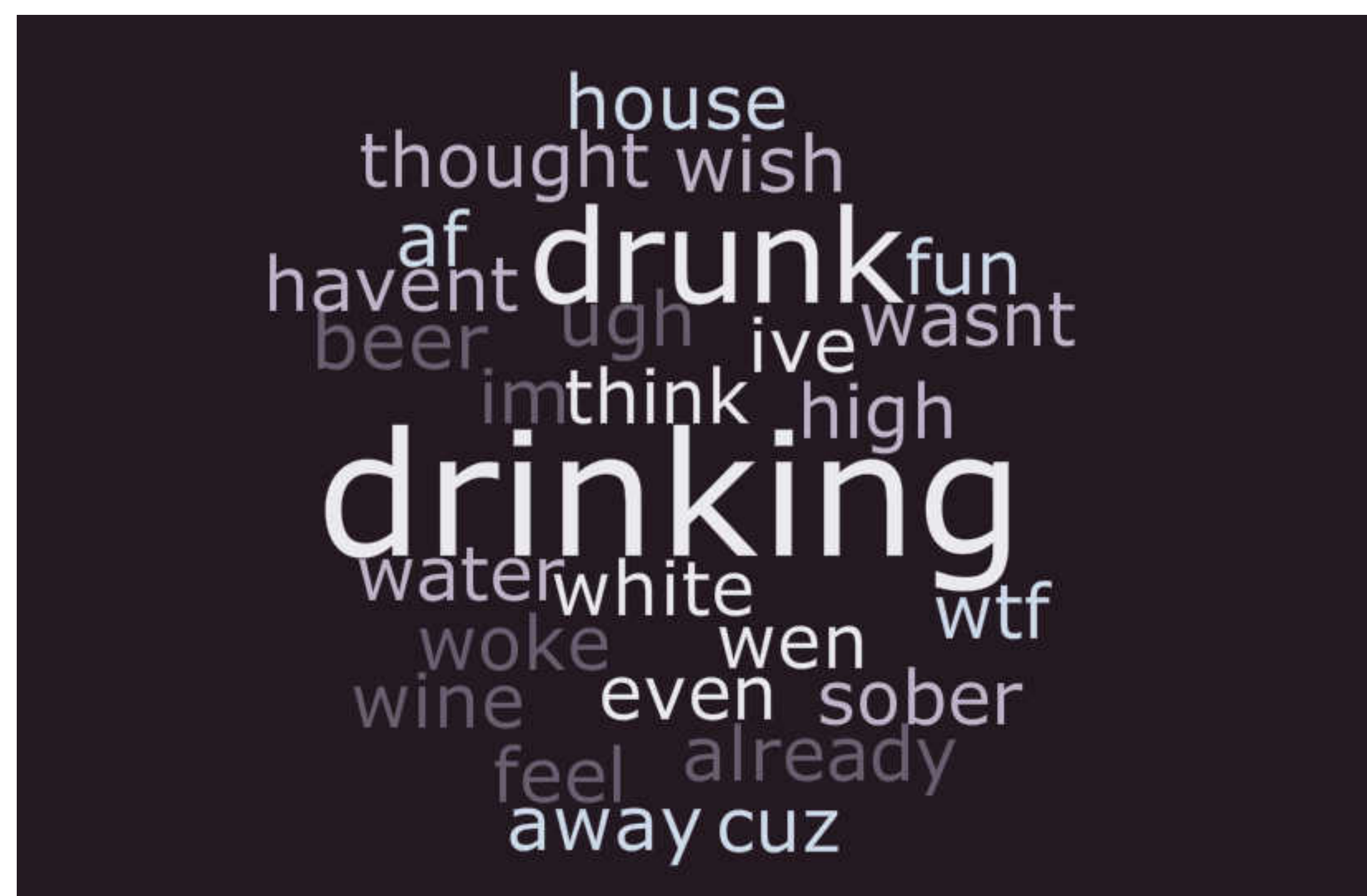


Assessing Alcohol Abuse Statistics Through Data Analysis Of Social Networking Sites

Jesse Gomer, Adam Maus
Center for Health Enhancement System Studies

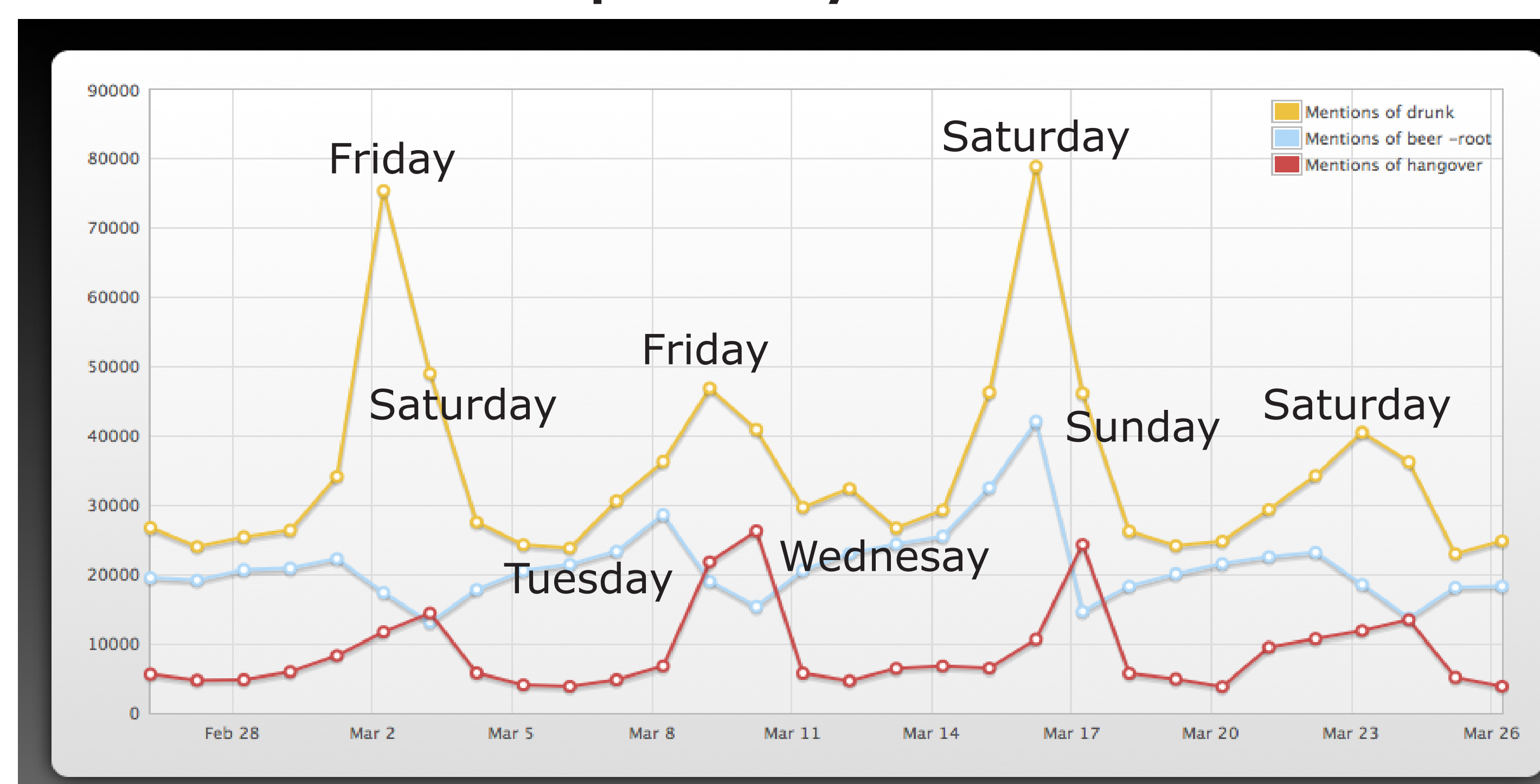
Most Relevent Features



Methods

We used data taken from Twitter via Datashift. Using Python, we filtered the data and made it suitable for Naive-Bayes analysis done by the Natural Language Toolkit and Weka. From the Naive-Bayes model, we were able to identify the most important words.

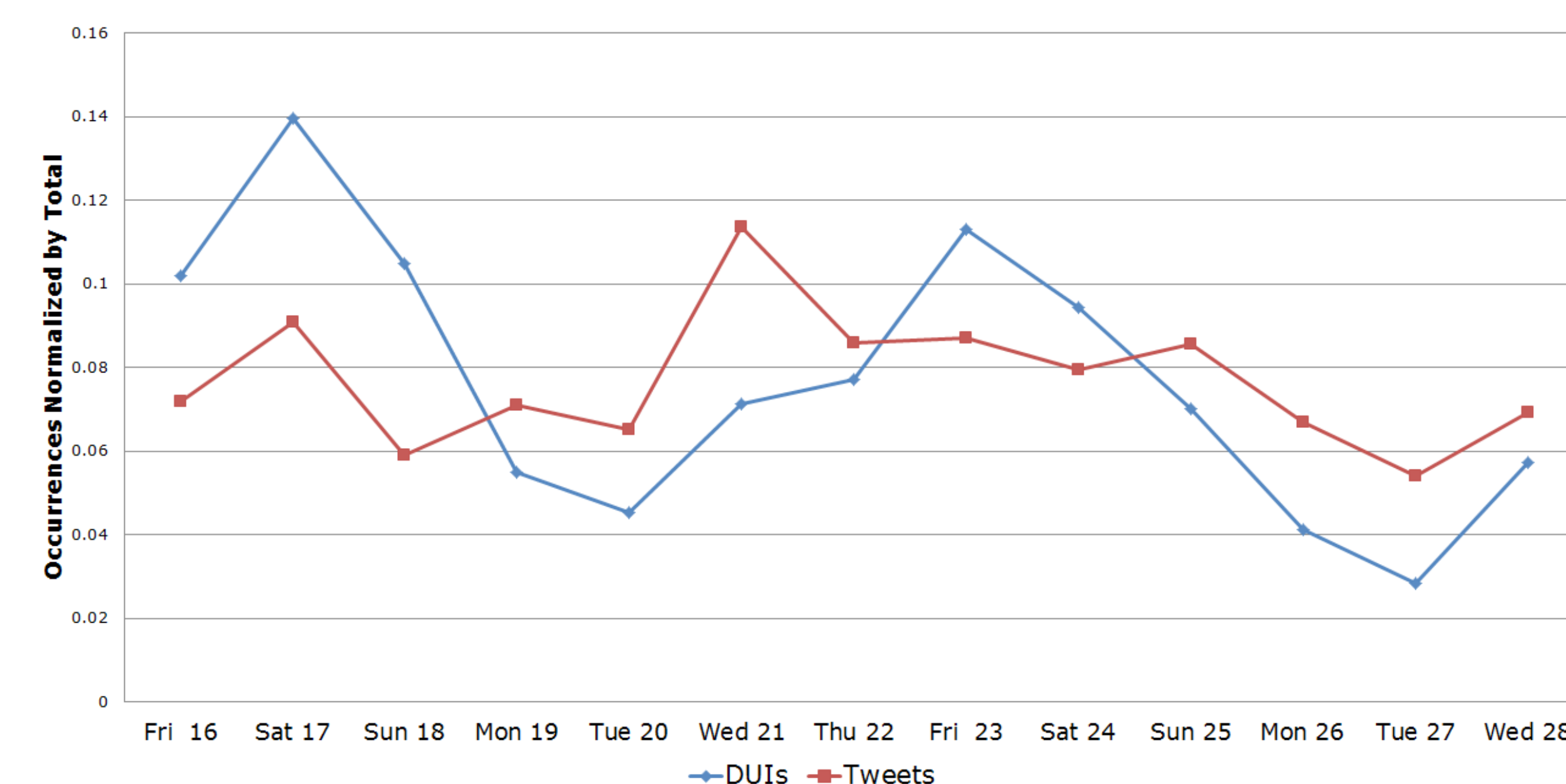
Frequency Chart



Abstract

We are developing a data mining system that uses social media combined with data analysis techniques to investigate how Twitter data, related to alcohol abuse, correlate with regional statistics such as the number of alcohol related car accidents in a specific region. We use publicly available Twitter data alongside natural language processing algorithms to develop a strategy for classifying, aggregating, and testing correlations between Twitter data and statistical data on the number of DUIs in the Los Angeles area. Correlations between social networking and statistical data could be used to identify locations where other statistical data are not accurate.

Comparing Drinking Tweets with DUIs



Results and Discussions

- Alcohol related tweets spike on the weekends
- Only socially acceptable alcohol behavior is posted on twitter
- There is a slight correlation between DUIs and alcohol related tweets
- Twitter is a valuable source of behavior data, but there are these problems:
 1. Lack of freely available Twitter data archives (the only archives are third party services)
 2. A large amount of noise in the data (people sending ambiguous tweets)
 3. Small percentage of alcohol related tweets