_yn = y → yeartxt = 1992-1999*	*0.063*	*0.340*	*3.098*
#6	*miles = lt 10000 + anti_brakes_yn = n → yeartxt = 2010-2021*	*0.051*	*0.697*	*3.096*
#7	*miles = gt 40000 + anti_brakes_yn = y → yeartxt = 1992-1999*	*0.061*	*0.340*	*3.092*
#8	*miles = gt 40000 + anti_brakes_yn = y + cruise_cont_yn = y → yeartxt = 1992-1999*	*0.053*	*0.334*	*3.038*
#9	*miles = 10000-40000 + anti_brakes_yn = n + cruise_cont_yn = n → yeartxt = 2010-2021*	*0.081*	*0.612*	*2.722*
#10	*miles = 10000-40000 + cruise_cont_yn = n → yeartxt = 2010-2021*	*0.083*	*0.558*	*2.481*
#11	*miles = 10000-40000 + anti_brakes_yn = n → yeartxt = 2010-2021*	*0.082*	*0.547*	*2.431*
#12	*miles = 10000-40000 + cruise_cont_yn = y + num_cyls = 5-10 cy → anti_brakes_yn = y*	*0.091*	*0.915*	*2.360*
#13	*yeartxt = 2000-2009 + cruise_cont_yn = y + drive_train = 4wd → anti_brakes_yn = y*	*0.060*	*0.915*	*2.360*
#14	*cruise_cont_yn = y + num_cyls = 5-10 cy + drive_train = 4wd → anti_brakes_yn = y*	*0.076*	*0.915*	*2.359*
#15	*cruise_cont_yn = y + drive_train = 4wd → anti_brakes_yn = y*	*0.079*	*0.914*	*2.355*
#16	*anti_brakes_yn = y + num_cyls = 5-10 cy + drive_train = fwd → cruise_cont_yn = y*	*0.073*	*0.936*	*2.353*
#17	*anti_brakes_yn = y + num_cyls = 5-10 cy + drive_train = 4wd → cruise_cont_yn = y*	*0.076*	*0.934*	*2.348*
#18	*yeartxt = 2000-2009 + cruise_cont_yn = y + num_cyls = 5-10 cy → anti_brakes_yn = y*	*0.201*	*0.906*	*2.337*
#19	*cruise_cont_yn = y + num_cyls = 5-10 cy + veh_speed = lt 35 mph → anti_brakes_yn = y*	*0.113*	*0.906*	*2.336*
#20	*cruise_cont_yn = y + num_cyls = 5-10 cy + drive_train = rwd → anti_brakes_yn = y*	*0.057*	*0.906*	*2.336*

Note: S: Support; C: Confidence; L: Lift; unk = Unknown.

cylinders with anti brake system. With all complaints that have vehicle milage over 40000 and 5-10 cylinders with anti brake systems, 36.1% of them are manufactured between 1992 and 1999. Rule #3 is similar to Rule #2, except with the anti brake system replaced with having a cruise control system. Rule # 4 is similar to Rule #1, and Rule #5 is similar to Rule #3. Generally, non-HEV vehicles manufactured between 2010 and 2021 that do not have anti brake systems nor cruise control with milage less than 10000 miles frequently appear in the crash-related complaint dataset. Non-HEV vehicles manufactured between 1992 and 1999 that have cruise control systems and anti brake systems and with 5-10 cylinders also frequently appeared in the crash-related complaint dataset.

4.1.2. Non-crash-related complaint data

Table 4 lists the top 20 rules that contain highly associated characteristics in the non-crash-related HEV complaint data. Figure 1b shows the visual display of the measures of support, confidence, and lift values for the top 20 rules. It shows that around 20% of the rules have low confidence measures. The rule that has the highest lift is *cruise_cont_yn = y, veh_speed = lt 35 mph* → *anti_brakes_yn = y*. This rule indicates that among all no-crash-related vehicle complaints, 5% of the total complaints are associated with vehicles with cruise control, that have speeds less than 35 mph, and have anti-lock brakes. 89.7% of the non-crash-related vehicle complaints with cruise control and speeds less than 35 mph also have anti-lock brakes. In general, the proportion of vehicles that have anti-lock brakes is 4.748 times higher if they have cruise control and have speeds less than 35 mph. Rules

Table 4. Top 20 rules sorted by the lift values (non-crash related HEV complaint data).

No	Rules	S	C	L
#1	*cruise_cont_yn = y, veh_speed = lt 35 mph → anti_brakes_yn = y*	0.05	0.897	4.748
#2	*cruise_cont_yn = y, num_cyls = 0-4 cy, drive_train = fwd → anti_brakes_yn = y*	0.084	0.889	4.71
#3	*cruise_cont_yn = y, drive_train = fwd → anti_brakes_yn = y*	0.099	0.887	4.699
#4	*yeartxt = 2000-2009, anti_brakes_yn = y → cruise_cont_yn = y*	0.12	0.785	4.698
#5	*yeartxt = 2000-2009, cruise_cont_yn = y, drive_train = fwd → anti_brakes_yn = y*	0.081	0.88	4.658
#6	*cruise_cont_yn = y, num_cyls = 0-4 cy → anti_brakes_yn = y*	0.108	0.875	4.631
#7	*yeartxt = 2000-2009, anti_brakes_yn = y, num_cyls = 0-4 cy → cruise_cont_yn = y*	0.088	0.774	4.628
#8	*cruise_cont_yn = y → anti_brakes_yn = y*	0.145	0.87	4.604
#9	*yeartxt = 2000-2009, cruise_cont_yn = y → anti_brakes_yn = y*	0.12	0.865	4.581
#10	*anti_brakes_yn = y, drive_train = fwd → cruise_cont_yn = y*	0.099	0.76	4.543
#11	*anti_brakes_yn = y, num_cyls = 0-4 cy → cruise_cont_yn = y*	0.108	0.756	4.52
#12	*anti_brakes_yn = y, num_cyls = 0-4 cy, drive_train = fwd → cruise_cont_yn = y*	0.084	0.75	4.484
#13	*anti_brakes_yn = y, veh_speed = lt 35 mph → cruise_cont_yn = y*	0.05	0.733	4.384
#14	*drive_train = fwd, veh_speed = unk → miles = unk*	0.056	0.463	3.569
#15	*cruise_cont_yn = n, veh_speed = unk → miles = unk*	0.091	0.451	3.478
#16	*anti_brakes_yn = n, cruise_cont_yn = n, veh_speed = unk → miles = unk*	0.083	0.44	3.395
#17	*anti_brakes_yn = n, veh_speed = unk → miles = unk*	0.086	0.438	3.378
#18	*miles = unk, cruise_cont_yn = n → veh_speed = unk*	0.091	0.852	3.368
#19	*miles = unk, anti_brakes_yn = n, num_cyls = 0-4 cy → veh_speed = unk*	0.059	0.837	3.309
#20	*miles = unk, drive_train = fwd → veh_speed = unk*	0.056	0.835	3.301

Note: S: Support; C: Confidence; L: Lift; unk = Unknown.

Table 5. Top 20 rules sorted by the lift values (non-crash related non-HEV complaint data).

No	Rules	S	C	L
#1	*miles = gt 40000 + cruise_cont_yn = y + num_cyls = 5-10 cy → yeartxt = 1992-1999*	0.058	0.381	3.404
#2	*miles = gt 40000 + anti_brakes_yn = y + num_cyls = 5-10 cy → yeartxt = 1992-1999*	0.053	0.374	3.337
#3	*miles = gt 40000 + cruise_cont_yn = y → yeartxt = 1992-1999*	0.068	0.369	3.296
#4	*miles = gt 40000 + anti_brakes_yn = y → yeartxt = 1992-1999*	0.061	0.359	3.206
#5	*miles = gt 40000 + anti_brakes_yn = y + cruise_cont_yn = y → yeartxt = 1992-1999*	0.055	0.358	3.199
#6	*yeartxt = 2000-2009 + cruise_cont_yn = y + drive_train = 4wd → anti_brakes_yn = y*	0.053	0.901	2.975
#7	*cruise_cont_yn = y + num_cyls = 5-10 cy + drive_train = 4wd → anti_brakes_yn = y*	0.069	0.898	2.966
#8	*cruise_cont_yn = y + drive_train = 4wd → anti_brakes_yn = y*	0.071	0.894	2.954
#9	*yeartxt = 2000-2009 + cruise_cont_yn = y + num_cyls = 5-10 cy → anti_brakes_yn = y*	0.163	0.885	2.924
#10	*cruise_cont_yn = y + num_cyls = 5-10 cy + veh_speed = lt 35 mph → anti_brakes_yn = y*	0.086	0.882	2.912
#11	*miles = 10000-40000 + cruise_cont_yn = y + num_cyls = 5-10 cy → anti_brakes_yn = y*	0.055	0.881	2.910
#12	*anti_brakes_yn = y + num_cyls = 5-10 cy + drive_train = fwd → cruise_cont_yn = y*	0.071	0.941	2.902
#13	*cruise_cont_yn = y + num_cyls = 5-10 cy → anti_brakes_yn = y*	0.227	0.871	2.876
#14	*yeartxt = 2000-2009 + cruise_cont_yn = y + veh_speed = lt 35 mph → anti_brakes_yn = y*	0.076	0.870	2.874
#15	*anti_brakes_yn = y + num_cyls = 5-10 cy + drive_train = 4wd → cruise_cont_yn = y*	0.069	0.929	2.865
#16	*yeartxt = 2000-2009 + anti_brakes_yn = y + drive_train = 4wd → cruise_cont_yn = y*	0.053	0.928	2.862
#17	*yeartxt = 2000-2009 + anti_brakes_yn = y + num_cyls = 5-10 cy → cruise_cont_yn = y*	0.163	0.927	2.857
#18	*anti_brakes_yn = y + num_cyls = 5-10 cy + veh_speed = 35-60 mph → cruise_cont_yn = y*	0.077	0.926	2.856
#19	*anti_brakes_yn = y + drive_train = 4wd → cruise_cont_yn = y*	0.071	0.926	2.856
#20	*anti_brakes_yn = y + num_cyls = 5-10 cy + veh_speed = lt 35 mph→ cruise_cont_yn = y*	0.086	0.926	2.856

#4 and #5 show that vehicles manufactured between 2000 and 2009, with cruise control, anti-lock brakes, and front-wheel drive, are highly associated in the non-crash-related vehicle complaints. Like crash-related vehicle complaint data, *cruise_cont_yn = y* and *anti_brakes_yn = y* also appear together in many rules with high lift values.

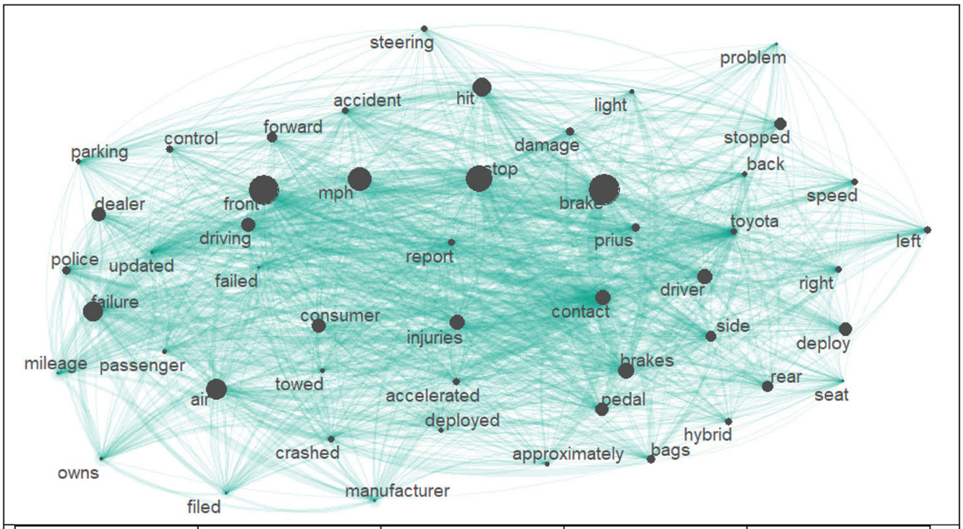
Table 5 lists the top 20 rules that contain highly associated characteristics in the non-crash-related non-HEV complaint data. The top rule with the highest lift is *miles = gt 40000 + cruise_cont_yn = y + num_cyls = 5-10 cy → yeartxt = 1992-1999*. This can be represented as among all complaints, 5.8% of them have vehicle mileage greater than 40000 miles and have cruise control systems with 5-10 cylinders. Among these complaints, 38.1% of them have vehicles manufactured between 1992 and 1999. The proportion of vehicles that are manufactured between 1992 and 1999 is 3.404 times higher if they have a milage over 40000 miles and have cruise control systems with 5-10 cylinders. Rules #1 to #5 all have similar meanings, which indicates that non-HEV vehicles manufactured during 1992 and 1999 with vehicle mileage greater than 40000 and have both cruise control system and anti-brake system with 5-10 cylinders frequently appear in the non-crash related dataset. Rules #6 to #9 are similar, which shows that 4-wheel drive non-HEV vehicles with 5-10 cylinders manufactured between 2000 to 2009 with both anti brake systems and cruise control systems frequently appear in non-crash-related complaints.

Results from the association rules mining indicate that patterns of association factors in the complaint database differ if the complaint is related to crash or no-crash for both HEV and non-HEV vehicles.

4.2. Complaint narrative analysis of HEV vehicles

Text mining is one of the popular data mining methods. It is a branch of NLP that provides exploratory analysis of unstructured datasets. Text mining can be extended to complicated statistical analysis based on the research question or research needs. Text network analysis is a tool of text mining which can provide important insights into hidden trends of unstructured text data. This method consists of three major processes:

- Step 1 - Text cleaning and token development: The first steps require some text mining steps to prepare the data for analysis. These steps include text cleaning by using i) stop word removal and removal of study-specific needs such as the removal of redundant punctuation, number, special text, or URLs, ii) token or phrase generation, and iii) performing stemming (i.e., word root generation. For example, the stem of 'injury' and 'injuries' is 'injur') or lemmatization (i.e., word generation by morphological analysis. For example, the lemma of 'injury' and 'injuries' is 'injury') based on the study needs.
- Step 2 - Node and edge generation: This step involves a word search algorithm. The algorithm will find at first two words or *n* words sentence by sentence in a corpus (i.e., a document with words and sentences). For the first combination, nodes with a word and edge between two words or *n* words will be created. The search will continue to get such a pair or pairs for the following sentences. If such a pair or pairs exist, the weight on the node and edge will crease by a unit weight.



Contact	Toyota	Updated	Failure	Brakes
12280	7289	5913	4689	4558
Prius	Mileage	Accident	Side	Manufacturer
4479	4474	3923	3703	3663

Figure 2. Text network from crash complaint reports.

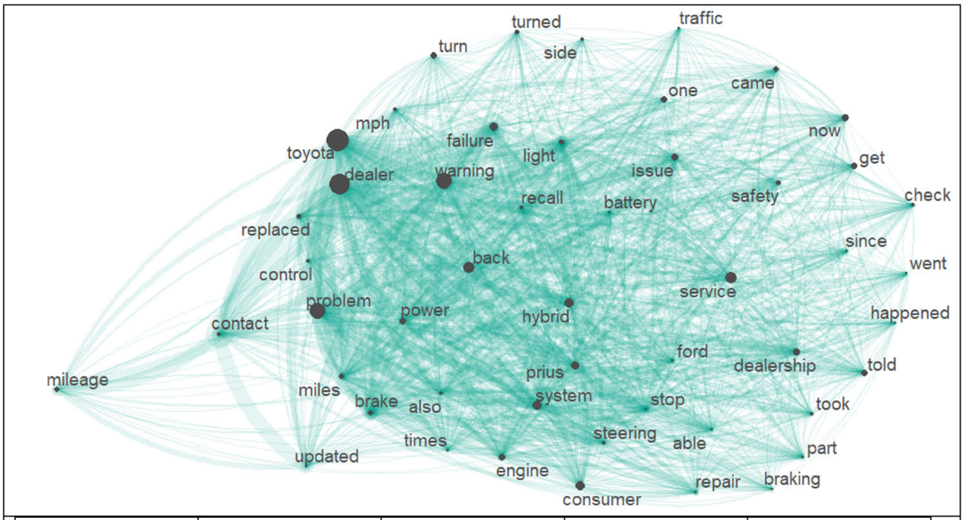
With the presence of a new word or new pair, the network will grow and follow the steps of weighting throughout the corpus.

- Step 3 - Network analysis: Network analysis involves the explanation of the nodes and edges based on the network plot. Usually, a large node indicates the presence of a keyword that is associated with many other words with some specific high edge association with particular words in the form of higher co-occurrence in the corpus. The presence of such nodes can be considered as clusters or key topics determined by the word with large nodes and higher edges for different combinations of word pairs.

In this study, complaint data are divided into two groups for text network analysis. The following section explains the results of the text network analysis.

4.2.1. Complaints associated with crash

Figures 2 and 3 show text networks from crash complaints and from no crash complaints, respectively. 'Contact,' 'Toyota,' 'update,' 'failure,' 'brakes,' 'Prius,' 'mileage,' 'accident,' 'side,' and 'manufacturer' are listed as key features in Figure 1. Visually, the words 'injuries,' 'report,' 'contact,' and 'consumer' appear closest to the center, whereas words such as 'problem,' 'steering, 'owns,' and 'left' appear furthest from the center. The focus in Figure 1 is associated with topics such as injuries, contact, reports,



Brake	Light	Stop	Contact	Ford
41899	36311	31117	30358	27740
Power	Battery	Recall	Failure	Updated
23896	23627	22802	22520	21881

Figure 3. Text network from non-crash complaint reports.

and accidents. The words with the largest nodes are 'brake,' 'stop,' 'from,' 'mph,' 'air,' and 'failure.'

Collocation indicates a set of words occurring together in a text corpus (i.e., a document with texts). Collocation helps to understand unambiguous meaning or connotation for word groups in the form of *n*-grams (*n*-gram indicates *n* number of word combinations; for example, bigram is a combination of two words) gathered from unstructured datasets by extracting information. This study extracts collocation in the form of rigid collocations by extracting *n*-grams that always occur side by side by keeping the same order. The lambda (λ) computed for an n-word collocated group is the coefficient for the n-way interaction parameter in the saturated log-linear model fitted to n-word counts. High z-values of the collocated word groups indicate the statistical significance of occurrences at the 95 percent confidence level. Table 6 shows the top collocations (20 bigrams, 10 trigrams, and 10 tetragrams) from the crash complaint narratives, organized by count, length, lambda, and z-value. '*Brake pedal*' has both the highest z-score of 39.505 and the highest count of 174. '*Pursuant freedom information act*' has the highest lambda score, at 11.187. After '*foot brake,*' which has a z-score of 23.370, there is a sudden drop in z-score, with '*curb front hit*' having a z-score of 3.178 (a drop in score of 20.192), which is due to the change in n-grams (bigram to trigram/tetragram). Air bad-related terms are frequently present in the crash-related complaints due to their deployment.

Table 6. Collocation from crash complaint narratives.

collocation	count	length	lambda	Z
brake pedal	174	2	5.190	39.505
police report	162	2	7.532	39.058
steering wheel	76	2	6.650	32.369
bags deploy	68	2	5.228	30.323
passenger side	58	2	4.869	26.848
suddenly accelerated	51	2	5.194	26.760
failed deploy	52	2	4.836	26.505
bags deployed	48	2	5.035	26.098
rear ended	42	2	6.234	23.705
foot brake	57	2	4.238	23.370
curb front hit	6	3	3.502	3.178
incident brake failure	6	3	6.246	3.041
instead accelerated stop	6	3	5.243	2.959
similar incident brake	6	3	7.306	2.933
brake instead accelerated	7	3	3.093	2.897
foot still brake	5	3	6.032	2.897
hit back seat	8	3	4.727	2.798
failure occurred parking	5	3	5.580	2.678
bags failed deploy	28	3	4.162	2.634
pedal nothing happened	8	3	5.353	2.588
owns Toyota Camry hybrid	6	4	8.764	3.066
independent mechanic diagnostic testing	6	4	9.191	2.228
pursuant freedom information act	59	4	11.187	2.193
steering wheel air bag	5	4	8.243	1.930
failure mileage current mileage	9	4	6.036	1.807
contact owns Hyundai sonata	5	4	8.099	1.784
document redacted protect personally	16	4	8.524	1.727
consumer states unintended acceleration	8	4	6.332	1.710
emanating underneath tried press	6	4	7.407	1.561
protect personally identifiable information	16	4	7.511	1.522

4.2.2. Complaints associated with no crash

Figure 3 shows the text network associated with no crash reports. The key features listed include *'brake,' 'light,' 'stop,' 'contact,' 'Ford,' 'power,' 'battery, 'recall,' 'failure,' and 'updated'.* The words that visually appear closest to the center of the network include *'hybrid,' 'back,' 'battery,'* and *'service.'* The words that appear furthest from the center include *'mileage,' 'updated,'* and *'contact.'* Regarding the key features, Figures 2 and 3 both focus on things such as *'contact,' 'brakes' or 'brake,' and 'failure.'* Interestingly, Figure 2 includes *'Prius'* (manufactured by Toyota) as a key feature, whereas Figure 3 includes *'Ford'* (a different manufacturer). This shows a slight difference in manufacturing between the two, although the general focus of the two seems to be similar. The words with the largest nodes are *'Toyota,' 'dealer,' 'warning,'* and *'problem.'*

The top 40 collocation word groups (20 bigrams, 10 trigrams, and 10 tetragrams) from the no crash complaint narratives, organized by count, length, lambda, and z-value, are shown in Table 7. *'Failure mileage'* (vehicle's mileage when the failure happens) has the highest z-value (83.635) and the highest count (637). *'Shells air conditioner intake'* has the highest lambda (9.960). Between *'hybrid system'* and *'stop safely now'*, there

Table 7. Collocation from non-crash complaint narratives.

collocation	Count	length	lambda	z
failure mileage	637	2	5.608	83.635
power steering	462	2	4.733	75.706
brake pedal	412	2	4.964	69.653
stop safely	371	2	5.930	67.846
lost power	350	2	5.567	66.141
safety issue	349	2	4.374	64.856
steering wheel	286	2	5.924	64.682
warning light	433	2	3.732	64.370
without warning	268	2	4.724	58.873
hybrid system	316	2	3.681	55.801
stop safely now	323	3	5.138	8.015
check engine light	143	3	1.925	7.196
also park light	5	3	4.696	5.523
dashboard still power	6	3	7.127	5.444
engine check light	5	3	4.078	5.195
immediately service brake	5	3	5.386	4.939
way back light	5	3	3.424	4.873
warning safely stop	5	3	4.104	4.771
service brake system	15	3	2.251	4.636
front left tire	13	3	3.749	4.629
warning came engine shut	13	4	8.120	3.105
repaired failure current mileage	13	4	5.558	2.543
warning light appeared dash	6	4	4.108	2.361
brake booster assembly replaced	6	4	6.290	2.260
shells air conditioner intake	5	4	9.960	2.110
able pull onto shoulder	5	4	5.146	2.071
protective lubricant gasket lubricated	5	4	9.581	1.941
brake pedal went floor	14	4	3.248	1.909
current failure mileages updated	29	4	5.621	1.872
contact owns tesla model	23	4	7.074	1.872

Table 8. Sentiment scores of complaint narratives.

Crash	Words	Standard Deviation	Average Sentiment Score
N	778,288	0.217	−0.087
Y	105,067	0.218	−0.074

is a sharp decrease in z-value (from 55.801 to 8.015, a drop of 47.786), which is due to the change in n-grams (bigram to trigram/tetragram). Air bag-related issues are not seen in non-crash complaint data.

Another way of understanding the insights of unstructured text is to perform sentiment analysis. This study used the open-source R package ‘sentimentr’ to perform sentiment analysis. The complaint narratives were analyzed to develop individual-level sentiment scores (Rinker, 2021). The sign and measure indicate the nature of the sentiments (‘+’ indicates positive sentiment, and ‘–’ indicates negative sentiment). Table 8 lists the measures of each group based on the involvement of crashes. The average sentiment of crash-related complaint reports is higher than non-crash-related complaint reports, which is intuitive as crash is associated with more negative sentiments. Figure 4 shows the distribution of sentiment scores for each report based on their involvement in traffic crashes.

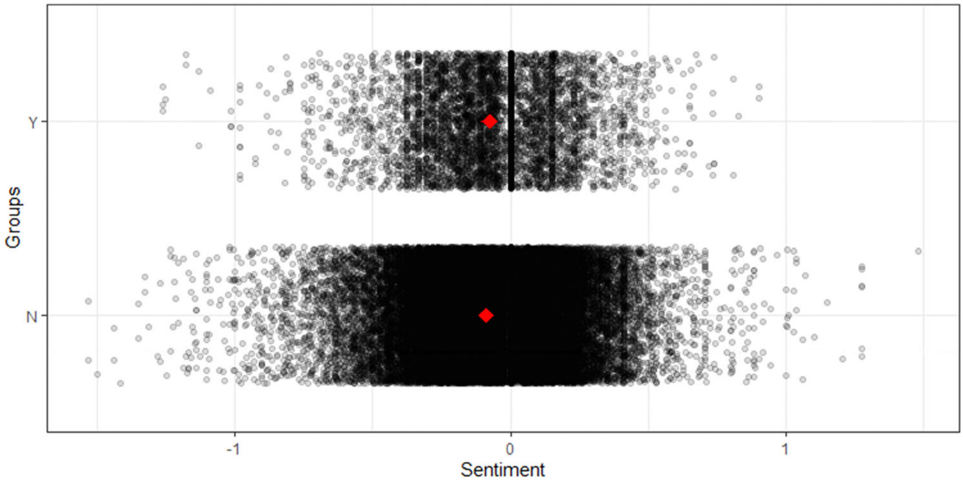


Figure 4. Distribution of sentiment scores in complaint narratives.

This study identified several key patterns in crash-involved complaints. The HEV vehicle crash-related complaint datasets show several key patterns. Vehicles manufactured between 2010 and 2021, all-wheel drive, without anti-lock brake, and without cruise control are highly associated. New technologies like anti-lock brakes are useful in avoiding collisions. Vehicles with mileage over 40,000, with anti-lock brakes, and with cruise control are also highly associated. This may indicate that high mileage HEVs with anti-lock brakes and cruise control are more likely to be complained about by consumers due to crash-related problems. Consumer complaints regarding vehicle speed control are associated with vehicles that have cruise control and anti-lock brakes. Front-wheel drive vehicles with cruise control and anti-lock brakes also appear frequently in the crash-related vehicle complaints data. The key features of crash-associated complaints are 'contact,' 'update,' 'failure,' 'brakes,' 'mileage,' 'accident,' 'side,' and 'manufacturer'. Brake pedal-related problems are the most significant in crash-related complaints. Non-HEV vehicle crash-related complaint datasets also identified several key patterns. These vehicles, manufactured between 2010 and 2021, do not have an anti-lock braking system or cruise control and have fewer than 10,000 miles on the odometer. Non-HEV vehicles, manufactured between 1992 and 1999, with cruise control and anti-braking systems, as well as 5-10 cylinders, appeared frequently in the crash-related complaint dataset.

HEV vehicle non-crash-related complaint datasets show several key patterns. Vehicles with operating speeds less than 35 mph, with anti-lock brakes, and with cruise control are the highest associated rules in the non-crash-related vehicle complaint data. Front-wheel drive vehicles manufactured between 2000 and 2009 with cruise control and anti-lock brakes appear more frequently in consumer complaint data. The key features of non-crash-associated complaints are ‘brake,’ ‘light,’ ‘stop,’ ‘contact,’ ‘power,’ ‘battery, ‘recall,’ ‘failure,’ and ‘updated’. Failure milage is the most significant in non-crash-related complaints. This shows that the failure mileage of HEVs is a major concern for consumers. Air bag-related issues are less likely to be seen in non-crash-related complaint data. Non-HEV vehicle crash related complaint datasets also identified several key frequent variables such as *miles = gt 40000, cruise_cont_yn = y, num_cyls = 5-10 cy,* and *yeartxt = 1992-1999.*

The key findings of this study highlighted some important features that are closely related to HEV consumer complaints. The identification of these features can provide HEV manufacturers with guidelines on how to improve their vehicles to improve consumer experiences.

5. Conclusions

Recently, electric vehicles have become the most significant disruptive technology, and the interest surrounding electric vehicles has generated more government spending on infrastructure improvement for electric vehicles, such as frequent charging stations and improvements in battery technology. Although electric vehicles are generating more interest among the public, electric vehicle enthusiasts and companies introduced electric vehicles decades ago. Vehicle complaint-related safety analysis is rare, and so are the complaint-related safety studies for HEVs. This study collected HEV-related complaint data from NHTSA’s vehicle complaint data. Among other concerns, safety is the key issue for any disruptive technology such as electric vehicles. Understanding the patterns of crash and non-crash-related complaint data can provide insights, which can later be transcribed into policy implications. This study aimed to answer two key research questions (**RQ1:** what are the patterns of associated factors in the complaint dataset when an HEV was either involved with crash or non-crash events? **RQ2**: How do complaining patterns differ in the complaint narratives when an HEV was either involved with crash or non-crash events?). The association rules mining results show a difference between association patterns in the crash and non-crash related HEV complaints. Non-HEV related complaints were also analyzed to identify the patterns among conventional vehicle complaint datasets. Non-HEV vehicles, manufactured between 1992 and 1999, with cruise control and anti-braking systems, as well as 5-10 cylinders, appeared frequently in the crash-related complaint dataset. The top rules in the HEV crash database are associated with the absence of anti-brake and cruise control systems in newer model cars (2010-2021). Recent advancements in road vehicle safety technologies such as antilock braking and collision avoidance systems are associated with crash reduction. The findings indicate that these technologies assist in avoiding collisions. Older vehicles (2000-2009) and mileage-related issues are dominant in non-crash-related complaint datasets. In this dataset, anti-brake and cruise control are mostly present in the top rules. The text network analysis and collocation word pairs provided latent trends of the unstructured complaint narrative data. Specific risk factors associated with vehicle defects can provide information regarding defect-related safety issues. The sentiment analysis results show higher negative sentiments in crash-related complaint narratives compared to non-crash-related complaint narratives.

Unlike regression models, data mining approaches do not ignore subgroups with different contributing factors. Thus, the interventions are frequently geared toward unique trends of various subgroups in the form of rules. This paper demonstrates the utility of using rule-based analysis methods to identify patterns of contributing factors without imposing assumptions on the subgroups. There are other methods, such as correspondence analysis and cluster correspondence analysis, that could be used to identify the patterns. However, rules mining provides performance measures such as lift, which can provide significance of the rules. On the other hand, correspondence analysis is mostly associated with dimension reduction and interpretation of the results from a two-dimensional plot. A combination of attributes is most likely to collectively influence the outcome of a complaint data. Given the critical nature of this data, a combination of attributes can contribute to a complaint-related event. Parametric models are usually not able to identify these groups of risk factors. Identifying these clusters of attributes using rules mining would provide a better understanding of the contributing factors and their patterns and will provide insights to vehicle manufacturers and policy makers.

This study has several limitations. First, the current analysis is limited to the top 20 association rules. Additional generalized findings can be acquired by using dimension reduction methods such as correspondence analysis. Second, the complaint narrative data are not detailed like police-reported crash narratives. The current analysis applied three NLP methods (text network analysis, collocation analysis, and sentiment analysis). Advanced NLP tools such as the topic model can provide more in-depth findings on the topics as clusters of risk factors. Third, vehicle models and years can provide additional insights. These two variables can be explored in future studies. Fourth, the counts of crash versus non-crash-related reports are unbalanced. Future studies can use Synthetic Minority Oversampling Technique (SMOTE) or similar techniques to resolve this issue. Fifth, the analysis is limited to HEVs as the current database does not provide information on EVs.

Acknowledgment

The authors like to thank three anonymous reviewers who provided critical and rigorous comments. The current version of the paper is much improved due to the suggested changes that are made to respond the queries of the reviewers.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Subasish Das http://orcid.org/0000-0002-1671-2753
Zihang Wei http://orcid.org/0000-0002-1790-022X

References

The Seattle Times. (2021). *A Tesla Model S Erupted 'like a Flamethrower.' It Renewed Old Safety Concerns about the Trailblazing Sedans*. Washington Post. Retrieved July 29, 2021, from https://www.seattletimes.com/business/a-tesla-model-s-erupted-like-a-flamethrower-it-renewed-old-safety-concerns-about-the-trailblazing-sedans/

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, *22*(2), 207–216. doi:10.1145/170036.170072

Bloomberg. (2021). *Auto News: Hyundai recalls electric cars including Kona, Ioniq on battery defect—Bloomberg*. Retrieved July 29, 2021, from https://www.bloomberg.com/news/articles/2021-02-24/hyundai-to-recall-82-000-electric-cars-globally-in-latest-blow.

Das, S., Geedipally, S. R., Dixon, K., Sun, X., & Ma, C. (2019). Measuring the effectiveness of vehicle inspection regulations in different states of the U.S. *Transportation Research Record: Journal of the Transportation Research Board*, *2673*(5), 208–219. doi:10.1177/0361198119841563

Das, S., Mudgal, A., Dutta, A., & Geedipally, S. R. (2018). Vehicle consumer complaint reports involving severe incidents: Mining large contingency tables. *Transportation Research Record: Journal of the Transportation Research Board*, *2672*(32), 72–82. doi:10.1177/0361198118788464

Ghazizadeh, M., & Lee, J. D. (2012). Consumer complaints and traffic fatalities: Insights from the NHTSA vehicle owner's complaint database. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 2256–2260. doi:10.1177/1071181312561475

Ghazizadeh, M., McDonald, A. D., & Lee, J. D. (2014). Text mining to decipher free-response consumer complaints: Insights from the NHTSA vehicle owner's complaint database. *Human Factors*, *56*(6), 1189–1203. doi:10.1177/0018720813519473

IndustryWeek. (2021). Audi recalls first electric vehicle in US on battery fire risk. Retrieved July 29, 2021, from https://www.industryweek.com/technology-and-iiot/article/22027722/audi-recalls-first-electric-vehicle-in-us-on-battery-fire-risk..

Kokkinos, K., & Nathanail, E. (2020). Exploring an ensemble of textual machine learning methodologies for traffic event detection and classification. *Transport and Telecommunication Journal*, *21*(4), 285–294. doi:10.2478/ttj-2020-0023

Korea Herald (2021). *EV defect claims multiply 46 times in 4 years*. Retrieved July 29, 2021, from http://www.koreaherald.com/view.php?ud=20201228000830.

Li, R., Pereira, F. C., & Ben-Akiva, M. E. (2015). Competing risk mixture model and text analysis for sequential incident duration prediction. *Transportation Research Part C: Emerging Technologies*, *54*, 74–85. doi:10.1016/j.trc.2015.03.009

Liu, F., & Wang, S. (2021). Predicting subway incident delays using text analysis based accelerated failure time model. *Journal of Transportation Safety & Security*, *13*(3), 340–356. doi:10.1080/19439962.2019.1638474

Pereira, F. C., Rodrigues, F., & Ben-Akiva, M. (2013). Text analysis in incident duration prediction. *Transportation Research Part C: Emerging Technologies*, *37*, 177–192. doi:10.1016/j.trc.2013.10.002

Ren, D., Liu, X., Feng, X., Lu, L., Ouyang, M., Li, J., & He, X. (2018). Model-based thermal runaway prediction of lithium-ion batteries from kinetics analysis of cell components. *Applied Energy*, *228*, 633–644. doi:10.1016/j.apenergy.2018.06.126

Rinker. (2021). T. R Package 'sentimentr.' https://github.com/trinker/sentimentr

Salas, A., Georgakis, P., & Petalas, Y. (2017 *Incident detection using data from social media* [Paper presentation]. Presented at the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). doi:10.1109/ITSC.2017.8317967

Schulz, A., Schmidt, B., & Strufe, T. (2015). *Small-scale incident detection based on microposts* [Paper presentation]. New York, NY. doi:10.1145/2700171.2791038

The Associated Press. (2021). Chevy bolts are recalled for a 2nd time over batteries that can set the cars on fire. *NPR*, Jul 23.

United States Department of Transportation Bureau of Transportation Statistics (USDOT). (2019). *National transportation statistics from Bureau of Transportation Statistics*. doi:10.21949/1503663

Walton, R. O., & Marion, J. W. (2020). A textual analysis of dangerous goods incidents on aircraft. *Transportation Research Procedia*, *51*, 152–159. doi:10.1016/j.trpro.2020.11.017

Yang, R., Xiong, R., Ma, S., & Lin, X. (2020). Characterization of external short circuit faults in electric vehicle li-ion battery packs and prediction using artificial neural networks. *Applied Energy*, *260*, 114253. doi:10.1016/j.apenergy.2019.114253

Zhao, Y., Liu, P., Wang, Z., Zhang, L., & Hong, J. (2017). Fault and defect diagnosis of battery for electric vehicles based on big data analysis methods. *Applied Energy*, *207*, 354–362. doi:10.1016/j.apenergy.2017.05.139