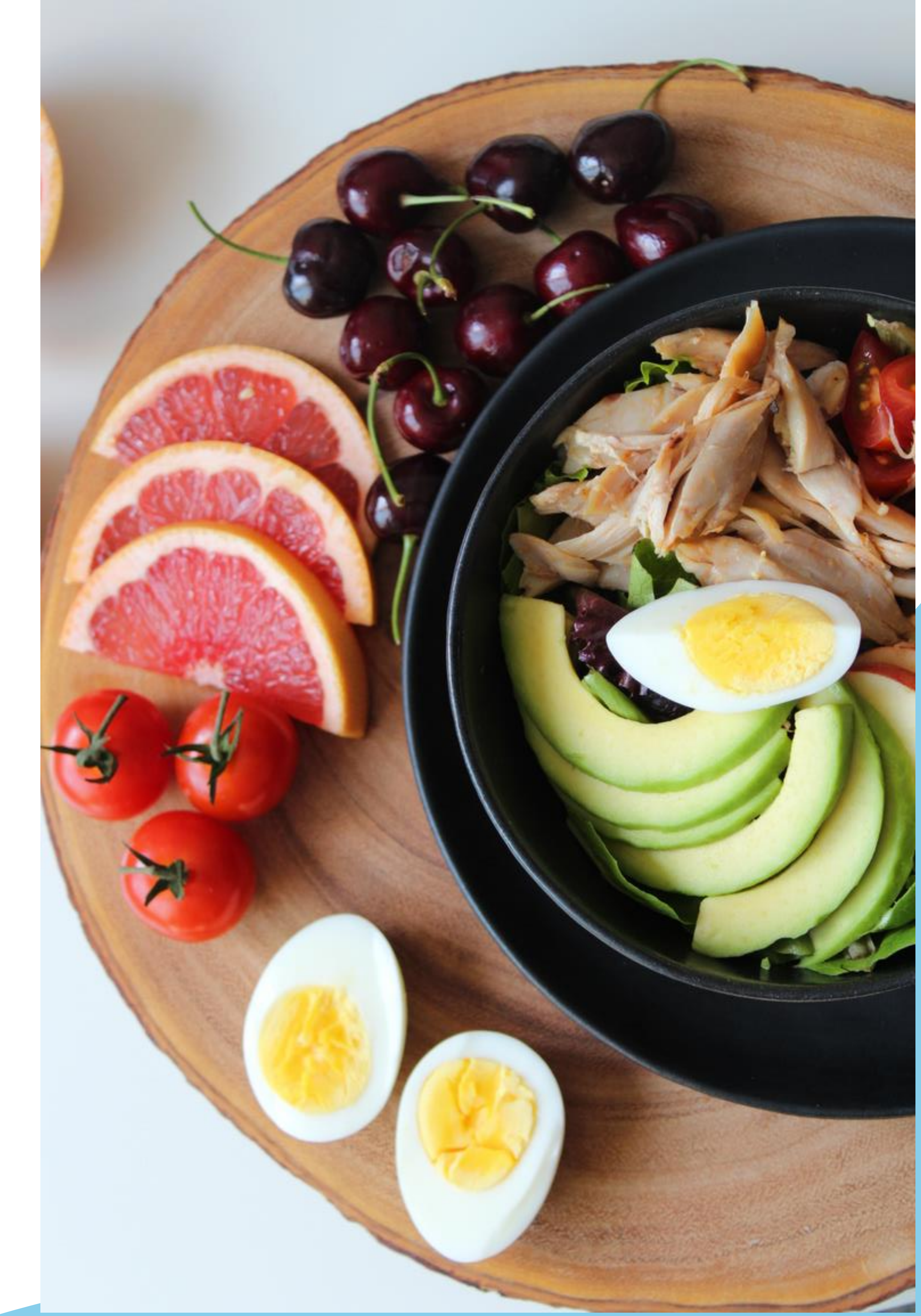
Project 3 Classifying Subreddit Posts

By: Rohazeanti

Contents

- What Do Our Posts Look Like?
- Problem Statement
- Gather & Clean Data
- Explore the Data
- Frequent Words
- Model
- Model Evaluation
- Top Features
- False Positive Posts
- Improvement and Recommendation



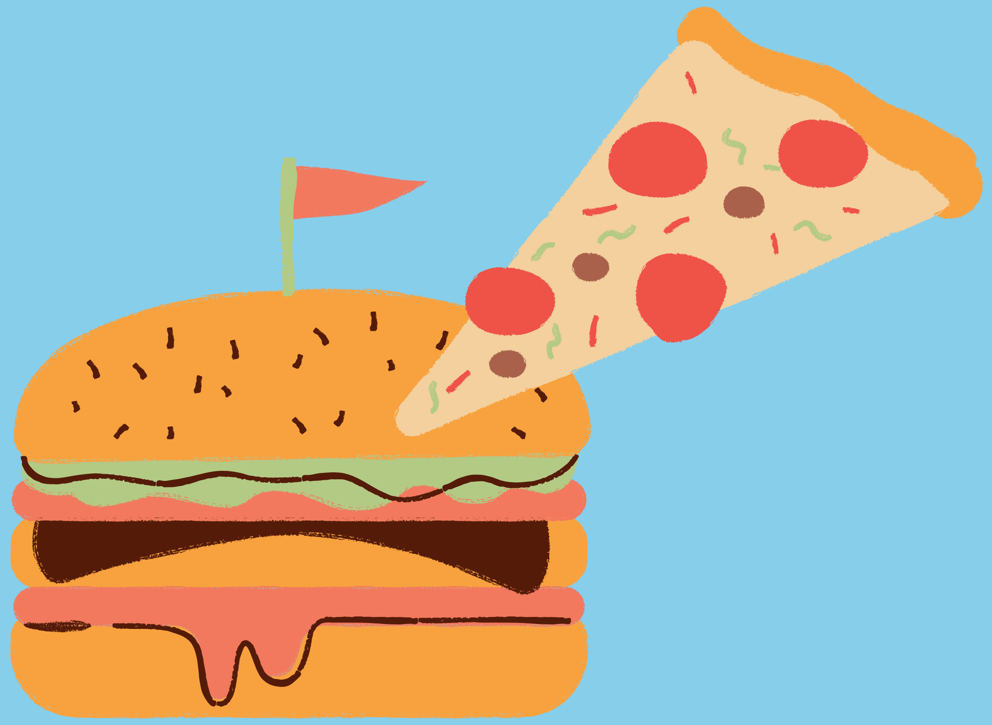
Can you guess which subreddit the posts below came from?

r/exercise or r/diet

How long will it take me to lose 8% body fat?

How to lose weight fast?

Is it normal to feel muscle tension in my quads when I squat?



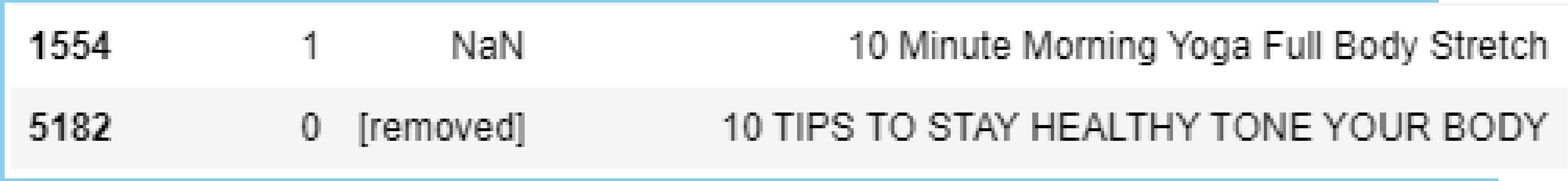
The Problem

Creating a classification model that can distinguish which of two subreddits a post belongs to.

Scrape posts from 2 subreddit forums and a Natural Language Processing model that can accurately identify which Subreddit forum a post belongs to.

Gather & Clean Data

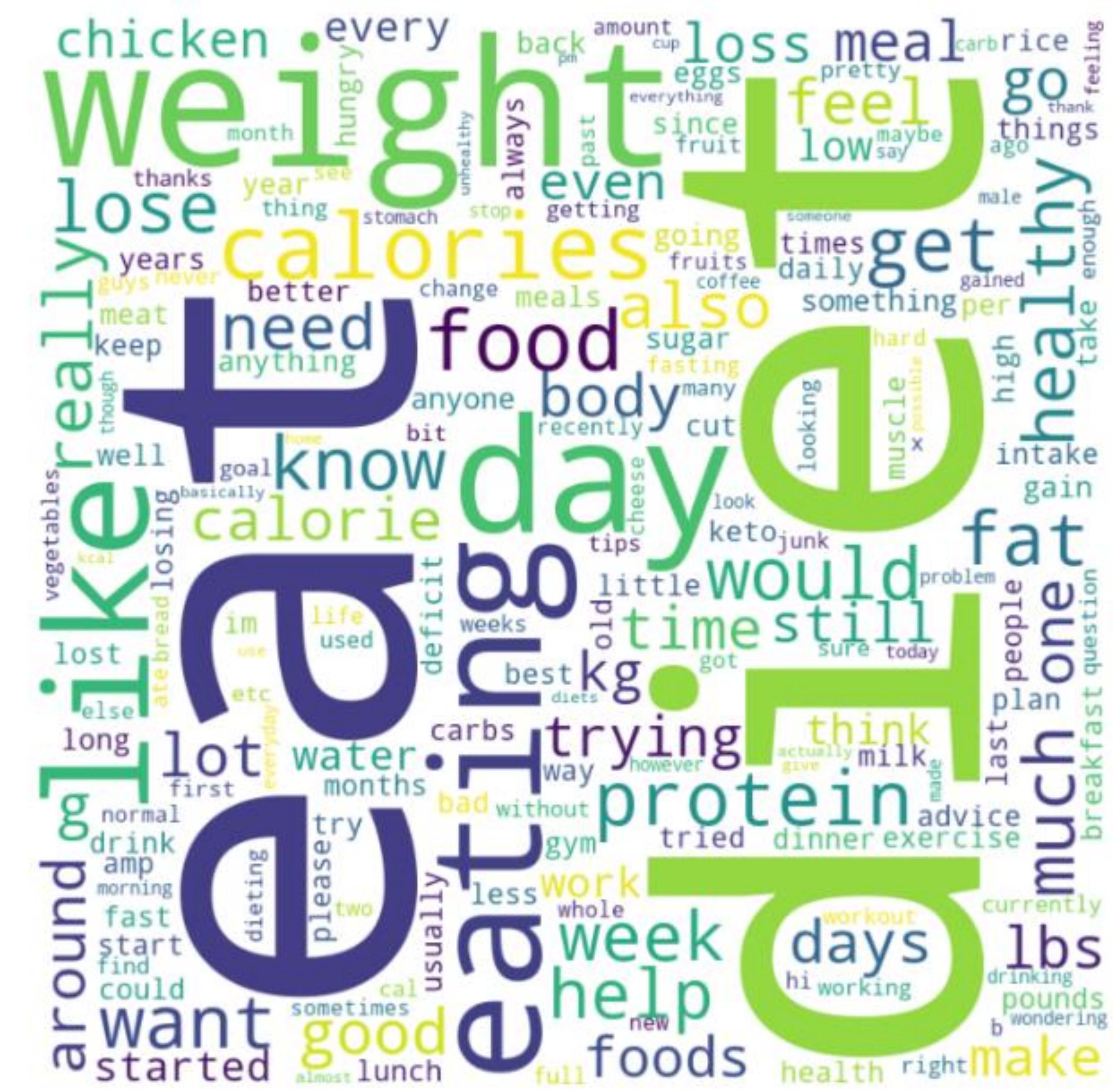
- Pull 3000 posts data using Reddit's API for each subreddit
- Missing values and data imputed by Reddit [removed],[deleted]



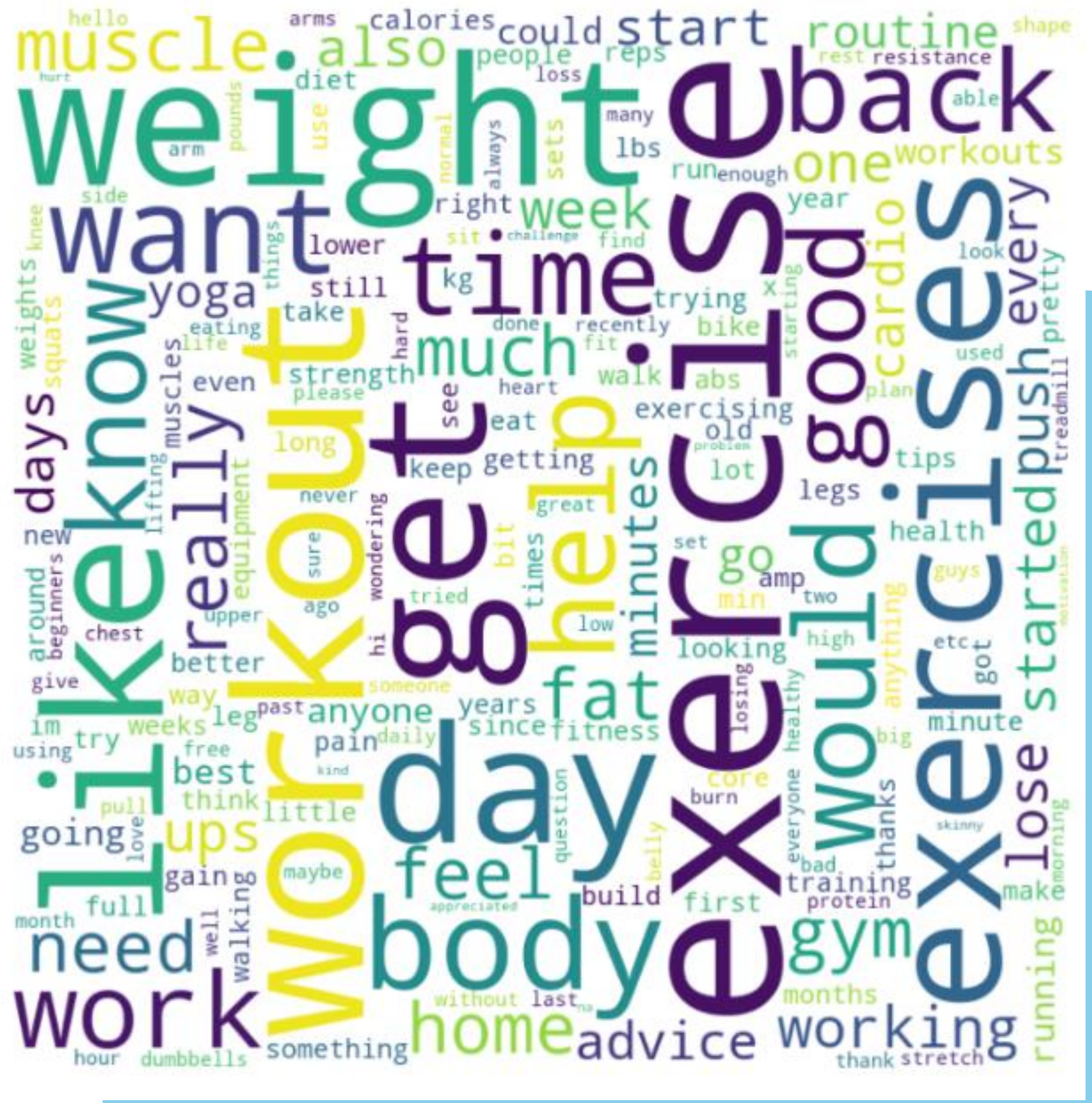
1554	1	NaN	10 Minute Morning Yoga Full Body Stretch
5182	0	[removed]	10 TIPS TO STAY HEALTHY TONE YOUR BODY

- Combine "title" and "selftext", create new feature "title_text"
- Drop duplicates
- Remove hyperlink, non-letters
- Create new features (number of characters, number of words)
- Drop rows with <2 number of characters
- Final number of rows: 5817

Explore The Data

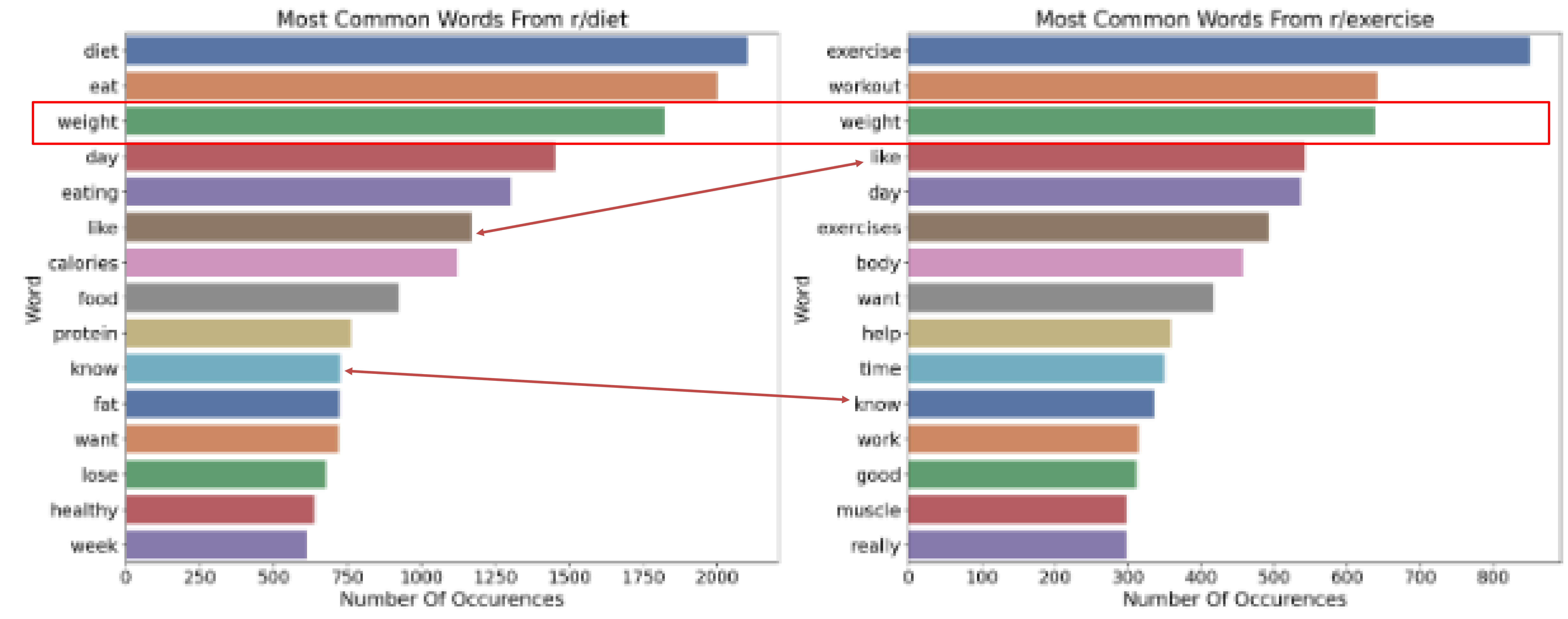


r/diet



r/exercise

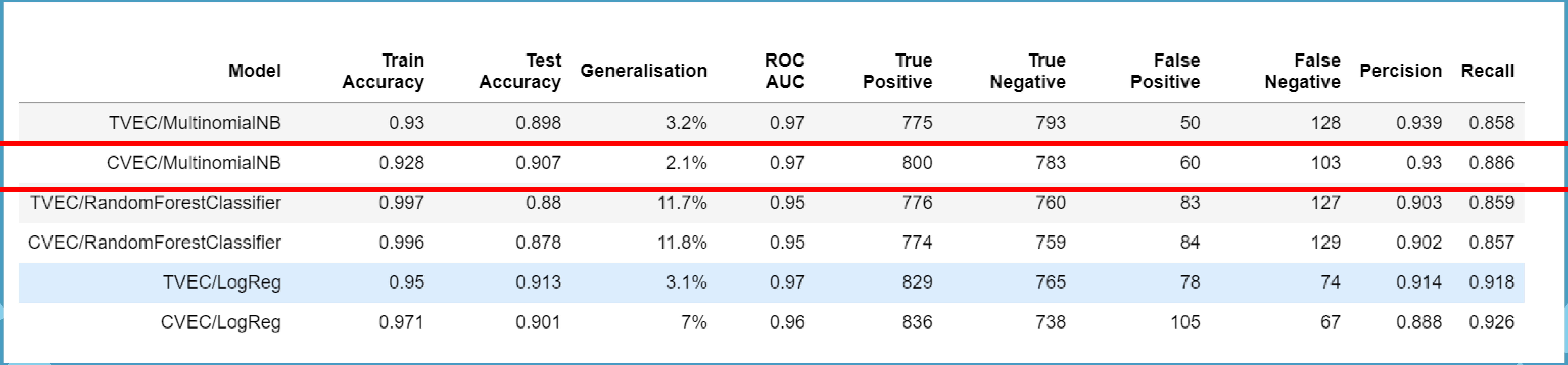
Most Common Words



Model the Data

- Splitting X and y
- Two Vectorizers
  - Countvectorizers
  - TfidfVectorizer
- Classifiers
  - Naive Bayes, Random Forest, LogisticRegression
- Hyperparameters arguments
  - 'stop_words': [None, 'english']
  - 'ngram_range': [(1, 1), (1, 2)]
  - 'max_df': [.85, .9, .95]
  - 'min_df': [2, 4, 6]
  - 'tf__max_features': [1000, 2000, 3000]
- GridSearchCV & Cross-Validation

Summary of Classification Results

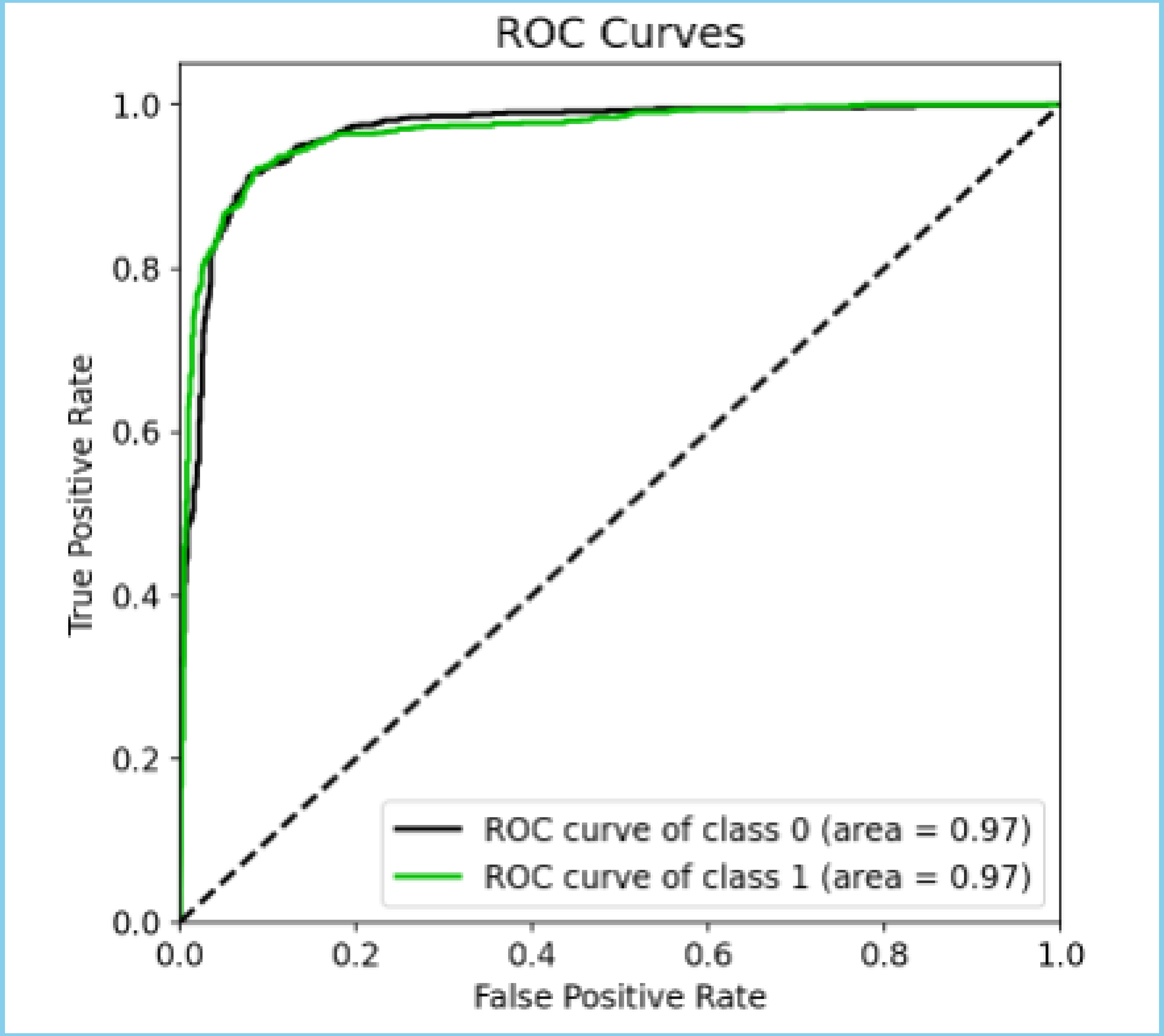


Model	Train Accuracy	Test Accuracy	Generalisation	ROC AUC	True Positive	True Negative	False Positive	False Negative	Percision	Recall
TVEC/MultinomialNB	0.93	0.898	3.2%	0.97	775	793	50	128	0.939	0.858
CVEC/MultinomialNB	0.928	0.907	2.1%	0.97	800	783	60	103	0.93	0.886
TVEC/RandomForestClassifier	0.997	0.88	11.7%	0.95	776	760	83	127	0.903	0.859
CVEC/RandomForestClassifier	0.996	0.878	11.8%	0.95	774	759	84	129	0.902	0.857
TVEC/LogReg	0.95	0.913	3.1%	0.97	829	765	78	74	0.914	0.918
CVEC/LogReg	0.971	0.901	7%	0.96	836	738	105	67	0.888	0.926

- CVEC/MultinomialNB
- Slightly overfit but acceptable
- Minimise False Positive, Optimise Precision
- Minimise False Negative, Optimise Recall

- stop_words='english’
- ngram_range=(1, 1)
- max_df=0.85
- min_df=2
- max_features=3000

ROC Curves



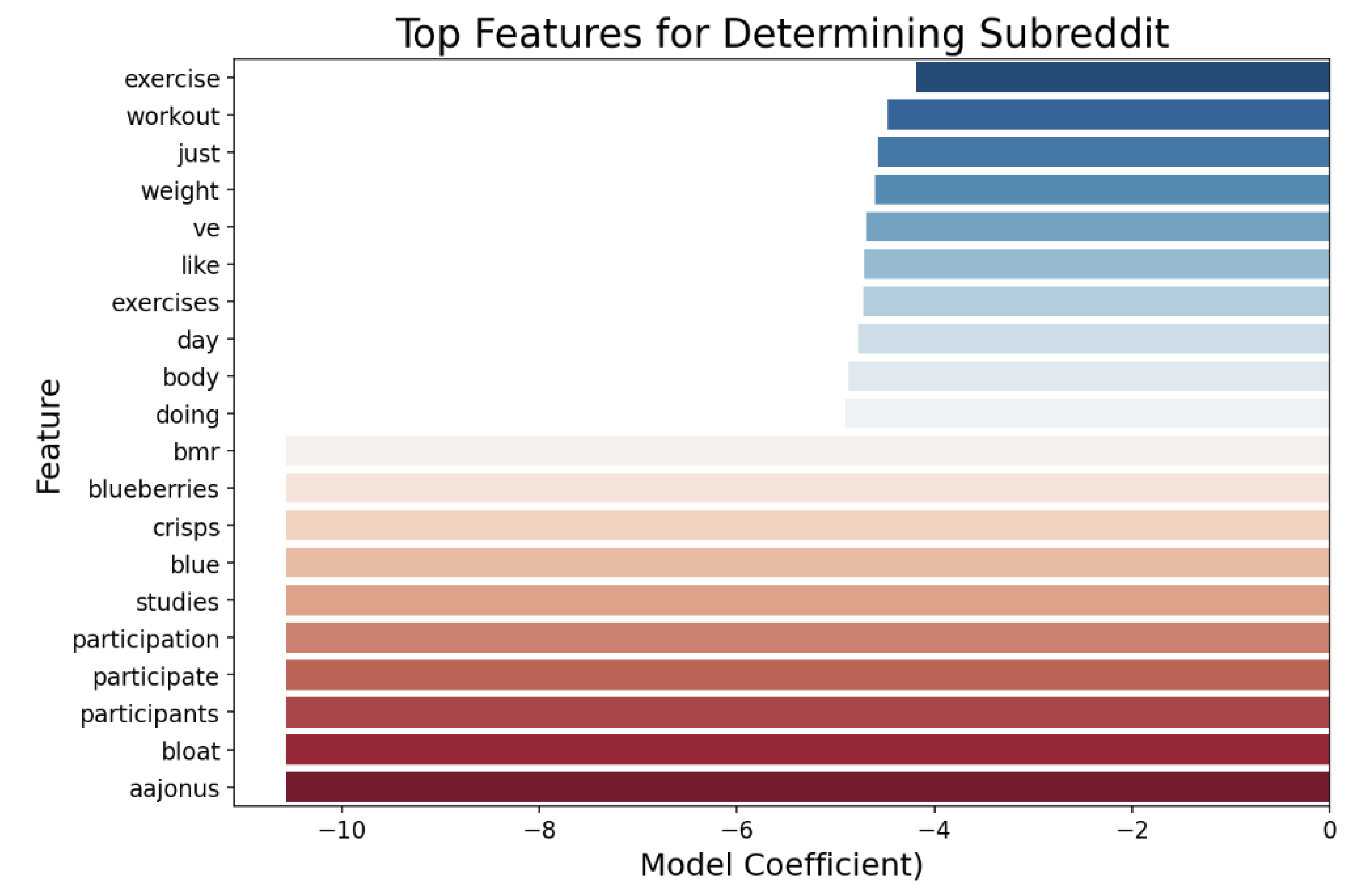
- AUC ROC score 97%

- **Sensitivity/Recall**: Of all the posts that were truly "diet", how many did we label?
- Recall value : 0.886

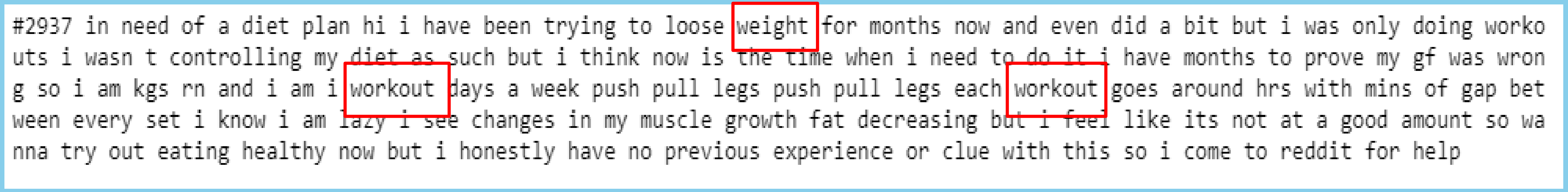
- **Precision**: Of all the post labelled as "diet", how many are actually “diet”? This precision rate relates to the low false positive rate.
- Precision value: 0.93

Top Features

- Words and jargons unique to topic post:
- "ve" = ventilation
- "bmr" = basal metabolic rate,
- "aajonus" = American alternative nutritionist and food-rights activist who focused on raw foods



False Positive Posts



```
#2937 in need of a diet plan hi i have been trying to loose weight for months now and even did a bit but i was only doing worko
uts i wasn t controlling my diet as such but i think now is the time when i need to do it i have months to prove my gf was wron
g so i am kgs rn and i am i workout days a week push pull legs push pull legs each workout goes around hrs with mins of gap bet
ween every set i know i am lazy i see changes in my muscle growth fat decreasing but i feel like its not at a good amount so wa
nna try out eating healthy now but i honestly have no previous experience or clue with this so i come to reddit for help
```

r/diet posts were identified as r/exercise

Improvement & Recommendation

- Look into misclassified posts
- Include more stopwords
- Try other advanced model

Other uses

- Overcome evolution of subreddits over time
  - Using probabilities of each post to understand the overall direction of subreddit
  - Observe new languages and topics
- Increase diversity of topics or refocus conversations

Thank you