Pedestrian Collisions with Bicyclist: Emotion Mining Using YouTube Data

Subasish Das, Ph.D. Associate Transportation Researcher, Texas A&M Transportation Institute (s-das@tti.tamu.edu)
Xiaoqiang Kong, Research Assistant, Texas A&M Transportation Institute (X-Kong@tamu.edu)
Ruihong Wang, Research Assistant, Texas A&M Transportation Institute (Ruihong.wang@tamu.edu)
Ahmadreza Mahmoudzadeh, Research Assistant, Texas A&M Transportation Institute (A.Mahmoudzadeh@tamu.edu)

Research conducted by

Texas A&M Transportation Institute — THE TEXAS A&M UNIVERSITY SYSTEM

Problem Statement

- **Pedestrians and cyclists' crashes could lead to serious social crisis. However, this issue is rarely studied.**
- The antagonism between vehicles and bicyclists is well-known. In recent years, a less well-known antagonism between pedestrians and bicyclists has gained more attention through social media like mainstream media websites.
- The topmost viewed YouTube videos on 'pedestrian collision with bicyclist' have 60.9 million views and contain around 25,000 comments.

Data

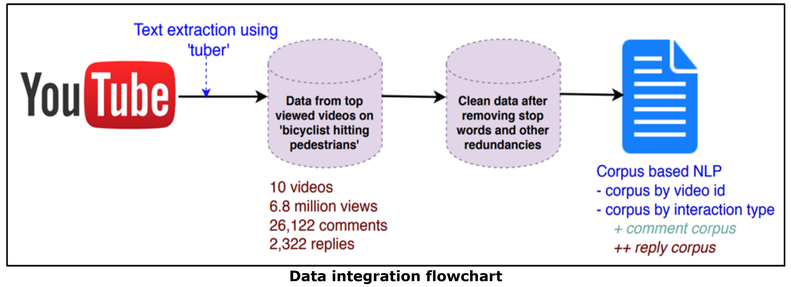
- To collect the 'bicycle hitting pedestrian' related videos in YouTube, the researchers automated the data collection (extract the video information as well as related comments) process by using an open-source R software package called "tuber."
- Another online YouTube comment scrapper has also been used.
- 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."

Methodology

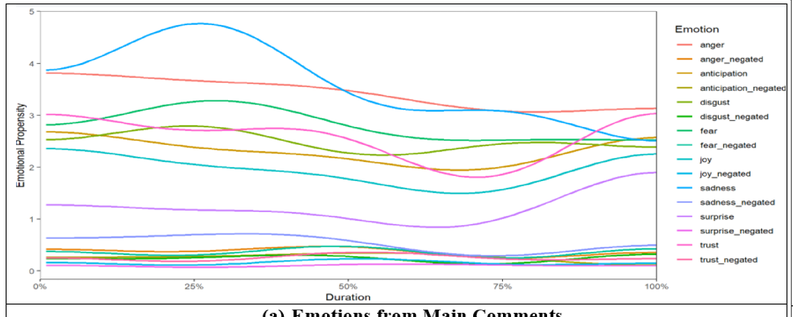
- Performed natural language processing (NLP) tasks— tokenization, lemmatization, parts of speech tagging and dependency parsing
- Performed term frequency-inverse document frequency (tf-idf):

$$IDF(term) = \ln\left(\frac{N}{d_i}\right) ; TF - IDF(w_i) = f_i \times \log\left(\frac{N}{d_i}\right)$$

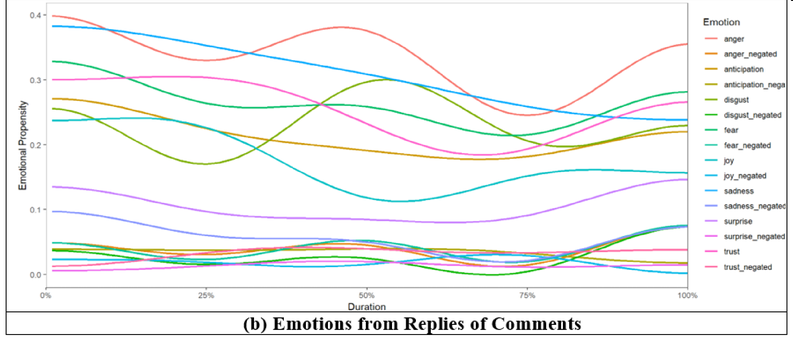
- Conducted sentiment analysis



Data integration flowchart

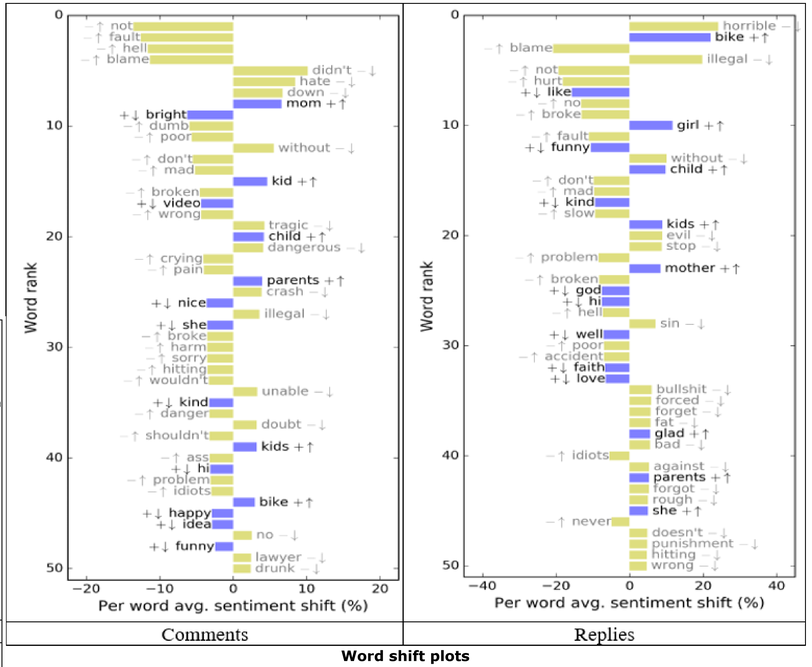


(a) Emotions from Main Comments



(b) Emotions from Replies of Comments

Emotion distribution over time



Word shift plots

Conclusions

- 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 shifting of the emotions in the replies and comments.
- The co-occurrence plots show the logic behind the negative emotions generated in the replies and comments.