Twitter Analysis on Marijuana

Submitted by,

Guru Prasad Natarajan

A20344932

Problem

- We tried to examine the sentiments and themes of marijuana-related tweets.
- Infer the demographics of the users such as the gender and ethnicity.
- Determine the number of pro-marijuana tweets, anti-marijuana and neutral tweets.

Approach

- Manual labelling of tweets into pro, anti and neutral marijuana sentiments.
- Train on the labelled data and perform test on the remaining of the tweets.
- Name lists approach to classify gender based on the first name.
- Used k-neighbor classifier to predict the ethnicity based on the last name.

Data

- Collected roughly around 10,000 tweets over a period of two weeks using certain keywords. Text, description and user name are our fields of interest.
- Name lists were collected based on the year of birth.
- Collection of names from 1890 – 2010.
- Web scrapped data which tags last names with the ethnicities.

Results

- The methodology was able to classify all the tweets accurately based on the labels.
- Accuracy of the model is 78.4%
- It did work well in classifying the tweets as pro, anti and neutral marijuana.
- Most of the tweets did not have a proper user name.
- Gender and ethnicity determination proved to be challenge.

Conclusion

- The number of pro-marijuana tweets ranks first followed by anti and neutral tweets.
- A greater portion support marijuana for medicinal use.
- Males exceed female in tweeting about marijuana.
- Identity of the users are mostly unknown.