Reddit Post Classification

Using NLP to Classify Text

By Eric Bayless
January 22, 2020

The Problem

- Predict whether the given text of a submission came from the r/Fitness or r/Bodyweightfitness subreddit

Data Acquisition

Method

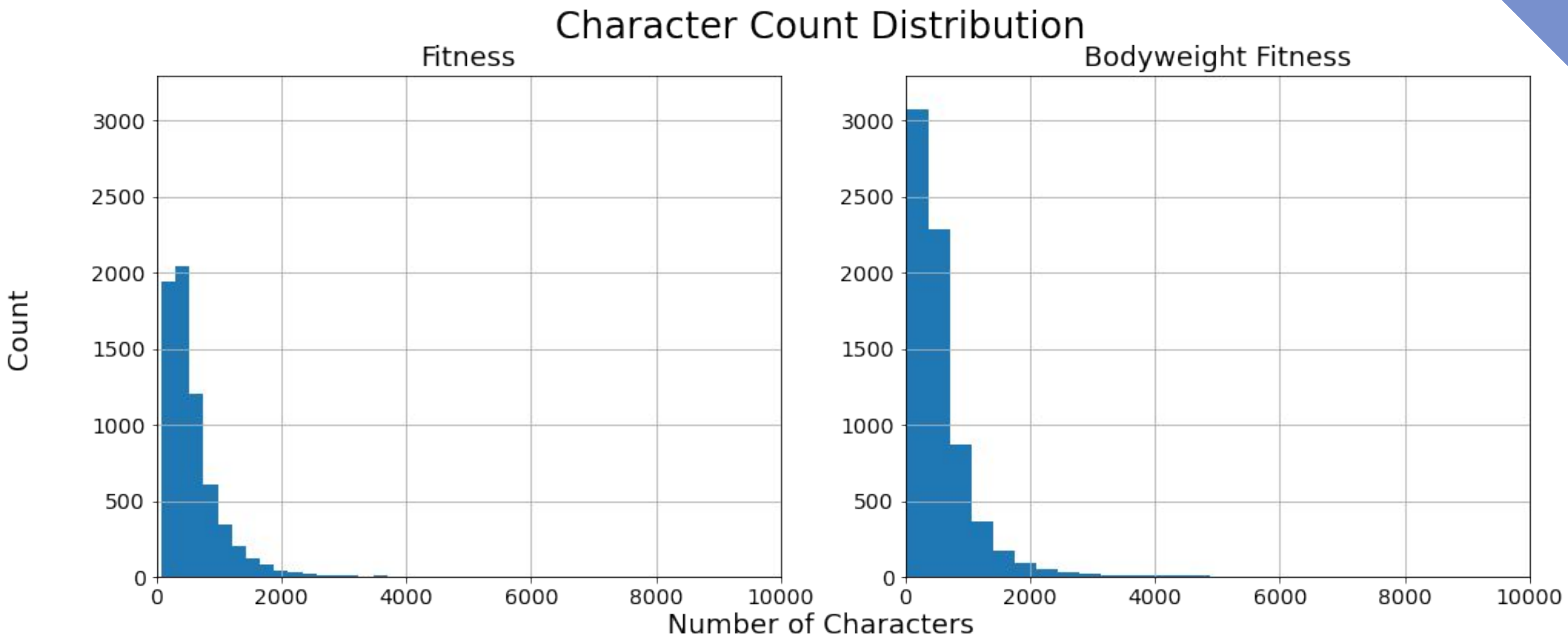
- Pushshift API
- Function
    - 100 posts per request
    - 10 seconds between
    - Starting with most recent

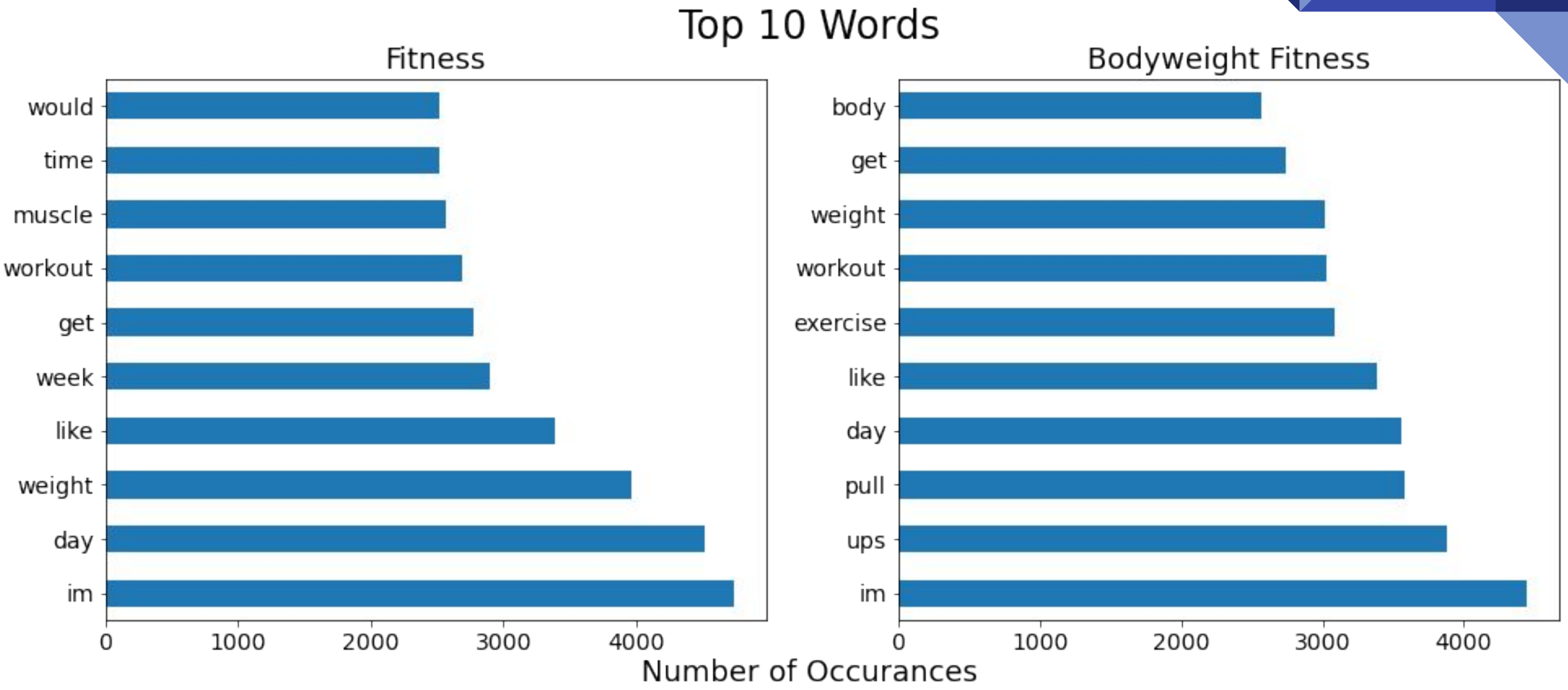
Data Cleaning

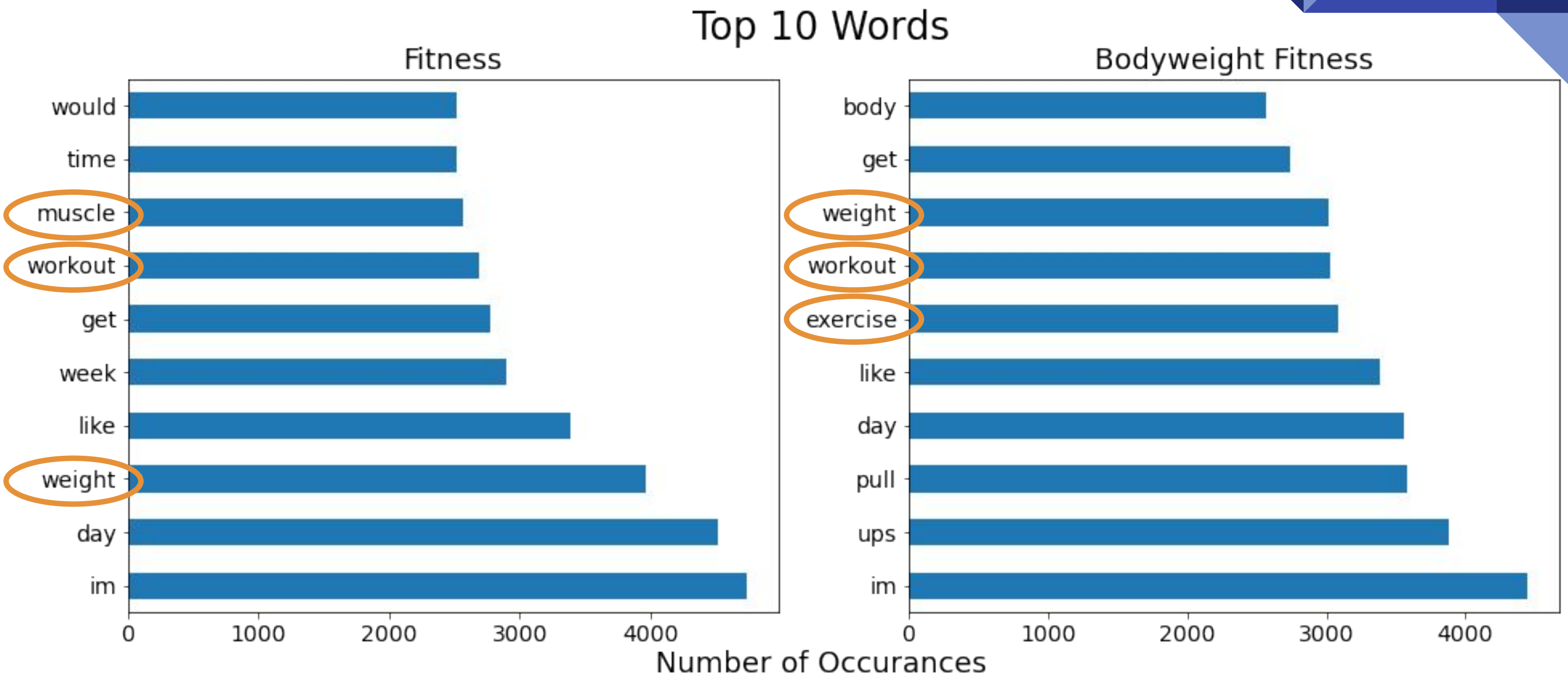
Method

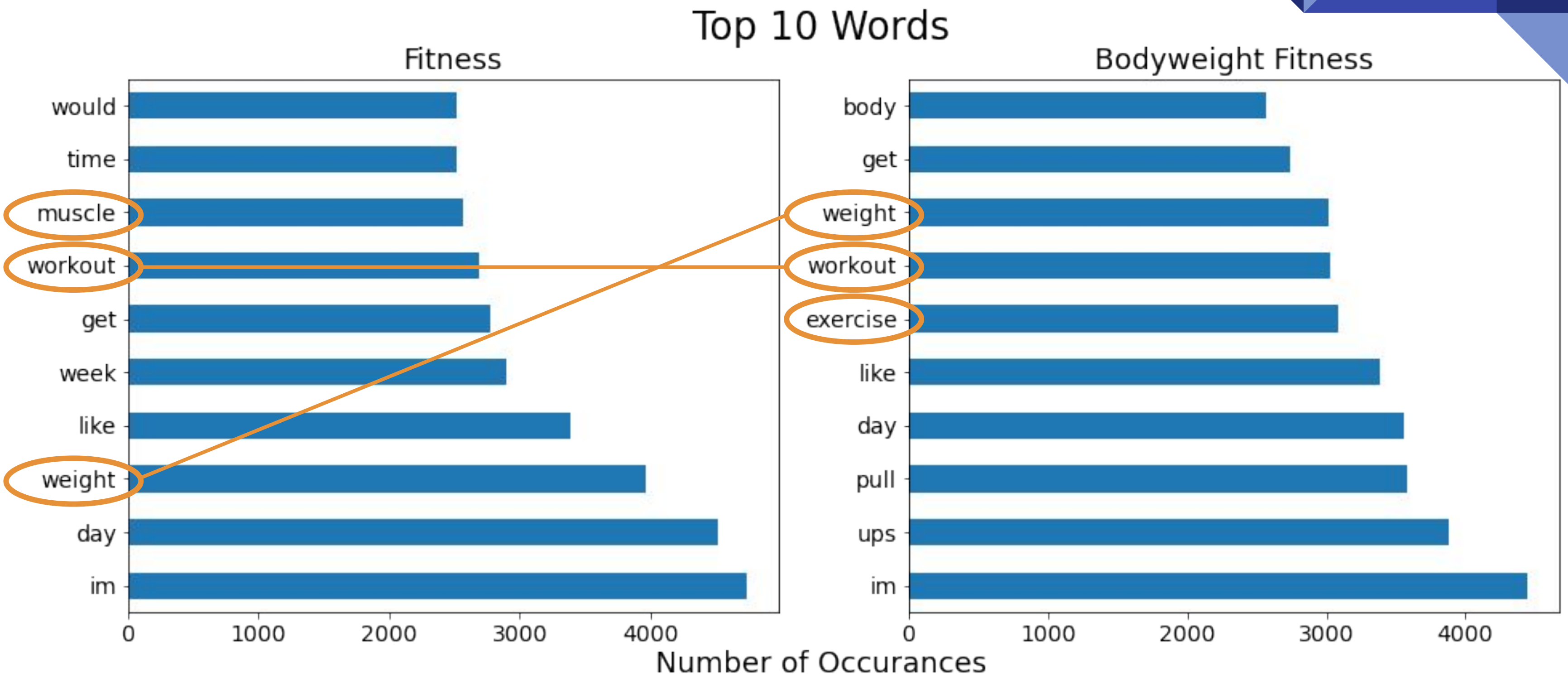
1. Drop null posts
2. Drop ‘deleted’ or ‘removed’
3. Drop posts containing subreddit name
4. Drop posts shorter than title

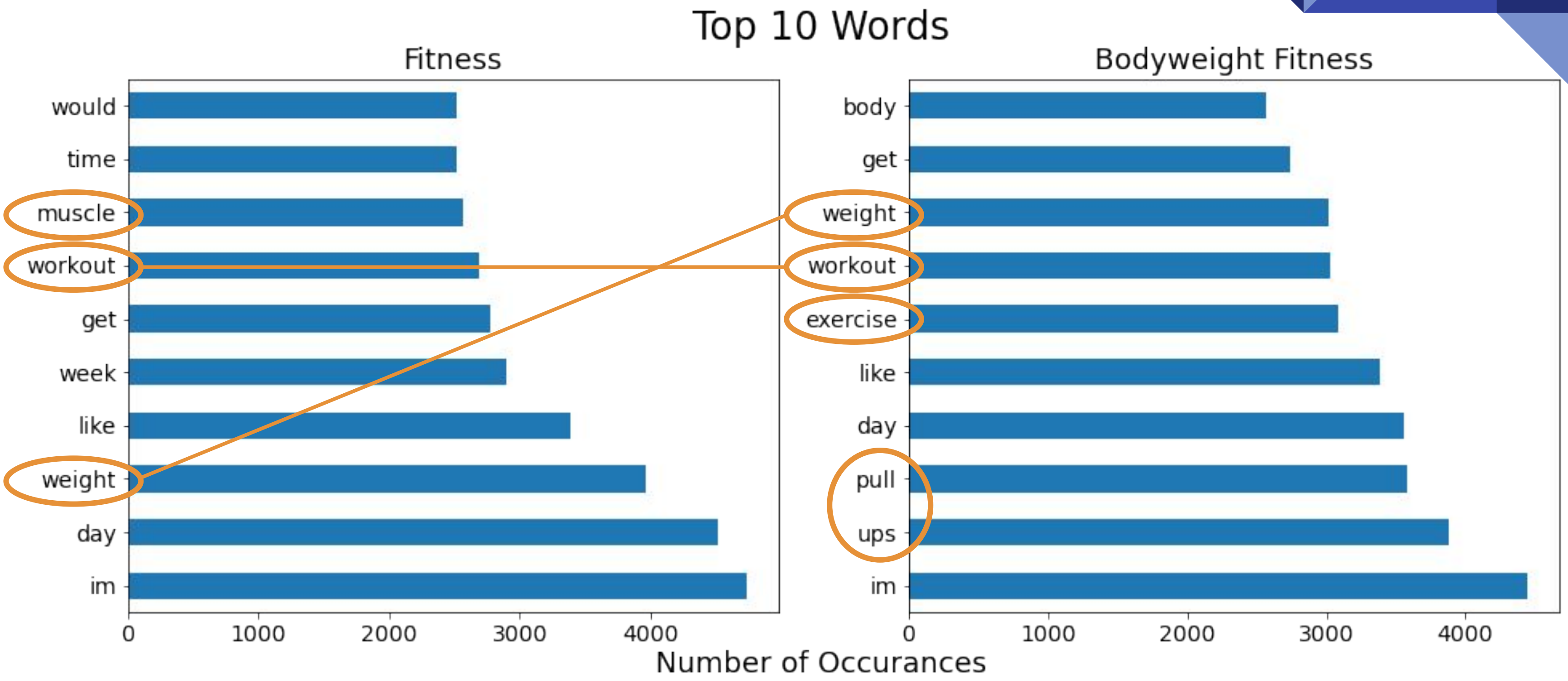
Exploratory Analysis


Character Count Distribution
Fitness
Bodyweight Fitness
Count
3000
2500
2000
1500
1000
500
0
0
2000
4000
6000
8000
10000
Number of Characters


Top 10 Words
Fitness
would
time
muscle
workout
get
week
like
weight
day
im
0
1000
2000
3000
4000
Bodyweight Fitness
body
get
weight
workout
exercise
like
day
pull
ups
im
0
1000
2000
3000
4000
Number of Occurances


Top 10 Words
Fitness
would
time
muscle
workout
get
week
like
weight
day
im
0
1000
2000
3000
4000
Bodyweight Fitness
body
get
weight
workout
exercise
like
day
pull
ups
im
0
1000
2000
3000
4000
Number of Occurances


Top 10 Words
Fitness
would
time
muscle
workout
get
week
like
weight
day
im
0
1000
2000
3000
4000
Bodyweight Fitness
body
get
weight
workout
exercise
like
day
pull
ups
im
0
1000
2000
3000
4000
Number of Occurances


Top 10 Words
Fitness
would
time
muscle
workout
get
week
like
weight
day
im
0
1000
2000
3000
4000
Bodyweight Fitness
body
get
weight
workout
exercise
like
day
pull
ups
im
0
1000
2000
3000
4000
Number of Occurances

Modeling

Baseline

- Most Frequent
- r/Bodyweightfitness
- Accuracy: 51%

Model Selection

- Multiple model types
- Accuracy 76-78%
- Custom preprocessing

Best Model

- Logistic Regression
- Accuracy: 79%
- Interpretable coefficients

Most Important Terms

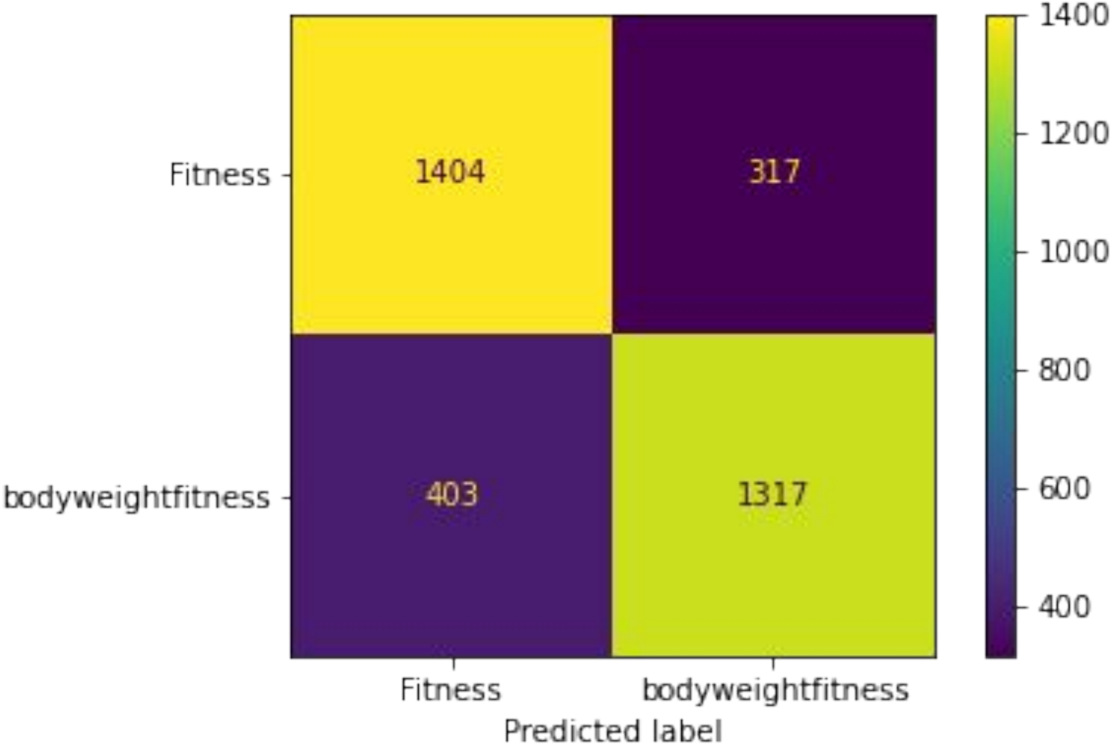
Bodyweight Fit.

1. rr
2. bodyweight
3. calisthenics
4. ring
5. covid
6. pull
7. dip
8. pushup
9. ups
10. pullups

Fitness

1. Bench
2. Gym
3. Lifting
4. Lift
5. Deadlift
6. Machine
7. Fitness
8. Bike
9. Deadlifts
10. dumbbell

Confusion Matrix



	Fitness	bodyweightfitness
Fitness	1404	317
bodyweightfitness	403	1317

Conclusions

Results

- 79% Accuracy
- Similarity of subreddits may prevent further improvement

Production Recommendation

- Logistic Regression
  - Highest accuracy
  - Interpretable

Future Steps

- Custom Vocabulary
- Advanced Neural Network
- More Preprocessing
- More Data

Questions?