Toxic Words

Dovid Burns

4/25/18

Springboard Capstone Two

1. Define the Problem

Toxic comments online are both damaging to people emotionally and prevent productive discussions about sensitive subjects. Many online platforms that allow for user comments such as Facebook, YouTube, Twitter, Wikipedia, Yelp and Instagram have difficulties ensuring that conversations are taking place in an appropriate way. This project will build a classifier using a dataset that contains comments from Wikipedia's talk page edits to classify them as toxic or benign. The model can then be used in many platforms to automatically detect—and possibly remove—these comments before they offend users or deter users from engaging in communications. Additionally, this can be used as a tool to find problematic users for serious warnings or banning from the site.

2. Identify the Client

*The Conversation AI team, a research initiative founded by Jigsaw and Google (both a part of Alphabet) are working on tools to help improve online conversation. One area of focus is the study of negative online behaviors, like toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). So far they've built a range of publicly available models served through the Perspective API, including toxicity. But the current models still make errors. This project will seek to correct these errors to create a stronger model. (***Need to change modify, taken from Kaggle competition page.** *)*

3. Describe the Data Set

This data set contains 159,571 comments of which more than ten percent are tagged toxic. The data set to be used is publicly hosted at https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge . Each comment has between six and 5000 characters.

4. Other Potential Data

- Comments from other types of website. Youtube, facebook, quora
- If we had the time frame from when the original post was made until when the comment was made… this variable could be strong in terms of predictable power.
- Being able to look at multiple comments by the same user might increase predictability, for example previous toxic comments will probably indicate future toxic comments.
- How many words were spelled incorrect—another possible feature.

5. Initial Findings

Disclaimer: the dataset contains text that may be considered profane, vulgar, or offensive.

After finding length of each comment and the percent of characters that are uppercase, analyzed these features. Found that by binning them there were vast difference in the length and percent uppercase to the percent of toxic comments. Did Chi-squared tests and found a p-value of zero for both variables thus these are statically sound. Found the top predicting words with their percent importance were:

fuck 0.042238

shit 0.024147

fucking 0.024142

bitch 0.019771

suck 0.014190

ass 0.013264

stupid 0.011506

asshole 0.010994

faggot 0.009861

idiot 0.009759

dick 0.007772

gay 0.007644

cunt 0.006265

hell 0.005792

cock 0.005212

article 0.004814

bastard 0.004449

shut 0.004268

pathetic 0.004139

bullshit 0.004071

6. Machine Learning

a. Data Pre-Processing

-

b. Naïve Bayes Classifier

c. Random Forest Classifier

7. Conclusion