

SUBREDDIT CLASSIFICATION
R/GAINIT & R/BODYWEIGHTFITNESS

PROJECT 3: WEB API & NLP

BY: MATTHEW LIO

SUBREDDITS

r/gainit
Weightlifting

VS

r/bodyweightfitness
Bodyweight

CONTENTS

- Introduction
- Problem Statement
- Cleaning & Pre-processing
- Modeling
- Performance Summary
- Conclusion & Model Limitations

INTRODUCTION

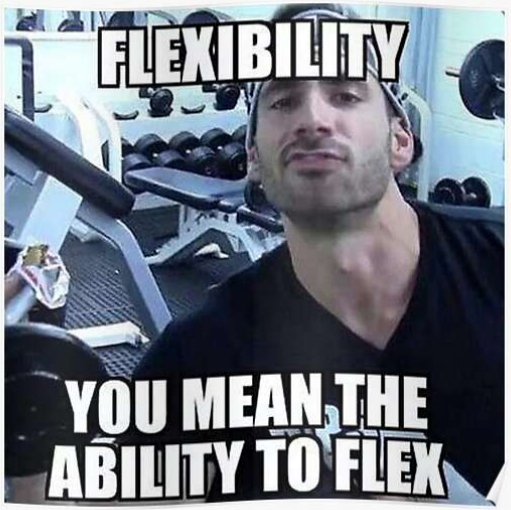
Fitness suffers from misinformation online

Diet craze

Wellness trends

Fitness fads

Broscience



PROBLEM STATEMENT

To classify fitness and health related questions or posts into the correct subreddit

- right experts can give proper advice and answers to these posts

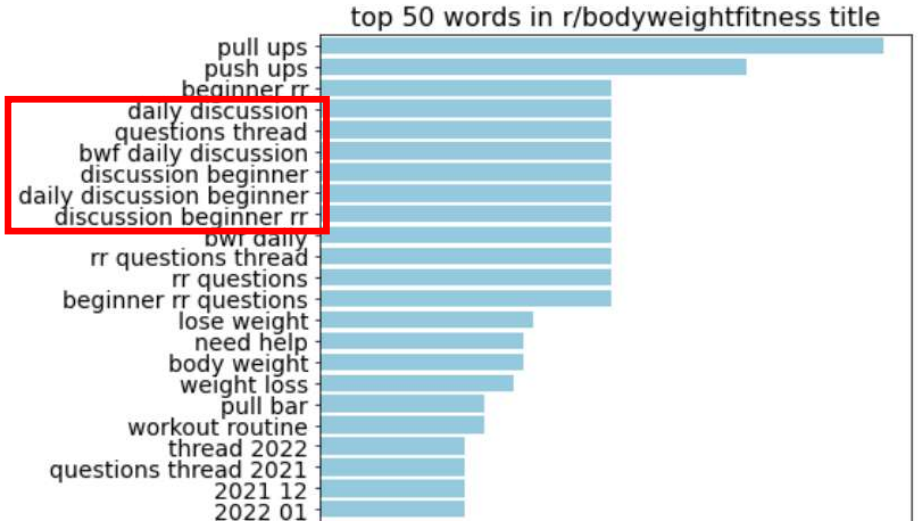
WRONG EXPERT = INCORRECT ADVICE = INEFFICIENT TRAINING/ DIET = EVEN INJURY

CLEANING & PRE-PROCESSING

- Moderator and Discussion Posts
- Common Posts
- Spams
- Duplicates
- Links
- Special Characters
- Tokenizing & Lemmatizing

CLEANING & PRE-PROCESSING

- Moderator and Discussion Posts



CLEANING & PRE-PROCESSING

- Moderator and Discussion Posts
- Common Posts

title
Beginner to push up Ultimate guide
Is ATHLEAN-X a good fitness YouTuber?

CLEANING & PRE-PROCESSING

- Moderator and Discussion Posts
- Common Posts
- Spams
- Duplicates

author	
ComplaintCapital2781	I barely have any chest mass. Is it possible f...
ComplaintCapital2781	I barely have any chest mass. Is it possible f...
ComplaintCapital2781	I barely have any chest mass. Is it possible f...

CLEANING & PRE-PROCESSING

- Moderator and Discussion Posts
- Common Posts
- Spams
- Duplicates
- Links

```
[Recommended
Routine](https://www.reddit.com/
r/bodyweightfitness/wiki/kb/reco
mmended_routine)
```

CLEANING & PRE-PROCESSING

- Moderator and Discussion Posts
- Common Posts
- Spams
- Duplicates
- Links
- Special Characters

REGEX

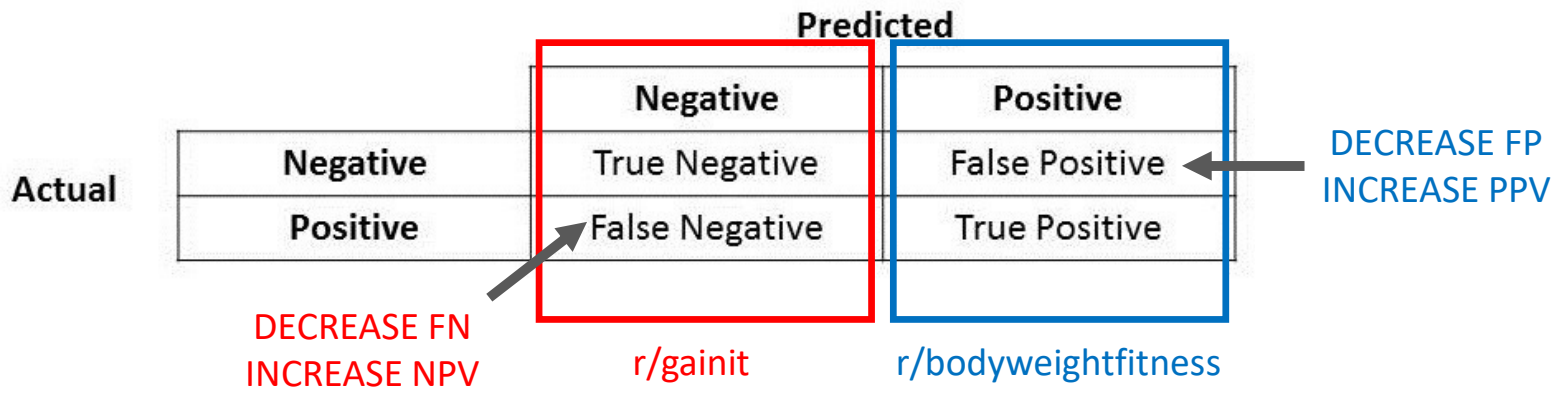
CLEANING & PRE-PROCESSING

- Moderator and Discussion Posts
- Common Posts
- Spams
- Duplicates
- Links
- Special Characters
- Tokenizing & Lemmatizing

	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

MODELING

- If wrongly classified, wrong experts would be answering the questions
- Incorrect advice -> inefficiency and even injuries
- Need to decrease False Positives and False Negatives!
- Increase PPV and NPV scores!

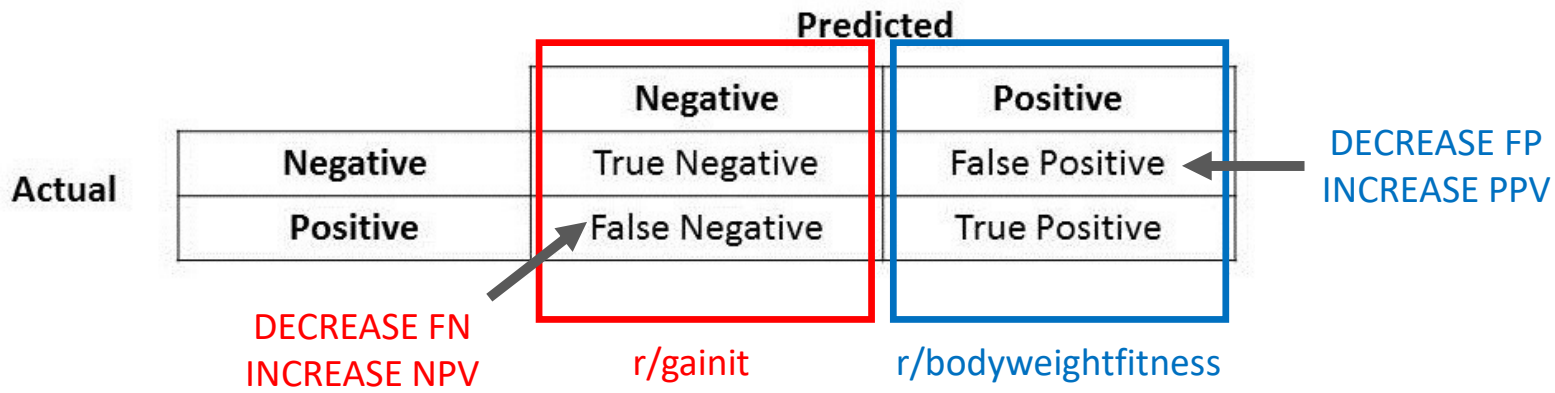


MODELING

Accuracy

PPV (Positive Predictive Value/Precision)

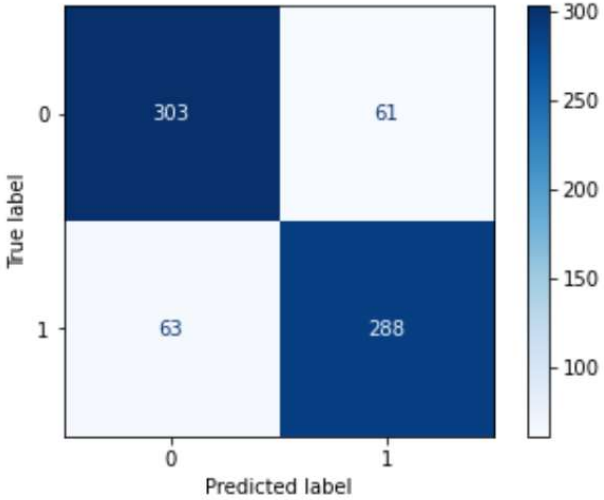
NPV (Negative Predictive Value)



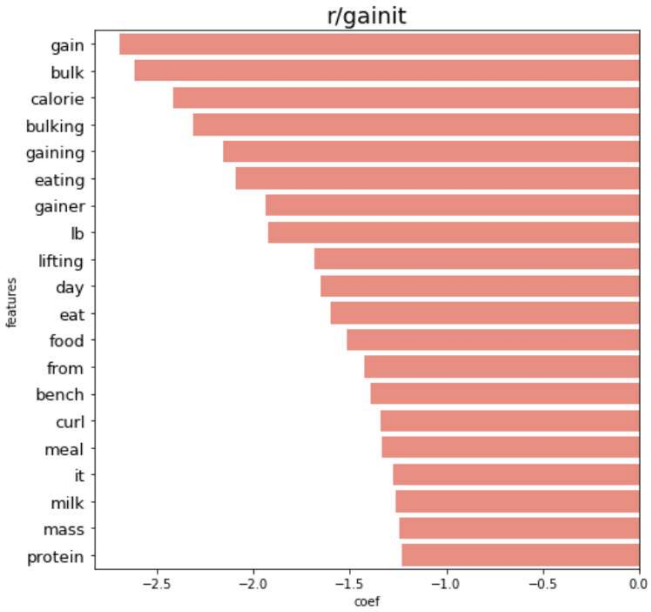
MODELING: LOGISTIC REGRESSION W TVEC

Primary words for each subreddit!

```
=============== Predicted Scores ==================

PPV (Positive Predictive Value/Precision): 0.825
NPV (Negative Predictive Value)          : 0.827
Accuracy    : 0.8266
```

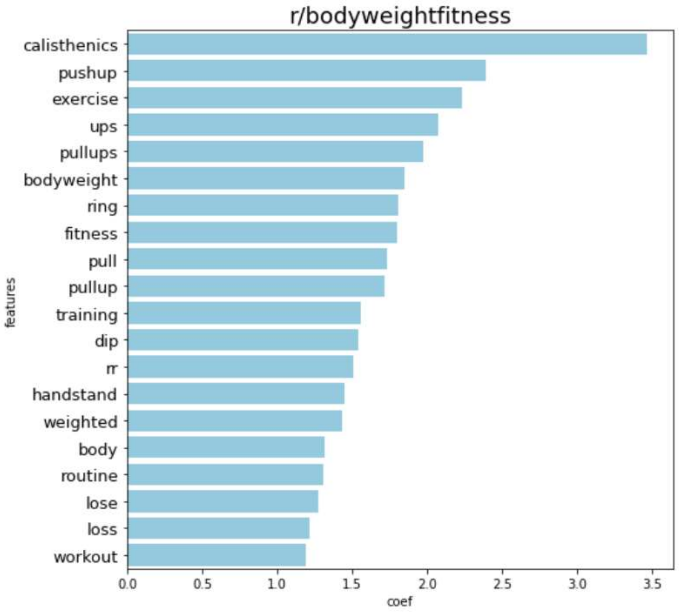


Mix bag



gain, calorie, lifting, protein

Mix bag

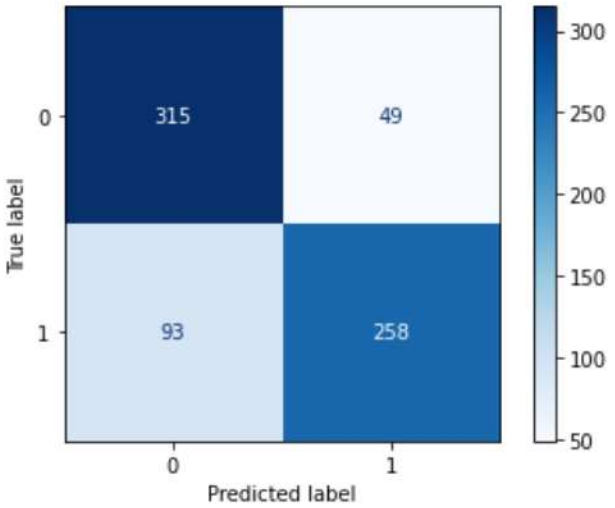


calisthenics, ring, handstand

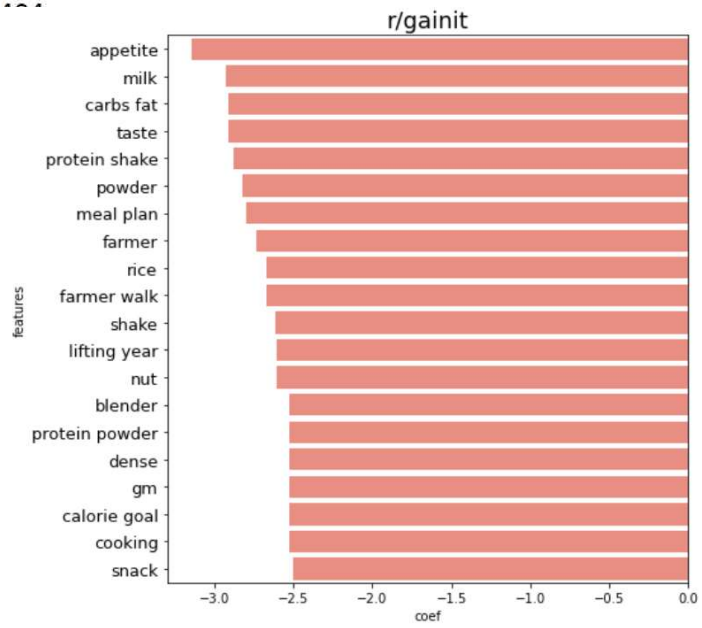
MODELING: NAÏVE BAYES W CVEC

Primary words for each subreddit!

```
=============== Predicted Scores ==================

PPV (Positive Predictive Value/Precision): 0.8[illegible]
NPV (Negative Predictive Value)          : 0.7
Accuracy    : 0.8014
```

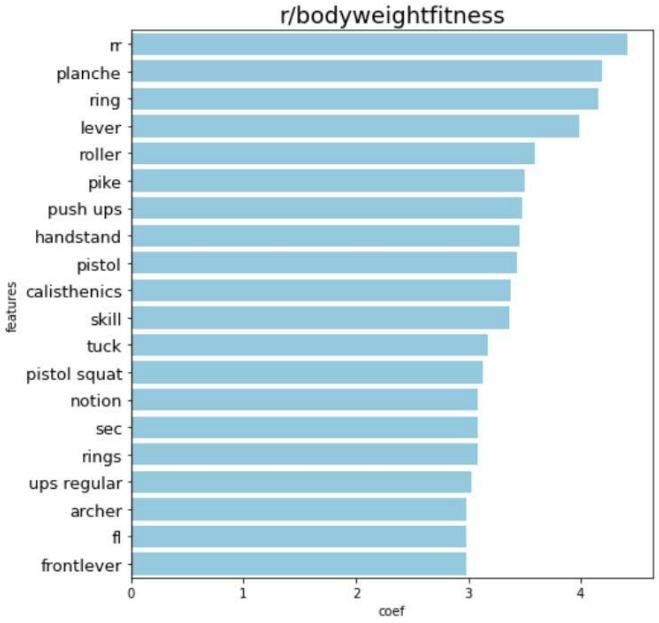


Diet & nutrition



Appetite, protein, meal plan

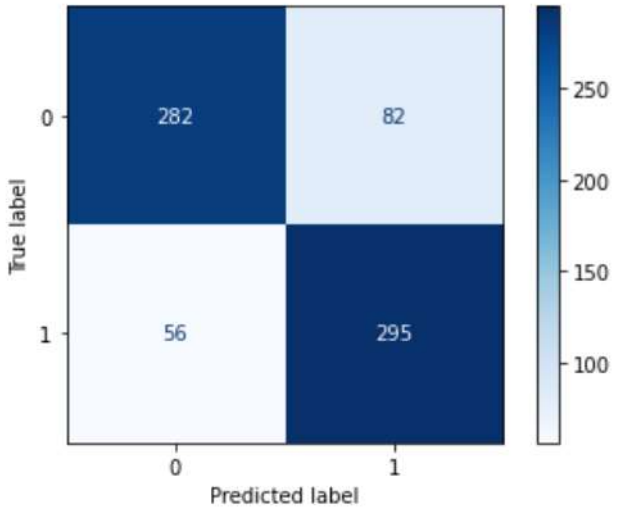
Advance techniques

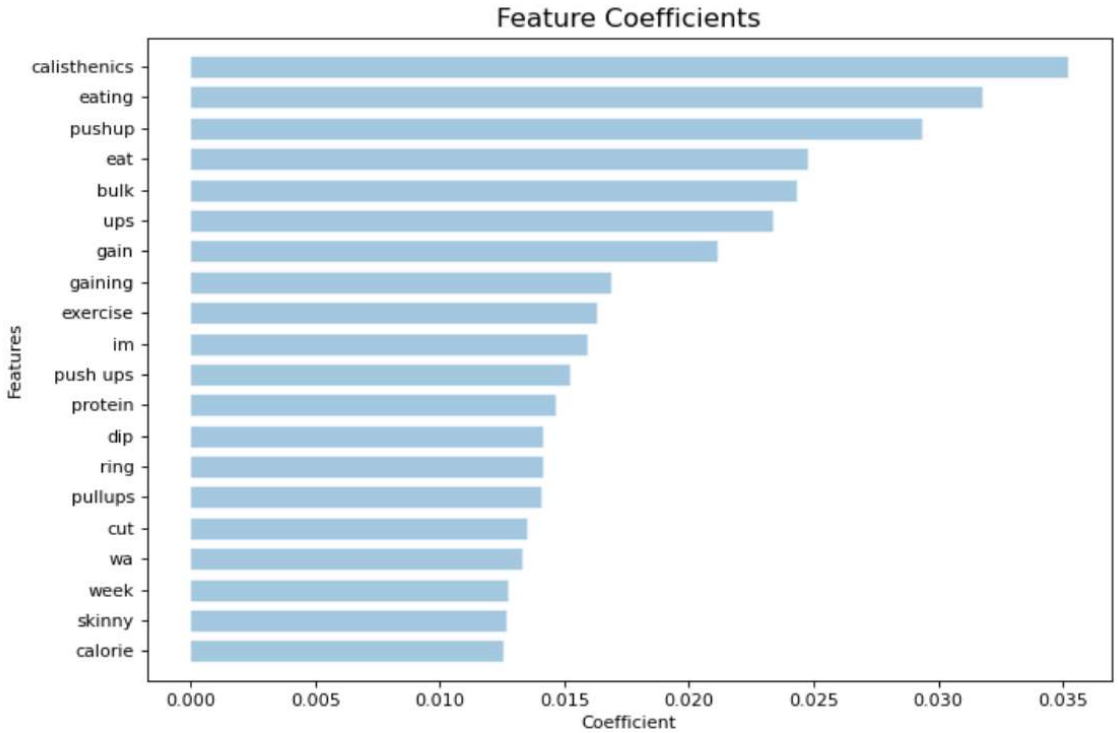


rr, planche, lever, pistol squat

MODELING: RANDOM FOREST W TVEC

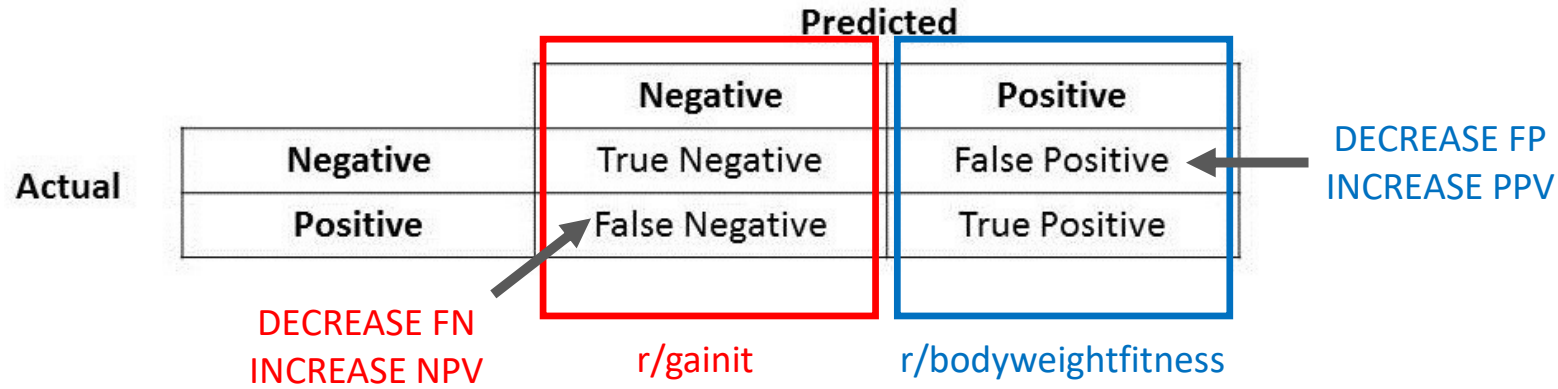
```
=============== Predicted Scores ==================

PPV (Positive Predictive Value/Precision): 0.7825
NPV (Negative Predictive Value)          : 0.8343
Accuracy   : 0.807
```





PERFORMANCE SUMMARY

	Estimator	Transformer	Train	Test	Best_score	PPV	NPV	Test_Acc
0	LogisticRegression()	cvec	0.9599	0.7930	0.8067	0.7544	0.8418	0.7930
1	LogisticRegression()	tvec	0.9174	0.8266	0.8184	0.8252	0.8279	0.8266
2	MultinomialNB()	cvec	0.8483	0.8014	0.7871	0.8404	0.7721	0.8014
3	MultinomialNB()	tvec	0.8880	0.8168	0.8016	0.8901	0.7691	0.8168
4	(DecisionTreeClassifier(max_depth=5, max_featu...	cvec	0.8296	0.8098	0.7941	0.7851	0.8373	0.8098
5	(DecisionTreeClassifier(max_depth=5, max_featu...	tvec	0.8385	0.8070	0.7937	0.7825	0.8343	0.8070



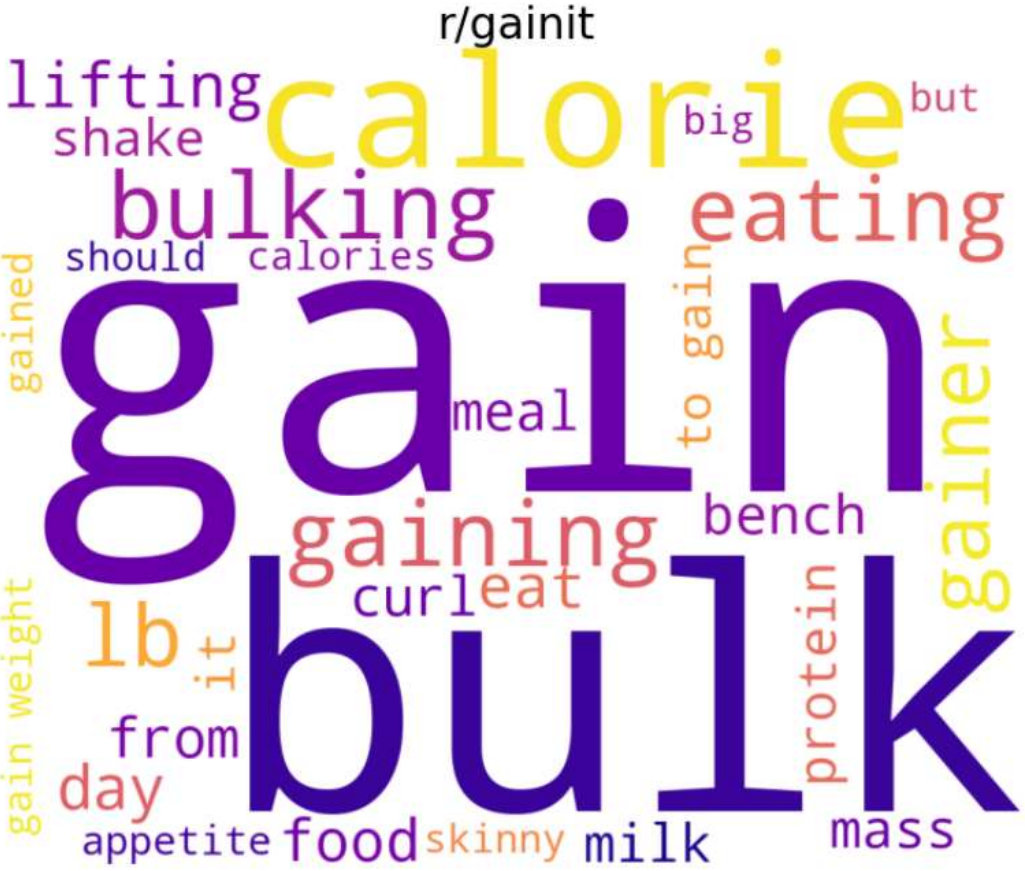
CONCLUSION

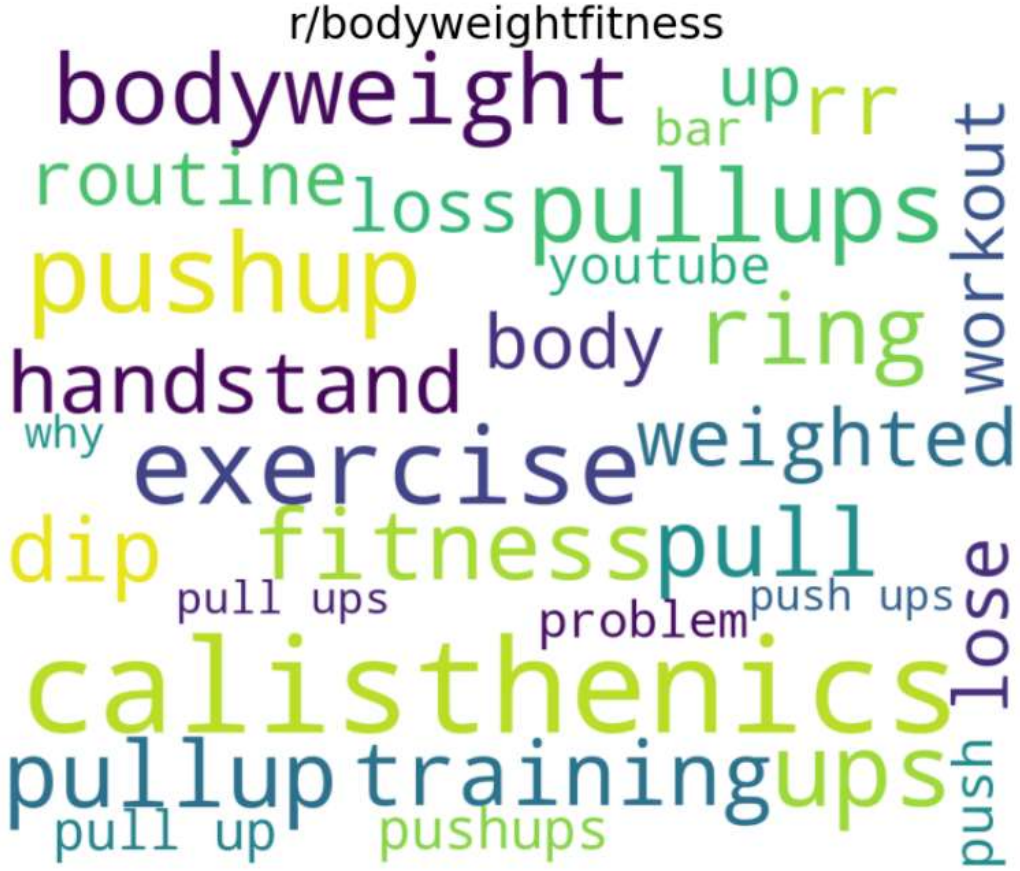
- Production model: Logistic Regression with TF-IDF Vectorizer
- Accuracy, PPV, NPV scores above 0.8
- Some posts could belong to either subreddits

MODEL LIMITATIONS

1. Fine-tune further on production model, avoid overfitting
2. Explore other complex estimators that might have the potential to be more accurate

THANK YOU


r/gainit
lifting
shake
calorie
big
but
bulking
eating
should
calories
gain
gained
meal
to gain
gainer
gaining
bench
bulk
curl
eat
lb
it
protein
from
day
appetite
food
skinny
milk
mass
gain weight


r/bodyweightfitness
bodyweight
up
rr
bar
routine
loss
pullups
pushup
youtube
body
ring
workout
handstand
why
exercise
weighted
dip
fitness
pull
pull ups
problem
push ups
lose
calisthenics
pullup
training
ups
push
pull up
pushups