Data collection

This study assumed that cannabis users in legalized states were inclined to drive across the state border under the drug's influence and potentially get into fatal accidents in border counties. To validate this assumption, the sentiment of cannabis users with regards to driving under influence would be mined from cannabis forums. Anonymous posts in forums were preferred compared to surveys and news articles as they were likely to give a more accurate depiction of the sentiments among cannabis users. Three cannabis forums that are popular among US cannabis users were chosen and all posts in relevant threads were scrapped based on 'driving stoned' and 'driving high' keyword searches. About 20,000 forum posts were scrapped in the end.

Assumption Validation

To validate the assumption that cannabis users were willing to drive under influence, the sentiment of cannabis users was mined using forum posts scrapped from cannabis forums. Further content filtering was performed on the scrapped posts during the textual pre-processing steps to filter highly relevant posts containing both driving and cannabis usage-related keywords (e.g. 'stoned', 'blazed', 'baked'). 4033 posts were selected for analysis and 403 posts (or 10%) were manually labelled with whether the user was supportive of driving under influence. Proponents of driving under influence claimed that they drive better under influence as they tend to drive slower and were less aggressive. Users against driving under influence were concerned about responsible road usage and safety, but many were just afraid of getting caught by law enforcement. Irrelevant or neutral posts were also labelled as not supportive.

A supervised learning model was built to label the remaining 90% of the posts. 6 machine learning algorithms were tested, including support vector machine, linear discriminant analysis, random forest and boosting, and the best model was selected using 5-fold cross-validation. Boosting using decision trees 'stumps' gave the best out-of-sample accuracy of 95.8%. After predicting the labels for the remaining posts, the results were aggregated with the manually labelled posts and summarized in **Figure 5**. A large majority (> 65%) of posts analysed from the cannabis forums indicated willingness to drive under influence and the assumption used for this study was validated.

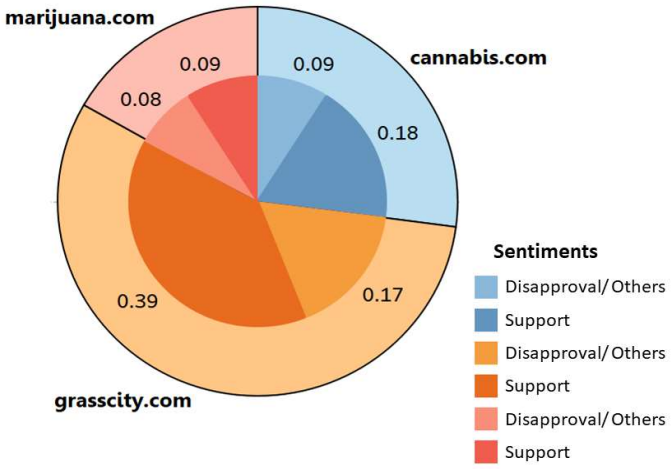

Figure 5: Sentiment of cannabis users regarding driving under influence mined from 3 cannabis forums. The outer pie chart depicted the source of posts analysed. The inner pie chart showed the breakdown between support and disapproval/others in each forum. The darker shade of colour represents support and the lighter shade represents disapproval/others. >65% of the posts indicated support for driving under influence.