# Project 3: Web Scraping & Classification

Anna Jobsis

# Problem Statement

The problem: What characteristics of a post on Reddit are most predictive of the overall interaction on a thread (as measured by number of comments)?

The solution: I will use Reddit data to build three models that will predict whether or not the number of comments on a Reddit post will be above or below the median in order to find the optimal predictors

# The Data

PRAW - Python Reddit API Wrapper:

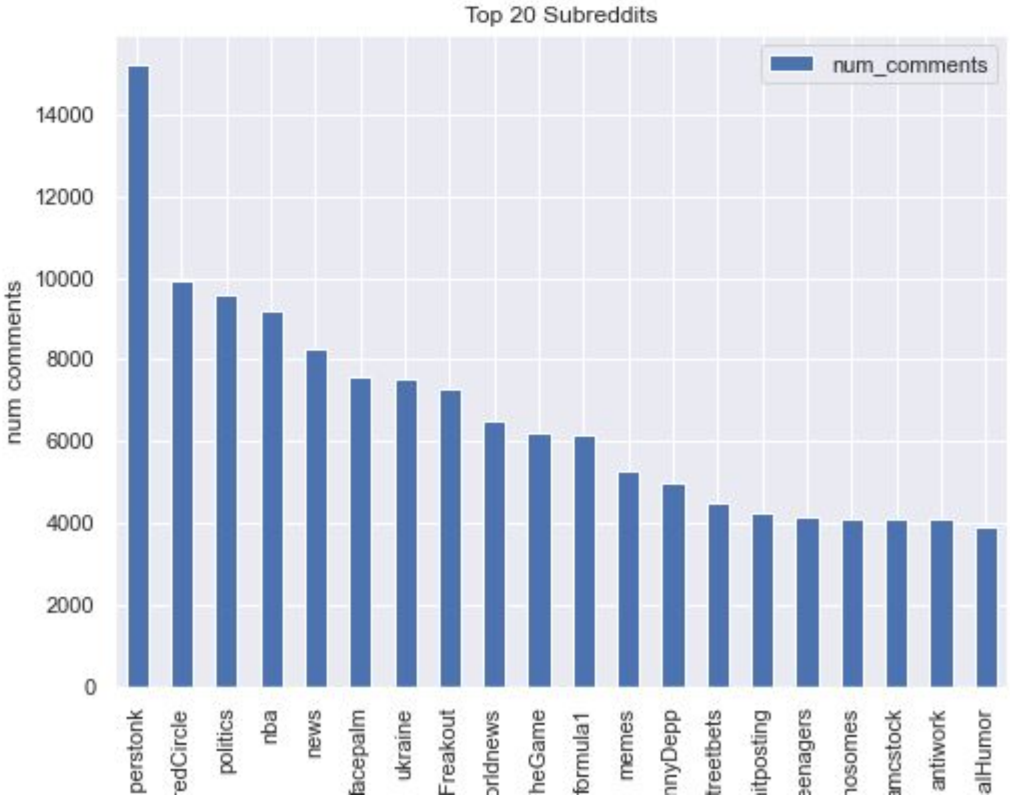
https://praw.readthedocs.io/en/latest/code_overview/models/submission.html

- Author
- Title
- Subreddit
- Created
- Number of comments
- Score (number of upvotes)
- Upvote ratio (percentage of upvotes from all the votes)

# Models

- Logistic Regression
- Random Forest
- Bagging

# Subreddits



- A way to short-list features to feed into the model
- Trends?
  - Politics
  - News
  - Current events
  - Entertainment

# Trends

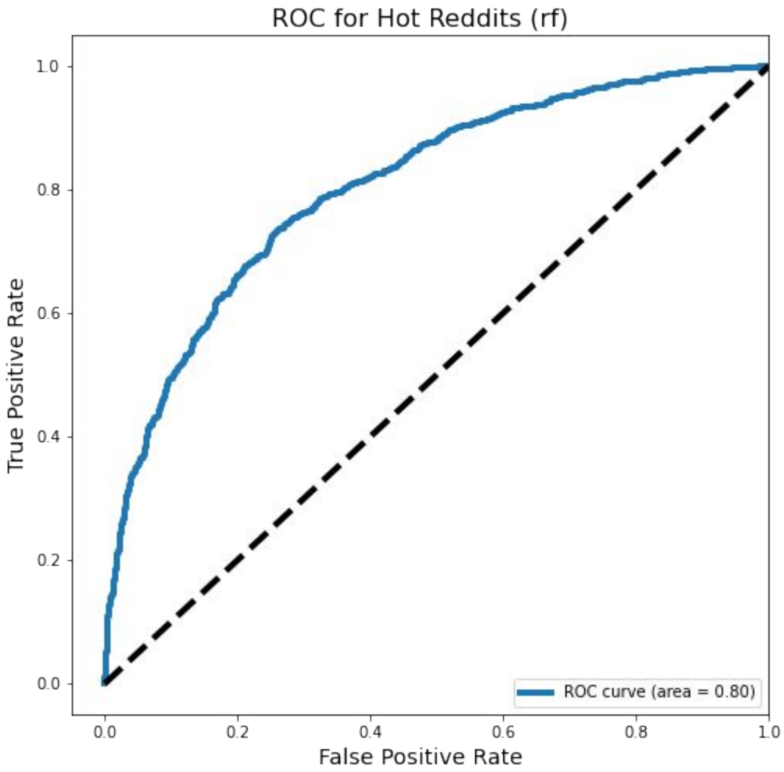
- Johnny Depp/Amber Heard
- Uvalde Shooting
- Ukraine/Russia
- Pride
- Temporal (year, today, 2020)
- Entertainment (season, spoiler, game)

# Words

The models tended to weight different words

- Logistic regression
  - Hate
  - Pride
  - Actually
  - Year
- Random Forest
  - New
  - Today
  - Dog
  - Pride
- Bagging
  - New
  - Today
  - Year
  - Pride

# Model Scoring



- Logistic Regression
  - Precision
  - Specificity
- Random Forest
  - Sensitivity
  - Accuracy
  - F1-Score
- Bagging
  - The worst performing model

# Conclusion

There were a few characteristics that seemed to have more predictive power

- Politics
- World events
- News
- Temporal (Year, 2020, Today)
- Entertainment

Further analysis:

- More data collected over a longer period of time
- How does cross-posting affect predictors
- Alternative vectorizers / stemming