Pedestrian Collisions with Bicyclist: Emotion Mining using YouTube Data

Subasish Das, Ph.D.
(Corresponding Author)
Texas A&M Transportation Institute
1111 RELLIS Parkway, Room 4414, Bryan, TX 77807
Email: s-das@tti.tamu.edu
ORCID: 0000-0002-1671-2753

Xiaoqiang "Jack" Kong
Department of Civil Engineering
Texas A&M University
400 Bizzell St, College Station, TX 77843
Email: X-Kong@tamu.edu

Ruihong Wang
Department of Electrical & Computer Engineering
Texas A&M University, Wisenbaker Engineering Building, College Station, TX 77843
Email: Ruihong.wang@tamu.edu

Ahmadreza Mahmoudzadeh
Zachry Department of Civil Engineering, Texas A&M University
3136 TAMU, College Station, TX 77843-3136
Email: A.Mahmoudzadeh@tamu.edu

Word Count: 5481 words + 3 table (250 words per table) = 6231 words

Submitted [08-01-2019]

ABSTRACT

In recent years, researchers have conducted many studies on vehicle-pedestrian incidents and vehicle-bicyclist crashes. However, there is a surprisingly limited amount of research focused on collisions between pedestrians and cyclists, and research about the tension between pedestrians and cyclists is even rarer. The lack of research on this subject is partly due to the limited number of pedestrian-cyclist crashes; it is also due to the fact that the consequences of these crashes are typically less severe than those of automobile-involved crashes. Despite the lack of research focus on this area, pedestrian-cyclist crashes could lead to a serious social crisis. The largest video sharing website, YouTube.com, contains many videos about pedestrian-cyclist crashes. The most viewed 'pedestrian collision with bicyclist' videos on YouTube have a combined 60.9 million views and contain around 25,000 comments in total. The application of content analysis and text mining to the comments from these videos can provide insight into potential interactions. The findings of this study show that the emotion patterns of comments and replies differ. This study also provides word shift plots that show the trend of the emotion used in comments and replies. Additionally, the co-occurrence plots show the reason behind the use of negative emotions. The findings of this study will provide additional insights into the ongoing debate on 'pedestrian collisions with bicyclists' issues.

Keywords: *autonomous vehicles, content analysis, text mining, sentiment analysis.*

INTRODUCTION

In recent years, non-motorized travel modes (walking and biking) have been gaining popularity in the U.S. with non-motorist travelers. To improve the safety of pedestrians and bicyclists, researchers have made multiple efforts to reduce vehicle-bicycle and vehicle-pedestrian crashes. A research area that is less explored is the collision between a bicyclist and pedestrian. A recent study examined the incidences of pedestrians injured by cyclists in California from 2005 to 2011 and in New York between 2004 to 2011 *(1).* The findings show that despite the increasing number of cyclists, there was a decline in the rate of pedestrians injured in collisions with cyclists. One of the primary reasons for the rate decrease is due to the cycling infrastructure improvements. In the absence of cycling infrastructures, cyclists often use pedestrian facilities for part or the entire journey. Pedestrians feel insecure in the presence of speedy bicyclists, especially in dense urban locations.

In recent years, antagonism between pedestrians and bicyclists has gained more attention through social media like mainstream media websites. It is easy to find videos and articles that contain furious comments about careless pedestrians or bicyclists online. The tensions between pedestrians and bicyclists are hard to resolve because both are vulnerable road users. Pedestrians may seem more vulnerable to most of the public, but this belief leads to a violent, negative image of bicyclists *(2).* Mainstream websites present contradicting arguments between the bicyclists and the rest of the public. However, the lack of related academic studies will bridge the gap by exploring the causes of tensions between bicyclists and pedestrians by performing topic modeling and text mining.

In recent years, the role that social media can have in shaping the opinions of individuals on various products and issues has gained attention. Because videos allow the visualization of information, concepts, and dialogues and permit user-generated communications, videos have developed public perceptions. YouTube.com has more than 1 billion users and is the largest online platform for open-access video content *(3).* As such, this platform plays a significant role in generating public opinion on many issues. YouTube contains the highest number of public opinions on video relevant interactions in comparison to other social media platforms. In comparison to conventional survey analysis, social media mining is efficient due to its capability of instantaneously capturing the most recent or real-time concerns, opinions, and sentiments. An analysis conducted on this unstructured and unexplored textual content related to consumer perception towards autonomous vehicles is necessary.

To perform knowledge discovery on the motives of user participation and consumption on YouTube videos associated with the ‘pedestrian collisions with bicyclists,’ this study applies the natural language processing (NLP) framework. The analysis can provide patterns and trends of the interactions between bicyclists and pedestrians.

EARLIER WORK AND RESEARCH CONTEXT

In recent years, researchers have conducted several studies on vehicle-bicyclist crashes and vehicle-pedestrian incidents. However, very little research has concentrated on the collision between cyclists and pedestrians. Moreover, any mention of tension between pedestrians and cyclists in research is rare because of the limited number of crashes; the exception is in some urban-core ped-bike active zones. More serious issues could be a result of the tension between the two groups *(1, 4, 5).*

Tuckel et al. conducted a study to examine the trend of pedestrian injuries in the bicyclist-pedestrian (bike-ped) collisions and investigate possible explanations. Based on the stats generated from the pedestrian and bicycle incident records in New York and California, the authors emphasized that the rate of pedestrians injured in the incidents with bicycles has decreased over time *(1).* However, some reports showed that bike-ped crashes are increasing in urban-core active areas. Even with a decreasing trend in the number of pedestrians injured in bicycle collisions, the total number of injured pedestrians in bicycle incidents is still significant. The New York Times published an article called "The Cyclist-Pedestrian Wars," detailing the rising tensions between bicyclists and pedestrians in the Upper East Side of New York. In the article, residents indicated that their main complaint was the disobedience of bicyclists, such as riding on the sidewalks. In 1996, the city increased the fine from $40 to $100 in an attempt to keep bicyclists off of sidewalks, but it did not result in the desired effect. In 1970, geometric design such as curb cuts was mandated by federal law to allow wheelchairs to roll onto sidewalks. This law also permitted bicyclists to weave effortlessly from the street to the sidewalks *(4).*

An Amsterdam study showed that traffic conflicts with bicycle paths are a specific safety problem for crossing pedestrians *(6).* In 2010, a published article in Reuters illustrated the reasons why there is always a tension between pedestrians and bicyclists and why bicyclists in New York received scattered public support. The article described four reasons: 1) the mindsets of bicyclists are as vulnerable as pedestrians when encountering motor vehicles, and they experience no guilt when riding on sidewalks; 2) unlike most motorists, there is no culture or feeling of obligation for bicyclists to yield to pedestrians; 3) in some locations, like over the Manhattan bridge, pedestrians walk on the bike-only path frequently and feel no obligation to give bicyclists' way back; and 4) pedestrian intuitively believe the bike lane is less dangerous than driveways and normally step into bike lanes without looking first. The article gives an example of a pedestrian wanting to cross the street in the middle of the block when cars are not moving, then the pedestrian walks into the street without looking and gets hit by a bicyclist. The bicyclist blames the pedestrian for stepping into the bike lane illegally, and the pedestrian blames the bicyclist for running too quickly and without looking *(7).* In Santa Monica, California, the tension between the pedestrians and bikers was explained in a published news article *(8).* One reason that bicyclists use sidewalks is the lack of comfortable area on bike routes. A critical root source of those tensions is the built environment. Without the resolution of the shortage of comfortable bike routes or sidewalks, strategies such as education, enforcement, or promoting good etiquette will not ease the tension. Based on 5,421 observations in Sydney, Australia, a recent study also found that environmental factors significantly affect the cycling speed on shared paths. Riders were significantly slower than the median speed on the shared path but faster than the median speed in the presence of a centerline or visual segregation between a bike lane and sidewalk *(5).* Werneke et al. found that pedestrians crossing the bicycle path without looking were the primary cause of the majority of bike-pedestrian incidents *(9).*

In the recent years, several studies have incorporated text mining in transportation engineering research: consumer complaint analysis *(10, 11),* social media mining *(12-15),* opinion mining on safety enhancement and bike-sharing *(16, 17),* topic modeling on transportation engineering conference papers and journals *(18-22)* and crash narrative investigation *(23-24).*

An investigation into the textual content related to YouTube videos on 'pedestrian collisions with bicyclists' has not yet been conducted. A debate on whether social media data is representative and unbiased enough for a robust study. This study contemplates if the collection of

25,000 comments and replies will provide an accurate representation of public opinion and sentiments on this issue.

METHODOLOGY

Natural Language Processing (NLP)

Natural language processing (NLP), a popular branch of computer science, tends to view the process of text and language analysis as being divided into a number of stages by conducting theoretical linguistic distinctions between syntax or parsing (relationship between words), semantics (meaning of a textual content), and pragmatics (process of extracting information from text). Typically, sentences retrieved from a text or document are first analyzed in terms of the associated syntax, which provides a procedure that is more suitable to an analysis in terms of semantics and other associated meaning *(25).* The popular branches of NLP include text mining, topic modeling, sentiment analysis, and emotion mining. Several NLP steps have been taken to accomplish the research goals *(25).* A brief introduction of these steps is described below:

- *Tokenization:* Tokenization is the process of changing a text into various segments called "tokens." In fact, this process is initiated by removing the capitals, punctuation, brackets, and other associated redundancies. Each token is usually a word, without any mark after it, or any capital letter. Finally, each word is defined as a string of alphanumeric characters located in white space.
- *Lemmatization:* Lemmatization is one of the significant preprocessing steps, which is used in NLP tasks. It is somehow similar to word-stemming; however, instead of generating a stem of the word, it replaces the suffix of the word with a different one for producing a normalized form of the word.
- *Parts of speech tagging:* It is the process of labeling the words with specific tags that represent the words' syntactic role. For example, adjectives, nouns, and adverbs are parts of speech. This process tags a word with its part of speech.
- *Dependency parsing:* It is the process of extracting a dependency parse from sentences that indicates their grammatical structure, as well as the relationships between "head" words and words modifying the heads. For example, dependency parser can distinguish which word is the object or subject of a sentence for the user, or it can modify the relationships and tell which words in that sentence are describing the subject.

Emotion Mining

Emotion mining, a similar method like sentiment analysis, detects, analyzes, and performs evaluations on humans' feelings towards different issues, topics, and scenarios. A specific direction of emotion mining includes text emotion mining, which refers to the examination of people's emotions based on their writing observations. Several steps are needed to be followed to perform emotion mining. First, a corpus (collection of texts; for example, one comment can be considered as a corpus; similarly all comments compiled for a video id can also be a considered as a corpus based on the study focus) of messages or comments or interactions that convey at least one of the following emotions: joy, anger, sadness, fear, love, surprise, guilt, disgust, and thankfulness is collected and cleansed. Then, several lexical and learning-based methods are proposed to categorize the emotion of test tweets and examine the effect of dimension reduction techniques, different feature sets, different learning algorithms, and configurations, in addition to addressing the problem of the input data sparsity. The results of the experiment show that a set of

Naive Bayes classifiers, each corresponding to one emotion, is the best-performing method for the task when using unigrams as features.

By manual annotation through Amazon's Mechanical Turk service, Mohammad and Turney *(26)* compiled a large English term–emotion association lexicon, named EmoLex. The lexicon focused on the emotions of joy, anger, sadness, fear, disgust, trust, surprise, and anticipation and has been argued by many to be the prototypical and basic emotions *(27-29).* Another challenge of the lexicon is how to handle the negation of emotions. For example, typically not sad does not mean happy, whereas not happy can often mean sad. However, humans are capable of expressing and distinguishing a few hundred different emotions, such as remorse, guilt, optimism, and enthusiasm (not just six). As shown through experiments, distinguishing these fine-grained emotions are beneficial to applications such as personality detection.

DATA DESCRIPTION

Data Collection

To collect the 'bicycle hitting pedestrian' related videos in YouTube, a detailed list of keywords was developed by using the following terms: "*walking biking collision*," "*biker hits ped*," "*bicyclist hit pedestrian*," "*pedestrian bike crash or incident or accident*," "*pedestrian bicyclist crash or incident or accident.*" The researchers automated the data collection (extract the video information as well as related comments) process by using an open-source R software package called "tuber" *(30).* Another online YouTube comment scrapper has also been used *(31).* The research team used several open-source R software packages to perform the analysis *(32-35).* The flowchart of data collection and analysis is shown in Figure 1.

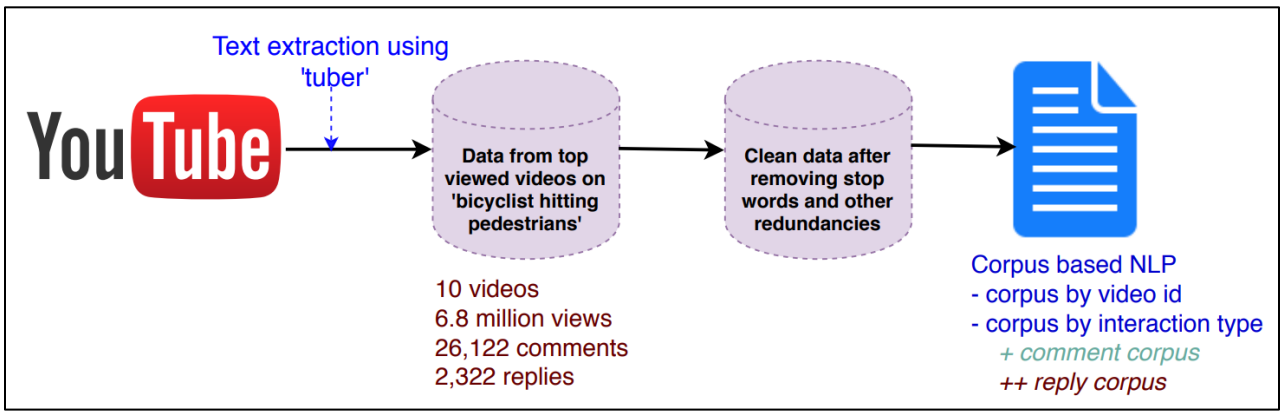


Figure 1 Flowchart of data collection and analysis.

Table 1 provides descriptive statistics of the top ten most viewed videos. After removing redundant and non-English comments, the final number of comments was 26,122. Among the top ten videos, one video was released earlier (in 2010). The overall number of views for all videos was 6,799,938 (mean: 679,994, standard deviation: 1,098,276). The number of likes on all videos was higher than dislikes (55,482 vs. 7,670). The number of comments were 26,122 (mean: 2,612; standard deviation: 4,901). In these videos, participants also replied to the comments. The replies to the comments are also collected and analyzed in this study. The corpora have around 2,000 replies, based on all replies.

Table 1 Clusters of Top-viewed 10 YouTube Videos on 'Bicyclists Hitting Pedestrians'

Video Id	Title	Publish Date	Duration (min.)	Views	Likes	Dislikes	Comments
zR4Okh23Zlo	Cyclist hits pedestrian	11-Aug-13	2:06	3,099,255	31,000	2,700	14,737
sYWPHHo0fPU	Pedestrian gets hit, cyclists talk, cops	26-Oct-16	6:33	2,255,886	12,000	4,300	7,939
Wq6rpVMcyas	Cyclist crashes into man full video	12-Nov-17	0:59	748,883	8,300	187	1,424
G4K8AjNIVPA	Angry pedestrian blocks cyclist as he races through zebra crossing	28-Sep-16	0:27	401,683	2,500	146	1,429
0Lm9TPym9A4	Man gets hit by bike	14-Aug-10	0:37	250,312	1,500	288	371
5Qurlf05YYI	Pedestrian and bicycle accident on Venice Beach	1-Apr-13	0:39	20,050	58	15	77
dnkErN9N8KY	Pedestrian hit by bicycle in San Francisco	15-Mar-17	1:44	9,989	51	7	41
uRoU826ywjw	Cyclist hits pedestrian	18-May-15	0:44	6,138	25	23	55
dXpmxmFW164	Accident on the bicycle lane	6-Apr-15	1:50	4,585	26	6`	23
s-PuD8fSI-I	Pedestrian almost hit by cyclist	15-Jul-14	0:34	3,157	22	4	26

METHODOLOGY

Term Frequency-Inverse Document Frequency

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 (*33*). One comment for video id can be considered as a document. Compilation of comments based on video id or any other specific clusters can be considered as a corpus. It considers the database size and term distribution in the database into account. It decreases the weight for commonly used words and increases the weight for words that are not used a lot in a collection of documents. The IDF for any given term is defined as (33):

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

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 term frequency to calculate a term's $TF - IDF$ (the two quantities multiplied together, $TF \times IDF$). This parameter is usually used to identify the important words within the content of each document. It does so by decreasing the weight of commonly used words and increasing the weight of words that are not used very much in a collection or corpus of

the document. Calculating $TF - IDF$ attempts to find the words that are important in a text, but not too common in all documents. The final parameter weight, w_i, for $TF - IDF$ can be written as *(33):*

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

Where,
f_i= frequency of the word i in the document.

Table 2 shows the TF-IDF values for the two categories based on interaction types. All comments or replies for each of these videos are combined by video ids for determining TF-IDF measures. As unigrams are not suitable in explaining the intent of the topics, bigrams are considered in this analysis. A threshold of 200 counts is considered as the baseline for comment corpora. For the reply corpora, this threshold was 20. The majority of the bigrams overlap in both categories. Intersection, signal phases, bike lanes, and lighting conditions are the most common bigrams in both categories. The bigrams 'the crosswalk' and 'parents fault' are present in the comment category analysis. In the reply categories, two unique bigrams are 'walk on' and 'walk in.'

Table 2 TF-IDF of the Top Bigrams from Comment and Reply Corpora

VID	Bigram	TF	IDF	TF-IDF
Comments				
sYWPHHo0fPU	the light	0.00331	1.20397	0.00398
zR4Okh23Zlo	bike lane	0.00567	0.69315	0.00393
sYWPHHo0fPU	the intersection	0.00197	1.60944	0.00317
sYWPHHo0fPU	yellow light	0.00124	2.30259	0.00286
sYWPHHo0fPU	red light	0.00232	1.20397	0.00279
sYWPHHo0fPU	light was	0.00152	1.60944	0.00245
zR4Okh23Zlo	bike path	0.00218	0.91629	0.002
zR4Okh23Zlo	parents fault	0.0007	2.30259	0.00162
sYWPHHo0fPU	the crosswalk	0.00146	0.91629	0.00134
sYWPHHo0fPU	slow down	0.00177	0.69315	0.00123
Replies				
sYWPHHo0fPU	the light	0.004216	1.504077	0.006342
zR4Okh23Zlo	walk on	0.001681	2.197225	0.003694
zR4Okh23Zlo	bike path	0.001639	2.197225	0.003602
sYWPHHo0fPU	the intersection	0.00233	1.504077	0.003505
sYWPHHo0fPU	light was	0.001498	2.197225	0.003291
sYWPHHo0fPU	was red	0.001387	2.197225	0.003047
zR4Okh23Zlo	cycle lane	0.001303	2.197225	0.002863
sYWPHHo0fPU	red light	0.002441	1.098612	0.002682
zR4Okh23Zlo	walk in	0.001177	2.197225	0.002586
sYWPHHo0fPU	slow down	0.002164	1.098612	0.002377

Sentiment Analysis

Mining on subjective texts containing opinion or sentiment can contribute to understanding perception towards a product. In other words, the objective of sentiment analysis is to determine which words or sentences express opinions, feelings, and sentiments. The sentiment score can be

easily calculated by using the number of positive words or sentences minus the number of negative words or sentences. The research team used 'udpipe' inbuilt functions to determine the sentiment scores *(34).* Boxplot boxes (shown in Figure 2) indicate the 25th percentile, median, and 75th percentile. Boxplot whiskers indicate the 5th percentile and the 95th percentile. The individual sentiment scores are overlaid on the boxplot as the dot-plot format. The values show that the median of the majority of the video comment groups is below zero, which indicates the nature of higher negative sentiments in these videos.

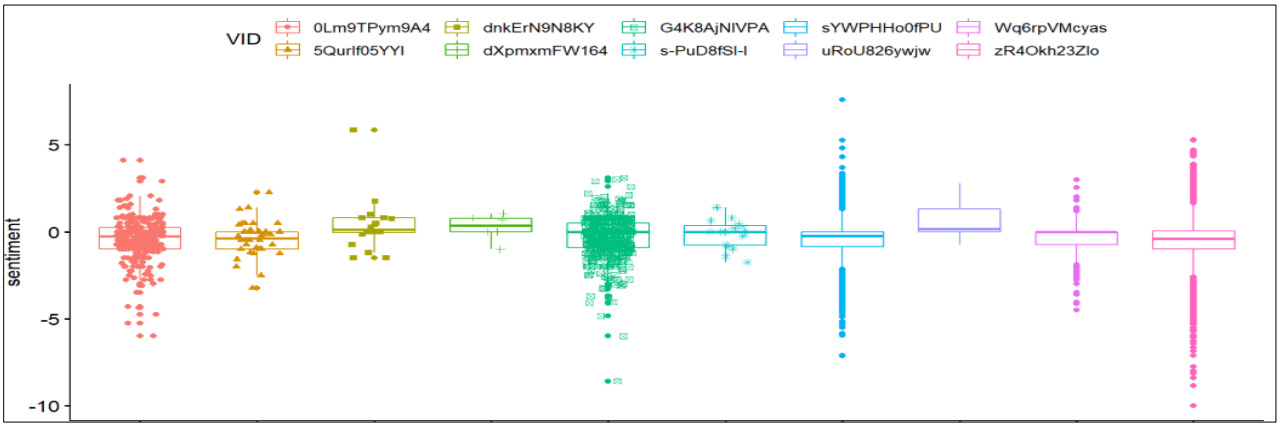


Figure 2 Boxplot of individual sentiment scores by video id

The descriptive statistics of the sentiment scores by the video ids are shown in Table 3. Each video id is listed with the maximum, minimum, mean, and standard deviation of each comment and reply. The video with the highest comment average is dnkErN9N8KY (Pedestrian hit by bicycle in San Francisco) with 0.39. It also has the highest maximum, minimum, and standard deviation.

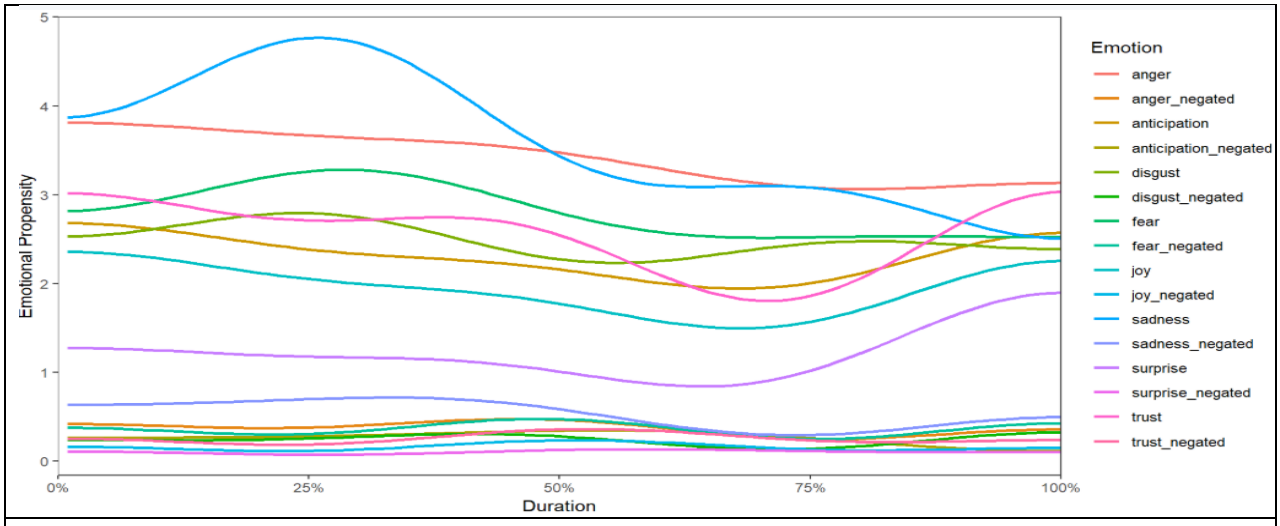
Table 3 Descriptive Statistics of Sentiment Scores by Videos

VID	Max		Min		Mean		STD	
	Comment	Reply	Comment	Reply	Comment	Reply	Comment	Reply
0Lm9TPym9A4	4.10	0.80	-6.00	-2.65	-0.43	-0.71	1.22	0.88
5Qurlf05YYI	2.25	2.80	-3.25	-0.75	-0.41	0.75	1.01	1.20
dnkErN9N8KY	5.85	5.85	-1.50	-2.50	0.39	-0.03	1.53	1.72
dXpmxmFW164	1.05	0.80	-1.00	-1.00	0.26	-0.07	0.75	0.90
G4K8AjNIVPA	3.10	3.10	-8.60	-4.85	-0.28	-0.30	1.13	1.22
s-PuD8fSI-I	1.40	1.40	-1.75	-1.40	-0.13	-0.04	0.84	0.89
sYWPHHo0fPU	7.60	--	-7.15	--	-0.37	--	1.06	--
uRoU826ywjw	2.80	4.80	-0.75	-7.15	0.68	-0.32	1.15	1.27
Wq6rpVMcyas	3.00	1.00	-4.50	-4.10	-0.33	-0.38	0.85	0.84
zR4Okh23Zlo	5.30	5.30	-10.00	-6.45	-0.49	-0.42	1.15	1.32

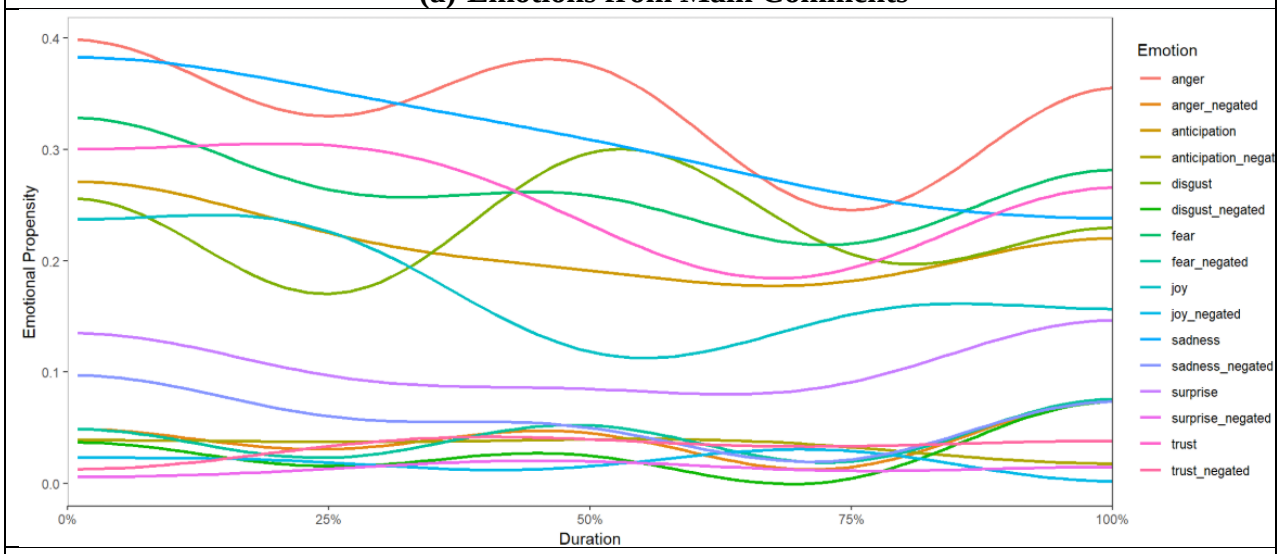
Emotion Mining

For the emotion mining tasks, this study considers eight major emotion types and their negations. The trends of the emotions are shown at the sentence level (shown in Figure 3). This method uses sentiment lexicons to find emotion words and then compute the emotion propensity per sentence

(34). The x-axis indicates the number of documents in percentage form. For example, if the analysis is conducted on 100 documents, 25 percent will indicate the 25th document, and if the vertical line is drawn on 25%, the intersecting points will be the emotion propensity score for that particular sentence. This visualization helps in understanding the overall trends of the emotions distributed by the participants. The general finding is that the negated terms are less in propensity scores than the main emotion-related words. Sadness and anger are the top two emotions in the 'comment' category. For the 'reply' category, anger shows the highest propensity. Sadness shows a declining trend over the duration of the sentences.



(a) Emotions from Main Comments



(b) Emotions from Replies of Comments

Figure 3 Individual tf-idf for texts categorized by content type.

Valence Shift Word Graphs

In their study, Dodds and Danforth *(36)* provided the importance of 'Valence Shift Word Graph.' Consider two texts T_{ref} (for reference) and T_{comp} (for comparison) with sentiment scores $s_{mean}^{(ref)}$ and $s_{mean}^{(comp)}$. Comparison of T_{comp} relative to T_{ref} can be expressed as *(37):*

$$s_{mean}^{(comp)} - s_{mean}^{(ref)} = \sum_{i=1}^{N} s_{mean}(w_i)\left[p_i^{(comp)} - p_i^{(ref)}\right] = \sum_{i=1}^{N}\left[s_{mean}(w_i) - s_{mean}^{(ref)}\right]\left[p_i^{(comp)} - p_i^{(ref)}\right] \quad (3)$$

where,

p_i= the $i-th$ distinct word's normalized frequency of occurrence and which we interpret as a probability, and

$$\sum_{i=1}^{N} s_{mean}^{(ref)}\left[p_i^{(comp)} - p_i^{(ref)}\right] = s_{mean}^{(ref)}\sum_{i=1}^{N}\left[p_i^{(comp)} - p_i^{(ref)}\right] = s_{mean}^{(ref)}(1-1) = 0, \quad (4)$$

w_i represents the word i in comparison text, p_i represents the percentage of word i in comparison text.

By introducing the term $-s_{mean}^{(ref)}$, the contribution of the ith word to the difference $s_{mean}^{(comp)} - s_{mean}^{(ref)}$ can be clear. Two major aspects in determining the sign of the ith word's contribution to the sentiment score are considered in Dodds et al. *(37)* study:

- Whether or not the ith word is, on average, more positive than text T_{ref}'s average, $s_{mean}^{(ref)}$.
- Whether or not the ith word is relatively more abundant in text T_{comp} than in text T_{ref}.

A word's sentiment is signified relative to text T_{ref} by $+$ (positive sentiment) and $-$ (negative sentiment), and its relative abundance in text T_{comp} versus text T_{ref} with $\uparrow$ (more prevalent) and $\downarrow$ (less prevalent). Combining these two binary possibilities leads to four cases *(37):*

- $+\uparrow$: Increased usage of relatively positive words– If a word is happier than text T_{ref} $(+)$ and appears relatively more often in text T_{comp} $(\uparrow)$, then the contribution to the difference $s_{mean}^{(comp)} - s_{mean}^{(ref)}$ is positive.
- $-\downarrow$: Decreased usage of relatively negative words– If a word is less happy than text $T_{ref}(-)$ and appears relatively less often in text T_{comp} $(\downarrow)$, then the contribution to the difference $s_{mean}^{(comp)} - s_{mean}^{(ref)}$ is also positive.
- $+\downarrow$: Decreased usage of relatively positive words– If a word is happier than text $T_{ref}(+)$ and appears relatively less often in text T_{comp} $(\downarrow)$, then the contribution to the difference $s_{mean}^{(comp)} - s_{mean}^{(ref)}$ is negative.
- $-\uparrow$: Increased usage of relatively negative words– If a word is less happy than text $T_{ref}(-)$ and appears relatively more often in text T_{comp} $(\uparrow)$, then the contribution to the difference $s_{mean}^{(comp)} - s_{mean}^{(ref)}$ is also negative.

The normalization of Equation 2 and conversion to percentages become *(37):*

$$\delta s_{mean,i} = \frac{100}{\left|s_{mean}^{(comp)} - s_{mean}^{(ref)}\right|}\underbrace{\left[s_{mean}(w_i) - s_{mean}^{(ref)}\right]}_{+/-}\underbrace{\left[p_i^{(comp)} - p_i^{(ref)}\right]}_{\uparrow/\downarrow}, \quad (5)$$

Where $\sum_i \delta s_{mean,i} = \pm 100$, depending on the sign of the difference in sentiment between the two texts, $s_{mean}^{(comp)} - s_{mean}^{(ref)}$, and the terms to which the symbols $+/-$ and $\uparrow/\downarrow$ apply have been indicated. The $\delta s_{mean,i}$ is called the per word sentiment shift of the ith word. Figure 4 can be interpreted by the following interpretations:

- Words on the right contribute to an increase in positive emotions in the corpus
- A yellow bar in the right with a down arrow indicates that a negative emotion was used less
- A purple bar in the right with an up arrow indicates that positive emotion was used more
- Words on the left contribute to a decrease in position emotions in the corpus
- A yellow bar in the left with an up arrow indicates that a negative emotion was used more
- A purple in the left with a down arrow indicates that positive emotion was used less

The word shift plots are not significantly different between the corpora (plural of corpus) developed for comments and replies. However, the degree of negative emotions is less used in the replies. Some of the terms, such as bike, are considered as positive emotion due to the reason of using conventional sentiment lexicons. There is a need for the development of transportation safety-related senti-lexicon to precisely capture the domain-specific sentiments and emotions, which is currently out of the scope of the present study.

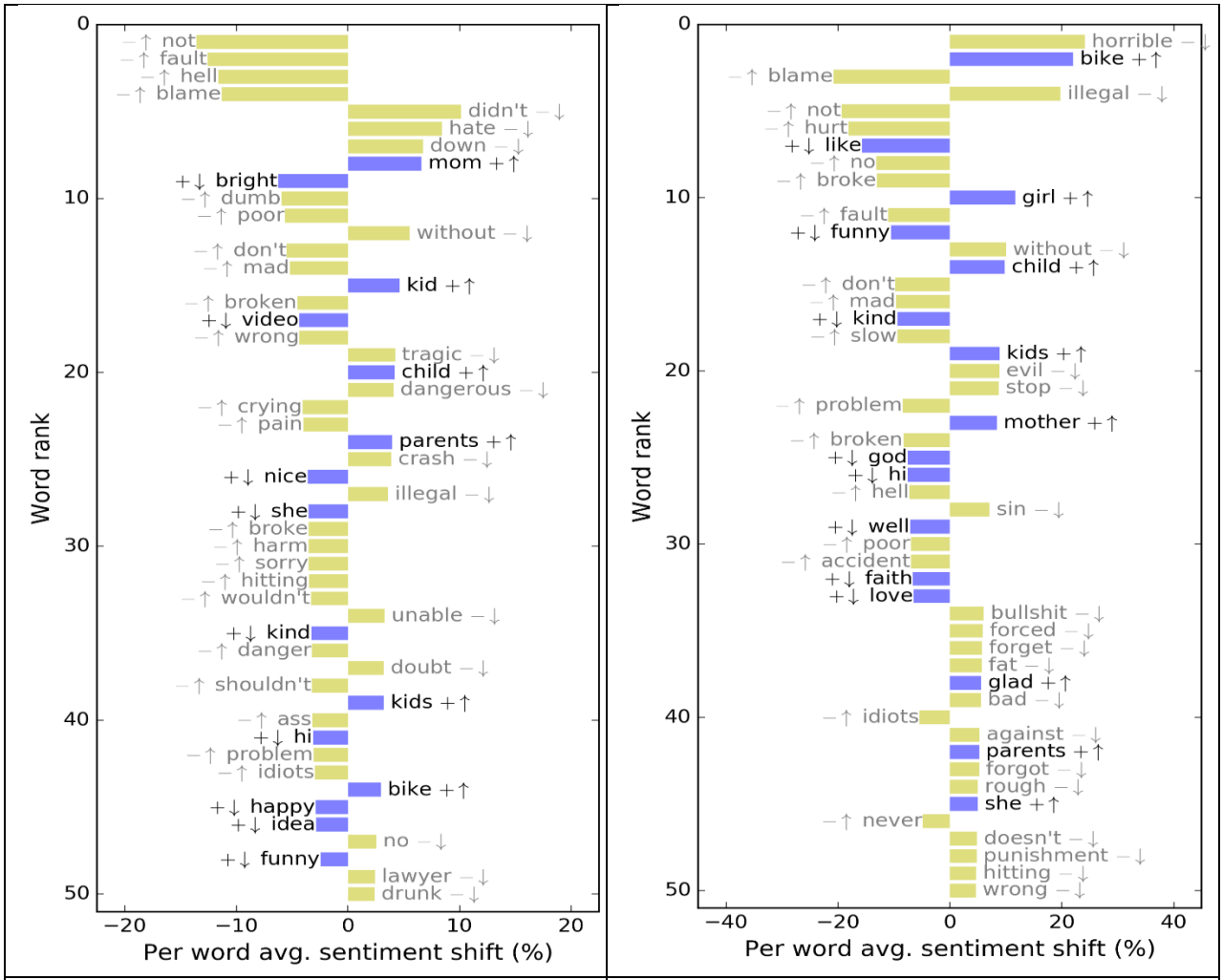
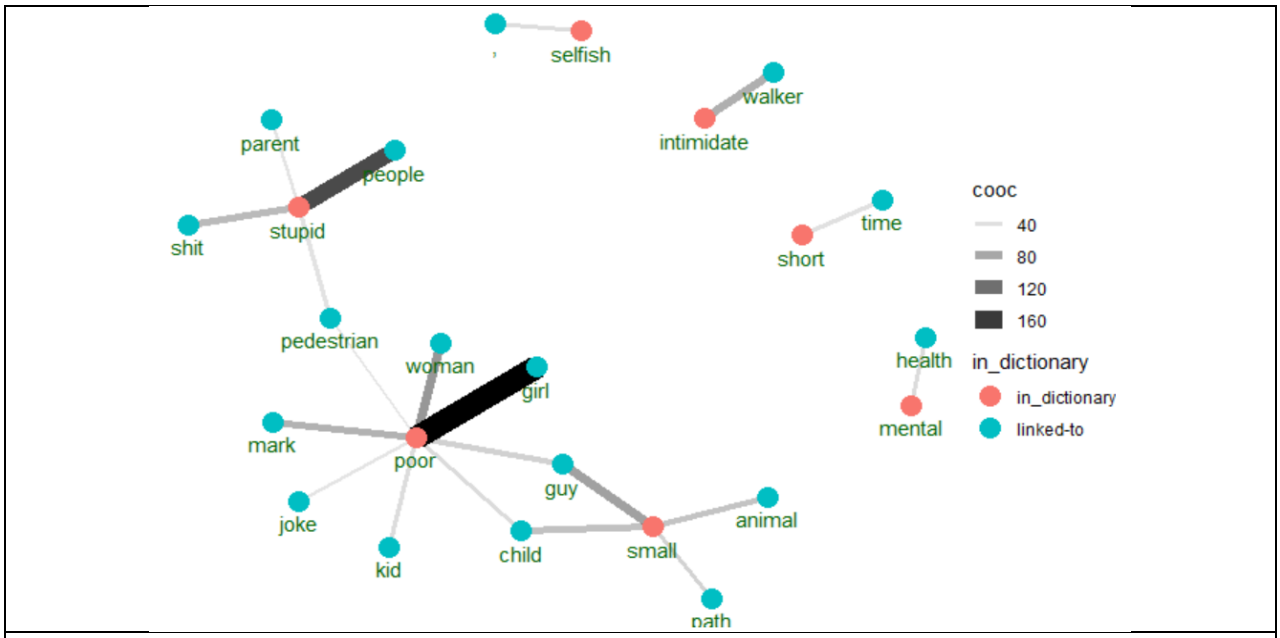


Figure 4 Valence shift word graphs based on comments and replies.

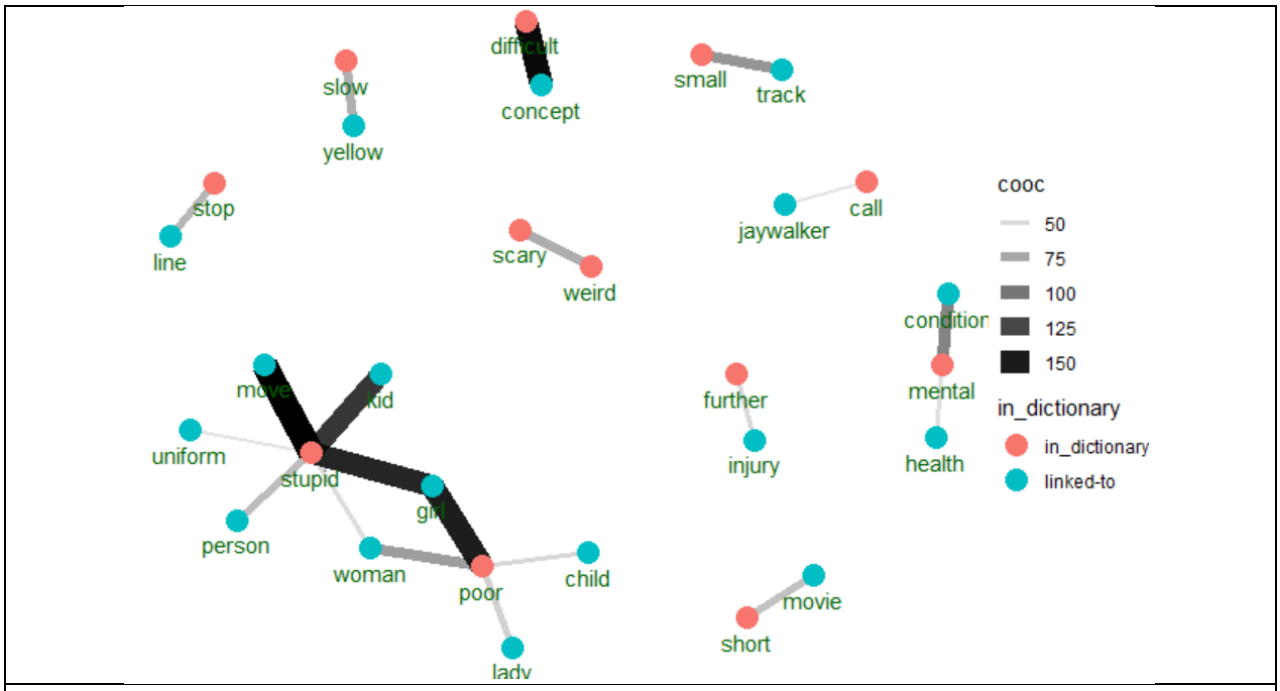
Co-occurrence of Negative Terms

The majority of the sentiment analysis and emotion mining studies perform only n-gram related studies to determine the sentiments and emotions over the corpus, document, or sentence level.

One area that is less explored is the investigation on determining the relation of other words with the negative sentiments and emotions. This approach will help answer what is causing a negative sentiment or negative emotion. The Stanford Dependencies (SD) representation *(38)* was originally developed as a practical representation of English syntax, aimed at natural language understanding (NLU) applications. This is deeply associated with grammatical relation-based syntactic concepts. This study used the dependency relationship output of 'udpipe' to find out which words are linked to negative words from the 'udpipe' sentiment dictionary *(34)*. Out of several parameters, this study used mainly the parameters associated with adjectives that modify a noun. Before conducting the dependency parsing, this study used the conventional NLP annotation (tokenization, lemmatization, parts of speech tagging). The lemma values of the negative words and the lemma values of the parent word are used to calculate the co-occurrence. The words' co-occurrence relationships for the datasets of comments and replies are shown in Figure 5.



(a) Words links with negative terms (Dataset of Comments)



(b) Words links with negative terms (Dataset of Replies)

Figure 5 Cooccurences of the negative terms.

CONCLUSIONS

The understanding of the public perception of 'bicyclist versus pedestrian collision' is important. Due to the lack of adequate and quick information collection and retrieval, conventional surveys are not sufficient in comprehending the tipping point of consumer perception. By performing an analysis of responses to pedestrians and cyclist's collision videos on YouTube, this study advances the general understanding of the communication between the bicyclists and pedestrians. The study concentrates on the content exploration of the YouTube videos by evaluating the behaviors of the end-users regarding disliking, liking, and commenting patterns.

The findings of this study include:

- A large number of comments, views, likes, and dislikes can indicate that the public is partaking in this debate.
- The tf-idf algorithm identifies multiple rare but significant words at various categorization levels.
- The valence shift word graphs provide a synopsis of the contexts of the words and the shifting of emotions in the replies and comments.
- The co-occurrence plots show the logic behind the negative emotions generated in the replies and comments.

The findings from this study can benefit the development of better communication between bicyclists and pedestrians. By providing a better understanding of people's attitudes towards the interactions between bicyclists and pedestrians, this research can be beneficial in advising practitioners as well as researchers. Additionally, this study developed a framework to determine consumer's adaptation needs that can be replicated for other transportation-related topics. There is a debate regarding a disadvantage to social media mining when considering a sample that is biased and not representative in the long run. Because social media posts are generated at a very high

frequency, they can produce a big textual data with a representative sample of holding honest sentiments of consumers – rarely can be captured in other conventional survey studies.

To build more successful organizations, better construct public policy, and more fully understand economic and social phenomena from a scientific perspective, the significance of quantifying the intensity and nature of emotional states at the population level is important. Although this study contributed positively to research, this current study has several limitations. One of the critical limitations is the small sample size. This study only evaluated the top ten viewed videos and related comments. Future studies should expand data or collect data from other video platforms like Facebook or Vimeo to understand the interactions between pedestrians and bicyclists.

ACKNOWLEDGMENT

The authors appreciate the assistance provided by the students on this manuscript preparation: Bita Maraghehpour and Ly-Na Tran.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm the contribution to the paper as follows: study conception and design: Subasish Das; data collection: Subasish Das; analysis and interpretation of results: Subasish Das draft manuscript preparation: Subasish Das, Xiaoqiang Kong, Ruihong Wang, and Ahmadreza Mahmoudzadeh. All authors reviewed the results and approved the final version of the manuscript.

REFERENCE

1. Tuckel, P., W. Milczarski, and R. Maisel. Pedestrian Injuries Due to Collisions with Bicycles in New York and California. Journal of safety research, Vol. 51, 2014, pp. 7–13.
2. IsolateCyclist. Cyclists Versus Pedestrians, 2012.
3. 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.
4. Goodman, J., and D. Spokes. The Cyclist-Pedestrian Wars. The New York Times, 2010.
5. Boufous, S., J. Hatfield, and R. Grzebieta. The Impact of Environmental Factors on Cycling Speed on Shared Paths. Accident Analysis & Prevention, Vol. 110, 2018, pp. 171–176.
6. Van der Horst, A. R. A., M. de Goede, S. de Hair-Buijssen, and R. Methorst. Traffic Conflicts on Bicycle Paths: A Systematic Observation of Behaviour from Video. Accident Analysis & Prevention, Vol. 62, 2014, pp. 358–368.
7. Salmon, F. A Unified Theory of New York Biking. REUTERS, 2010.
8. Kavanagh, G. Unraveling Ped & Bike Tension in Santa Monica, CA, 2012. https://la.streetsblog.org/2012/09/10/unraveling-ped-bike-tension-in-santa-monica/ Accessed: July 2019.
9. Werneke, J., M. Dozza, and M. Karlsson. Safety–critical Events in Everyday Cycling–Interviews with Bicyclists and Video Annotation of Safety–critical Events in a Naturalistic Cycling Study. Transportation Research Part F: Traffic Psychology and Behaviour, Vol. 35, 2015, pp. 199–212.

10. 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, 2014. 56(6): 1189–1203.
11. Mehrotra, S., and S. Roberts. Identification and Validation of Themes from Vehicle Owner Complaints and Fatality Reports using Text Analysis. The 97th TRB Annual Meeting Compendium Papers, Washington D.C., 2018.
12. Das, S., X. Sun, M. Zupancich, and A. Dutta. Twitter in Circulating Transportation Information: A Case Study on Two Cities. The 96th TRB Annual Meeting Compendium Papers, Washington D.C., 2017
13. 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, 2016. 67: 321–342.
14. Das, S., L. Minjares-Kyle, K. Dixon, A. Palanisamy, and A. Dutta. #TRBAM: Understanding Communication Patterns and Research Trends by Twitter Mining. The 97th TRB Annual Meeting Compendium Papers, Washington D.C., 2018.
15. Das, S., G. Medina, L. Minjares-Kyle, and Z. Elgart. Social Media Hashtags associated with Bike Commuting: Applying Natural Language Processing Tools. The 97th TRB Annual Meeting Compendium Papers, Washington D.C., 2018.
16. Das, S., X. Sun, and A. Dutta. Investigating User Ridership Sentiments for Bike Sharing Programs. Journal of Transportation Technologies, 2015. 5(2): 69–75.
17. 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.
18. 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. Volume: https://doi.org/10.3141/2552-07.
19. 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. Volume: https://doi.org/10.3141/2614-04.
20. 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. Volume: https://doi.org/10.3141/2614-01
21. Sun, L., and Y. Yin. Discovering themes and trends in transportation research using topic modeling. Transportation Research Part C: Emerging Technologies, 2017. 77: 49–66.
22. Das, S., and A. Dutta. Knowledge Extraction from Transportation Research Thesaurus. The 97th TRB Annual Meeting Compendium Papers, Washington D.C., 2018.
23. 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_VisibilityWeather_FinalReport.pdf. Accessed July 27, 2018.
24. Brown, D. E. Text Mining the Contributors to Rail Accidents. IEEE Transactions on Intelligent Transportation Systems, 2016. 17(2): 346–355.

25. Sojka, P., A. Horak, I. Kopecek, and K. Pala (Eds). Text, Speech, and Dialogue. 19th Internatuional Conference, TSD 2016, Czech Republic, Spetember, 2016 Procceedings. Lecture Notes in Artificial Intelligence, Springer, Switzerland, 2016.
26. Mohammad, S., and P. Turney. Crowdsourcing a Word-Emotion Association Lexicon. https://arxiv.org/abs/1308.6297. Accessed July 2019.
27. Plutchik, R. A general psychoevolutionary theory of emotion. Emotion: Theory, research, and experience, 1(3), 1980, 3–33.
28. Plutchik, R. On emotion: The chicken-and-egg problem revisited. Motivation and Emotion, 9(2), 1985, 197–200.
29. Plutchik, R. The psychology and biology of emotion. New York: Harper Collins, 1994.
30. Sood, G. tuber: Access YouTube from R. R package version 0.9.7, 2018.
31. Klostermann, P. A web client that scrapes YouTube comments. https://github.com/philbot9/youtube-comment-scraper Accessed: July 2019.
32. Feinerer, I., K. Hornik, and D. Meyer. Text Mining Infrastructure in R. Journal of Statistical Software 25(5), 2008, pp. 1-54.
33. Silge, J. and D. Robinson. Text Mining with R: A Tidy Approach. O'Reilly Media, Inc., 2017.
34. Wijffels, J. udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. R package version 0.8.2. Accessed: July 2019.
35. Rinker, T.sentimentr: Calculate Text Polarity Sentiment version 2.7.1. http://github.com/trinker/sentimentr Accessed: July 2019.
36. Dodds, P, and C. Danforth. Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. Journal of Happiness Studies, 11(4), 2010, pp 441–456.
37. Doss, P., K. Harris, I. Kloumann, C. Bliss, and C. Danforth. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. Plos One. Vol 6(12), 2011.
38. de Marneffe, M., T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, and C. Manning. Universal Stanford dependencies: A cross-linguistic typology. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), 2014.