Using NLP to Classify Reddit Posts:

Bodyweight Fitness or Weightlifting

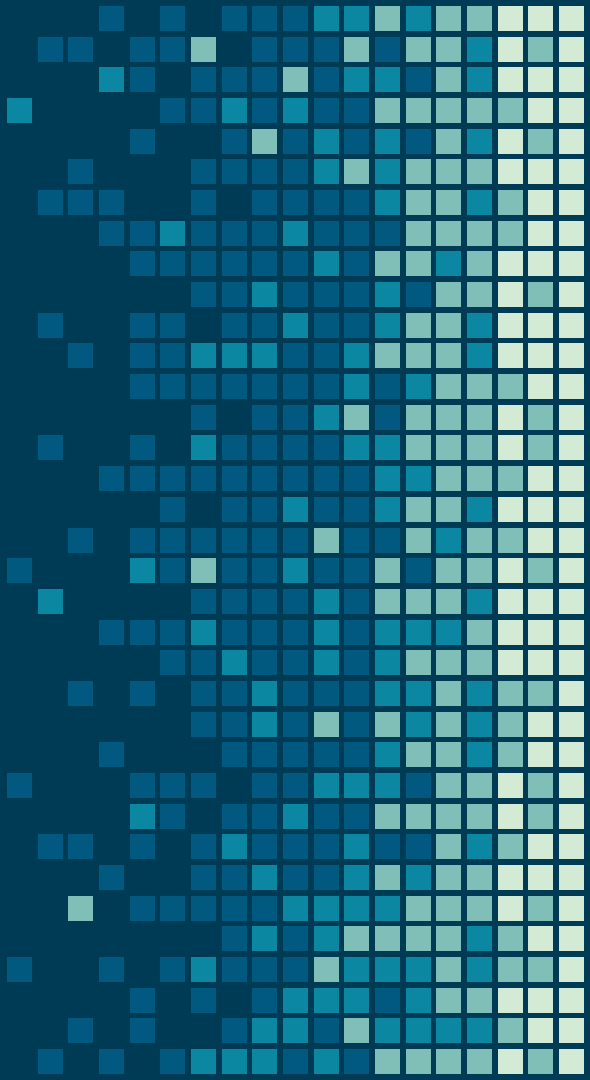
Jennifer Brown
General Assembly - DSIR-2-8
March 2021

Overview

- Introduction
- Background
- Findings
- Conclusions
- Recommendations and next steps

Introduction

- Problem Statement
  - Can we predict to what subreddit a post belongs? Specifically, can we predict the categorization of r/bodyweightfitness and r/weightlifting.
- Importance
- Audience



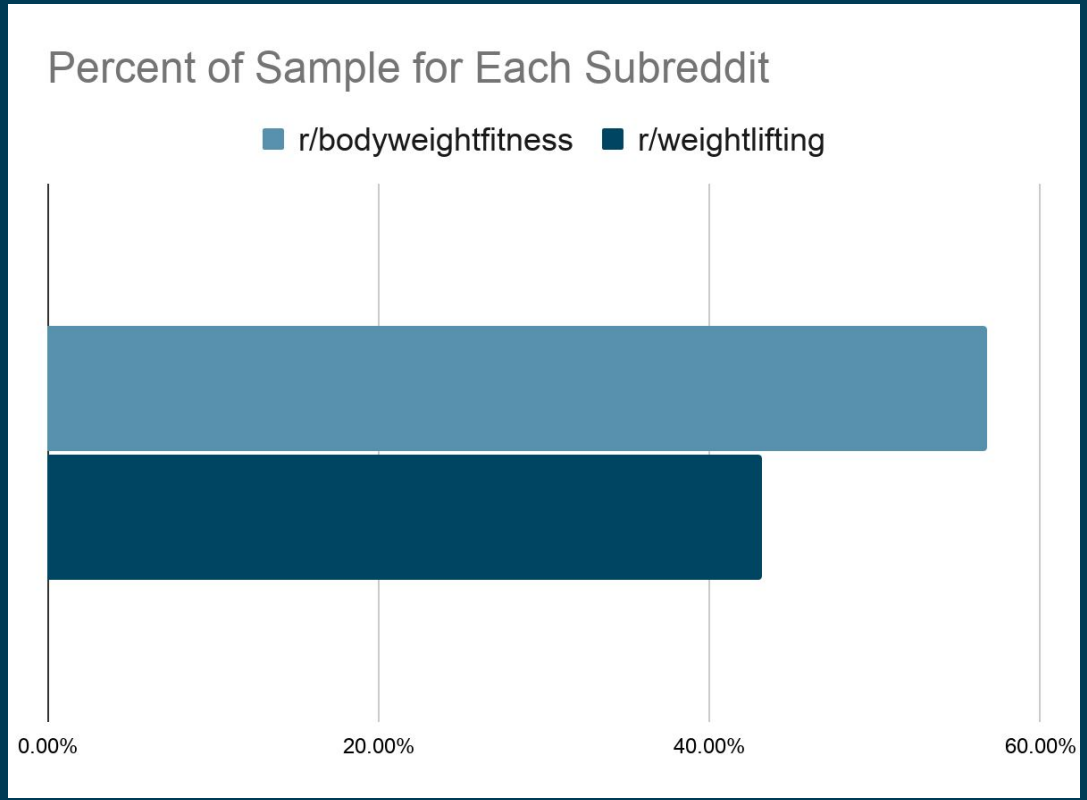
Background

- Data scraped from two subreddits
  - Pushshift's API

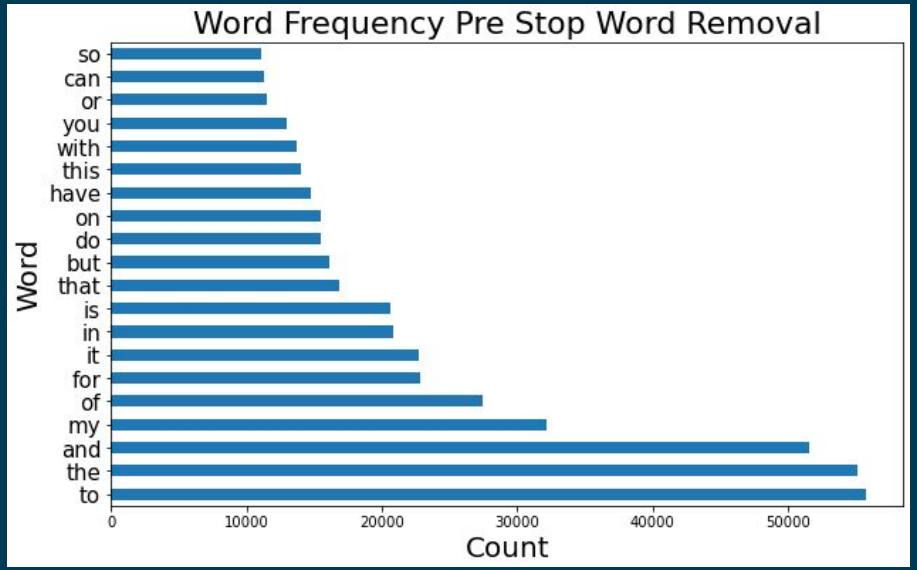
r/bodyweightfitness (1)	r/weightlifting (0)
Hey fit people, I've gotten a new pair of gymnastics rings and I was wondering, which good workout programs do you know for rings?I was looking especially for something like full body 3x per week	I'm 80 kg and clean and jerk 106 kg. At my gym they teach split jerk, but I feel like I am better in every way when I push jerk. I get more power, I get lower quicker and my catch feels more stable and powerful.

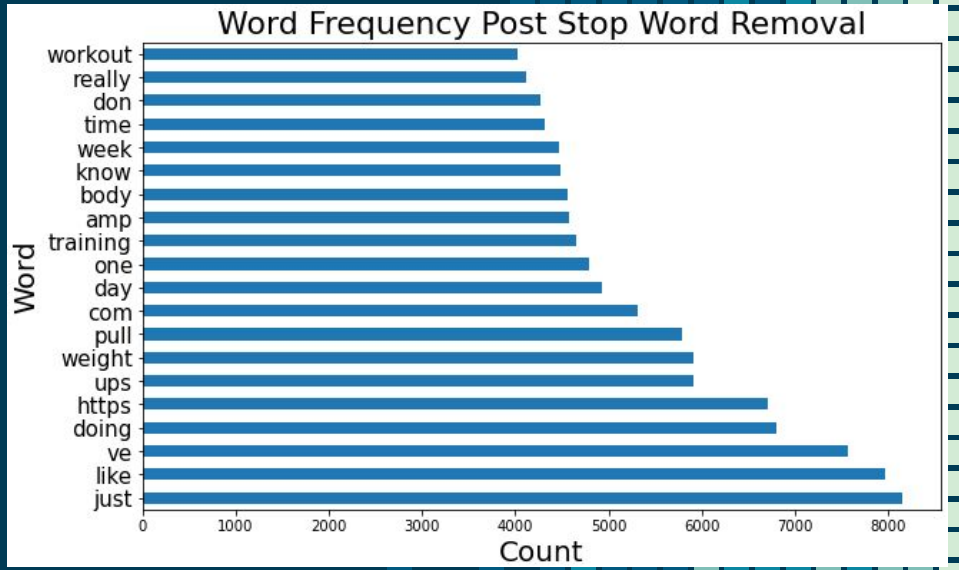
Background cont.

- 20,000 from each subreddit
- Data cleaning
  - No videos or automod created threads
  - No deleted rows (moderator, user deleted, etc.)
  - Text in ‘selftext’ field
  - Custom stop words list
- Modeling
  - Logistic regression, naïve bayes, random forest


Percent of Sample for Each Subreddit
r/bodyweightfitness
r/weightlifting
0.00%
20.00%
40.00%
60.00%

Stop Words


Word Frequency Pre Stop Word Removal
Word
so
can
or
you
with
this
have
on
do
but
that
is
in
it
for
of
my
and
the
to
0
10000
20000
30000
40000
50000
Count

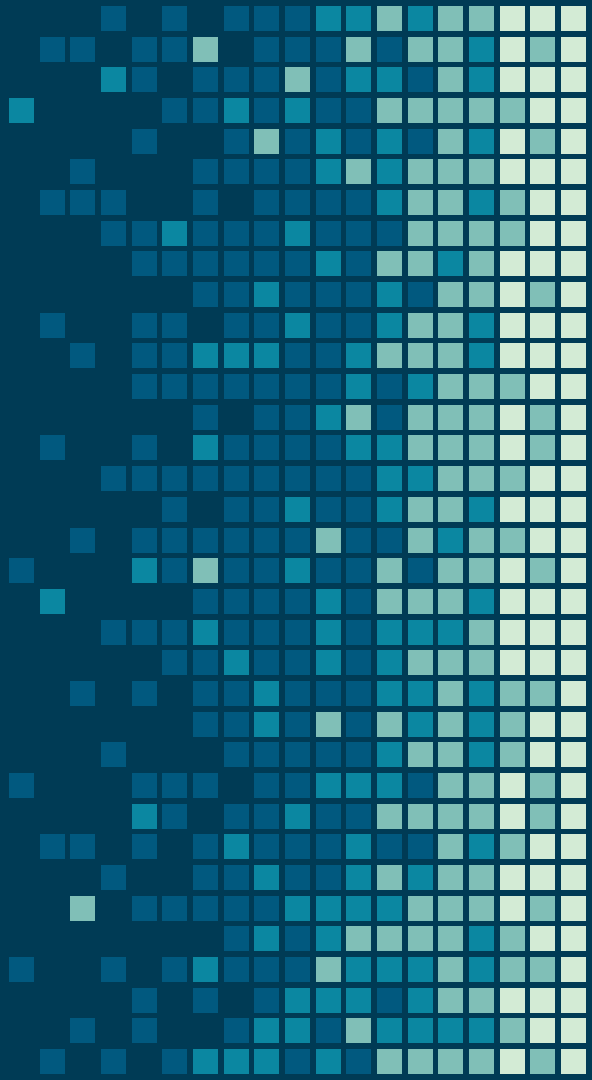

Word Frequency Post Stop Word Removal
Word
workout
really
don
time
week
know
body
amp
training
one
day
com
pull
weight
ups
https
doing
ve
like
just
0
1000
2000
3000
4000
5000
6000
7000
8000
Count

The Final Models

- Baseline / Null model
  - 0.567966
- Logistic Regression with TF-IDF
- Naïve Bayes with TF-IDF

Results

r/bodyweightfitness	r/weightlifting
pull	weightlifting
workout	squat
push	lifting
body	weight
weight	snatch
exercises	bar
pull ups	amp
muscle	gym
routine	clean
training	good



Results

	Log Reg w/ TF-IDF	Naïve Bayes w/ TF-IDF
Cross Val	0.900	0.885
Train Accuracy (R2)	0.933	0.897
Test Accuracy (R2)	0.902	0.888
Specificity	0.854	0.823
Sensitivity	0.940	0.937
Precision	0.894	0.874

Conclusions

- Posts can be accurately predicted with greater accuracy than the Null Model
- Final Models had similar performance.
- Evidence of some overfitting
- Difference in top words

Recommendations

- Build more models
- Refine stop words list
- Collect more data
- Examine post titles
- Apply to other fitness domains
    - r/running vs. r/slowjogging

THANKS!

Any questions?

You can find me at:

@username

user@mail.me