CROSSFIT

vs

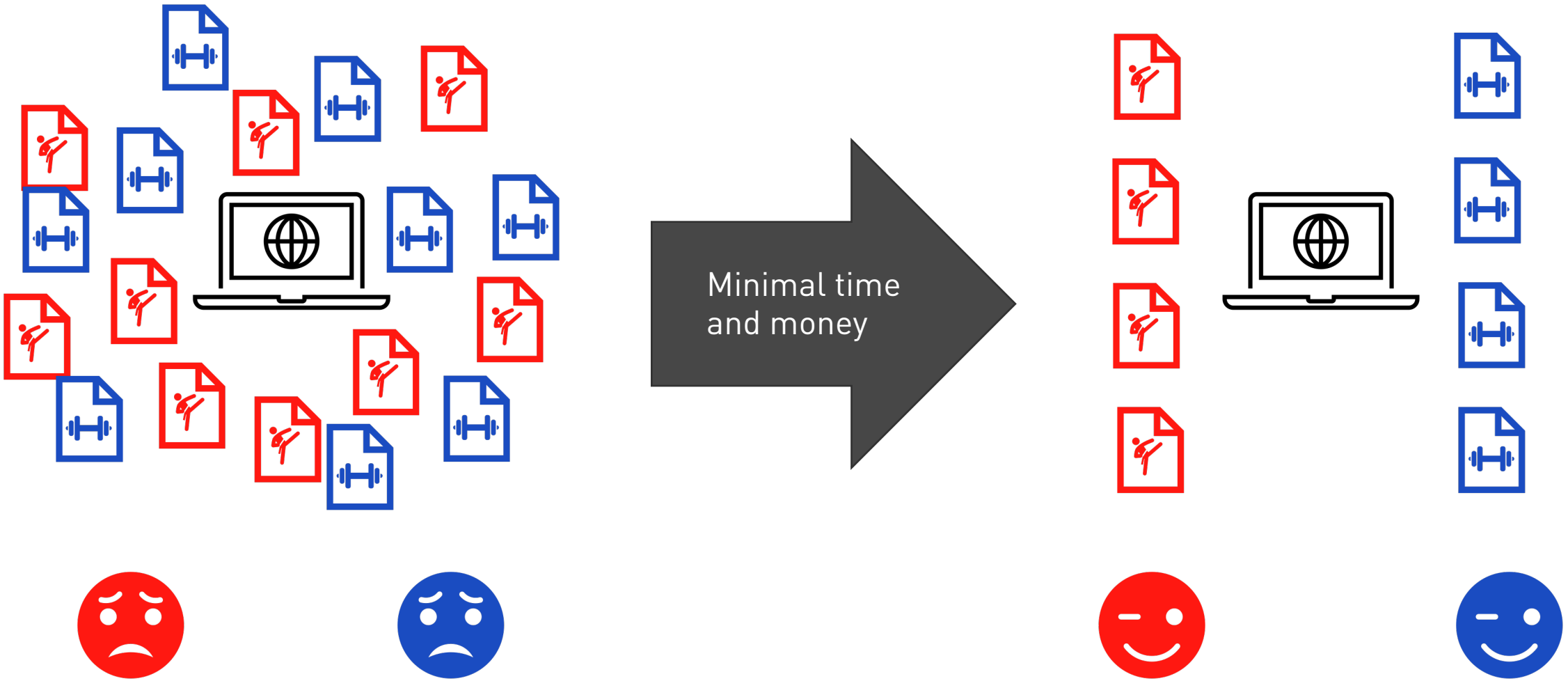
BRAZILIAN JIU-JITSU

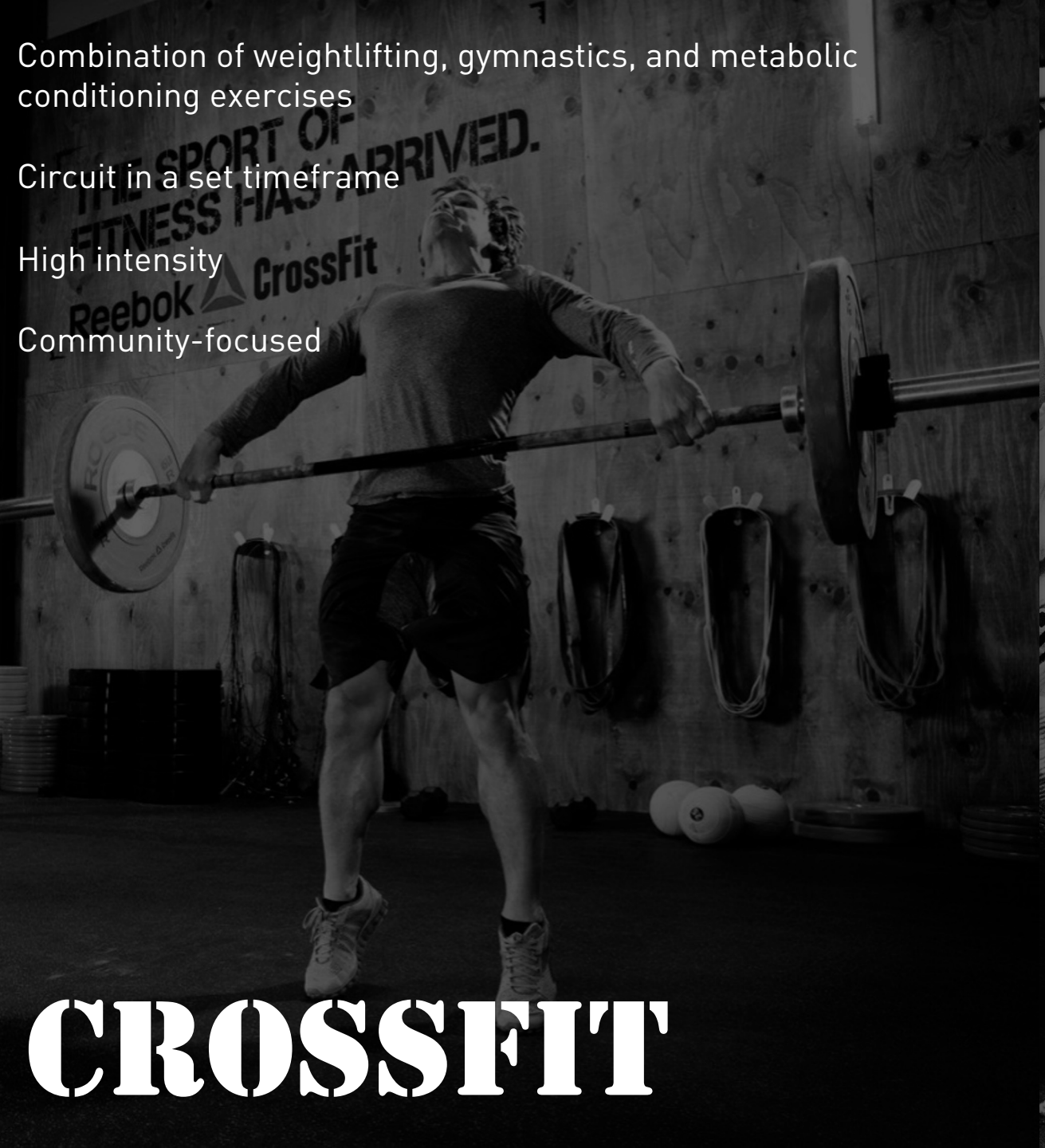
Using natural language processing to classify a gym's forum posts

CONTENT

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

THE BUSINESS PROBLEM





CROSSFIT

Combination of weightlifting, gymnastics, and metabolic conditioning exercises

Circuit in a set timeframe

High intensity

Community-focused

BRAZILIAN JIU-JITSU

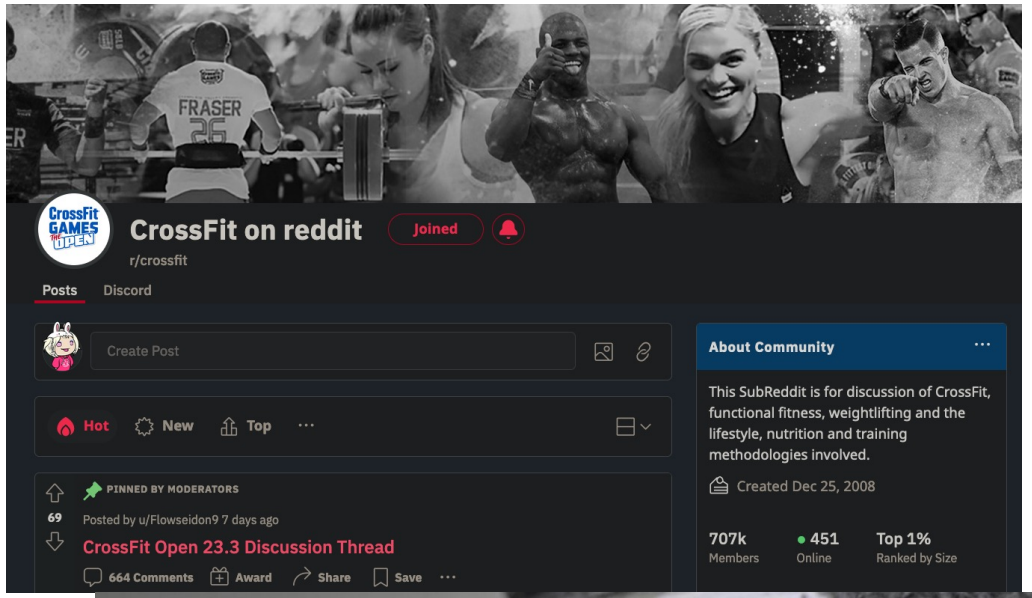
Martial art emphasizing leverage and technique over brute strength. Most spar with a gi as the uniform.

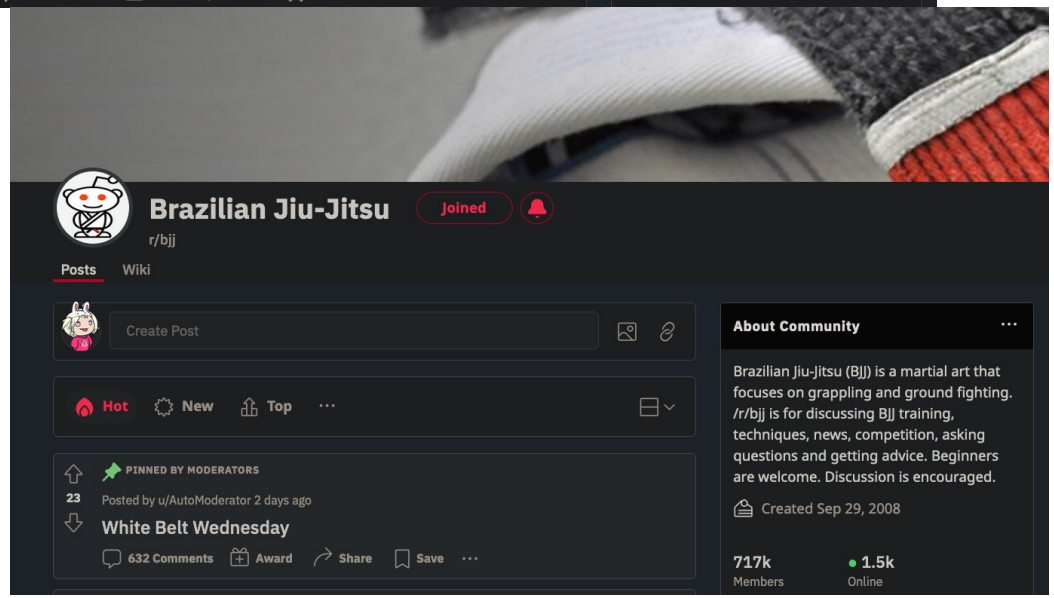
Aim to take opponents down to the ground and use submissions, joint locks, and chokes. Belt system for ranking by technical knowledge and skill

CROSSFIT

BRAZILIAN JIU-JITSU

DATA COLLECTION





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

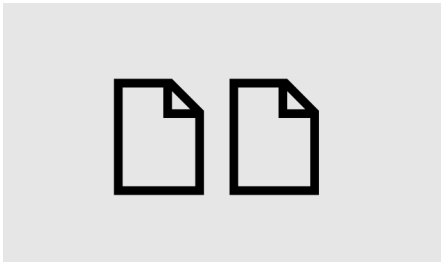
HOW MANY SUBMISSIONS?

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 punctation, digits and line breaks

“Views? TIA!”

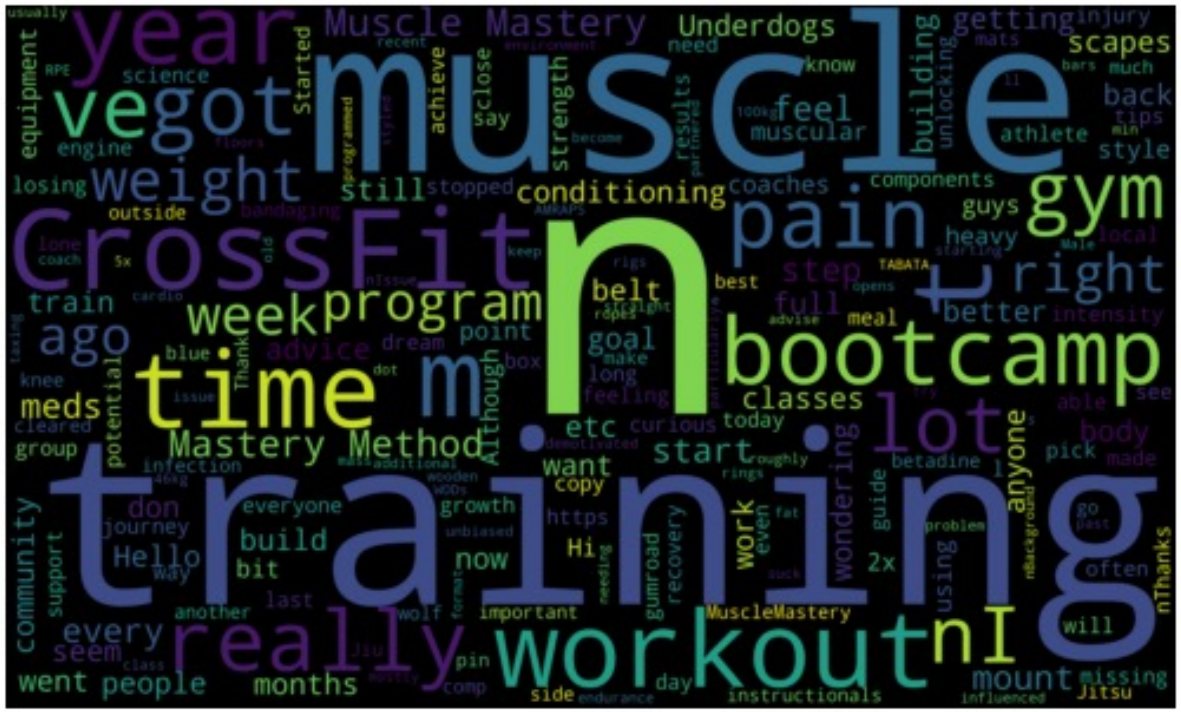
Excluded posts with fewer than 5 words

advice, gym, injury

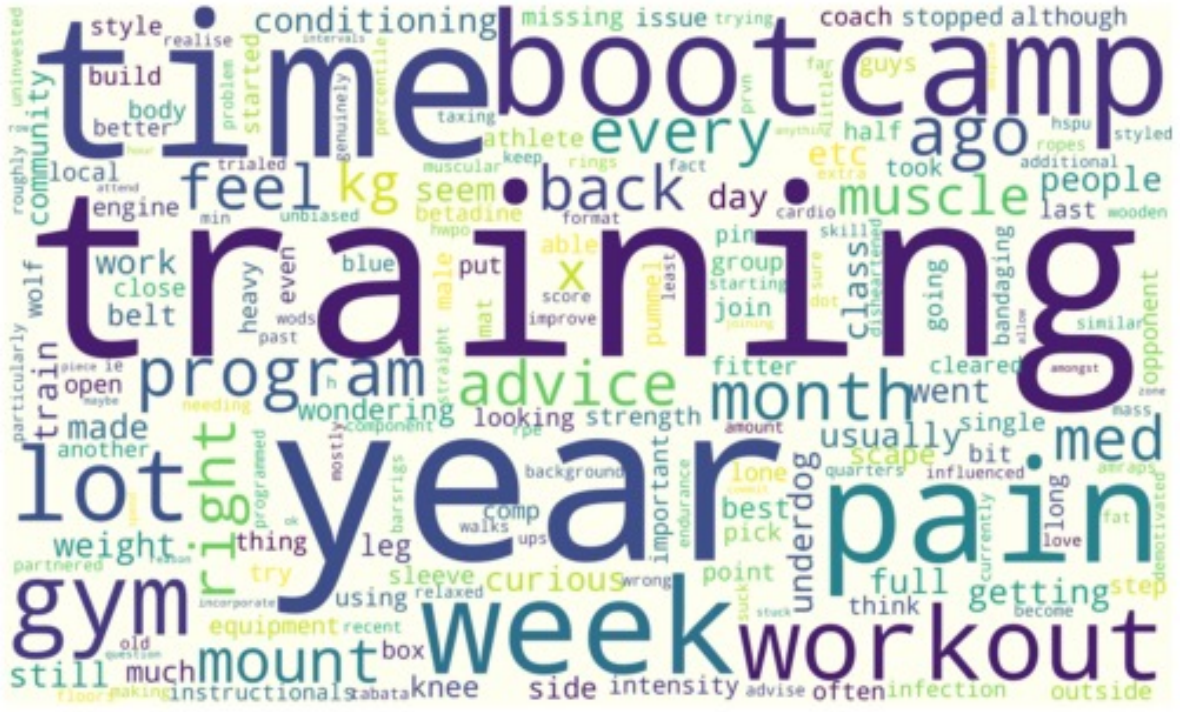
Convert sentences to words through tokenization

EXPLORATORY ANALYSIS

BEFORE CLEANING



AFTER CLEANING



EXPLORATORY ANALYSIS

WORD COUNT

98710

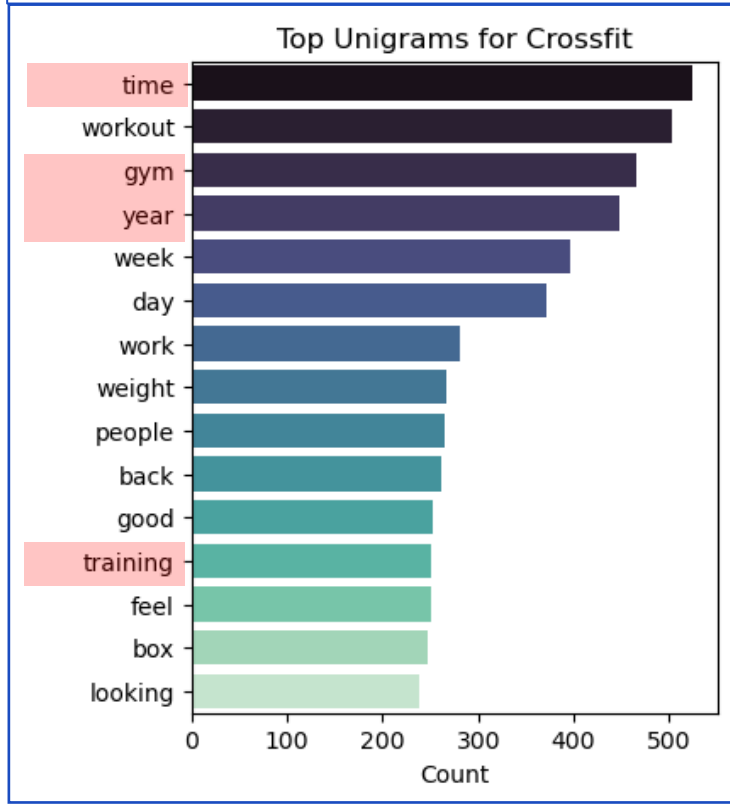
AVERAGE SUBMISSION LENGTH

36

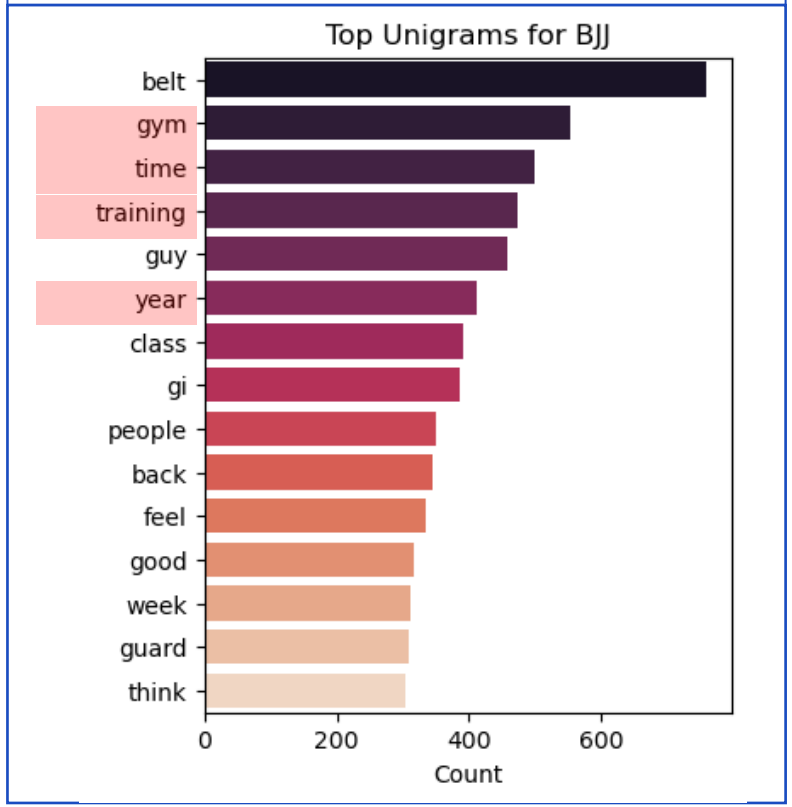
CLASS DISTRIBUTION

49% Crossfit vs 51% BJJ

CROSSFIT TOP 15 WORDS

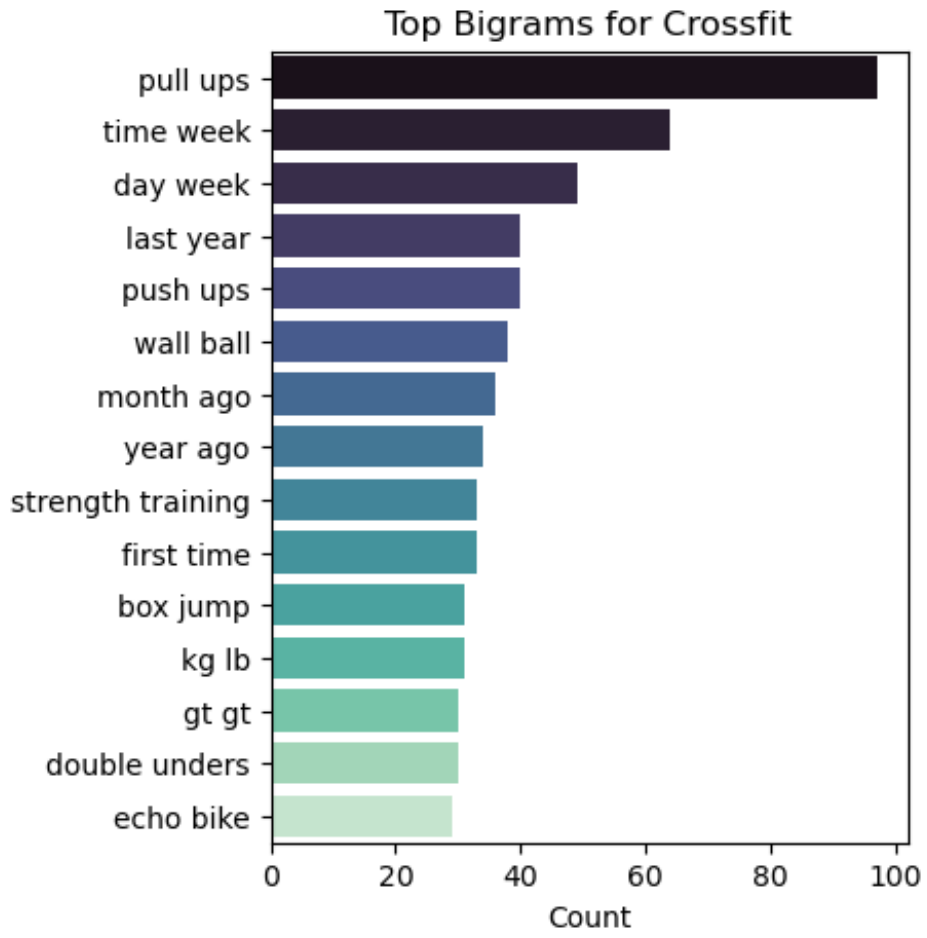


BJJ TOP 15 WORDS

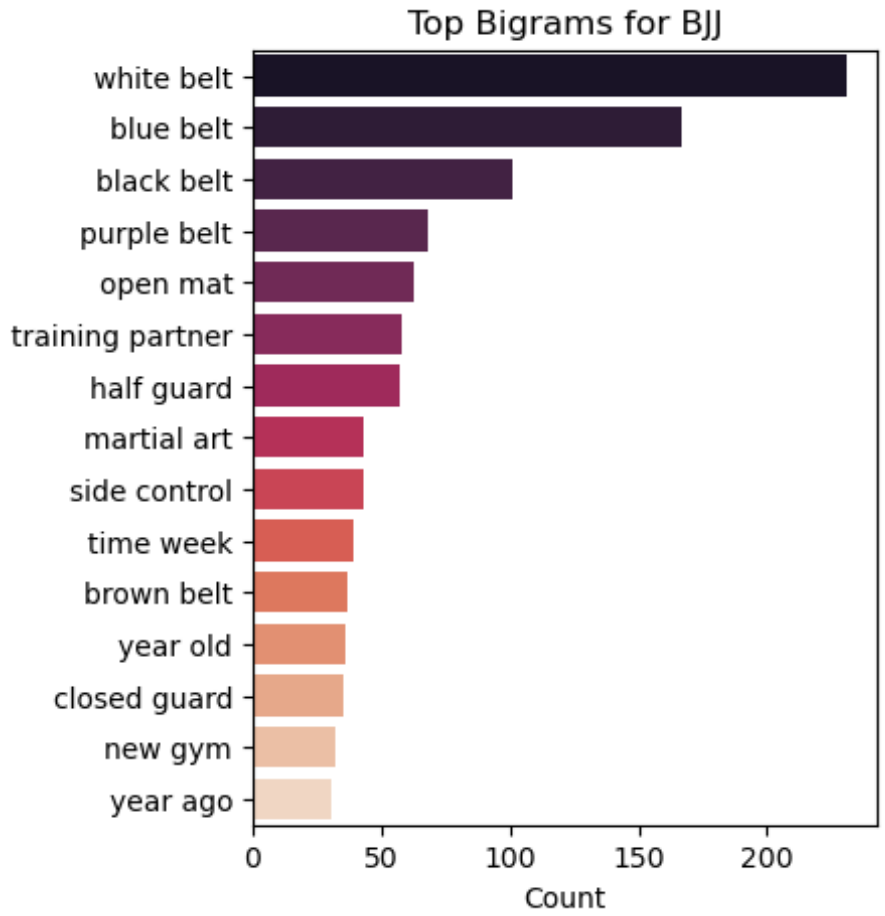


EXPLORATORY ANALYSIS

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

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

CONFUSION MATRIX

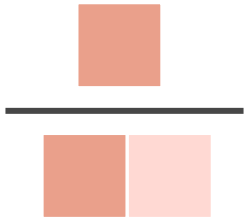
ACTUAL VALUES	PREDICTIONS: Crossfit	PREDICTIONS: BJJ
Crossfit	TN	FP
BJJ	FN	TP

Accuracy =

Suitable for balanced classes

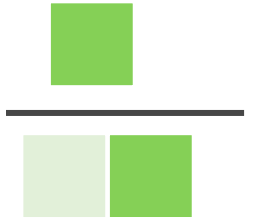
Ideally optimize to 1

Specificity =



True negative rate

Recall / Sensitivity =



True positive rate

As similar as possible

MODEL RESULTS

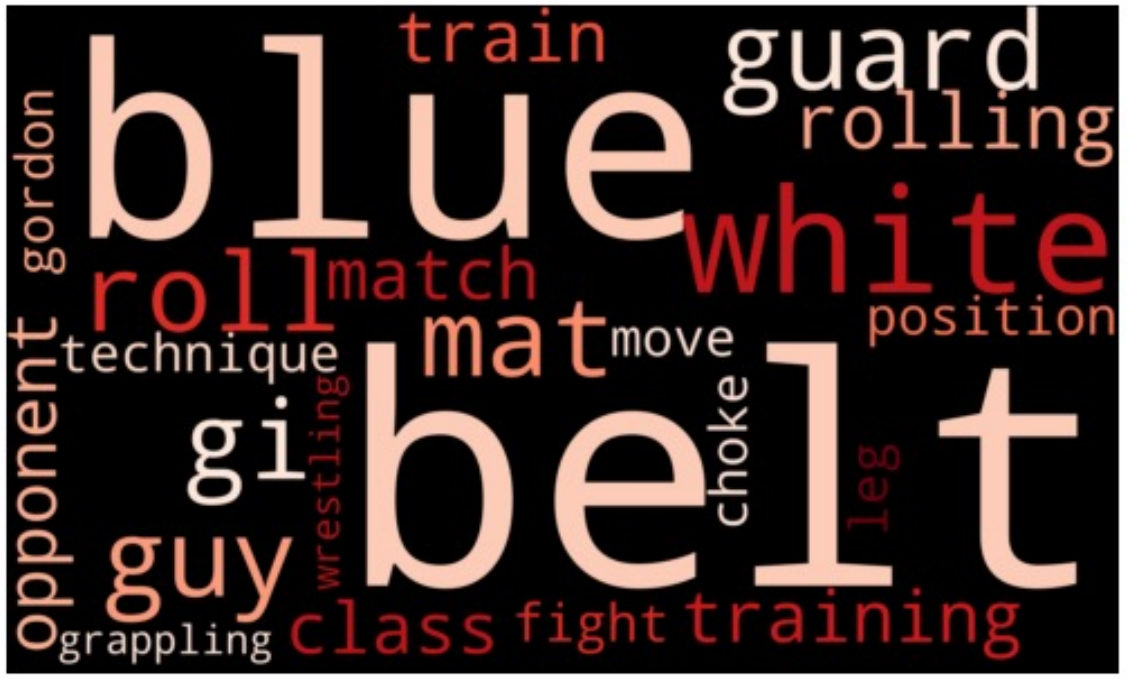
	Model 1: Naive Bayes with Count-Vectorizer	Model 2: Logistic Regression with Count-Vectorizer	Model 3: Support Vector Machines with Count-Vectorizer	Model 4: Random Forest with Count-Vectorizer	Model 5: Naive Bayes with TFIDF	Model 6: Logistic Regression with TFIDF	Model 7: Support Vector Machines with 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

WORDS LIKELY TO PREDICT BJJ



RECOMMENDATIONS & PROPOSED NEXT STEPS

Test production model

Achieved 93% accuracy on unseen data

Identified both Crossfit and BJJ posts correctly at least 91% of the time

Gym implements production model

Refine model by training with data containing Singlish

Topic modeling for upstream categorising

Try other models such as neural networks

THANK YOU


WHEN LIFE KNOCKS YOU DOWN
LIFE
BUT YOU SPECIALIZE IN THE GROUND GAME