Addressing Underage Drinking and Drunk Driving from Tweets

INTRODUCTION

Drinking has been a major source of concern in the World. Everyday there are various negative incidents happening around us as a result of drinking. It often results in abuse in the home, illness from over-drinking, insatiable addiction, and many more severe consequences, sometimes leading to death. Among the many drinking habits, underage drinking and drunk-driving are of major concern. Numerous measures have been taken to prevent these drinking habits, but as of now, such drinking continues to take place, and often the consequences are so severe that some stories even make it to the national news.

One big problem with preventing such drinking habits is that it is very difficult to determine, without the use of expensive equipment, whether a person is drunk. Drinking is ubiquitous and we cannot keep checking everyone around us to see if they are drunk. Further, people have individual rights and checking to see if a person is drunk without his/her approval, or without the permission from public safety department is unethical.

A much easier way to determine whether people are drunk is when they themselves acknowledge the fact. That is where social media comes in. Often, we see people bragging about drinking to others, and social media is a common ground for communicating such information. Numerous people send Tweets from bars and pubs saying that they are having a good time drinking. By looking at their tweets, and
locations from where they are tweeting, we can find out whether they are talking about being drunk at the moment.

Looking at their profile data, such as celebrities they are following, the friends they have, the type of products they are interested in, their tweeting habits and text content in tweets, people wishing them birthday, etc., we will be able to predict their age with good accuracy. Thus, social media makes a good prospect for eliciting who are drinking at the moment, and which of them are below the legal age for drinking.

In the world today, because cell phones and other mobile devices with GPS are very accessible, people are increasingly becoming inclined to checking-in at various locations from their social media. This results in their Tweets getting geo-tagged by the location they are tweeting from. Therefore, after searching Tweets and determining that a user is drunk, we can check to see if their next or previous tweet was sent within a very short time and whether that tweet was from a completely different location. This would be an indication that they were traveling drunk, and maybe driving drunk.

**a. Intellectual Merit**

The proposed research will give rise to plenty of new knowledge and findings in the field of Data Mining and Big Data Analytics. The project offers a broad area for collaboration - researchers from the social sciences, computer science, health informatics, and statistics can all be part of this research area. The topic is interdisciplinary and of national importance, and its applications towards diminishing drunk driving and preventing underage drinking will come at a very low cost. In other words, downloading Tweets and then applying data mining techniques using computers to identify drunk people is a much cheaper process compared to current approaches involving police traveling and investigating whether people are drunk.
b. Broader Impacts

New students will be trained in learning to handle big data, resulting in more advanced degrees such as MS or PhD. New courses will be taught at the department geared towards social media data analysis.

Further, the research, if undertaken, will greatly help improve the society as a whole. Underage drinking and drunk driving are two huge concerns for the society, and by rectifying these issues, the project will help make society safer and healthier. In addition, broader topics such as global drinking patterns, or towards finer granularity such as city or neighborhood based drinking can also be analyzed using the suggested research approach.

RESEARCH PLAN

A large proportion of the budget will be used recurrently used for data collection and labeling through human intelligence, and hiring graduate students to work in the project. Initially data will be collected from Twitter, then keyword filtering will be applied using words related to drinking, such as “drunk”, “wasted”, “shot”, etc, to narrow down from millions of tweets to those about drinking. These tweets will be sent to Amazon Mechanical Turkers for answering questions such as:

- is the tweet about user drinking?
- is the tweet about user drinking at the time of tweeting?
- does the user’s Twitter info suggest the user is underage?

Next, using the labeled tweets, classifiers such as SVMs will be trained to answer the above questions regarding new (unseen) tweets. After the SVMs answer questions regarding new tweets, these tweets and their corresponding SVMs’ answers will be provided to the Turkers again, this time to verify whether the SVM labelings were correct. This will serve as an evaluation procedure, judging the accuracy of the SVM classifiers. If the accuracy is not good, then further data collection, labeling and SVM training will be applied to get better results. After a few repeating of this process, and
analyzing SVM classifications and tuning their parameters, we expect to see good classification performance.

Then we will start visualizing the tweets by geo-tag, which will give rise to several information not visible to the local police. Tweets give us the opportunity to model human behavior as a large scale, and taking advantage of this, we can find out where the tweets are clustering. This can lead us to find underage drinking groups such as freshmen partying in college, teens going into parks to drink, etc. Drunk driving can be determined in real time, and we can develop an app to alert the police of the locations and movements of drunk drivers. We can also address demographics, such as male vs female drinking, drinking on weekdays vs weekends, drinking in high-school or college, partying, etc.