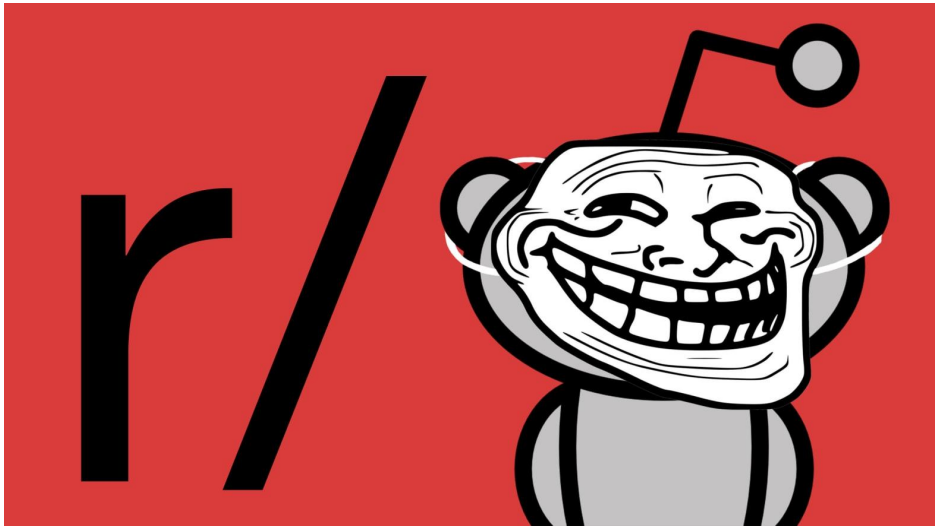
A tool for rapid detection of Reddit trolls



John Burt
Springboard Career Data Science Program
Capstone 1 project

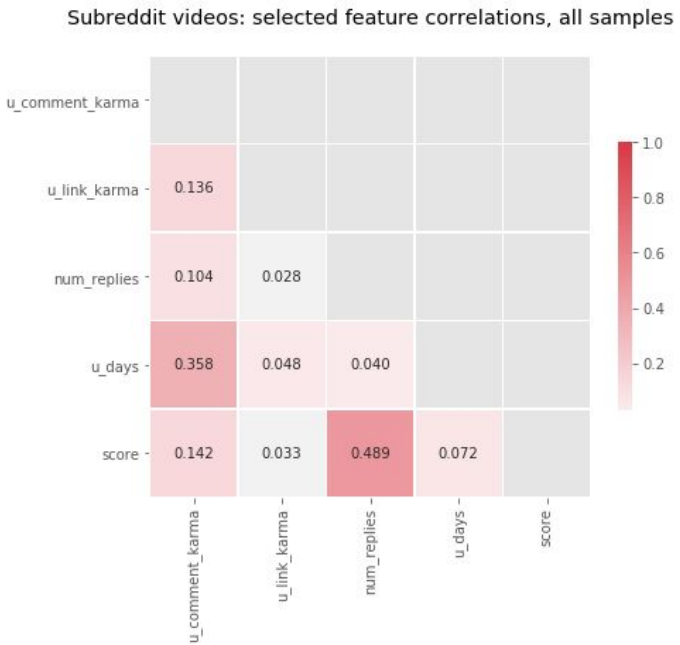
Troll definition:

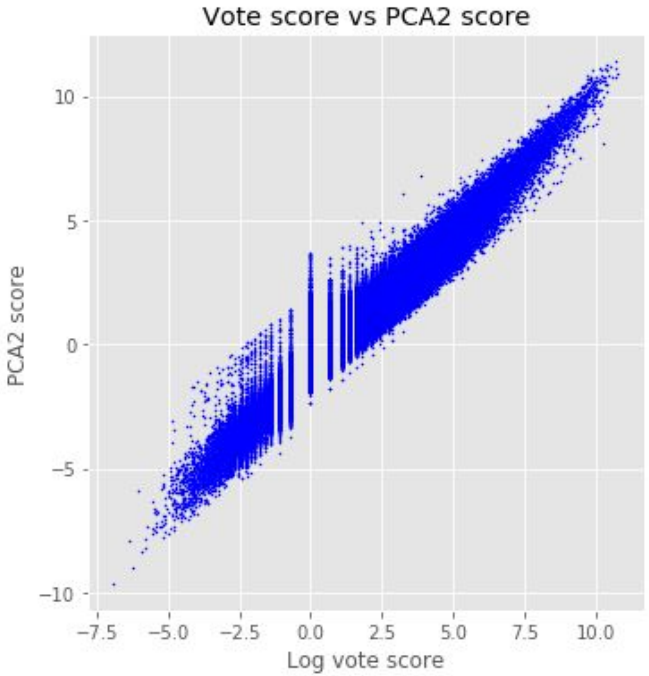
- Post offensive and toxic comments.
- Comments not welcome in community.
- Break site posting rules.

Detecting trolls:

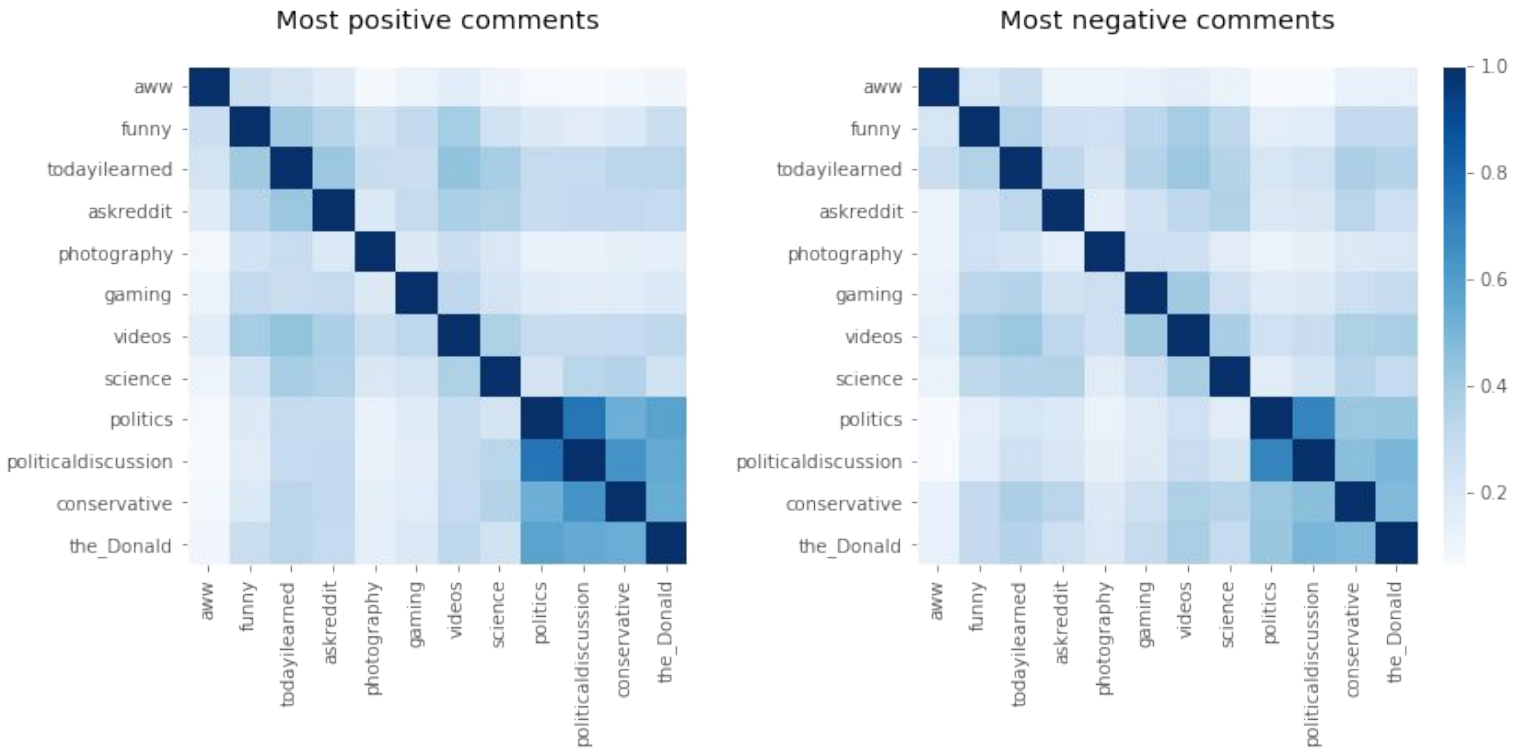
- Downvoting by other users.
- Reports to moderators, and mod policing.
- Auto-detection using ML.

Labelling comment data: troll or not-troll?

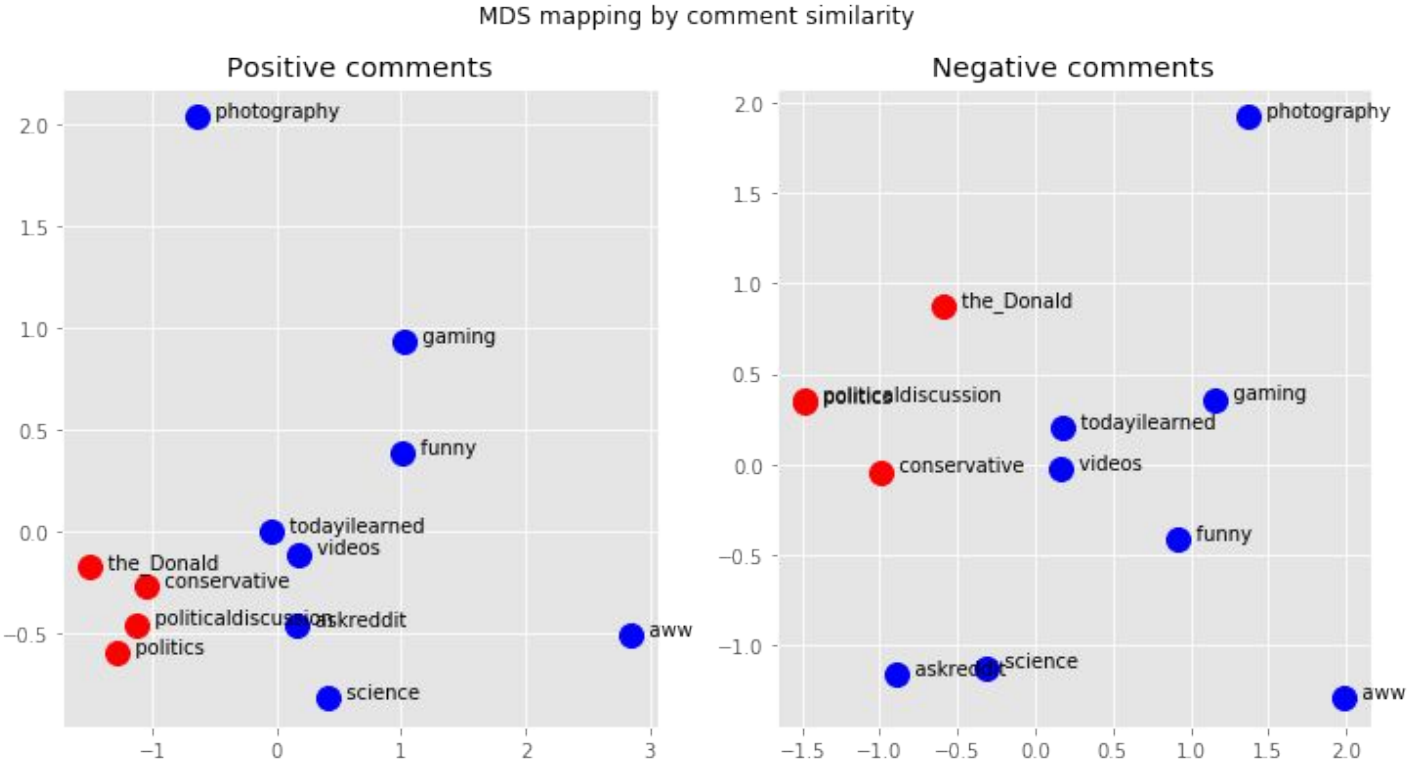




Are troll comments specific to each subreddit?

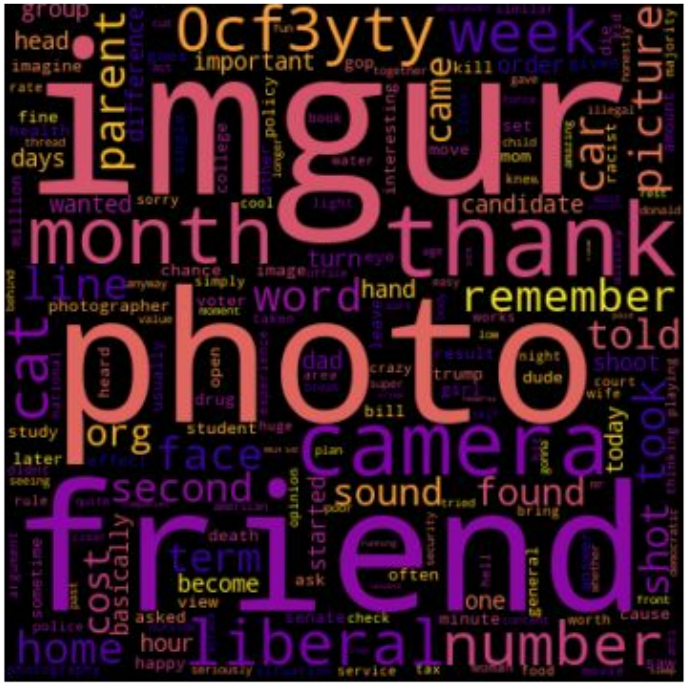


Are troll comments specific to each subreddit?

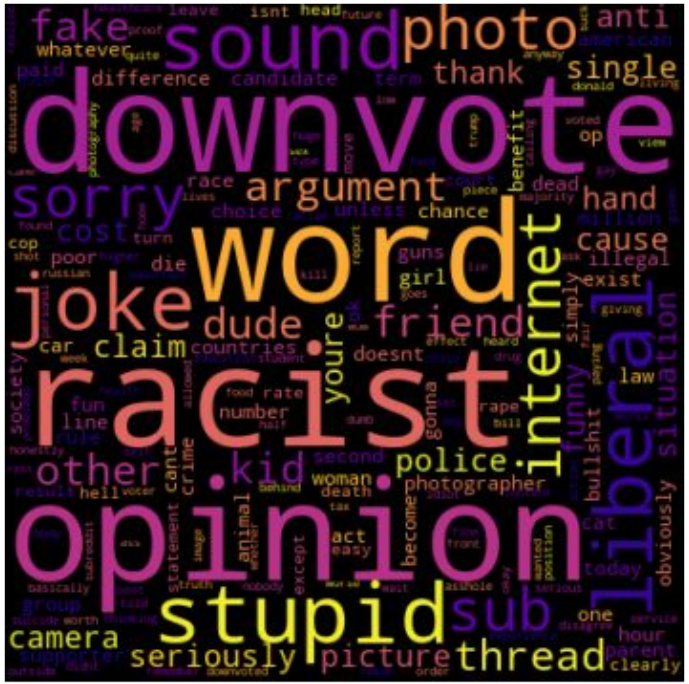


Troll words:

All positive comment text

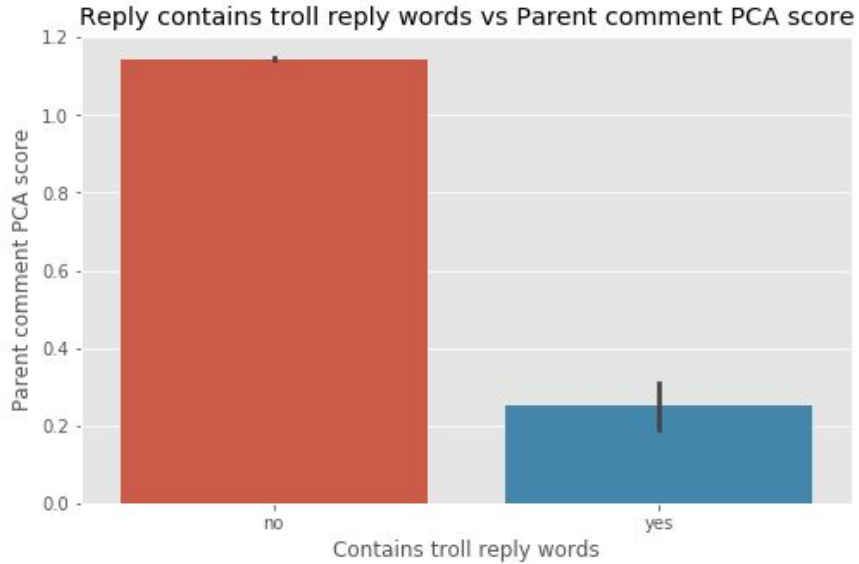


All negative coment text



Does toxicity score predict troll comments?

Toxicity score is lower when replies contain the word “troll”

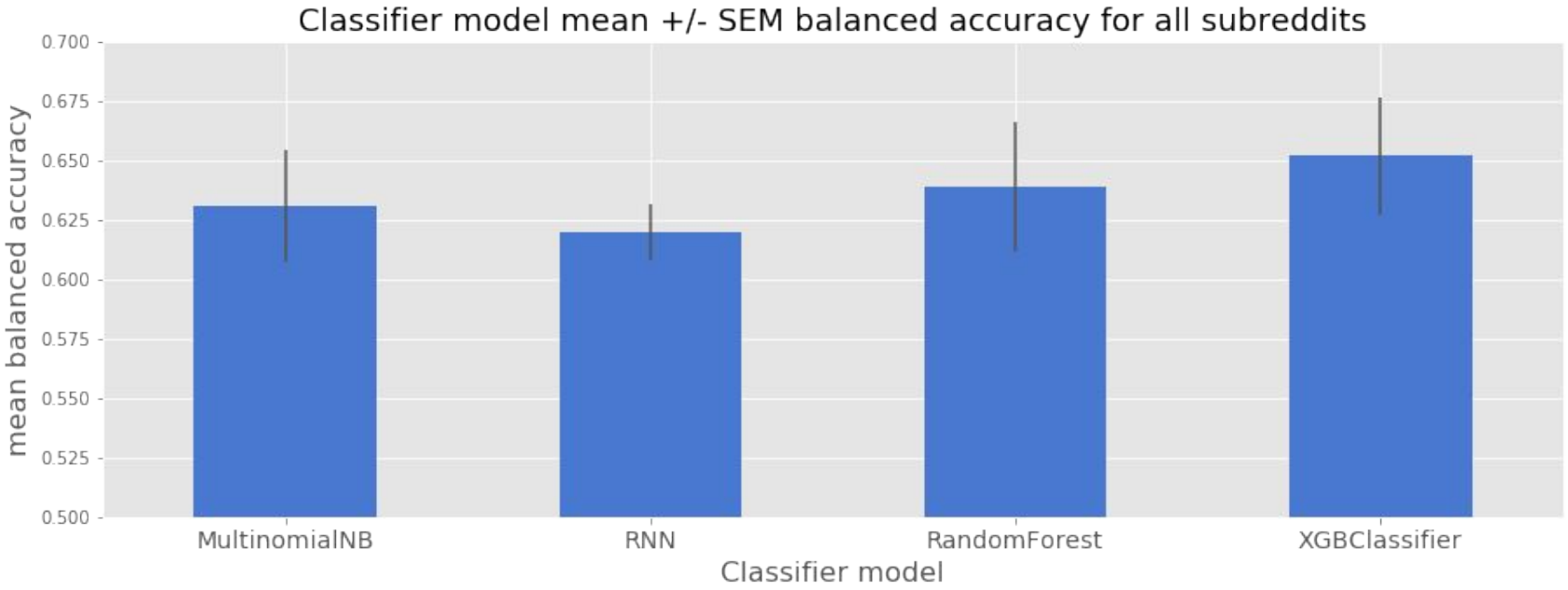


Predicting toxicity from comment text and metadata

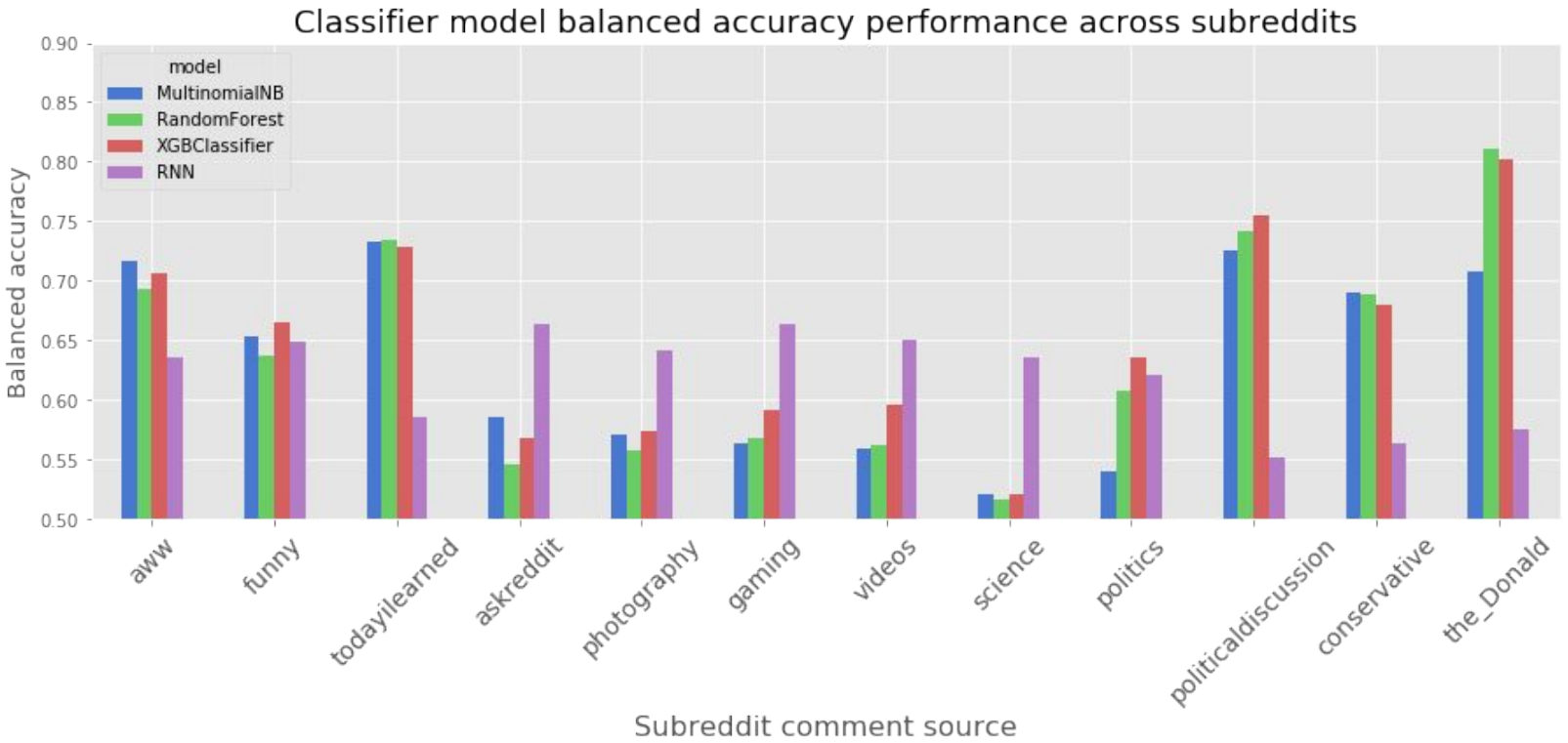
Classifier models tested

Multinomial Naive Bayes
Random Forest
XGBoost
Recurrent Neural Network

Classifier results:



Model performance across subreddits:



Results:

All models performed poorly overall.

Multinomial Naive Bayes, Random Forest and XGBoost had variable accuracy across subreddits.

Recurrent Neural Network had poorest overall, but was consistent.

Conclusions:

Develop Recurrent Neural Network model further.

Engineer new features to include in models.

Combine models for benefit of stacking effect.