

CROSSFIT

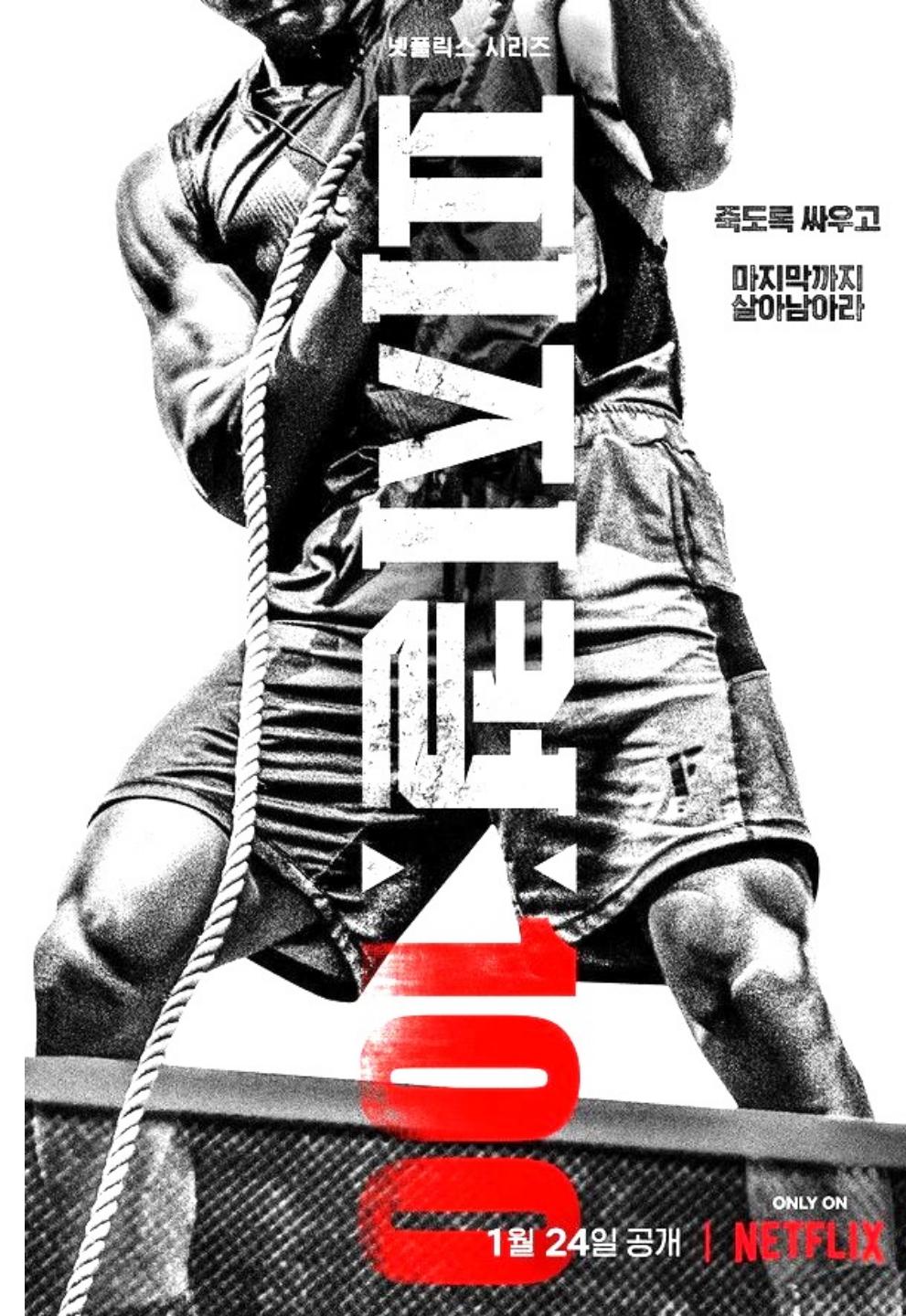
vs

BRAZILIAN JIU-JITSU

Using natural language processing
to classify a gym's forum posts

넷플릭스 시리즈

죽도록 싸우고
마지막까지
살아남아라

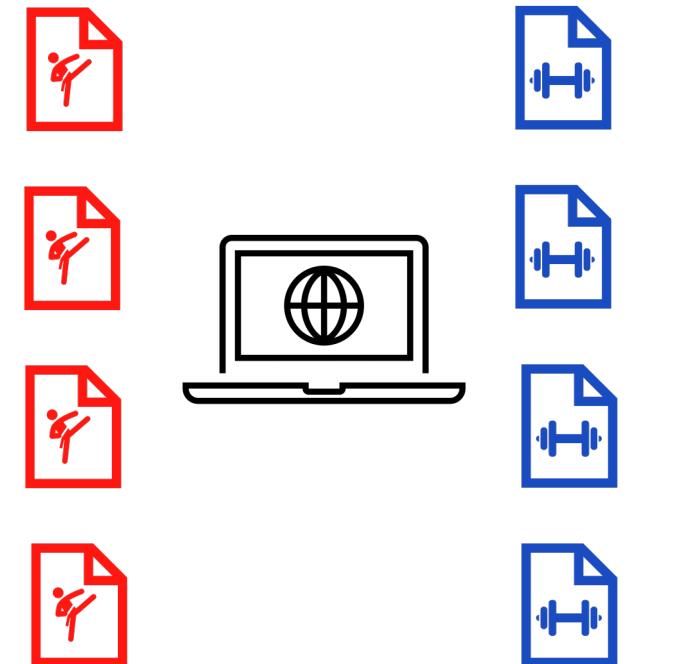
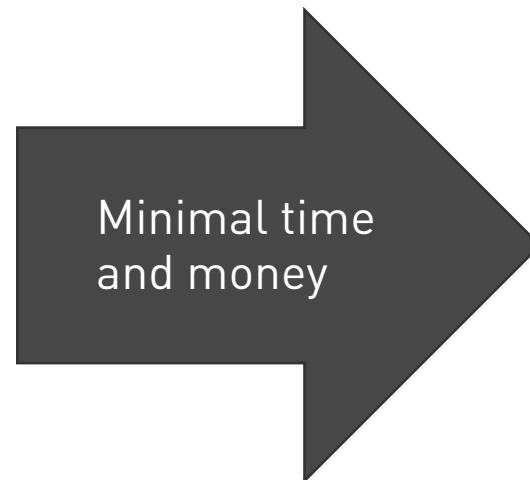
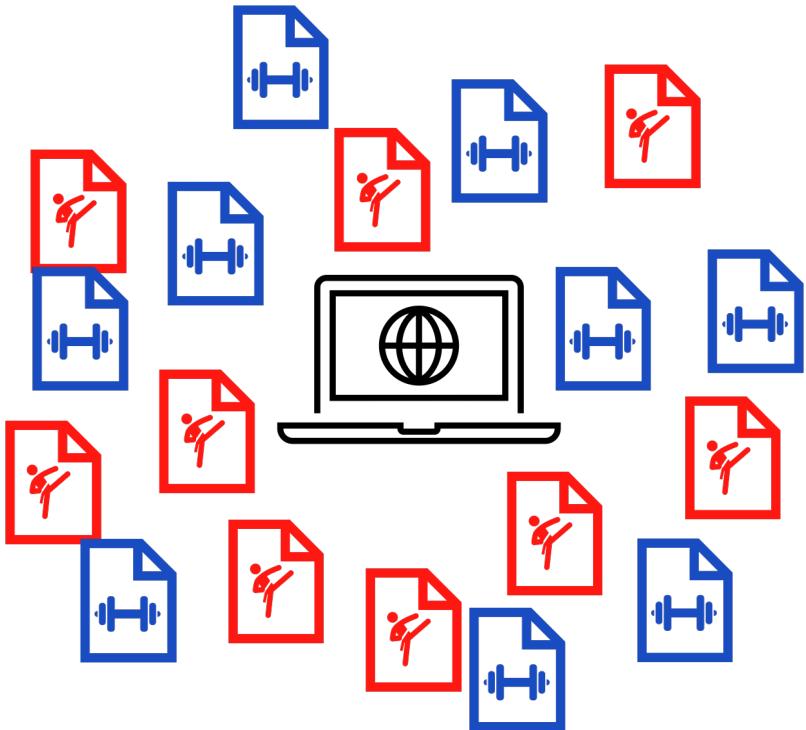




CONTENT

- The business problem
- What is Crossfit and BJJ?
- Data collection and preprocessing
- Modeling and evaluation
- Recommendations and next steps

THE BUSINESS PROBLEM





Combination of weightlifting, gymnastics, and metabolic conditioning exercises

Circuit in a set timeframe

High intensity

Community-focused

CROSSFIT

BRAZILIAN JIU-JITSU

Martial art emphasizing leverage and technique over brute strength. Most spar with a gi as the uniform.

Aim to take opponents down to the ground and use submissions, joint locks, and chokes. Belt system for ranking by technical knowledge and skill



BRAZILIAN JIU-JITSU

FULL MOUNT GUARD
OMOPLATA
HALF GUARD SIDE CONTROL
JIU-JITSU
LEG LOCK SIDE CONTROL
KIMURA GI TRIANGLE
NORTH SOUTH GI
SIDE CONTROL CHOKE NO-GI
REAR NAKED CHOKE NO-GI

WORLD CLASS FITNESS IN 100 WORDS
EAT MEAT & VEGETABLES
NUTS & SEEDS • SOME FRUIT
LITTLE STARCH • NO SUGAR.
KEEP INTAKE TO LEVELS THAT WILL SUPPORT EXERCISE BUT NOT BODY FAT.

PRACTICE & TRAIN
MAJOR LIFTS:
DEADLIFT
CLEAN | SQUAT
PRESSES
CLEAN & JERK
SNATCH
SIMILARLY,
MASTER THE BASICS
OF GYMNASTICS:
PULL-UPS
DIPS | ROPE CLIMBS
PUSH-UPS | SIT-UPS
PRESSES TO HANDSTANDS
PIROUETTES | FLIPS
SPLITS | HOLDS

BIKE, RUN, SWIM, ROW, ETC.
HARD & FAST

FIVE OR SIX DAYS PER WEEK,
MIX THESE ELEMENTS IN AS MANY COMBINATIONS & PATTERNS
AS CREATIVITY WILL ALLOW.

ROUTINE IS THE ENEMY.

KEEP WORKOUTS
SHORT & INTENSE. | REGULARLY LEARN
AND PLAY
NEW SPORTS.

COACH GREG GLASSMAN | DESIGN BY WWW.CECILYBREEDING.COM

CROSSFIT

DATA COLLECTION

The image contains two side-by-side screenshots of Reddit subreddits. The top screenshot shows the 'CrossFit on reddit' community (r/crossfit). It features a banner image of several CrossFit athletes, including one with 'FRASER 26'. Below the banner, the subreddit header includes the 'CrossFit GAMES OPEN' logo, 'Joined', and a bell icon. The main interface shows 'Posts' selected, a 'Create Post' button, and sorting options 'Hot', 'New', 'Top'. A pinned post by u/Flowseidon9 titled 'CrossFit Open 23.3 Discussion Thread' has 69 upvotes and 664 comments. The sidebar displays the 'About Community' section, which describes the subreddit as a place for discussion of CrossFit, functional fitness, weightlifting, and related topics. It notes 707k members, 451 online users, and a rank of Top 1% for size. The bottom screenshot shows the 'Brazilian Jiu-Jitsu' community (r/bjj). It features a banner image of a person in a red and white gi. The subreddit header includes the 'r/bjj' logo, 'Joined', and a bell icon. The main interface shows 'Posts' selected, a 'Create Post' button, and sorting options 'Hot', 'New', 'Top'. A pinned post by u/AutoModerator titled 'White Belt Wednesday' has 23 upvotes and 632 comments. The sidebar displays the 'About Community' section, which describes r/bjj as a platform for discussing BJJ training, techniques, news, competition, and advice. It notes 717k members, 1.5k online users, and a rank of Top 1% for size.

WHY REDDIT?

Open community of over 700K enthusiasts per sport, with daily discussion threads that mimic the gym's forum

GETTING DATA

Scraped via the PushShift API, collecting submissions (aka posts) from the past 4 months

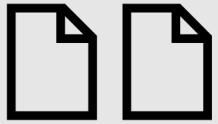
HOW MANY SUBMISSIONS?

r/crossfit: 1.4K (47%)
r/bjj: 1.6k (53%)

FINAL SCRAPED DATASET

3,110 posts

DATA CLEANING



Dropped
duplicated posts

[Removed]

Excluded
moderated posts
and spam

T ➤ t

Lowercase all text

Crossfit,
jujitsu

Removed relevant
keywords

“I”, “me”,
“your”, “our”

Removed commonly
used words (i.e.
stopwords)

😊 grinning
_ face

Converted emojis
to equivalent
descriptions

Caring ➤ Care

Reduced words to
root form using
lemmatization

/n
#

Removed
punctuation, digits
and line breaks

“Views? TIA!”

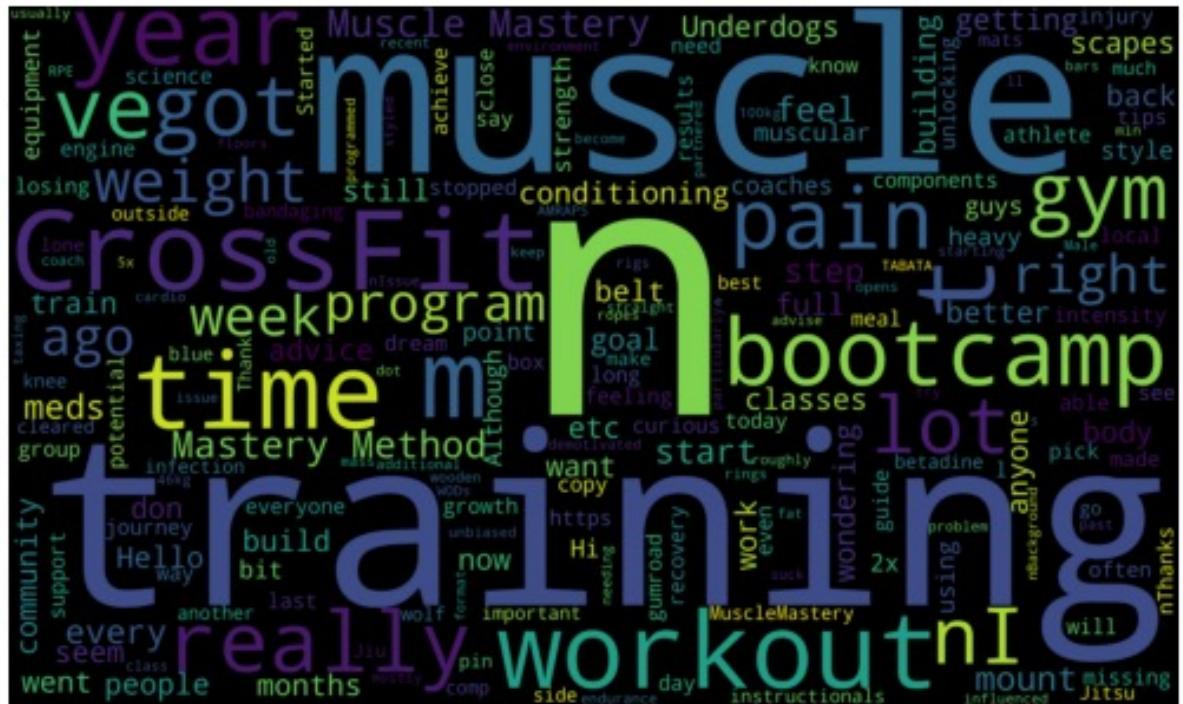
Excluded posts
with fewer than 5
words

advice, gym,
injury

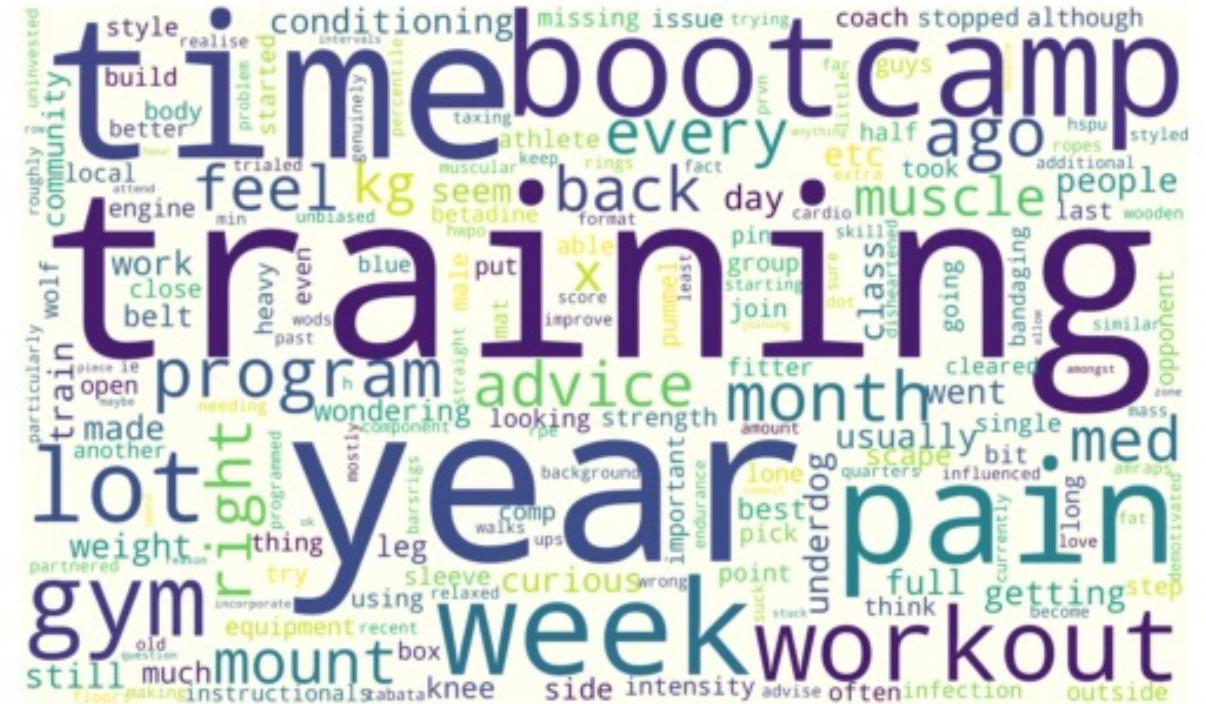
Convert sentences to
words through
tokenization

EXPLORATORY ANALYSIS

BEFORE CLEANING



AFTER CLEANING

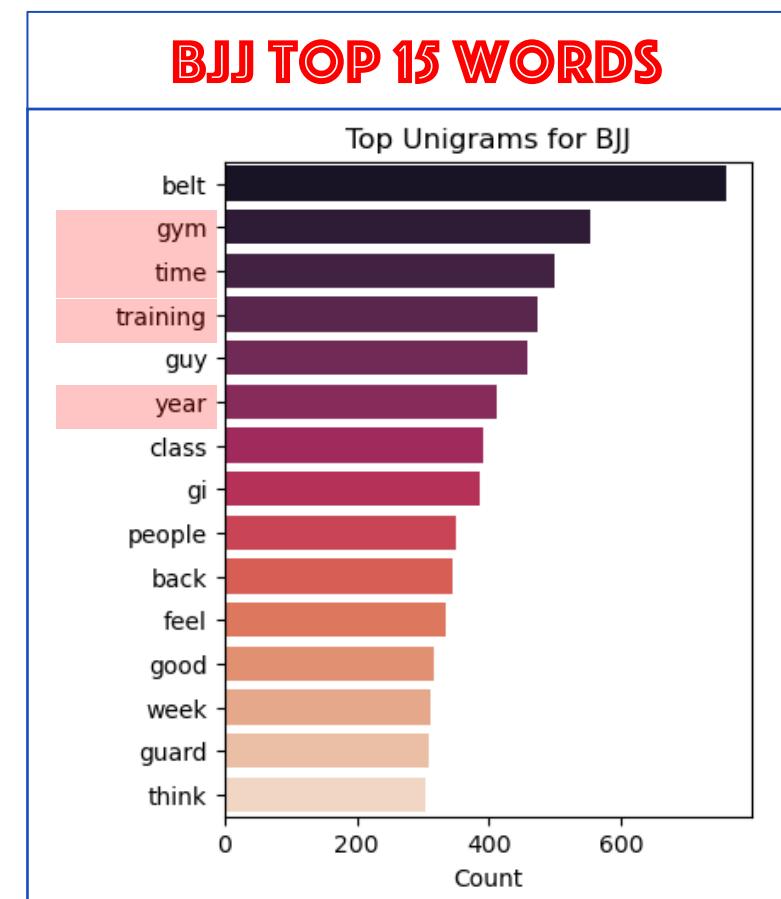
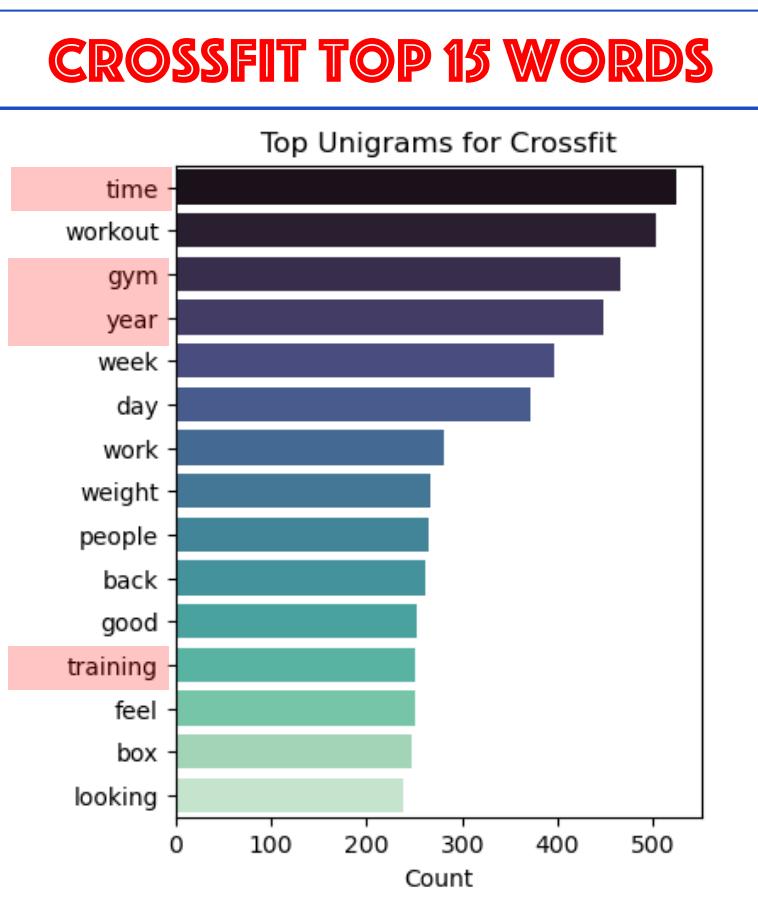


EXPLORATORY ANALYSIS

WORD
COUNT
98710

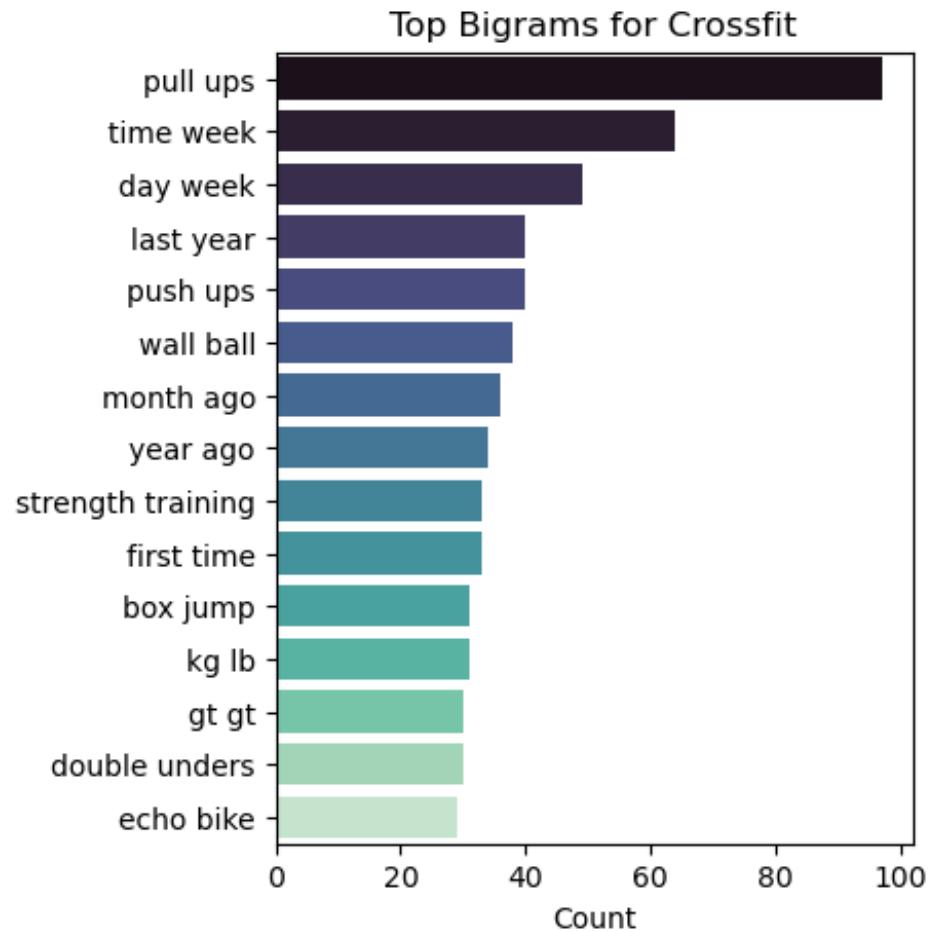
AVERAGE
SUBMISSION LENGTH
36

CLASS DISTRIBUTION
49% vs **51%**
Crossfit BJJ

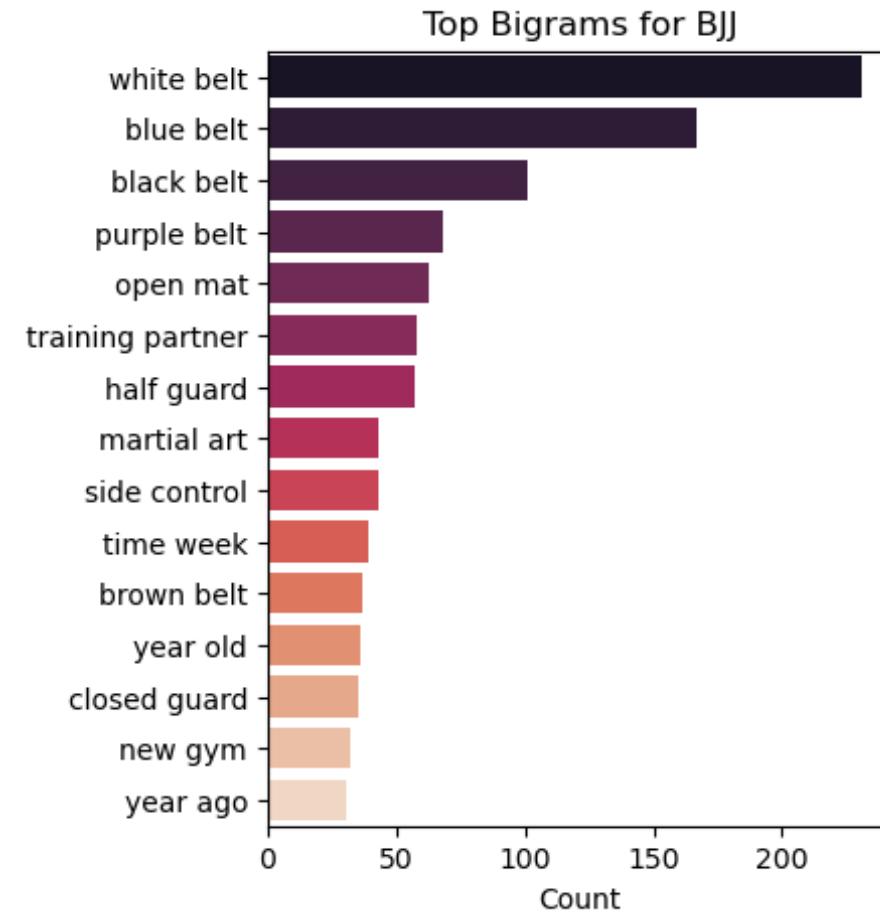


EXPLORATORY ANALYSIS

CROSSFIT TOP 15 PHRASES



BJJ TOP 15 PHRASES



PREPROCESSING AND MODELLING

CountVectorizer

Creates vocabulary based on count and treats every word equally

Term frequency-inverse document frequency (TFIDF)

Creates vocabulary by accounting for the occurrence of the word in each submission vs its rarity across the entire data set

1

2

3

4

5

6

7

8

Naïve Bayes

Probability-based
Quick
History of good performance
Assumes feature independence

Logistic Regression

Common algorithm
Good for binary classification
Allows penalizing of model complexity (aka regularization), improving the model's performance on unseen data.

Support Vector Machines

Handles high dimensionality data like text well
Finds a boundary that separates two classes
Has built-in regularization

Random Forest

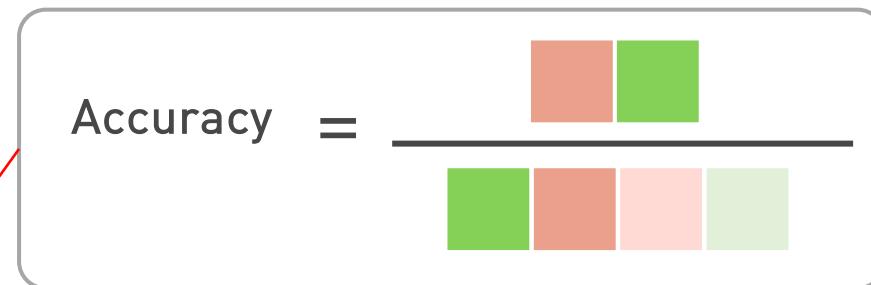
Can manage a high number of features
Uses multiple decision trees
Minimizes overfitting

EVALUATION METRICS

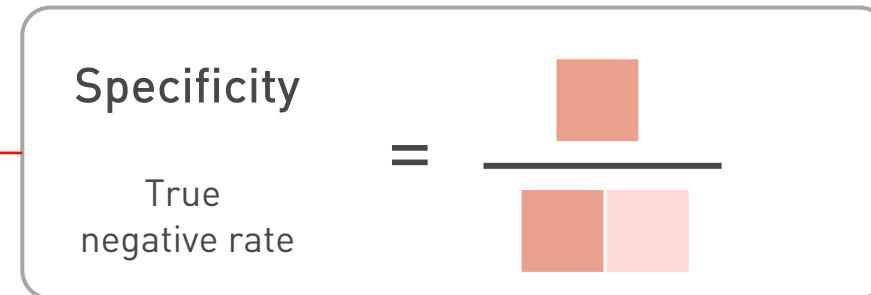
1 Business problem indicates that both classes are equally important

		PREDICTIONS	
		Crossfit	BJJ
ACTUAL VALUES	Crossfit	TN	FP
	BJJ	FN	TP

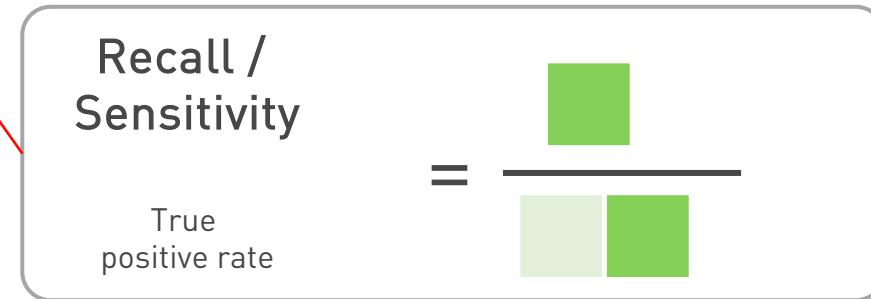
2 Our positive and negative classes have a 50-50 distribution



Suitable for balanced classes
Ideally optimize to 1



As similar as possible



MODEL RESULTS

	Model 1: Naive Bayes with Count- Vectorizer	Model 2: Logistic Regression with Count- Vectorizer	Model 3: Support Vector Machines with Count- Vectorizer	Model 4: Random Forest with Count- Vectorizer	Model 5: Naive Bayes with TFIDF	Model 6: Logistic Regression with TFIDF	Model 7: Support Vector Machines with TFIDF	Model 8: Random Forest with TFIDF
Accuracy	90%	85%	87%	90%	90%	90%	88%	90%
Specificity Correctly predicted Crossfit posts out of actual Crossfit posts	90%	88%	90%	94%	91%	91%	89%	93%
Recall Correctly predicted BJJ posts out of actual BJJ posts	90%	83%	85%	85%	88%	89%	87%	86%
Difference in specificity and recall	0%	5%	5%	9%	4%	2%	3%	7%
Degree of overfitting								

Production model

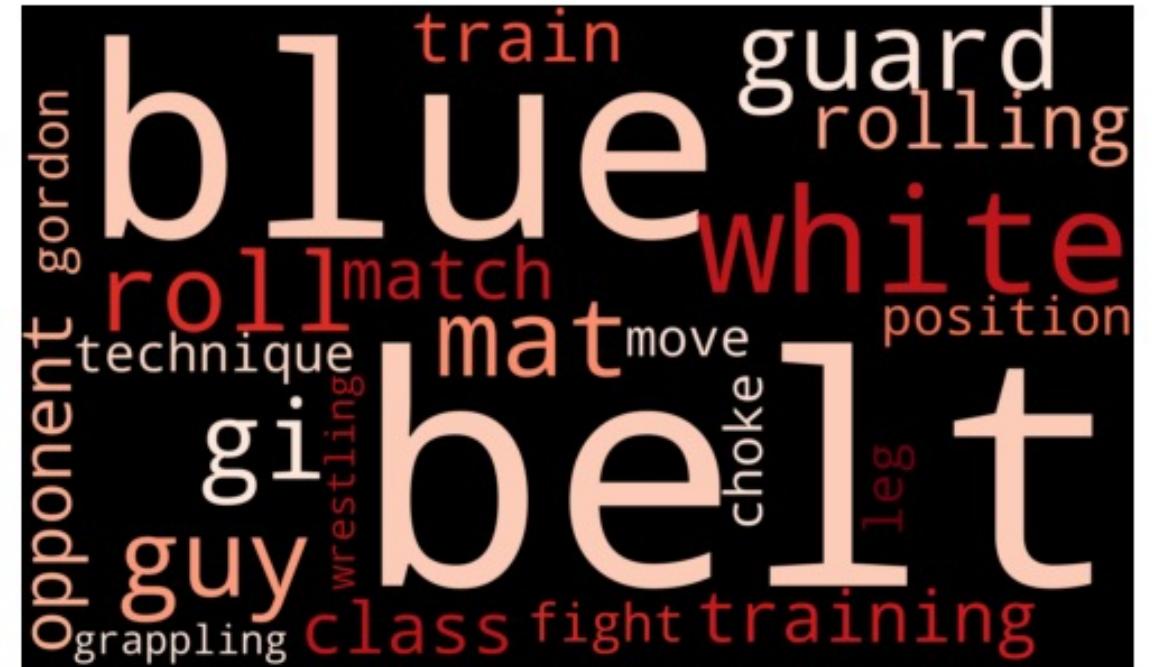
MODEL PREDICTIONS

WORDS LIKELY TO PREDICT CROSSFIT



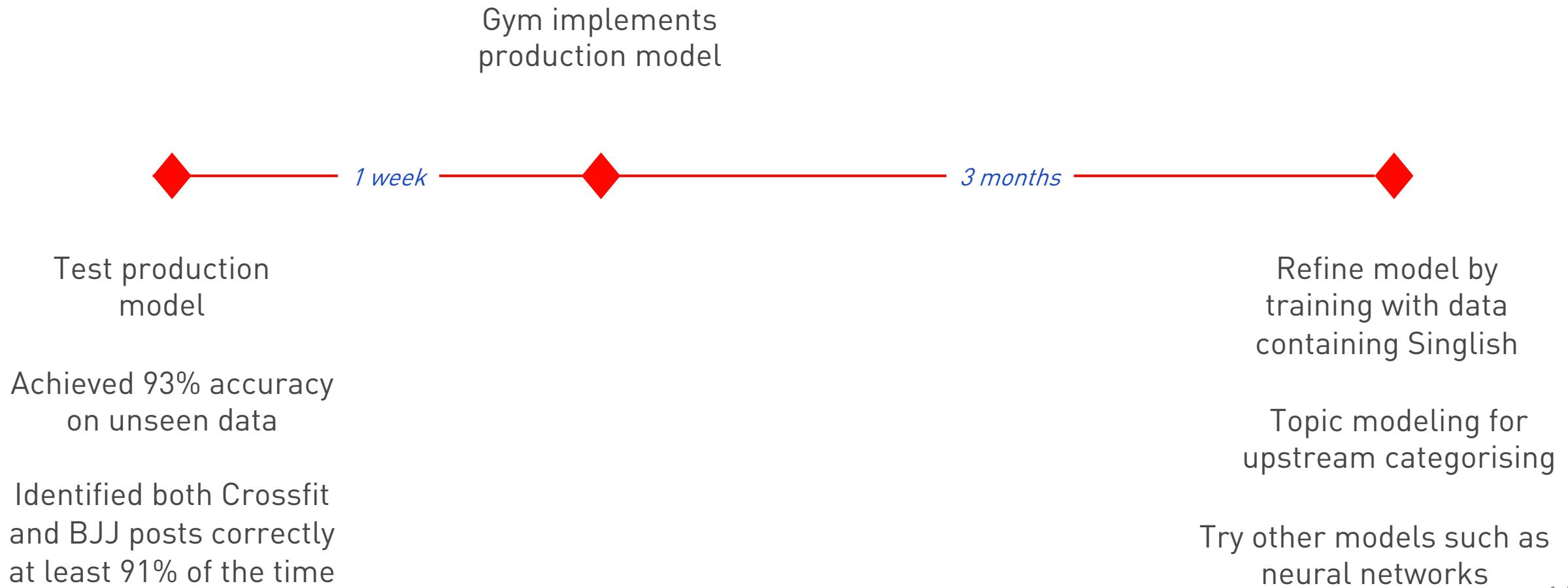
A word cloud visualization for CrossFit predictions. The words are primarily in blue and light blue, set against a black background. The most prominent words are "open", "app", "week", and "bar". Other visible words include "fitness", "movement", "box", "score", "home", "program", "wod", "squat", "lift", "muscle", "rope", "shoe", "ups", "cf", "rep", "bike", "looking", "programming", "athlete", and "rep".

WORDS LIKELY TO PREDICT BJJ



A word cloud visualization for BJJ predictions. The words are primarily in pink, red, and white, set against a black background. The most prominent words are "blue", "white", and "belt". Other visible words include "train", "guard", "rolling", "gordon", "roll", "match", "mat", "move", "position", "opponent", "gi", "guy", "wrestling", "class", "fight", "training", "choke", "leg", "technique", "move", and "oppo".

RECOMMENDATIONS & PROPOSED NEXT STEPS



THANK YOU

