Reddit Post “Hot” Classification

Andy Deemer

Problem Statement: What characteristics of a post are most predictive of the overall interaction on a thread?

Web Scraping & Datasets

“Hot” Posts from the Reddit Homepage - 10,000 in this dataset

Classification:

“Not Hot” - 51% had fewer comments than the median (Majority Class)
“Hot” - 49% had comments equal to or above the median

Median Number of Comments = 53

Features Ultimately Used in Models

- Original Content
- Text Only
- Video
- Over 18
- Spoiler

Natural Language Processing on Title and Subreddit

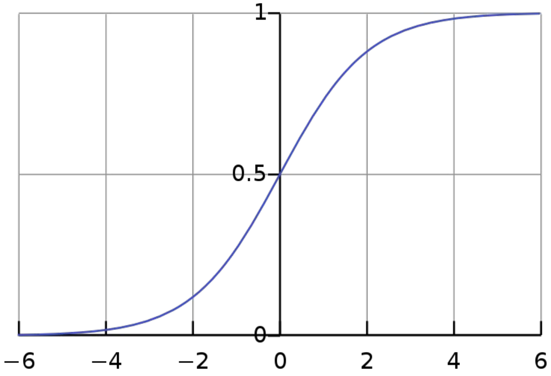
Classification Models

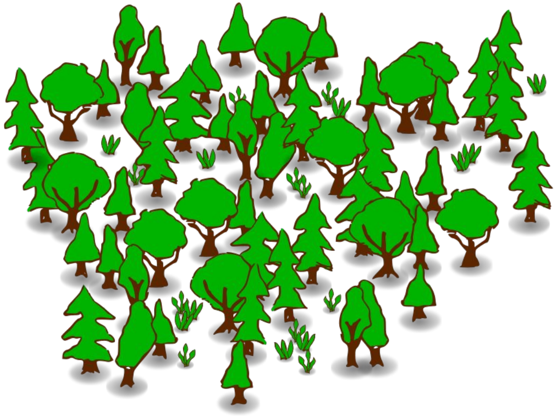
Logistic Regression

Accuracy - 73%

Random Forest

Accuracy - 75%





Top Indicative Subreddits

15-20 posts each of these in Dataset

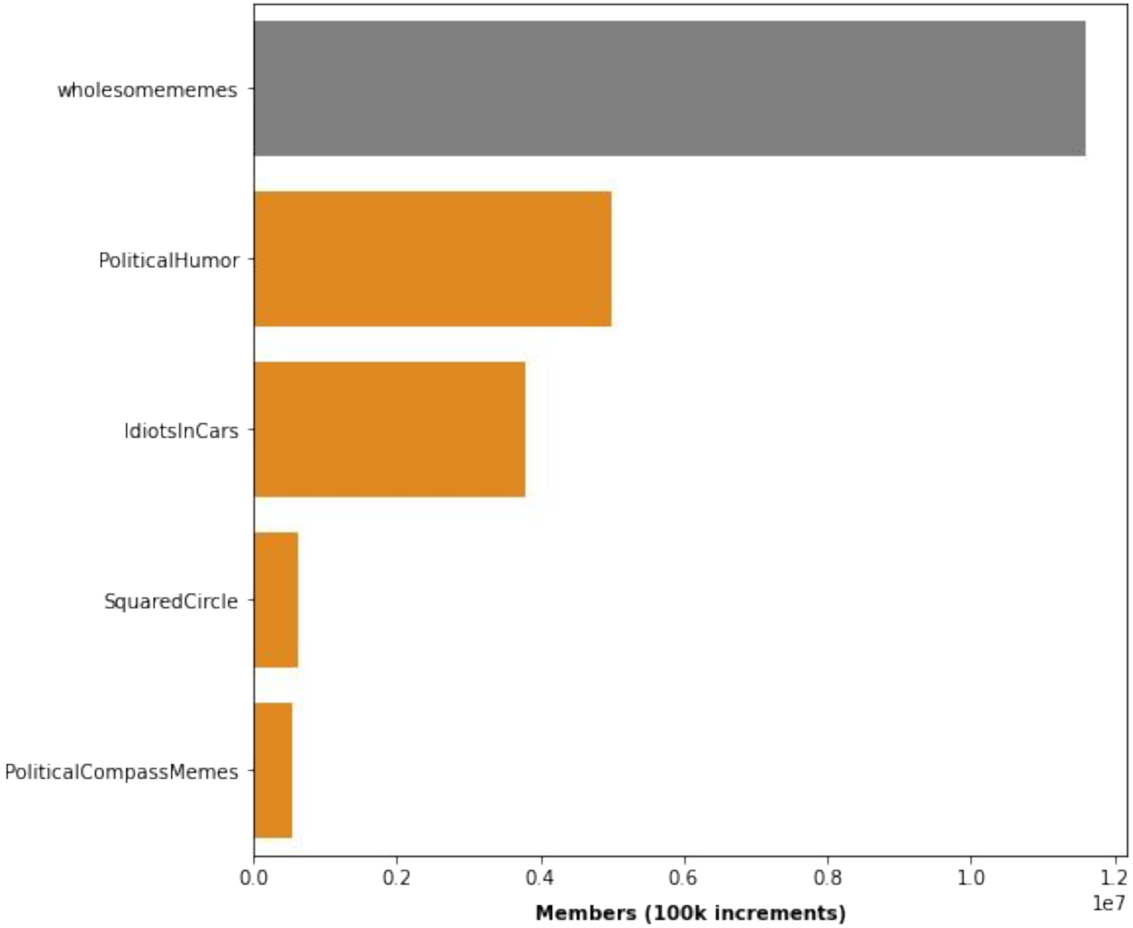
wholesomememes: -0.8
- Most Members, but not “Hot”
- don’t comment on funny memes

PolitcalCompassMemes: 15.1
- Fewer members, higher interaction
- This type of meme elicits more response

SqauredCircle: 9.8
- WWE subreddit
- fewer members relative to others
- super engaged members

PoliticalHumor: 3.5
- opinion based posts
- dependent on current events

IdiotsInCars: 3.1
- all videos and images
- everyone likes making fun of drivers



Top Word Indicators

These words in Reddit posts were the words that most indicated if a post is "Hot".

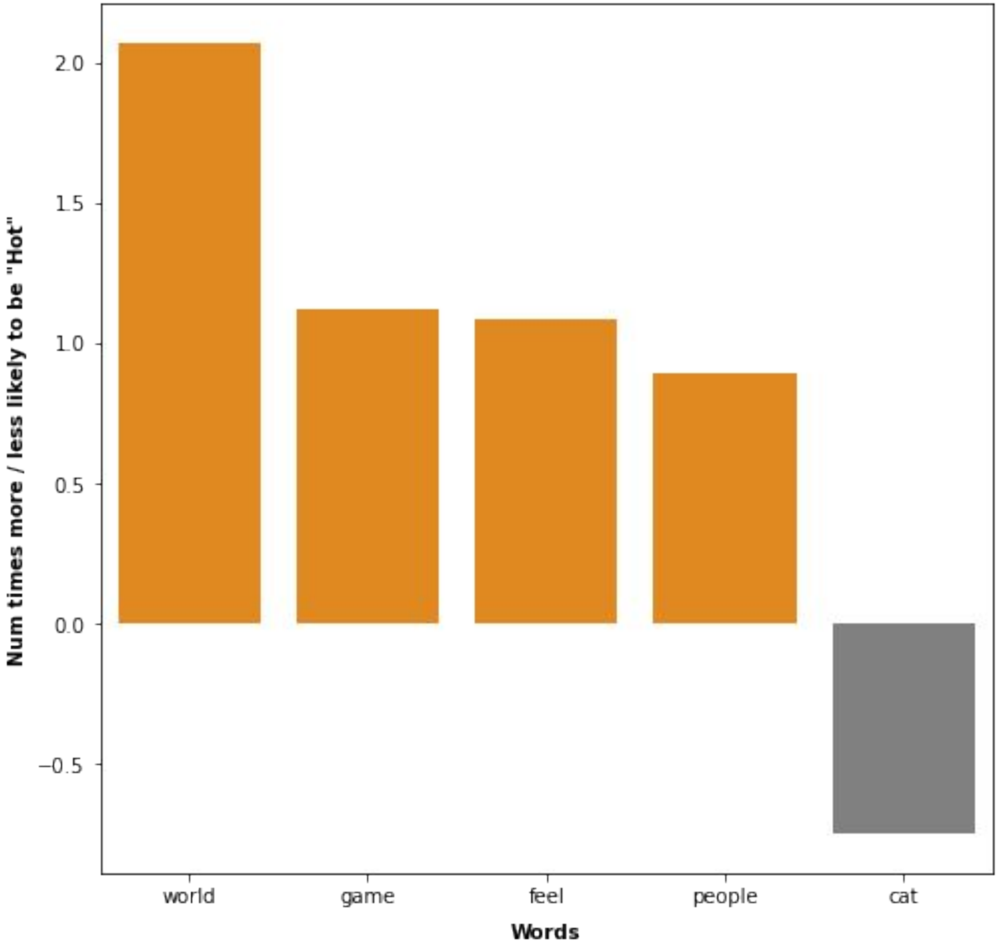
"World"
- posts about current events

"Game"
- lots of sports fans active on Reddit

"Feel"
- opinionated posts elicit response

"People"
- opinionated posts about people

"Cat"
- posts about cats are not "hot"



Is Self

Text Only Posts 4 times more likely to be "Hot"

Distribution:

95%	More Than Text
5%	Text Only

Top 5 Text Only Posts:

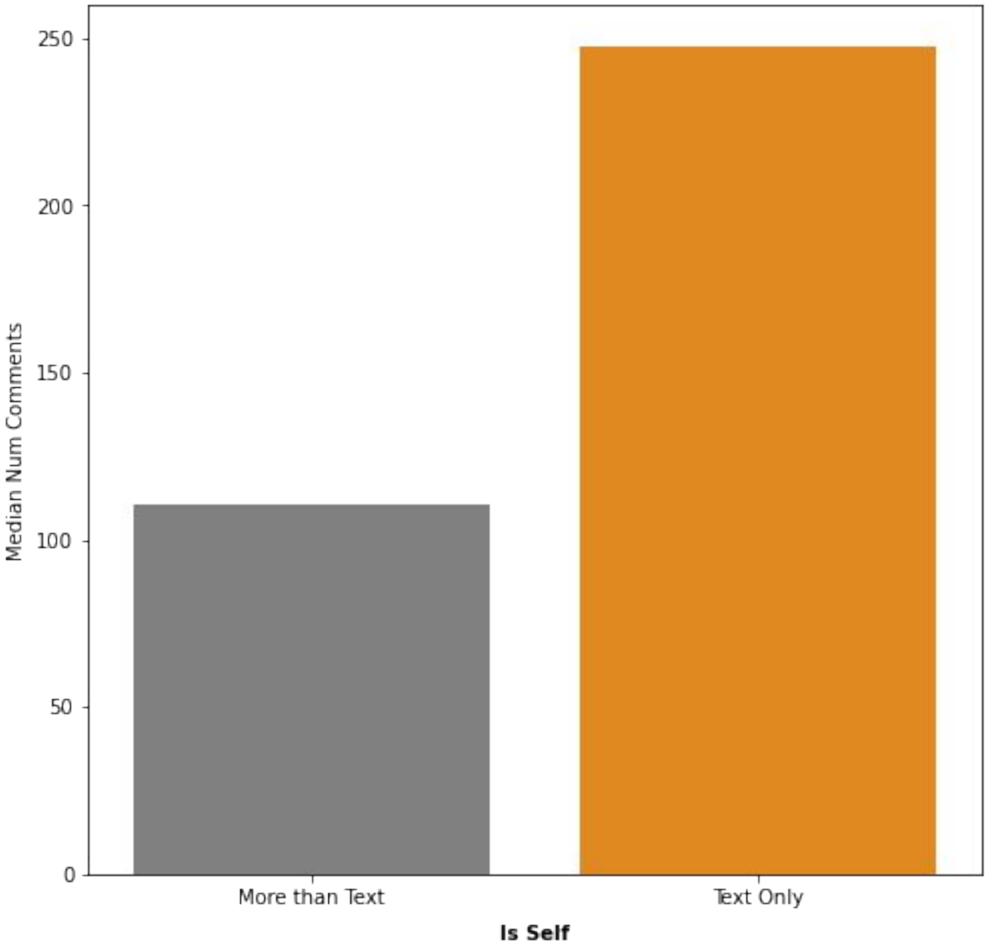
Anyone elses street doing sod all for the jubilee?

Tifu by eating ramen before inviting a guy over

I wish tower damage was kept as it is right now

Post game thread on the boston celtics defeat the miami heat 100 96

If a girl hits you are you going to hit her back?



Conclusions and Recommendations

- Random Forest Model
  - most accurate
  - least intelligible
- Pay attention to what subreddit a post originates from
- subreddits with more members doesn’t always coincide with high interaction
- Titles with with opinions or controversial topics have high interaction
- There are keywords that generally result in higher interaction, but these factors are less important than the two aforementioned features.