Research Article

Transportation Research Record
1–12
© National Academy of Sciences:
Transportation Research Board 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0361198119842110
journals.sagepub.com/home/trr

YouTube as a Source of Information in Understanding Autonomous Vehicle Consumers: Natural Language Processing Study

Subasish Das[1], Anandi Dutta[2], Tomas Lindheimer[3], Mohammad Jalayer[4], and Zachary Elgart[5]

Abstract

The automotive industry is currently experiencing a revolution with the advent and deployment of autonomous vehicles. Several countries are conducting large-scale testing of autonomous vehicles on private and even public roads. It is important to examine the attitudes and potential concerns of end users towards autonomous cars before mass deployment. To facilitate the transition to autonomous vehicles, the automotive industry produces many videos on its products and technologies. The largest video sharing website, YouTube.com, hosts many videos on autonomous vehicle technology. Content analysis and text mining of the comments related to the videos with large numbers of views can provide insight about potential end-user feedback. This study examines two questions: first, how do people view autonomous vehicles? Second, what polarities exist regarding (a) content and (b) automation level? The researchers found 107 videos on YouTube using a related keyword search and examined comments on the 15 most-viewed videos, which had a total of 60.9 million views and around 25,000 comments. The videos were manually clustered based on their content and automation level. This study used two natural language processing (NLP) tools to perform knowledge discovery from a bag of approximately seven million words. The key issues in the comment threads were mostly associated with efficiency, performance, trust, comfort, and safety. The perception of safety and risk increased in the textual contents when videos presented full automation level. Sentiment analysis shows mixed sentiments towards autonomous vehicle technologies, however, the positive sentiments were higher than the negative.

Advances in vehicle technology have made autonomous vehicles a matter of common public interest and this interest has increased significantly. An online survey of adults over 18 years old conducted in June 2016 found that 87% of respondents had heard about autonomous vehicles, a significant increase from 74% reported in 2013 (*1*).

Companies and researchers have conducted surveys that investigate the public's opinion, acceptance, and preferences, among other topics, related to self-driving technology and autonomous vehicles. Despite the increasing use of online videos about autonomous vehicles, there has not yet been a systematic content analysis of the videos in this arena. Increased attention has been paid in recent years to the role that social media can play in shaping perceptions of individuals on various issues and products. Videos have been used for developing public perceptions because they permit the visualization of concepts, information, and dialogues and allow user-generated communications. YouTube.com is the largest online platform for open access video content and has more than one billion users (*2*). As such, this platform plays a key role in creating public opinion on autonomous vehicles. The top 15 YouTube videos on autonomous vehicles have a total of 60.9 million views and

[1] Texas A&M Transportation Institute, Texas A&M University System, College Station, TX
[2] Computer Science and Engineering Department, Texas A&M University, College Station, TX
[3] City of College Station, College Station, TX
[4] Department of Civil and Environmental Engineering, Rowan University, Glassboro, NJ
[5] Texas A&M Transportation Institute, Texas A&M University System, Houston, TX

Corresponding Author:
Address correspondence to Subasish Das: s-das@tti.tamu.edu

Table 1. Studies on Perceptions of Autonomous Vehicles

No.	Region studied	Concerns regarding autonomous vehicles	Finding	Study
1	U.K., U.S.A. and Australia	• Safety • Security • Loss of control	• People who know about autonomous vehicles have a favorable opinion towards them	(*4*)
2	Berkeley, California, U.S.A	• Loss of control • Liability	• Respondents found enhanced safety to be the most attractive feature • Wealthier people are more likely to be interested in self-driving cars	(*3*)
3	Austin, Texas, U.S.A.	• Equipment failure	• Respondents believed that self-driving cars would relieve congestion and improve safety • Younger males familiar with this technology are more likely to pay a higher price for it	(*8*)
4	U.S.A.	• Loss of control	• A majority of respondents did not know much about self-driving cars. • Level 4 automation more appealing compared to Level 5	(*5*)
5	109 countries	• Safety • Security • Legal liability	• Respondents from developed nations were more concerned about personal data transmitting	(*9*)
6	U.S.A.	• Lack of trust	• Majority of respondents feel less safe sharing the road with self-driving cars	(*6*)
7	U.S.A.	• Lack of trust	• Consumer interest in self-driving cars had increased from 2014 to 2016 • Interest in safety feature of autonomous vehicles has increased also	(*10*)
8	U.S.A.	• Lack of trust	• Younger adults are comfortable with fully autonomous vehicles • Older adults are more comfortable with automation systems that assist the driver	(*11*)
9	U.S.A.	• Lack of trust • Loss of control	• Respondents preferred manually driven vehicles over self-driving vehicles • Respondents were more comfortable with partially self-driving cars as compared with completely self-driving cars	(*7*)

contain around 25,000 comments. YouTube allows users to watch videos without logging in to a user account, which makes it distinct from other social media applications. In this respect, commenting on YouTube videos is different from other social media applications. Social media mining can be considered as an alternative to conventional surveys due to its capability to capture the most recent or real-time opinions, concerns, and sentiments instantaneously. There is a need to conduct analysis on this unexplored and unstructured textual content associated with consumer perceptions, in the present case, toward autonomous vehicles.

This study collected comments and additional information related to the 15 most-viewed YouTube videos on autonomous vehicles to perform this analysis. Natural language processing (NLP) methods including text mining, sentiment analysis, and polarity mapping were employed to accomplish the research goals.

Earlier Work and Research Context

Researchers have conducted many surveys over recent years to measure public perception of autonomous vehicle technology. In 2013, researchers investigated the attitudes of residents in Berkeley, California, toward self-driving vehicles (*3*). The survey found that safety and convenience were the most attractive features of self-driving cars. Liability, cost, and loss of control were among the main concerns. Most participants (46%) in the Berkeley study believed that autonomous cars should drive with normal traffic. In 2014, Schoettle and Sivak developed a questionnaire which addressed the expected benefits, concerns, and overall interests of self-driving vehciles, among other topics. These researchers conducted the survey online in the U.S.A., U.K., and Australia through a web-based company (*4*). Of the respondents, 56.8% had a very positive to a somewhat positive opinion, 29.4% had a neutral opinion, and 13.8% had a somewhat negative to very negative opinion. Kelley Blue Book (KBB) surveyed U.S. residents during May 2016 (*5*) to learn about the level of concern about fully autonomous vehicles and loss of control. The survey found that 49% prefer a safer roadway, even if it means having less control of the vehicle, and 51% prefer having full control of the vehicle. The KBB survey also found that one in three people said they would never buy a fully autonomous (level 5) vehicle. Concerns about fully autonomous vehicles were confirmed by surveys conducted by the AAA Foundation. The survey found that three out of four drivers in the U.S.A. feel “afraid” to ride in a self-driving car, and only one in five respondents would trust a vehicle to drive itself (*6*). A study by Schottle et al. (*7*) also found that people are more comfortable with partially self-driven vehicles as compared with completely self-driving cars. Table 1 summarizes studies measuring public perception of autonomous vehicles in the U.S.A. and around the world. Many of the studies found that people are mostly concerned about the lack of control, security, and safety. Studies also showed that attitudes toward autonomous vehicles are much more favorable among younger people and people who are familiar with autonomous vehicles. These studies are crucial for understanding public perception.

Studies on perceptions of autonomous vehicles do not include assessment of the information provided to inform the public about autonomous vehicles, however, information about the response to different types of information on autonomous vehicle technology is required to understand how public perception is formed. Content analysis is one of the key methods for carrying out this type of study. It has been extensively used to analyze data from YouTube and other social networking sites related to health, politics, and marketing (*12–16*). Hawkins and Filtness used content analysis to study different attributes of driver sleepiness videos on YouTube (*17*). The study team watched 442 videos and classified them according to the following themes: tone, outlook on driver sleepiness, and portrayal of driver sleepiness. The tone was coded as “humorous,” “neutral,” or “serious.” The outlook was coded as “dangerous,” “amusing,” “does not impact driving,” or “can be overcome.” It was found that humorous videos had significantly more views; amusing videos received the most views per video, comments, and likes. The study helped researchers understand what categories of videos capture the public’s interest (*17*).

In recent years, several studies have incorporated text mining in transportation engineering research: consumer complaint analysis (*18*, *19*), social media mining (*20–23*), opinion mining on safety enhancement and bike sharing (*24*, *25*), topic modeling on transportation engineering conference papers and journals (*26–30*), and crash narrative investigation (*31*, *32*).

Investigation of the textual content related to YouTube videos on autonomous vehicles has not yet been conducted. It is always questioned whether social media data are representative and unbiased enough for a robust study. This study contemplates that an aggregation of seven million words could be a low-cost approach to understanding public opinion and sentiment about autonomous vehicle technologies. It is important to note that a larger dataset does not reduce bias. An extension of the current study incorporating sampling methods like systemic and snowball sampling could be considered as a viable alternative in understanding public opinion on autonomous vehicles. This current study applies the NLP framework to perform knowledge discovery on the motives of user participation and consumption of YouTube videos associated with autonomous vehicles. The analysis will not only capture the perception of people towards autonomous vehicles but also provide insight into the future of the adaptation of autonomous vehicles on public roadways.

Data Collection and Data Processing

Classification of the selected videos into clusters is vital to understanding the knowledge pattern in each of these clusters. Videos were categorized based on the title and the content within the video.

A detailed list of keywords was developed by using the following terms: “self-driving cars,” “self-driving vehicles,” “autonomous car,” “autonomous vehicle,” “automated vehicle,” “automated car,” and “driverless car.” The researchers automated the data collection process (extracting the video information as well as related comments) by using an open-source R software package called “tuber” (*33*). For NLP tasks, two R packages (*tm*, and *tidytext*) were used (*34*, *35*). Initially, the researchers collected a list of 107 related video data and associated comments. The top 15 videos (by the number of views) were selected to accomplish the research goals. Initially, these videos had 38,746 associated comments.

Table 2 lists the numbers of channel subscribers and views of the 15 most-viewed YouTube videos on self-driving cars. It is seen that most of these videos are posted on channels maintained by the autonomous car companies. It can be argued that these videos do not present the technical barriers, safety concerns, infrastructure issues, and other problems associated with autonomous vehicles. However, there is a likelihood that the viewers are aware of some of these crucial concerns, as revealed in the later part of the study. Additionally, exploration of the 15 videos shows that around 20% of them are for marketing purpose; the rest of the videos are either proof of concept or comparison. The median number of subscribers per channel was 171,547 (inter quartile range [IQR] = 48,057–1,614,232). The median number of views was 1,851,729 (IQR = 1,113,255–7,223,283). Additionally, the per-year-view data shows an exponential growth of comments:

- 2014: 144 comments
- 2015: 653 comments
- 2016: 9,066 comments
- 2017 (until June 30, 2017): 15,766 comments

Representative probability samples are difficult to acquire in NLP studies. However, some studies have attempted systemic sampling and snowball sampling in social media mining studies, which are not conducted in this study.

Table 2. Subscribers and Views of 15 Most-viewed YouTube Videos on Autonomous Vehicles

No.	Video ID	Video name	Channel	Uploaded	Subscribers	Views
1	EPTIXIdrq3Q	Hyundai: The Empty Car Convoy	HyundaiWorldwide	2014	100,177	12,468,869
2	CqSDWoAhvLU	A First Drive	Google Self-driving car project	2014	48,057	10,776,991
3	cdgQpa1pUUE	Self-Driving Car Test: Steve Mahan	Google	2012	5,164,918	7,952,431
4	IWB4xj7EILg	A Driving Experience of a Different Kind – the F 015 − Mercedes-Benz Original	Mercedes-Benz	2015	385,167	7,223,283
5	XZxZC0lgOlc	Mercedes Self Driving Truck Driving Itself Mercedes Future Truck 2025 Commercial	CARJAM TV	2015	171,547	4,630,723
6	0DS9PY6iaxE	BMW Vision Self Driving Car World Premiere 2016 New BMW Vision Concept	CARJAM TV	2016	171,547	3,906,134
7	tP7VdxVY6UQ	Testing Tesla's Autopilot System At 70 mph	Car Throttle	2015	1,614,232	3,518,868
8	_CdJ4oae8f4	Self-Driving Uber Running Red Light	Charles Rotter	2016	344	1,851,729
9	WsnKzK6dX8Q	New Self-Driving BMW 330i	Phone and a Drone	2011	1,475	1,575,418
10	uCezICQNgJU	Ready for the Road	Google Self-driving Car Project	2015	48,057	1,456,678
11	TsaES—OTzM	A Ride in the Google Self Driving Car	Google Self-driving Car Project	2014	48,057	1,277,175
12	KTrgRYa2wbI	Meet the 26-Year-Old Hacker Who Built a Self-Driving Car... in His Garage	Bloomberg	2015	594,891	1,113,255
13	WBjY3QGNdAw	The Real Moral Dilemma of Self-Driving Cars	Veritasium	2017	4,190,807	1,083,117
14	MO0vdNNzwxk	Tesla Test Drive: Model P85D, Autopilot, Zero to 60	Bloomberg	2014	594,891	1,042,394
15	XIzimkcuEuk	Self Driving Mercedes: Behind the Wheel!	Marques Brownlee	2017	4,645,688	1,045,301

Natural Language Processing

NLP is a scientific method which tends to view the process of language analysis as being decomposable into a number of stages, mirroring the theoretical linguistic distinctions drawn between syntax, semantics, and pragmatics. The simple approach is that the sentences of a text are first analyzed in relation to the associated syntax, which provides a procedure that is more suitable to an analysis in relation to semantics and other meanings. A short description of the steps is given below:

- *Tokenization*: Segmentation of texts is known as tokenization. For computational linguistics purposes, the words thus identified are frequently referred to as tokens, and word segmentation is also known as tokenization.
- *Lexical analysis*: In the realms of computational morphology, this analysis aims to uncover information that may require additional text processing (for example, synonym clustering) and analysis.
- *Syntactic analysis*: Extracting the meaning of a sentence is a key issue. Sentences are not just linear sequences of words. This step involves the determination of the syntactic or grammatical structure of each sentence.
- *Semantic analysis*: This step provides a structured data frame that is more amenable to further analysis and subsequent interpretation.
- *Pragmatic analysis*: This analysis involves analysis of the coherent placement of words and their meaning.
- *Knowledge discovery*: The final stage provides knowledge discovery or important findings from the complex unstructured text data.

This study mainly focuses on two NLP techniques: (i) using popular text mining algorithm TF-IDF to identify rare but significant trends, and (ii) performing sentiment analysis on texts segregated by content type and automation levels.

Descriptive Statistics

The key text categorization task was performed based on two broader categories:

- Categorization by content types of videos
- Categorization by different automation levels

The U.S. National Highway Traffic Safety Administration (NHTSA) developed six different

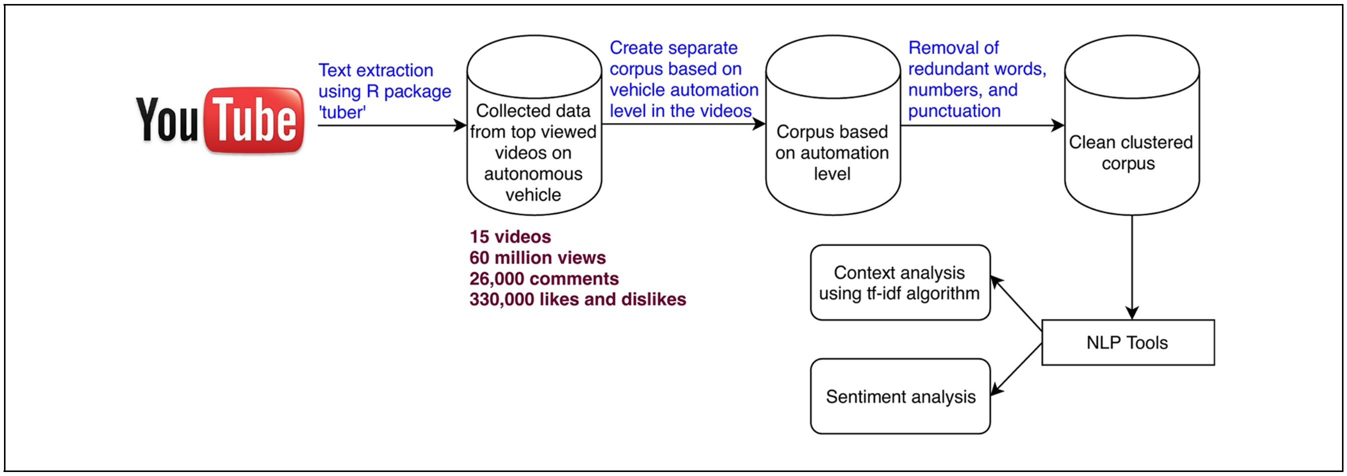
Text extraction using R package 'tuber'
Collected data from top viewed videos on autonomous vehicle
15 videos
60 million views
26,000 comments
330,000 likes and dislikes
Create separate corpus based on vehicle automation level in the videos
Corpus based on automation level
Removal of redundant words, numbers, and punctuation
Clean clustered corpus
NLP Tools
Context analysis using tf-idf algorithm
Sentiment analysis

Figure 1. Flowchart of the study's workflow.

Table 3. Clusters of Most-viewed 15 YouTube Videos on Autonomous Vehicles

No.	Video ID	Category	Level of automation	Comments	Likes	Dislikes	Ratio of likes and dislikes
1	EPTIXldrq3Q	Marketing	Level 2 (L2)	594	14,265	809	17.63
2	CqSDWoAhvLU	Marketing	Level 5 (L5)	4653	49,948	3,272	15.27
3	cdgQpa1pUUE	Proof of concept	Level 5 (L5)	6039	44,882	2,797	16.05
4	IWB4xj7EILg	Proof of concept	Level 5 (L5)	2400	38,507	2,927	13.16
5	XZxZC0lgOlc	Proof of concept	Level 4 (L4)	500	9,199	1,329	6.92
6	0DS9PY6iaxE	Marketing	Level 3 (L3)	700	12,883	1,080	11.93
7	tP7VdxVY6UQ	Comparison/test	Level 3 (L3)	2100	31,918	1,005	31.76
8	_CdJ4oae8f4	Violation	Level 3 (L3)	1455	5,157	3,960	1.30
9	WsnKzK6dX8Q	Comparison/test	Level 2 (L2)	288	4,802	187	25.68
10	uCezlCQNgJU	Proof of concept	Level 5 (L5)	500	0	0	–
11	TsaES—OTzM	Proof of concept	Level 5 (L5)	600	5,177	342	15.14
12	KTrgRYa2wbI	Proof of concept	Level 3 (L3)	500	8,528	203	42.01
13	WBjY3QGNdAw	Comparison/test	Level 3 (L3)	2800	37,262	3,900	9.55
14	MO0vdNNzwxk	Proof of concept	Level 2 (L2)	300	2,952	152	19.42
15	XIzimkcuEuk	Comparison/test	Level 2 (L2)	2200	41,787	551	75.84

automation levels (*36*). NTHSA adopted the levels of automation outlined by SAE International in 2016. The levels draw a distinction between levels 0 to 2 and 3 to 5. The distinction is based on whether the human operator or the automated system is responsible for monitoring the driving environment. Levels 0 and 1 indicate almost no automation. A vehicle of level 2, 3 or 4 could have one or multiple systems that perform a specific function under certain conditions (i.e., freeway driving, self-parking). Level 5 has an automated system that is capable of performing under all conditions.

In this study, the textual features of autonomous vehicle-related videos on YouTube (video title, description, comments, likes, and dislikes) were collected. The data were classified based on two categories: (i) content type (proof of concept, comparison, marketing, and violation), (ii) automation level (level 2, level 3, level 4, or level 5). The classifications were done to identify the main intent of the videos. For example, the "violation" category includes only one video, which shows an automated Uber car violating a red signal. The current data contains a larger sample of data (bag of seven million words). This study performed two key analyses: context analysis by text mining and sentiment analysis (see Figure 1).

Table 3 provides descriptive statistics of the top 15 most-viewed videos. This study has limited the analysis to 15 videos due to computational/memory issues. After removing redundant and non-English comments, the final number of comments was 25,629. Among the top 15 videos, one video was released earlier (in 2011). The overall number of views for all videos was 60,922,366 (mean: 4,061,491, standard deviation: 3,807,296). The hourly views are around 1,200. The number of likes on all videos was higher than dislikes (307,267 versus 22,514). The number of comments was 25,629 (mean: 1,709; standard deviation: 1,720). This study determined the like-dislike ratio (likes versus dislikes ratio) statistics for two broad text categorizations (content type based and automation level based). The statistics show that comparison/test videos have the highest like-dislike ratio of 35.71 among all the categories. However, this category also has the highest variability (standard deviation: 28.35) thus indicating that the reception of the four comparison/test videos was quite different. The violation category has one video, and this category has the lowest like-dislike ratio. Regarding the level of automation, it was found that level 2 automation had the highest like-dislike ratio (34.64) with the highest standard deviation of 27.68. Level 4 had the lowest like-dislike ratio (6.92). In general, the reception of lower levels of automation in videos was better than the reception of higher automation levels. The authors acknowledge that the perception of "like-dislike" buttons is not the best measure of understanding public opinion, however, these statistics provide the extent of people's engagement in social media.

Text Mining

Most Frequent Terms. Term frequency (TF) is the first step in understanding patterns from any unstructured text data. By considering all comments as a single corpus, a general TF analysis was performed. The words with the highest frequencies were the names of self-driving car companies or words associated with human driving. Words related to vehicle technology or parts ("wheel," "steering," "system," "technology," "computer," and "control") are also highly frequent. The word "trust" is one of the words with the highest frequencies, which implies a lack of confidence in the functionality of autonomous vehicles.

Context Analysis using Term Frequency-Inverse Document Frequency (TF-IDF)

Instead of using a word or word group frequencies, another approach is to look at a term's inverse document frequency (IDF). Spark Jones first introduced this concept in 1972 (*19*). This concept accounts for the database size and term distribution in the database in determining the weights. It decreases the weight for commonly used words and increases the weight for words that are used infrequently in a collection of documents (for example, one can consider one comment as one document; one document can also contain all comments of the videos of a particular category). This approach has provided robust usefulness in NLP. The IDF for any given term is defined as

$$IDF(term) = \ln\left(\frac{N}{d_i}\right) \tag{1}$$

where,

$N =$ number of documents in a database

$d_i =$ number of documents containing the word i in the entire database

This can be combined with TF to calculate a term's TF-IDF (the two quantities multiplied together,$TF \times IDF$). The concept behind TF-IDF is to find the important words for the content of each document by decreasing the weight for commonly used words and increasing the weight for words that are not used very much in a collection or corpus of documents (*20*). Calculating TF-IDF attempts to find the words that are important in a text, but not too common. The final term weight, wt_i, for TF-IDF can be written as:

$$TF - IDF(wt_i) = f_i \times \log\left(\frac{N}{d_i}\right) \tag{2}$$

where $f_i =$ frequency of the word i in the document.

Example: For instance, consider the word "the." It appears in all four documents. Thus,

$$\begin{aligned} &IDF(\text{'the'}) = \ln\left(\frac{4}{4}\right) = 0 \text{ and} \\ &TF - IDF\left(\text{'the'}, doc_{comparison}\right) = TF - IDF\left(\text{'the'}, doc_{marketing}\right) \\ &= TF - IDF(\text{'the'}, doc_{violation}) = TF - IDF\left(\text{'the'}, doc_{concept}\right) \\ &= \text{TF} \times \text{IDF} = \text{TF} \times 0 = 0 \end{aligned} \tag{3}$$

The word "Tesla" is considered as another example. The count numbers considered in this example do not indicate real calculated counts. Consider that "Tesla" appears in only two documents out of four documents. So,

$$\text{IDF}(\text{'tesla'}) = \ln\left(\frac{4}{2}\right) = 0.301 \tag{4}$$

The TF-IDFs for the four documents are:

$$\begin{aligned} &TF - IDF\left(\text{'tesla'}, doc_{comparison}\right) = \frac{2,100}{15,000} {*}log\frac{4}{2} = 0.042 \\ &TF - IDF\left(\text{'tesla'}, doc_{marketing}\right) = \frac{50}{20,000} {*}log\frac{4}{2} = 0.0008 \\ &TF - IDF\left(\text{'tesla'}, doc_{concept}\right) = 0^{*}log\frac{4}{2} = 0 \\ &TF - IDF(\text{'tesla'}, doc_{violation}) = 0^{*}log\frac{4}{2} = 0 \end{aligned} \tag{5}$$

Figure 2 shows the TF-IDF for the four categories. The comparison category has mainly neutral words and words related to morality such as "dilemma," "moral,"

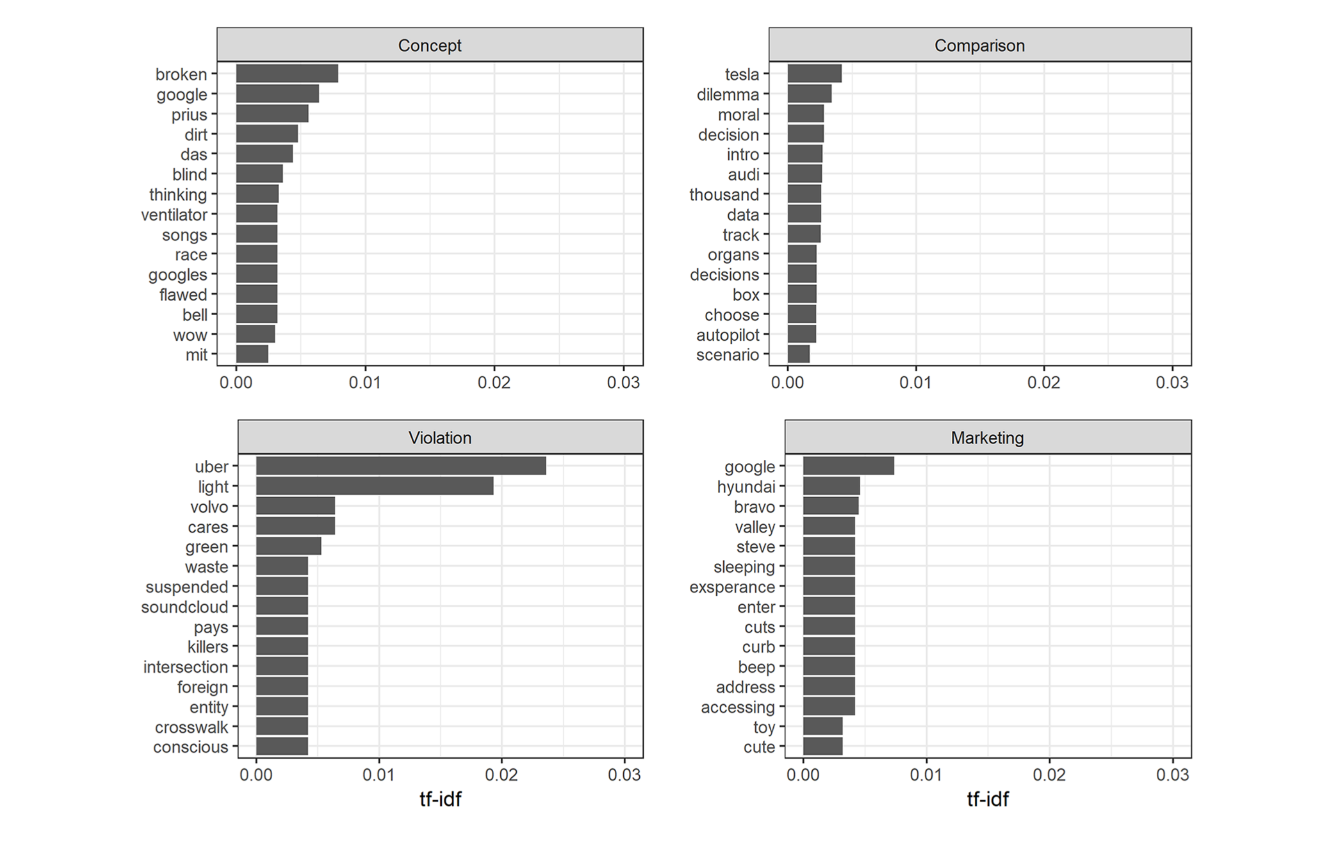

Figure 2. Individual TF-IDF for texts categorized by content type: (*a*) concept, (*b*) comparison, (*c*) violation, (*d*) marketing.

and "punishment." The marketing category has mostly positive or neutral words such as "bravo," "toy," and "cute." The concept category has a high frequency of the term "broken." This category has both positive words ("wow") and negative words ("dirt," "broken," and "flawed"). It also contains a technical term "Data Acquisition System (DAS)." The violation category has high TF-IDF for "uber" and "light." This is because there was only one video in the violation category and it was about an Uber autonomous car missing a red light. This category has some strong negative words like "killers" and "waste."

Figure 3 shows the TF-IDF for the top 15 words. The color indicates the higher presence of that word in that automation level. For example, "autopilot" has a TF-IDF value of 0.008. The color indicates that the presence of this word is more likely to be in Level 2 compared with other automation levels. The top words in Level 2 are "autopilot," "Tesla," "Hyundai," "class," "Musk," and "German." For Level 5, the highly associated words are "blind," "Google," "broken," "innovation," and "dirt."

Co-occurrence of Terms. It is also important to investigate the use of terms for various automation levels (see Figure 4). Medium automation (Level 3 [conditional automation] and Level 4 [high automation]) and full automation (Level 5 [full automation]) keywords are compared with Level 2 (partial automation) keywords. The axes represent a percentage of word frequencies in each dataset. Words that are close to the dotted line in these plots have similar frequencies in both sets of texts (for example, in the Level 2 versus Level 3–4 plot, set 1 includes Level 2 word percentages, and set 2 includes Level 3–4 word percentages). Words that are far from the line (gray or light green colored texts) are words that are found more in one set of texts than the other. For example, in the Level 2 versus Level 3/4 panel, words with similar frequencies are "automatic," "cab," "accident," and "future". "Mercedes," "Tesla," "people," and "trust" are examples of words with different proportions in these automation levels. The second plot (Level 2 versus Level 5 panel) shows similar and disproportionate frequencies of different sets of words. "Google," "Tesla," and "Mercedes" show a significant difference in frequency proportions. "Tesla" and "Mercedes" show high proportions on the Level 2 axis, while "Google" shows high proportions on the Level 5 axis. Additional observations can be made from this graphic to identify trends in different automation levels. For both plots, the term "accidents" occurs more frequently in Levels 3 to 5 than

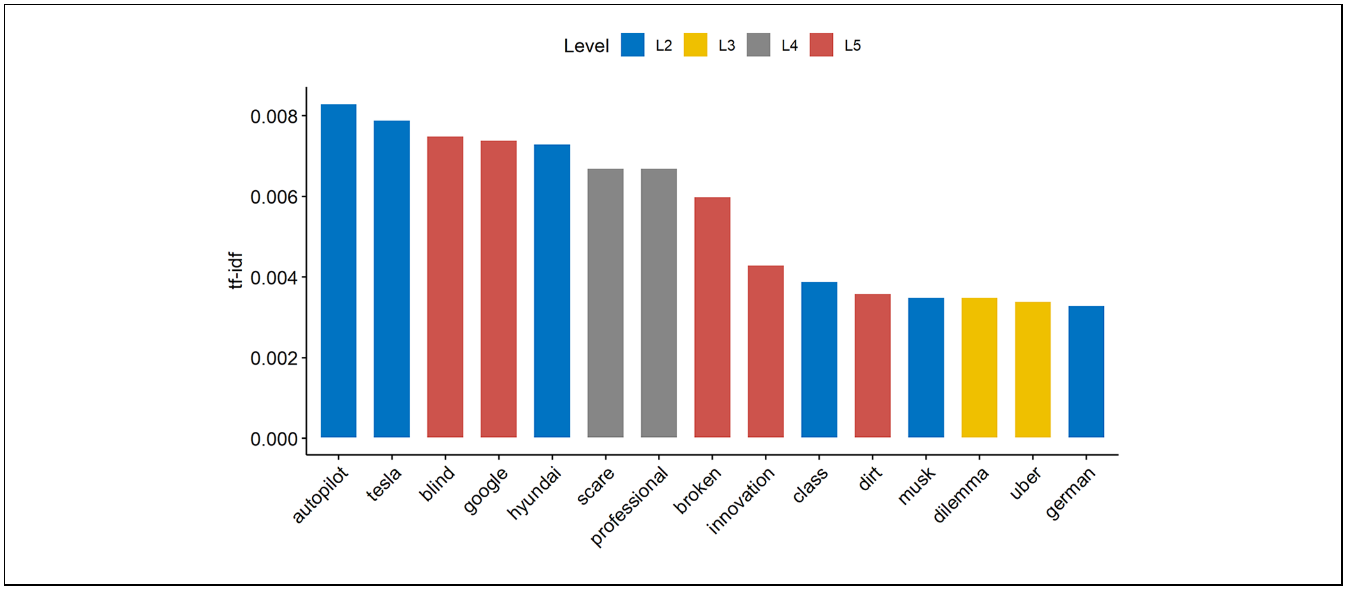

Figure 3. Combined TF-IDF plot for text categorized by different automation levels.

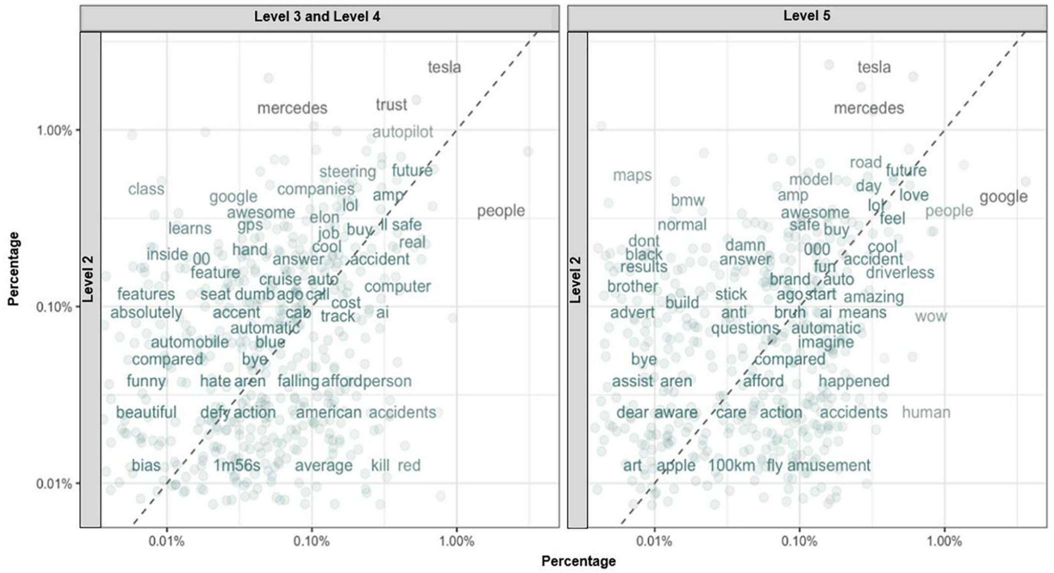

Figure 4. Co-occurrence of words for text categorized by different automation levels: (*a*) Level 3 and Level 4, (*b*) Level 5.

in Level 2. For example, "accidents" contributes 0.15% of the complete textual content in Level 2–3 text content but only 0.04% in Level 2 textual contents. This information suggests that the consumers' concerns about crash and safety increase significantly with an increase in automation level.

Sentiment Analysis

Mining of subjective texts containing opinion or sentiment can contribute to understanding perceptions towards a product. In other words, the objective of sentiment analysis is to determine which words or sentences express opinions, feelings, and sentiments. For example, "amazing" contains a positive connotation, whereas "worst" carries a negative one. Similarly, "maybe" expresses uncertainty and "court" carries a litigious meaning. The concept of sentiment analysis is not described in this study. Interested readers are referred to the study by Das et al. (*24*) for details of the sentiment analysis procedure. By using a sentiment score algorithm, words/terms were tagged in four sentiment

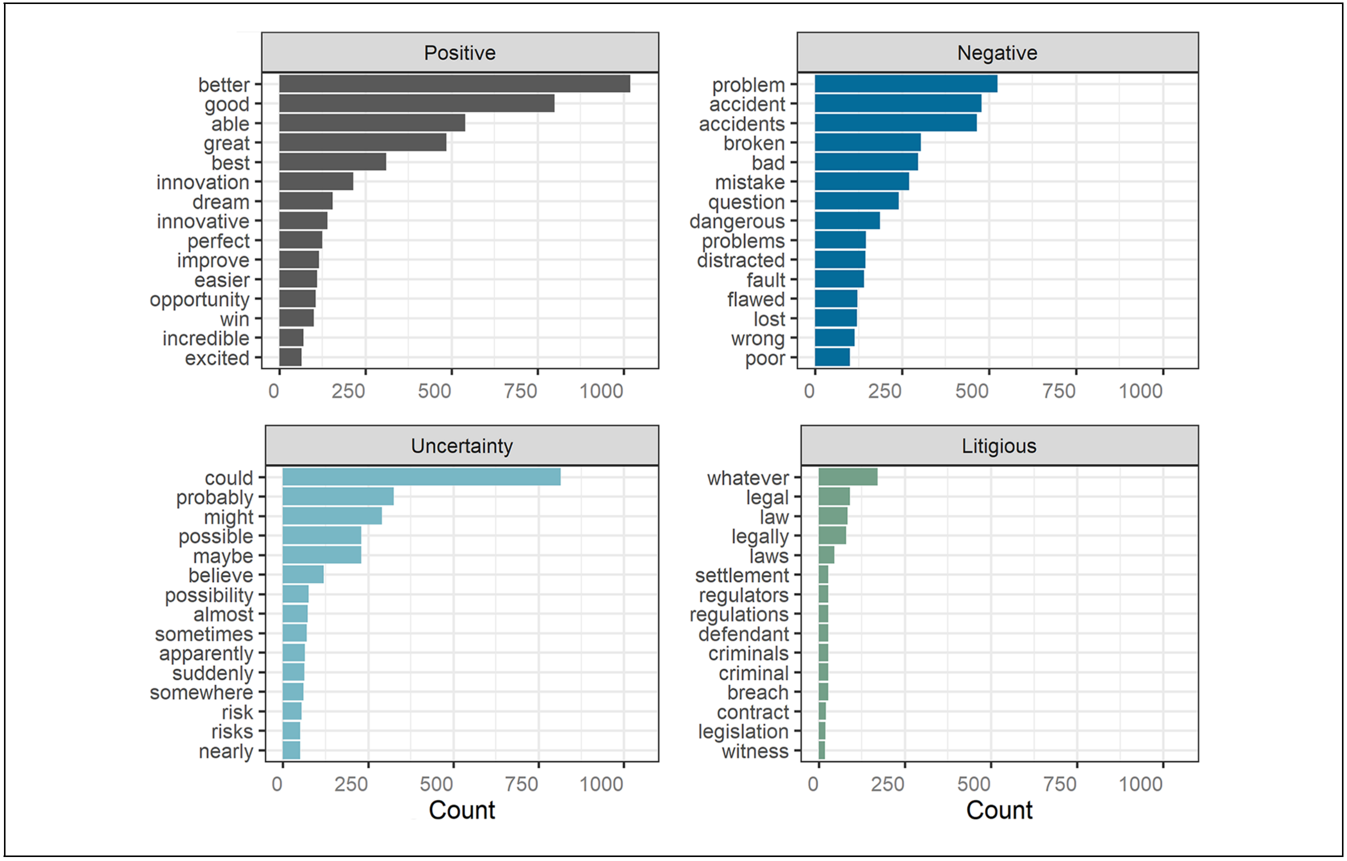

Figure 5. Top keywords associated with different sentiment classifications: (*a*) positive, (*b*) negative, (*c*) uncertainty, (*d*) litigious.

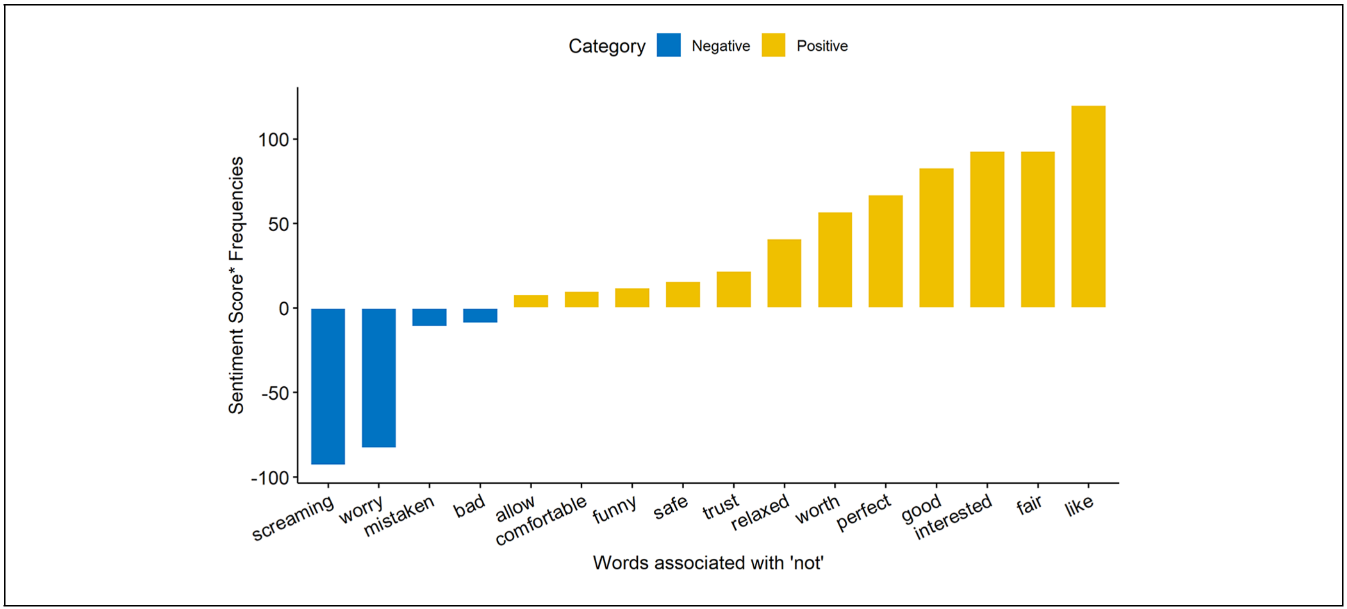

Figure 6. Sentiment score-frequencies versus words associated with "not."

classifications (as shown in Figure 5): (i) positive, (ii) negative, (iii) uncertain, and (iv) litigious. Figure 6 shows that negative comments revolve around the concern of potential problems with the automated system being flawed or breaking, and potential accidents. Positive comments used words "better," "innovation," and "able," showing that there is an outlook that automated vehicles will improve transportation, mobility, and other aspects of daily commuting. The figure also shows that positive words were used more frequently than terms that portray uncertainty or negative attitudes. The quantification of all words (with sentiment classification) shows that positive terms are 35% higher than negative, uncertain, and litigious terms.

Let a document with opinion be *t*, which can be a term that evaluates or expresses on a subject or a group of subjects. In the most general case, *t* consists of a sequence of words or sentences $t = w_1, w_2, \ldots, w_n$. The definition of a sentiment passage on a feature is as follows: "A sentiment on a feature *f* of an object *o* evaluated in *t* is a group of consecutive words or sentences in *t* that expresses a positive or negative opinion on *f*" (*37*). Additionally, sentiments also contain subjectivity. Objective sentence presents some information about the world, and a subjective sentence expresses some personal feelings or beliefs. Document-level sentiment classification involves a definite task with assumptions. These are stated below:

- *Task*: Given a set of opinionated terms *t*, it determines whether each term $t \in T$ expresses a positive/negative/uncertain/litigious sentiment on an object. Given an opinionated document *t* that comments on an object *o*, determine the orientation/subjectivity *oo* of the opinion expressed on *o*, that is, discover the opinion orientation/subjectivity *oo* on feature *f* in the quintuple (o, f, oo, h, p), where $f = o$ and *h, p*, and *o* are assumed to be known or irrelevant.
- *Assumption*: The opinionated term $t =$ express opinions on a single object *o* and the opinions are from a single opinion holder *h*.

At times, TF does not help in understanding the inherent meaning. A group of words can express subjectivity better than a single word. For example, "like" presents an affirmation message, on the other hand, "not like" represents a completely opposite meaning. Figure 6 shows sentiment scores (multiplied by frequencies) of the words associated with "not." This plot shows that higher percentages of people are not comfortable with autonomous technologies. The words associated with "not" are "like," "interested," "fair," "good," "perfect," "worth," "relaxed," "trust," "comfortable," "allow." These findings are in line with other studies (*3–6*, *8*, *10*, *11*).

It is important to note that negations can appear in different forms, altering or reversing not only the meaning of a single word or word groups but also the inherent meaning. Future studies may consider rigorous analysis on the two major categorizations by involving the scope of each negation word or the phrase preceding or succeeding it because the polarities can be altered using negation words or phrases.

Conclusions

Understanding people's perceptions and ideas about autonomous vehicles is a very important area of research. Consumer adaptation and awareness is a key area of focus for the autonomous vehicle industry. To make consumers know about their products, the autonomous vehicle industry develops many videos and shares them online. This study advances the general understanding of the perceptions and barriers of end users towards autonomous vehicles by performing an analysis of reactions to videos about the autonomous vehicle on YouTube. The study focuses on the content exploration of these videos by examining the attitudes of the end users regarding liking, disliking, and commenting patterns. Conventional surveys are not without limitation in understanding consumer perceptions due to the lack of prompt and effective information collection and retrieval. This study can be considered an alternative approach to assess consumer response to autonomous vehicle technology. The findings of this study include:

- A large number of views, comments, likes, and dislikes can indicate that the public is engaging with autonomous vehicle technologies.
- Two major categorizations (content-based and automation level-based) show different perception trends.
- The likelihood of the perception of safety increases with the increase in automation level.
- The TF-IDF algorithm identifies several rare but significant words for different categorization levels. The quantification of these terms/words generally contains added values for interested focus groups like car industries, investors, and potential buyers.
- Positive sentiments towards autonomous vehicles are more frequent than negative, uncertain, or litigious sentiments. Sentiment analysis was performed at the disaggregate level to show word-level association with different frequencies. The finding can be related to bias in two important ways: young people are more likely to post comments on social media, and young people are more positive about vehicle automation technologies. This can be considered as a limitation of the current study.
- The subjectivity analysis was done partly by considering words associated with "not." The findings show that people express stronger feelings when using the term "not." Words that are associated with "not" are "like," "interested," "fair," "good," "perfect," "worth," "relaxed," "trust," "comfortable," and "allow." These terms identify several

major topics associated with concerns about the adoption of vehicle automation.

Findings from this study can benefit car industries, policy makers, and finally consumers. This research can provide additional insights guiding practitioners as well as researchers by providing an understanding of the attitudes of people towards autonomous vehicle technology. Additionally, this study developed a low-cost framework to discover consumers' adaptation needs which can be replicable for other transportation-related topics. There is an argument that social media mining has a disadvantage in considering a sample which is, in the long run, biased and not representative. Social media posts have two key advantages: (i) the data analysis can be done in real time, and (ii) variety of views on the term of interest. However, social media posts usually contain substantial amounts of noise due to trolls and unrelated information.

Despite its positive contributions to the state of the industry, this current study has several limitations. First, the study only analyzed 15 most-viewed videos and their related comments. Future studies should consider expanded data or data collected from other video platforms like Vimeo or Facebook. Second, the sentiment lexicon used in this study is based on three established lexicons with the addition of an extra 200 words. An extended list that incorporates the transportation engineering sentiment lexicon is a potential research topic. MacEachren et al. argued that the majority of text mining studies focused on data extraction and text categorization but made limited efforts in transforming the findings into actionable knowledge (*38*). It is important to translate the findings into applicable knowledge/outcomes. The current study has developed a framework of analysis in understanding public perception of autonomous cars from YouTube comments. This can be considered as a starting point for performing more rigorous analysis in the future. The findings from the current exploratory analysis need to be interpreted with caution.

Acknowledgments

This project has not been funded by any research grant. The authors thank four anonymous reviewers for their valuable comments and guidelines which helped to improve the manuscript.

Author Contributions

The authors confirm their contributions to the paper as follows: study conception and design: AD, SD; data collection: SD; analysis and interpretation of results: SD, AD, TL; draft manuscript preparation: SD, AD, TL, ZE, MJ. All authors reviewed the results and approved the final version of the manuscript.

References

1. *Autonomous Vehicles*. State Farm Mutual Automobile Insurance Company, Strategic Resources Department. https://newsroom.statefarm.com/download/229160/2015autonomousvehiclesreport.pdf. Accessed July 27, 2018.
2. Naslund, J. A., S. W. Grande, K. A. Aschbrenner, and G. Elwyn. Naturally Occurring Peer Support through Social Media: The Experiences of Individuals with Severe Mental Illness Using YouTube. *PLOS One*, Vol. 9, No. 10, 2014, p. e110171.
3. Howard, D., and D. Dai. Public Perceptions of Self-driving Cars: The Case of Berkeley, California. Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014.
4. Schoettle, B., and M. Sivak. *A Survey of Public Opinion about Autonomous and Self-Driving Vehicles in the US, the UK, and Australia*. Report UMTRI-2014-21. The University of Michigan Transportation Research Institute, 2014.
5. *Future Autonomous Vehicle Driver Study*. Kelly Blue Book. 2018. https://mediaroom.kbb.com/future-autonomous-vehicle-driver-study. Accessed Feb 28, 2019.
6. *AAA Foundation. Three-quarters of Americans "Afraid" to Ride in a Self-Driving Vehicle*. AAA Newsroom. http://newsroom.aaa.com/2016/03/three-quarters-of-americans-afraid-to-ride-in-a-self-driving-vehicle/. Accessed July 27, 2018.
7. Schoettle, B., and M. Sivak. *Motorists' Preferences for Different Levels of Vehicle Automation: 2016*. Report SWT-2016-8. Sustainable Worldwide Transportation, University of Michigan, 2016.
8. Bansal, P., K. M. Kockelman, and A. Singh. Assessing Public Opinions of and Interest in New Vehicle Technologies: An Austin Perspective. *Transportation Research Part C: Emerging Technologies*, Vol. 67, 2016, pp. 1–14.
9. Kyriakidis, M., R. Happee, and J. C. de Winter. Public Opinion on Automated Driving: Results of an International Questionnaire among 5000 Respondents. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 32, 2015, pp. 127–140.
10. Giffi, C., J. Vitale, R. Robinson, and G. Pingitore. *The Race to Autonomous Driving: Winning American Consumers' Trust*. Deloitte Review. https://dupress.deloitte.com/dup-us-en/deloitte-review/issue-20/winning-consumer-trust-future-of-automotive-technology.html.
11. Abraham, H., C. Lee, S. Brady, C. Fitzgerald, B. Mehler, B. Reimer, and J. F. Coughlin. *Autonomous Vehicles, Trust, and Driving Alternatives: A Survey of Consumer Preferences*. Massachusetts Institute of Technology, Cambridge, Mass., 2016.
12. Paek, H. J., K. Kim, and T. Hove. Content Analysis of Antismoking Videos on YouTube: Message Sensation Value, Message Appeals, and their Relationships with Viewer Responses. *Health Education Research*, Vol. 25, No. 6, 2010, pp. 1085–1099.
13. Small, T. A. What the Hashtag? A Content Analysis of Canadian Politics on Twitter. *Information, Communication & Society*, Vol. 14, No. 6, 2011, pp. 872–895.
14. Smith, A. N., E. Fischer, and C. Yongjian. How Does Brand-related User-generated Content Differ across

YouTube, Facebook, and Twitter? *Journal of Interactive Marketing*, Vol. 26, No. 2, 2012, pp. 102-113.

15. Waters, R. D., and P. M. Jones. Using Video to Build an Organization's Identity and Brand: A Content Analysis of Nonprofit Organizations' YouTube Videos. *Journal of Nonprofit & Public Sector Marketing*, Vol. 23, No. 3, 2011, pp. 248–268.
16. Yoo, J. H., and J. Kim. Obesity in the New Media: A Content Analysis of Obesity Videos on YouTube. *Health Communication*, Vol. 27, No. 1, 2012, pp. 86–97.
17. Hawkins, A., and A. J. Filtness. Driver Sleepiness on YouTube: A Content Analysis. *Accident Analysis & Prevention*, Vol. 99, 2015, pp. 459–464.
18. Ghazizadeh, M., A. D. Mcdonald, and J. D. Lee. Text Mining to Decipher Free-Response Consumer Complaints. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 56, No. 6, 2014, pp. 1189–1203.
19. Mehrotra, S., and S. Roberts. Identification and Validation of Themes from Vehicle Owner Complaints and Fatality Reports using Text Analysis. Presented at 97th Annual Meeting of the Transportation Research Board, Washington, D.C., 2018.
20. Das, S., X. Sun, M. Zupancich, and A. Dutta. Twitter in Circulating Transportation Information: A Case Study on Two Cities. Presented at 96th Annual Meeting of the Transportation Research Board, Washington, D.C., 2017.
21. Gu, Y., Z. Qian, and F. Chen. From Twitter to Detector: Real-Time Traffic Incident Detection Using Social Media Data. *Transportation Research Part C: Emerging Technologies*, Vol. 67, 2016, pp. 321–342.
22. Das, S., L. Minjares-Kyle, K. Dixon, A. Palanisamy, and A. Dutta. #TRBAM: Understanding Communication Patterns and Research Trends by Twitter Mining. Presented at 97th Annual Meeting of the Transportation Research Board, Washington, D.C., 2018.
23. Das, S., A. Dutta, G. Medina, L. Minjares-Kyle, and Z. Elgart. Extracting Patterns from Twitter to Promote Biking. *IATSS Research*, 2018. doi:10.1016/j.iatssr.2018.09.002
24. Das, S., X. Sun, and A. Dutta. Investigating User Ridership Sentiments for Bike Sharing Programs. *Journal of Transportation Technologies*, Vol. 5, No. 2, 2015, pp. 69–75.
25. Chen, F., and R. Krishnan. *Transportation Sentiment Analysis for Safety Enhancement*. Technologies for Safe and Efficient Transportation, Carnegie Mellon University, Pittsburgh. utc.ices.cmu.edu/utc/CMU%20Reports%202013%202/Final%20Report%20Chen.pdf. Accessed July 27, 2018.
26. Das, S., X. Sun, and A. Dutta. Text Mining and Topic Modeling of Compendiums of Papers from Transportation Research Board Annual Meetings. *Transportation Research Record: Journal of the Transportation Research Board*, 2016. 2552: 48–56.
27. Das, S., K. Dixon, X. Sun, A. Dutta, and M. Zupancich. Trends in Transportation Research. *Transportation Research Record: Journal of the Transportation Research Board*, 2017. 2614: 27–38.
28. Boyer, R. C., W. T. Scherer, and M. C. Smith. Trends over Two Decades of Transportation Research. *Transportation Research Record: Journal of the Transportation Research Board*, 2017. https://doi.org/10.3141/2614-01.
29. Sun, L., and Y. Yin. Discovering Themes and Trends in Transportation Research Using Topic Modeling. *Transportation Research Part C: Emerging Technologies*, Vol. 77, 2017, pp. 49–66.
30. Das, S., A. Mudgal, A. Dutta, and S. Geedipally. Vehicle Consumer Complaint Reports Involving Severe Incidents: Mining Large Contingency Tables. *Transportation Research Record: Journal of the Transportation Research Board*, 2018. 2672: 72–82.
31. Das, S., B. Brimley, T. Lindheimer, and A. Pant. *Safety Impacts of Reduced Visibility in Inclement Weather*. Center for Advancing Transportation Leadership and Safety, University of Michigan, Ann Arbor. www.atlas-center.org/wp-content/uploads/2017/04/SafetyImpacts_Visibility-Weather_FinalReport.pdf. Accessed July 27, 2018.
32. Brown, D. E. Text Mining the Contributors to Rail Accidents. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17, No. 2, 2016, pp. 346–355.
33. Sood, G. *Tuber: Access YouTube from R. R package version 0.9.7.* https://github.com/soodoku/tuber. Accessed Feb 28, 2019.
34. Feinerer, I., K. Hornik, and D. Meyer. Text Mining Infrastructure in R. *Journal of Statistical Software*, Vol. 25, No. 5, 2008, pp. 1–54.
35. Silge, J., and D. Robinson. *Text Mining with R: A Tidy Approach*. O'Reilly Media, CA, 2017.
36. *Federal Automated Vehicles Policy*. Accelerating the Next Revolution in Roadway Safety, National Highway Traffic Safety Administration, U.S. Department of Transportation, Washington, D.C., 2016.
37. Zhang, J. *Visualization for Information Retrieval*. Springer Science & Business Media, WI, 2007.
38. MacEachren, A., A. Jaiswal, A. Robinson, S. Pezanowski, A. S. P. Mitra, X. Zhang, and J. Blanford. Senseplace2: Geotwitter Analytics Support for Situation Awareness. *Proc., IEEE Conference on Visual Analytics Science and Technology*, Providence, R.I., 2011.

The Standing Committee on Effects of Information and Communication Technologies (ICT) on Travel Choices (ADB20) peer-reviewed this paper (19-02467).