2022 Mid-Continent Transportation Research Symposium
Ames, Iowa
September 14–15, 2022
intrans.iastate.edu/events/midcon2022/

Apply Machine Learning (ML) to Crash Narratives in Order to Identify Potentially Misclassified Crashes

sudesh Bhagat, Iowa State University
Jonathan Wood, Iowa State University
Skylar Knickerbocker, Iowa State University
Anuj Sharma, Iowa State University

Abstract

Objective

Crash narratives are reports which are generated for a given crash that describe what happened prior to, during and after the crash which are typically written by the responding law enforcement officer. Analyzing crash data aids in the identification of the cause of crashes as well as crash types with the help of organized data. This, in turn, informs policy making decisions. Accurate analysis, however, requires stringent quality control, which helps eliminate inappropriate coding and disorganized data that leads to misinformation.Â Existing studies show that errors in the crash data coding exist due to a variety of factors that overall impact data quality. The objective of this project is to apply Machine Learning (ML) to crash narratives in order to identify potentially misclassified crashes.

Method

This crash narrative analysis is based on five years of data from 2015 to 2018. Data is downloaded from Iowa Department of Transportation (DOT). Two sets of files are used, one is crash data information which has crash, vehicle and person level data. The second is the crash narrative data file, which contains descriptive information pertaining to the events related to the crash. The first round of analysis will be looking at the distracted crashes whereas the second round will focus on work zone and other types of crashes. As part of the project, all the words from the crash narrative will be extracted to create a list that can be used to classify a narrative as likely to be related to distractive driving or work zones. Crashes with high probability (based on the ML models) of being related to distracted driving or work zones, which were not coded as such, indicate high likelihood of being misclassified and the full narratives can be manually reviewed to determine which of those crashes are actually misclassified.

Expected results

XGboost is expected to improve efficiency in terms of assigning crash coding. It uses binary prediction to develop a model that identifies patterns within a dataset with labels and features. The model then predicts the labels on a new dataset's features. It is anticipated that the results will provide a list of crashes that are likely to be miscoded (which will need to be manually reviewed) as well as a list of words that were the most likely to differentiate between crashes associated with or not associated with distraction or work zones.

The contents of this abstract reflect the views of the author(s), who are responsible for the facts and accuracy of the information presented herein.