

Selling Happiness:

A Sentimental Journey of Classification

Project 3

Dillon Diatlo | GA DSB-122

ABOUT ME



DILLON DIATLO

DATA SCIENTIST,
GOOD GUY

1 Time Lost to Harnish

DILLONDIATLO@GMAIL.COM

TABLE OF CONTENTS

01

CHALLENGE

02

EDA &
SENTIMENT
ANALYSIS

03

CLASSIFICATION
MODELS

04

RECS
& STEPS



01

CHALLENGE

CHALLENGE



Steve Huffman, money-gobbling Reddit CEO, can't stop thinking about gobbling money. And now with Reddit valued at \$6.5B (a fact he ends every sentence with) he wants even more.

His next venture? Well, let's just say Huffman has figured out a way to bottle happiness and wants to sell this happiness to sad athletes—specifically runners and swimmers.

The objective of this project is to:

- Perform sentiment analysis to see who is less happy: r/running or r/Swimming
- Use Natural Language Processing to build classification models that can accurately categorize Reddit posts so Huffman can sell, sell, sell to unhappy athletes

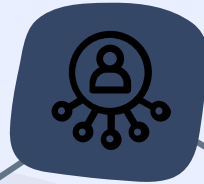


02

EDA &
SENTIMENT ANALYSIS

EDA

COLLECT
DATA
PRAW
Reddit's API
r/Swimming & r/running



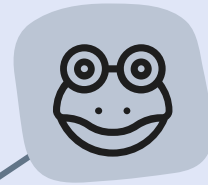
CLEAN

Reading
Erasing
all_texting

SENTIMENT
ANALYSIS
CountVectorize
Split by Subreddit



GO BACK
Start Modeling
Realize Lemmatize



EDA

**COLLECT
DATA**
PRAW
Reddit's API
r/Swimming & r/running

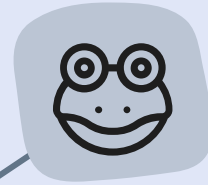


CLEAN
Reading
Erasing
all_text

**SENTIMENT
ANALYSIS**
CountVectorize
Split by Subreddit



GO BACK
Start Modeling
Realize Lemmatize



EDA

COLLECT
DATA
PRAW
Reddit's API
r/Swimming & r/running



CLEAN

Reading
Erasing
all_texting

SENTIMENT
ANALYSIS
CountVectorize
Split by Subreddit



GO BACK
Start Modeling
Realize Lemmatize

EDA

**COLLECT
DATA**
PRAW
Reddit's API
r/Swimming & r/running

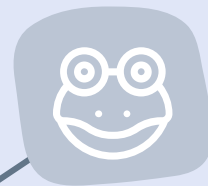


CLEAN
Reading
Erasing
all_texting

**SENTIMENT
ANALYSIS**
CountVectorize
Split by Subreddit
Hypothesis

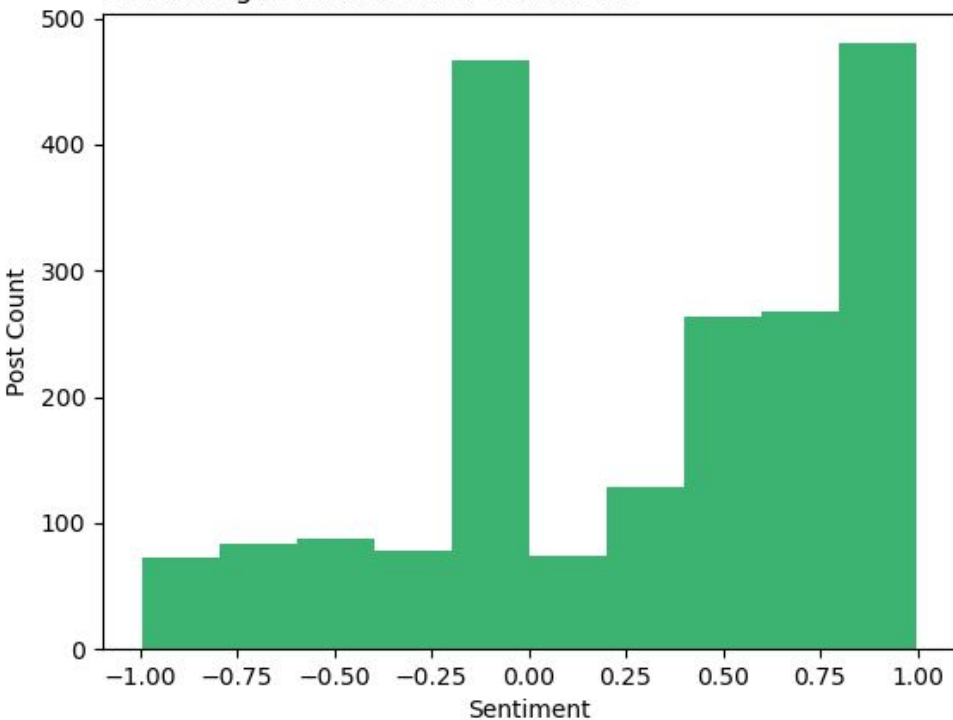


GO BACK
Start Modeling
Realize Lemmatize

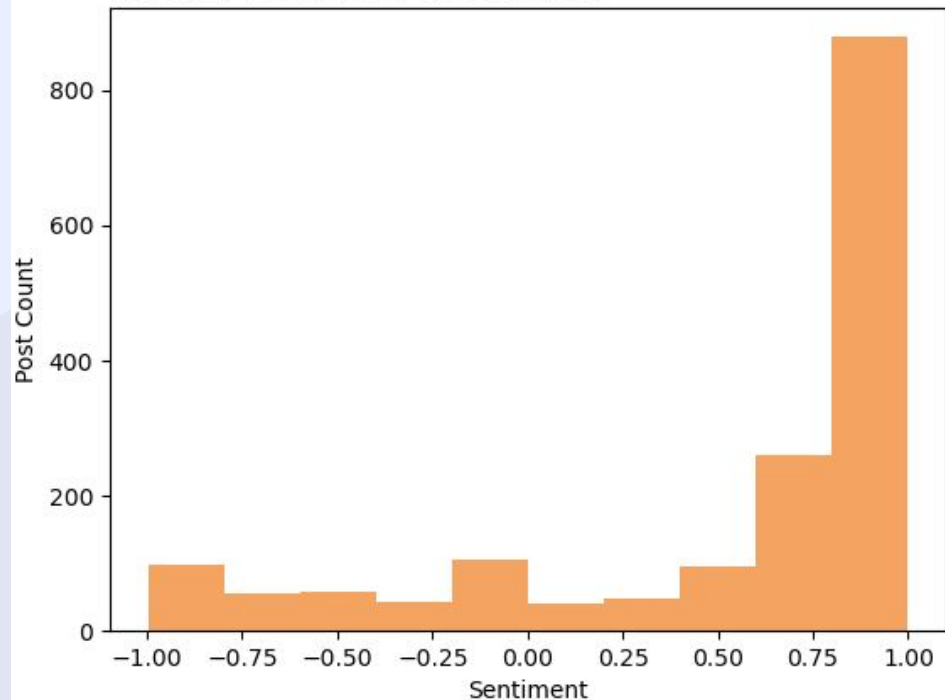


DISTRIBUTION OF POST SENTIMENT

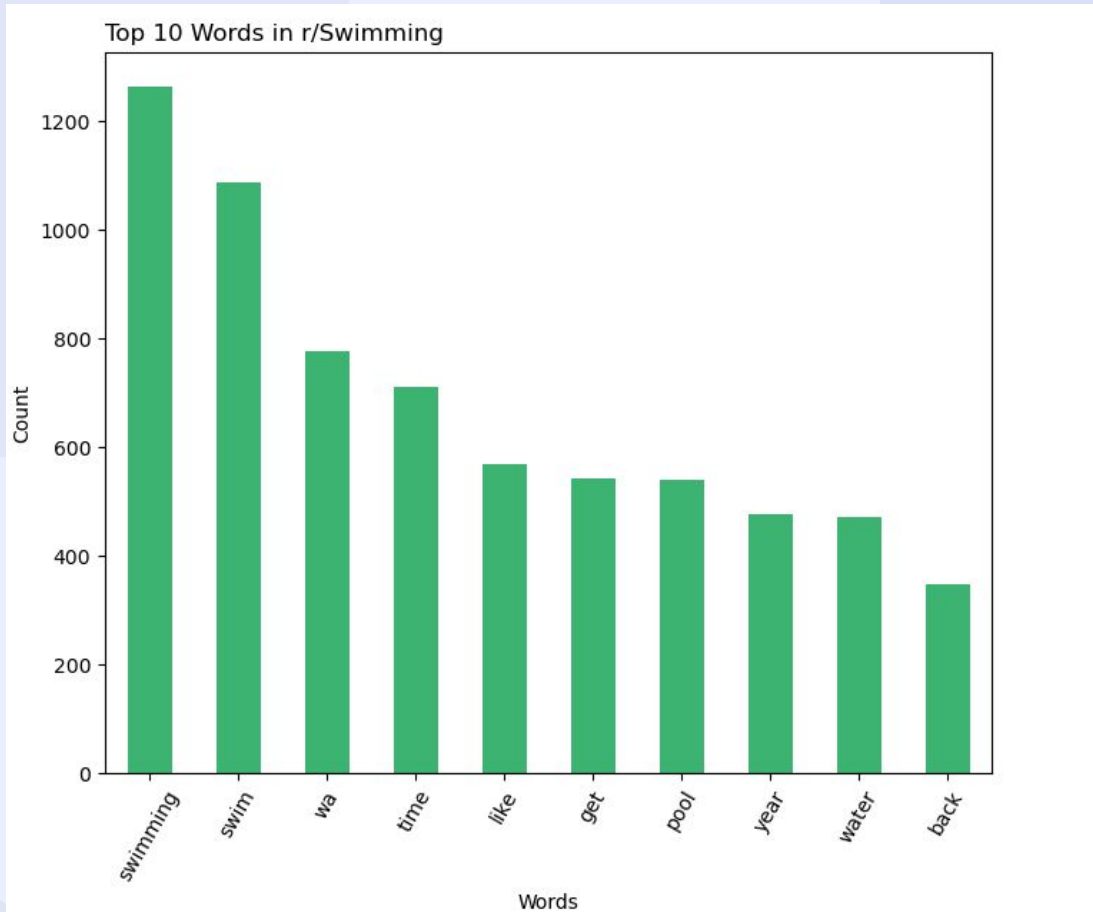
Swimming: Distribution of Sentiment



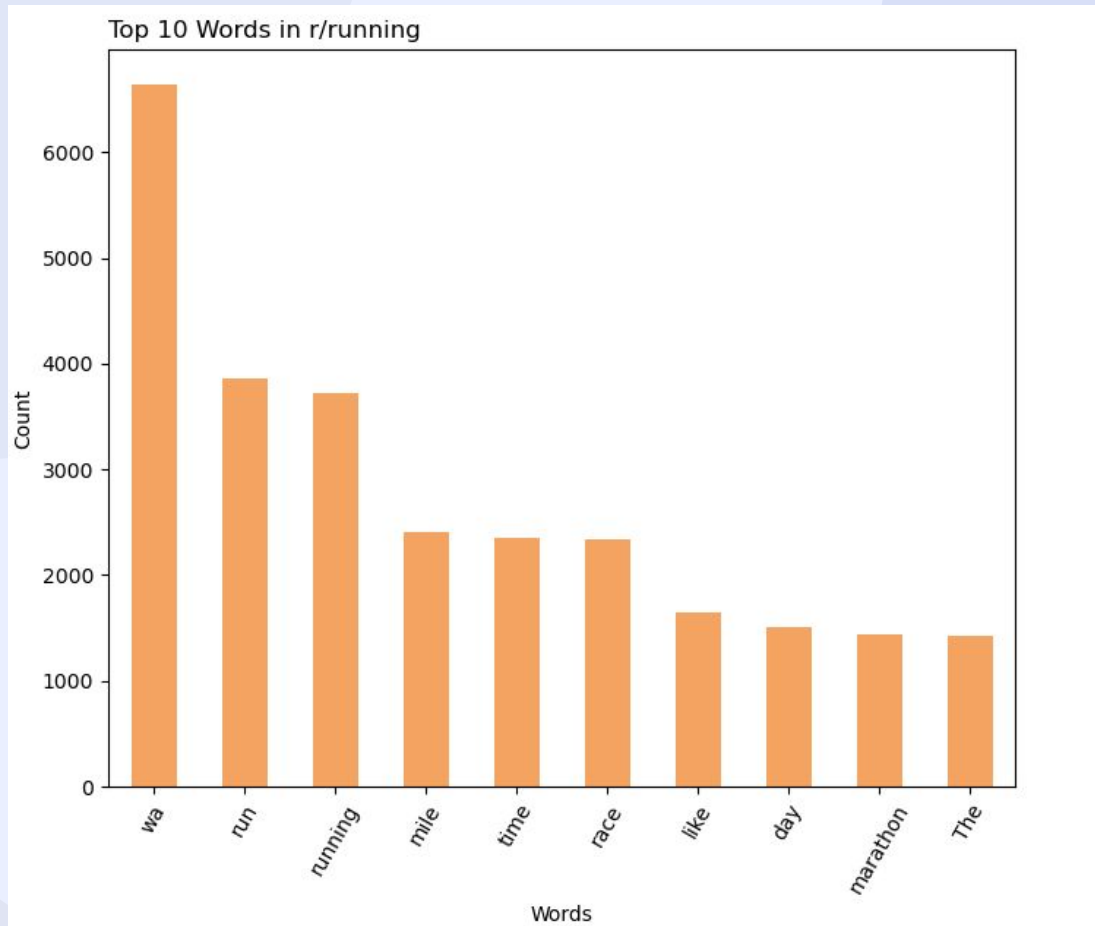
Running: Distribution of Sentiment



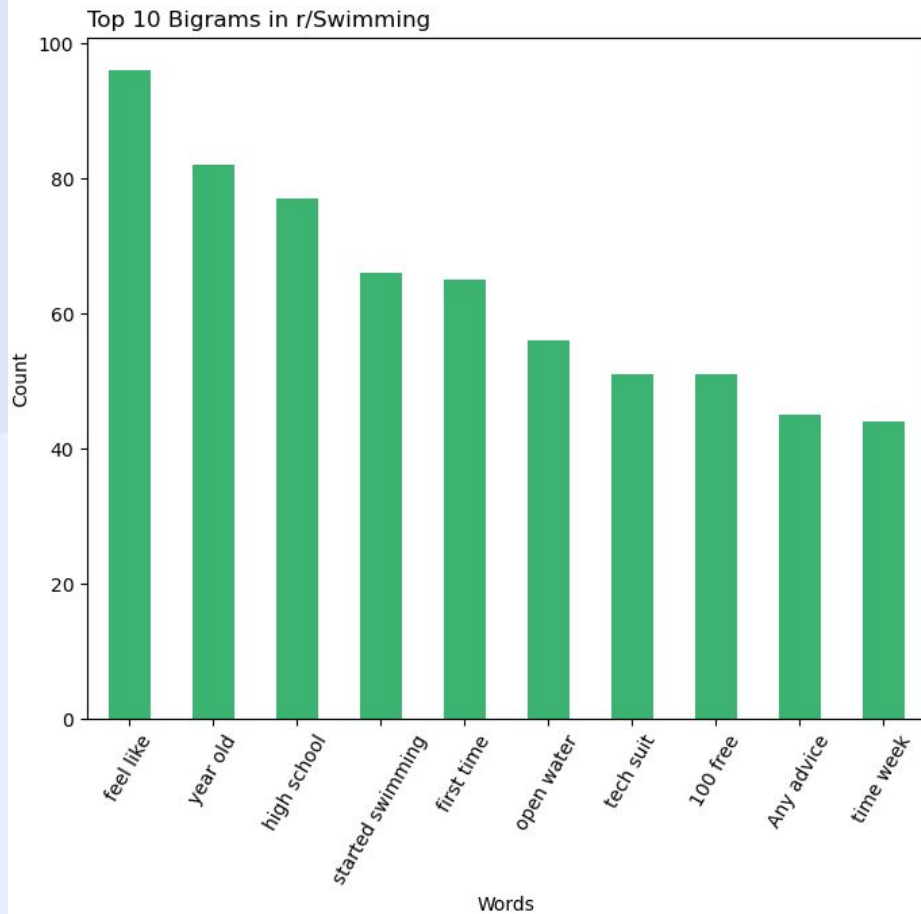
r/Swimming



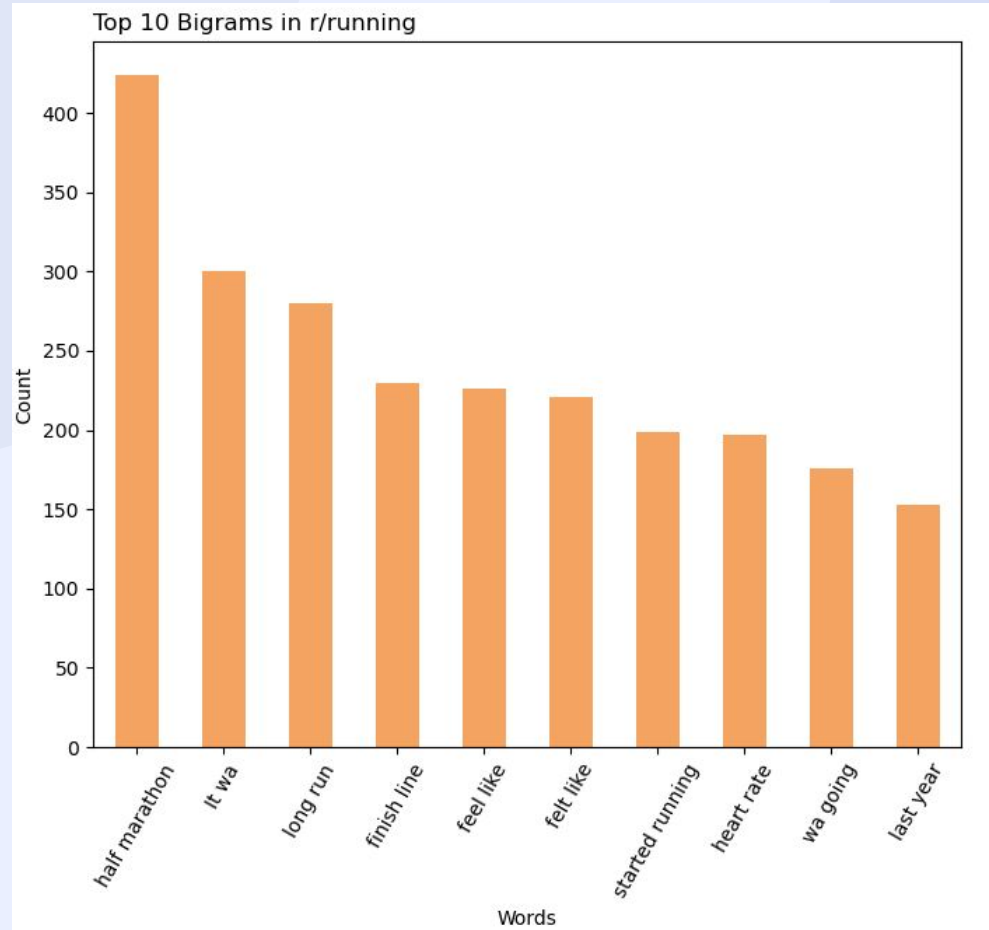
r/running



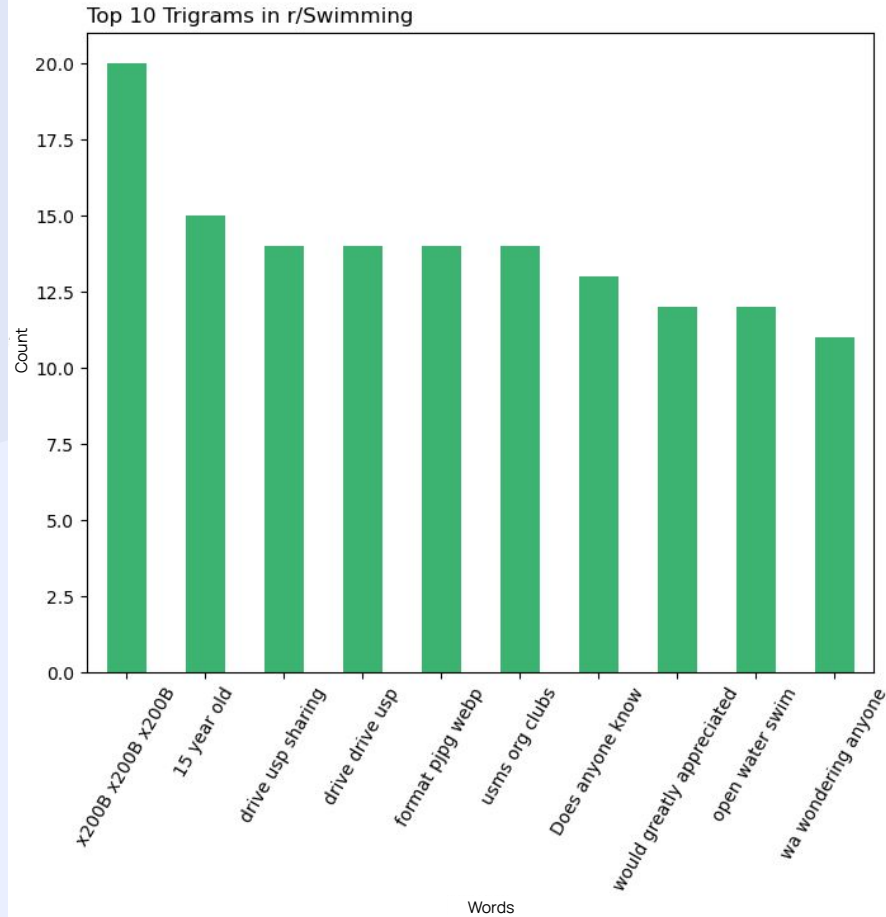
r/Swimming: BIGRAMS



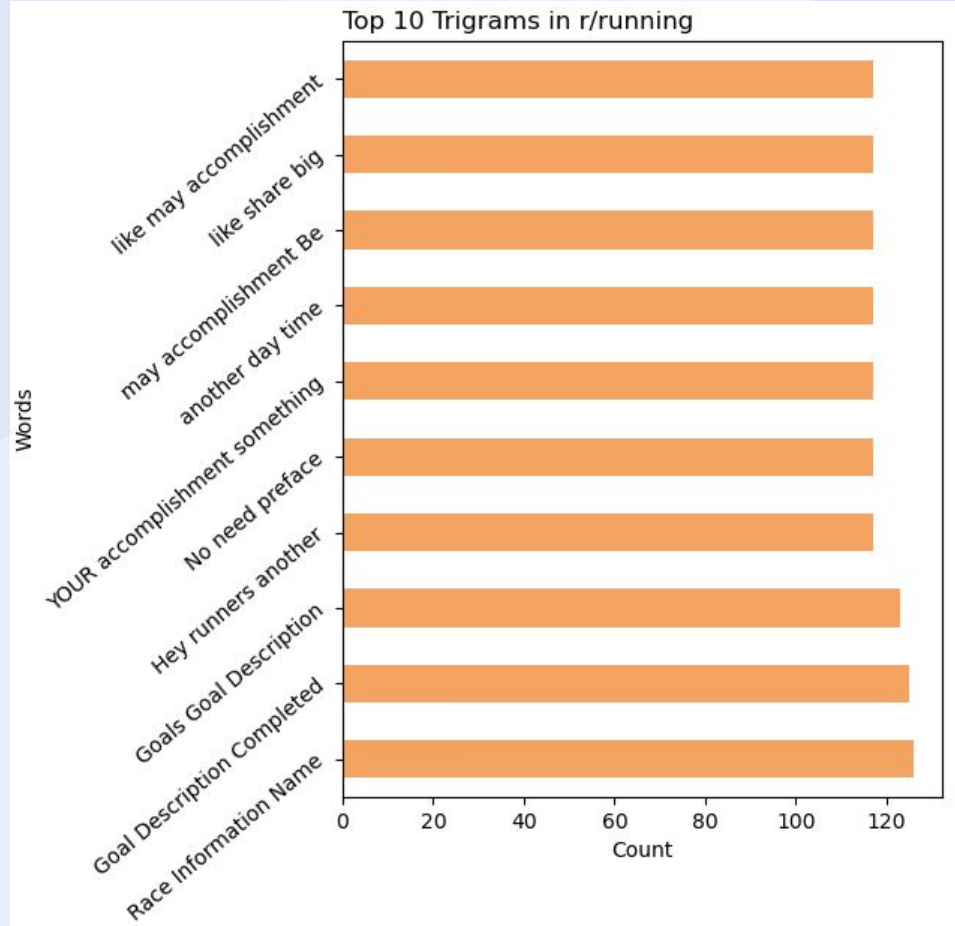
r/running: BIGRAMS



r/Swimming: TRIGRAMS



r/running: TRIGRAMS



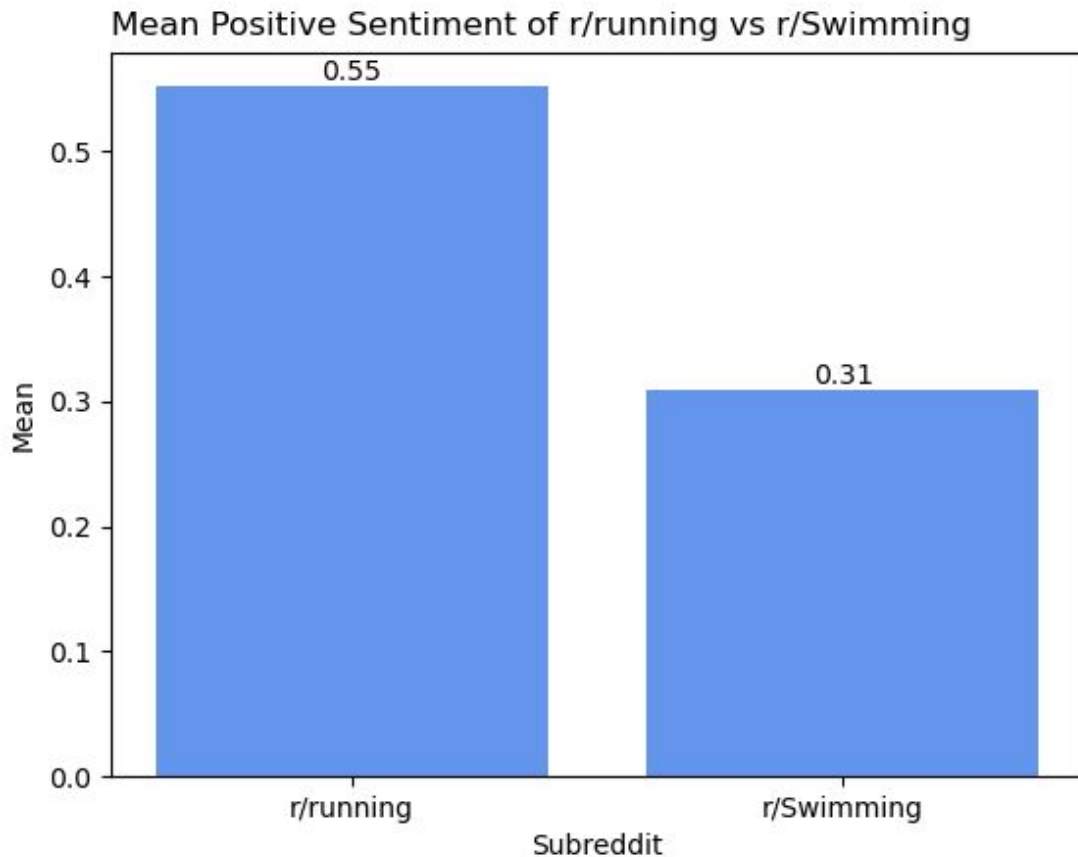
WHO'S HAPPIER?

Guess!

WHO'S HAPPIER?

r/running
Post
Sentiment:
0.55

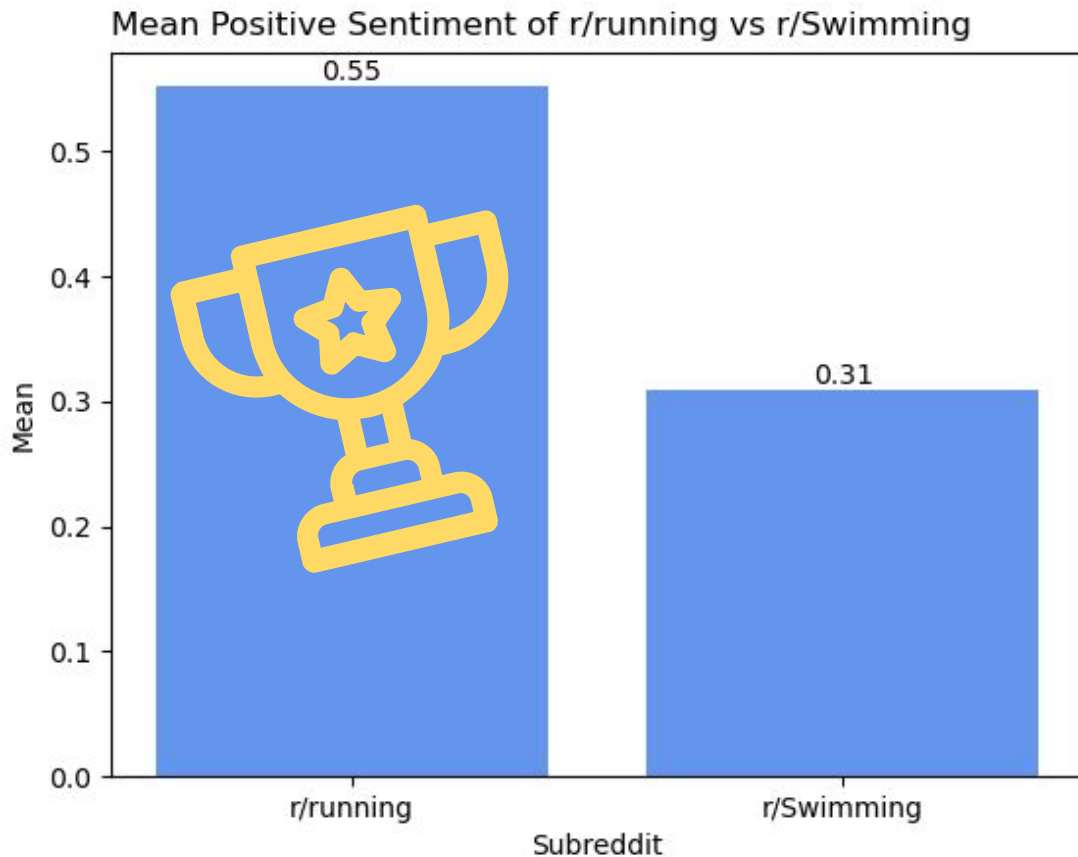
r/Swimming
Post
Sentiment:
0.31



WHO'S HAPPIER?

r/running
Post
Sentiment:
0.55

r/Swimming
Post
Sentiment:
0.31



KEEP IN MIND


It could just be the nature of the sports...

Running:

Lower barrier
Accomplishment

Swimming:

Multimodal nuance
Improvement focused
(makes sense)







03 PREDICTIVE MODELS








0.4568

Baseline Accuracy
Predicting Based on Mean

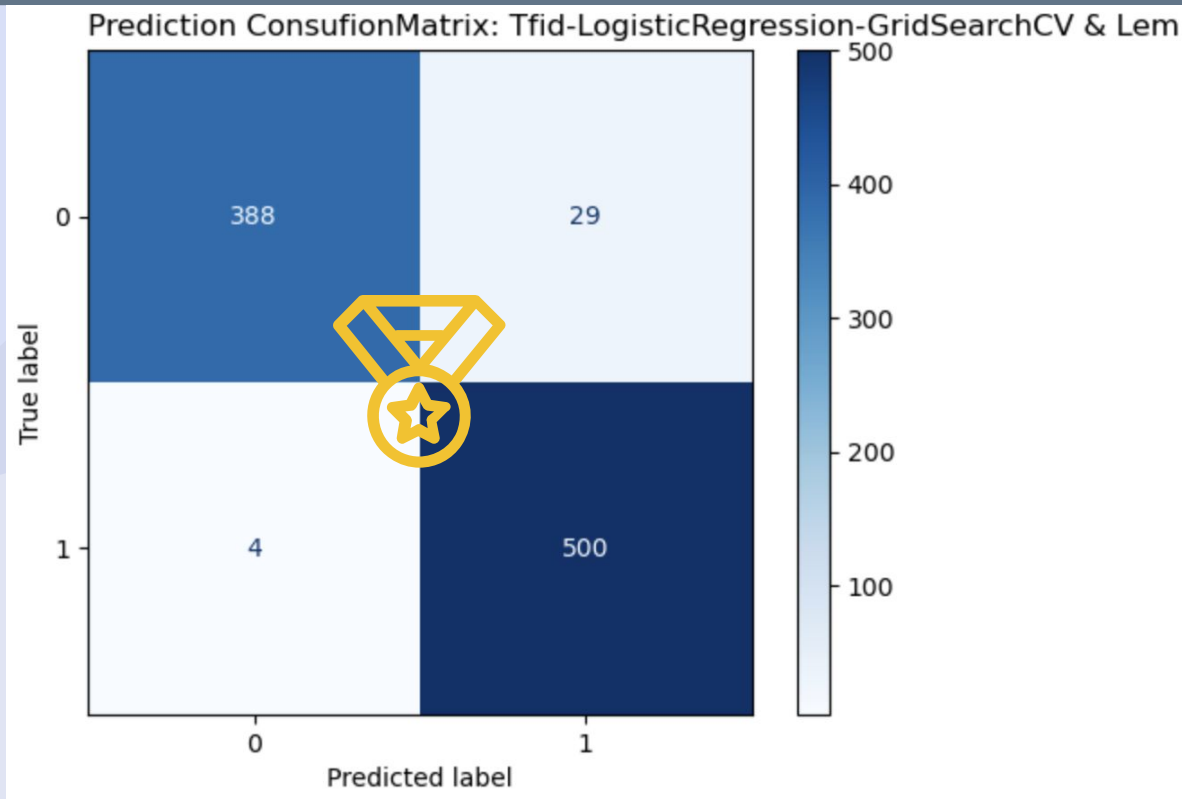
PREDICTIVE MODELS

MODEL	 TRANSFORMER	 ESTIMATOR	 TRAIN	 TEST
1	TfidfVectorizer (w/ Lemmatizer)	Logistic Regression w/ GridSearchCV	0.98	0.9685
1	TfidfVectorizer	RandomForest w/ GridSearchCV	1.0	0.9685
2	TfidfVectorizer (w/ Lemmatizer)	RandomForest w/ GridSearchCV	1.0	0.9641
3	CountVectorizer	MultinomialNB	.9699	0.9457
4	CountVectorizer	RandomForest	1.0	0.9381

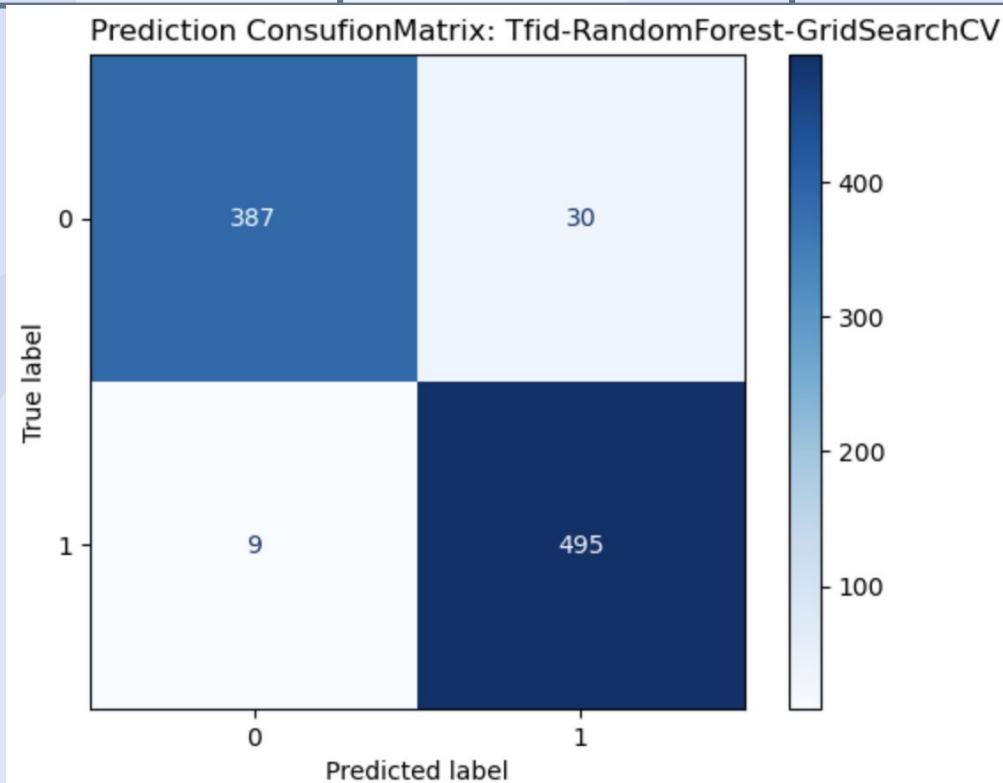
PREDICTIVE MODELS

MODEL	 TRANSFORMER	 ESTIMATOR	 TRAIN	 TEST
	TfidfVectorizer (w/ Lemmatizer)	Logistic Regression w/ GridSearchCV	0.98	0.9685
1	TfidfVectorizer	RandomForest w/ GridSearchCV	1.0	0.9685
2	TfidfVectorizer (w/ Lemmatizer)	RandomForest w/ GridSearchCV	1.0	0.9641
3	CountVectorizer	MultinomialNB	.9699	0.9457
4	CountVectorizer	RandomForest	1.0	0.9381

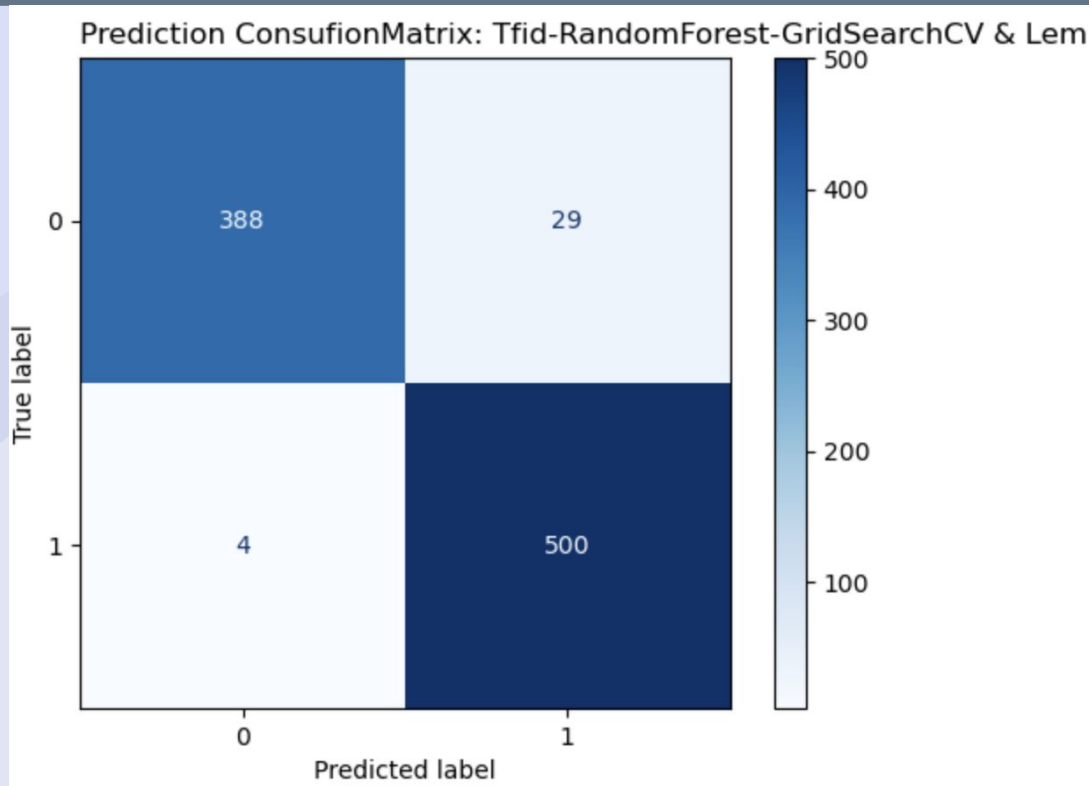
MODEL	 TRANSFORMER	 ESTIMATOR	 TRAIN	 TEST
1	TfidfVectorizer w/ Lemmatizer	Logistic Regression w/ GridSearchCV	0.98	0.9685



MODEL	TRANSFORMER	ESTIMATOR	TRAIN	TEST
1	TfidfVectorizer	RandomForest w/ GridSearchCV	1.0	0.9685



MODEL	TRANSFORMER	ESTIMATOR	TRAIN	TEST
2	TfidfVectorizer w/ Lemmatizer	RandomForest w/ GridSearchCV	1.0	0.9641



04

RECS & STEPS

RECS

RECS

96.8%

For now, use
this model

w/ Lemmatized Data, TfidfVectorizer,
RandomForest, GridSearchCV

RECS

96.8%

For now, use
this model

w/ Lemmatized Data, TfidfVectorizer,
RandomForest, GridSearchCV

r/Swimming

Advertise to
this subreddit

Is the less happy subreddit
(but maybe not sad)

NEXT STEPS

SENTIMENT
& DATETIME



STEMMING
& FILTER

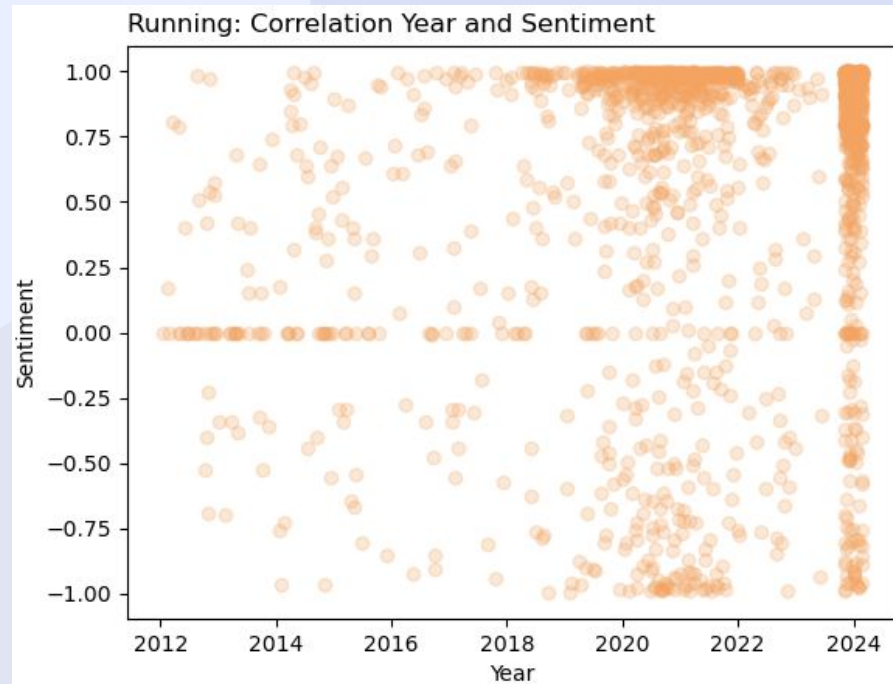
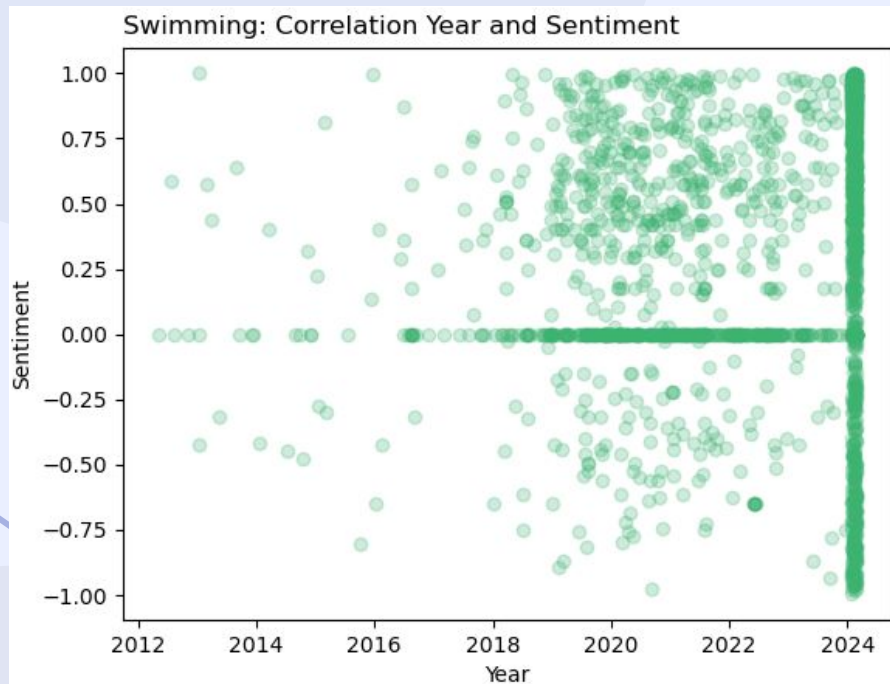


PINPOINT
USERS



BONUS:
Does year
impact sentiment?
Big time.

CORRELATION YEAR & SENTIMENT



Q&A



DILLON DIATLO
DATA SCIENTIST,
GOOD GUY

DILLONDIATLO@GMAIL.COM

Thank you