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

CHALLENGE

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

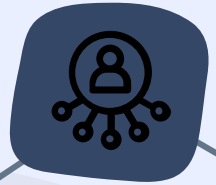
02

EDA &
SENTIMENT ANALYSIS

EDA

COLLECT DATA

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

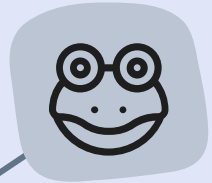


CLEAN

Reading
Erasing
all_texting

SENTIMENT ANALYSIS

CountVectorize
Split by Subreddit



GO BACK

Start Modeling
Realize Lemmatize

EDA

COLLECT DATA

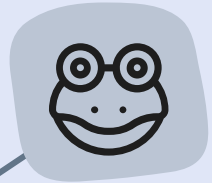
PRAW
Reddit's API
r/Swimming & r/running

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

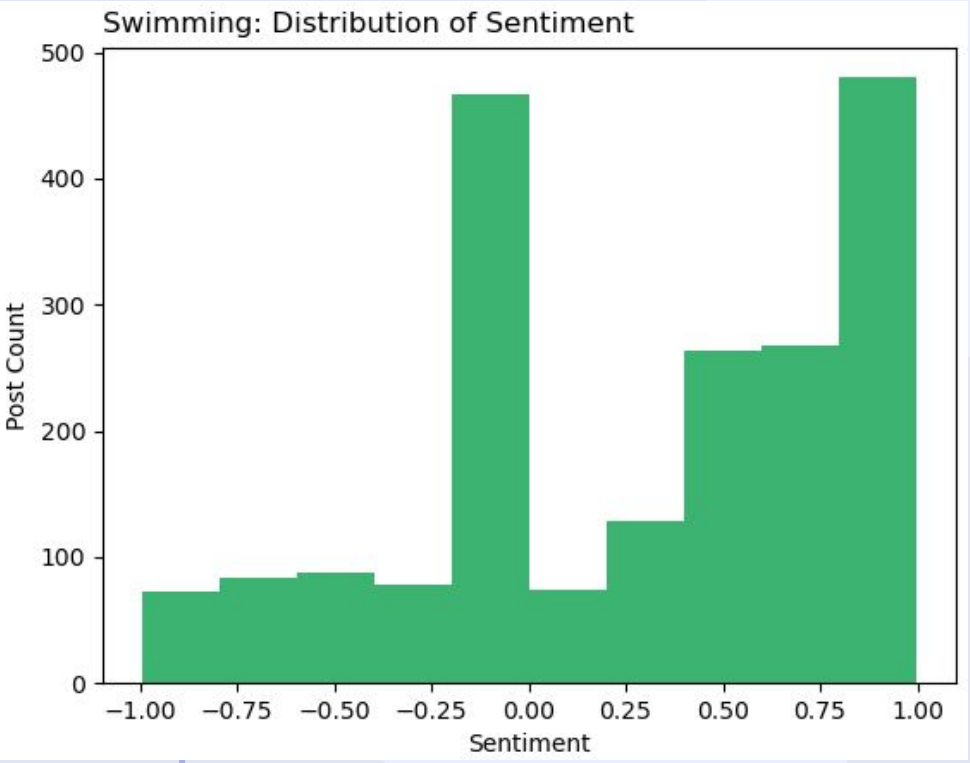
SENTIMENT ANALYSIS

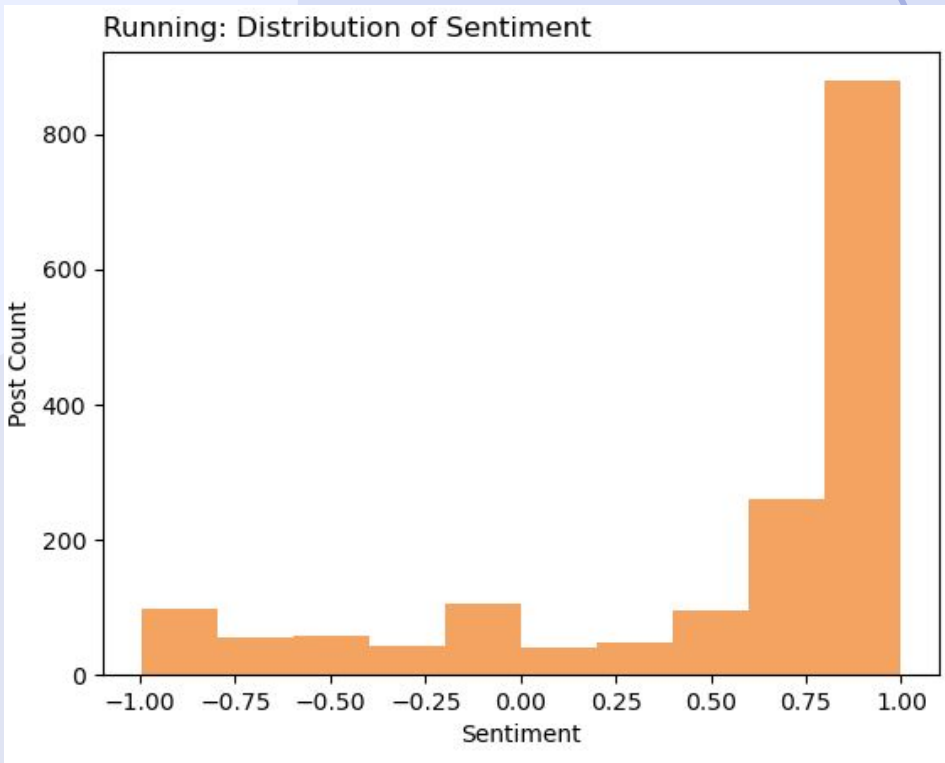
CountVectorize
Split by Subreddit
Hypothesis

GO BACK

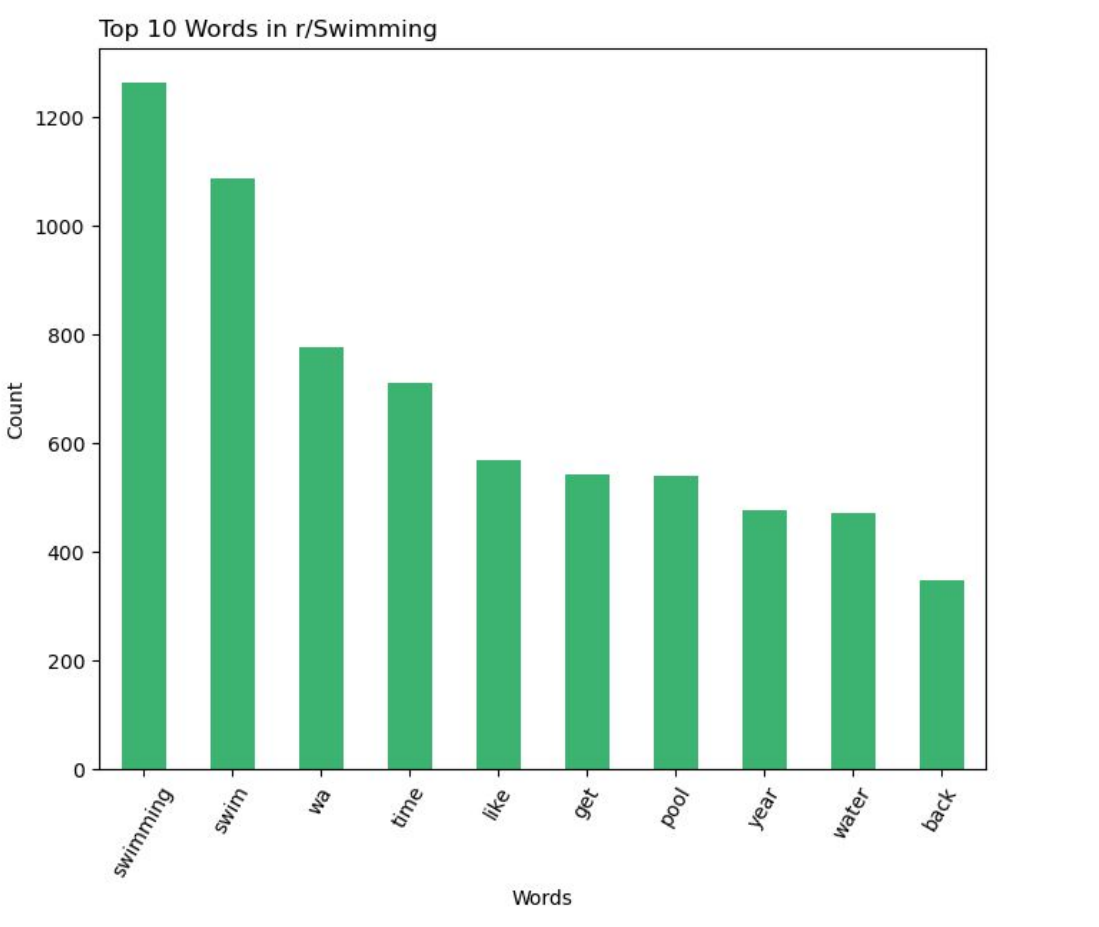
Start Modeling
Realize Lemmatize

DISTRIBUTION OF POST SENTIMENT

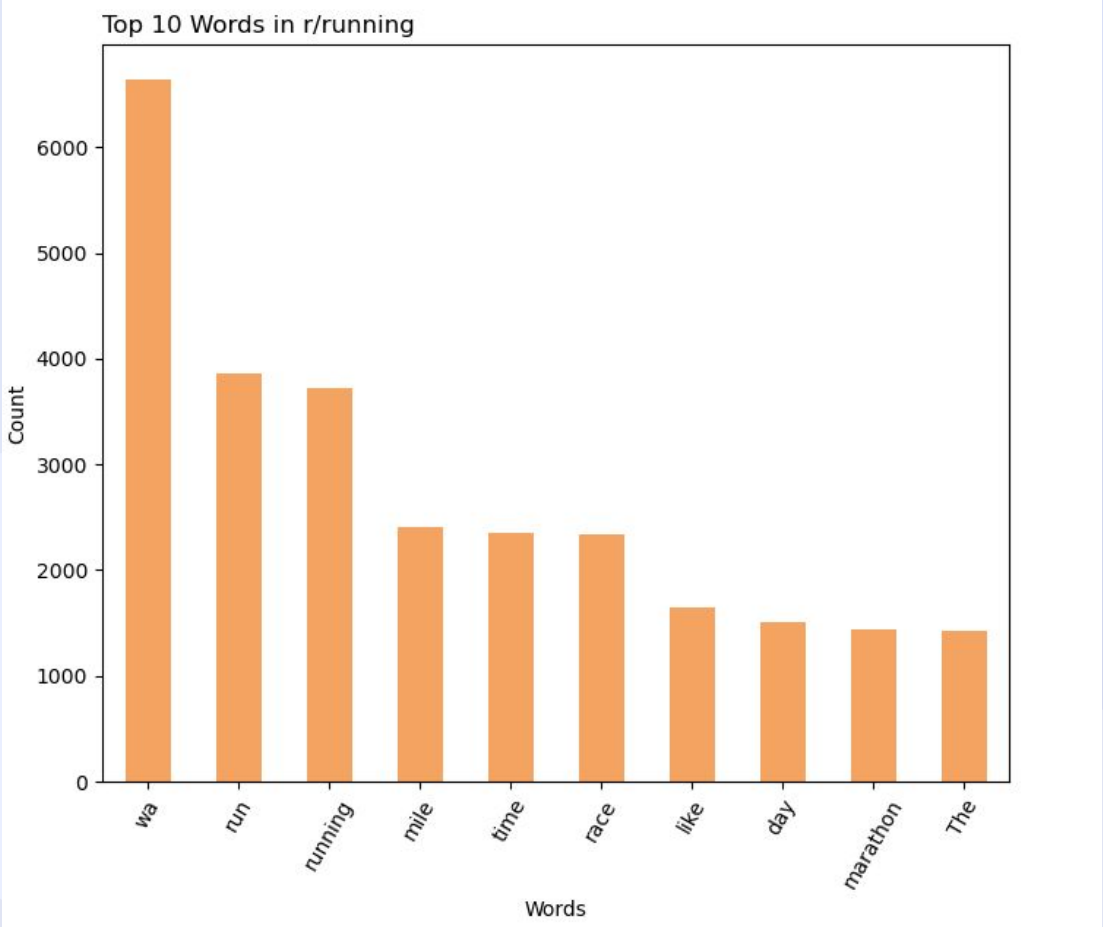




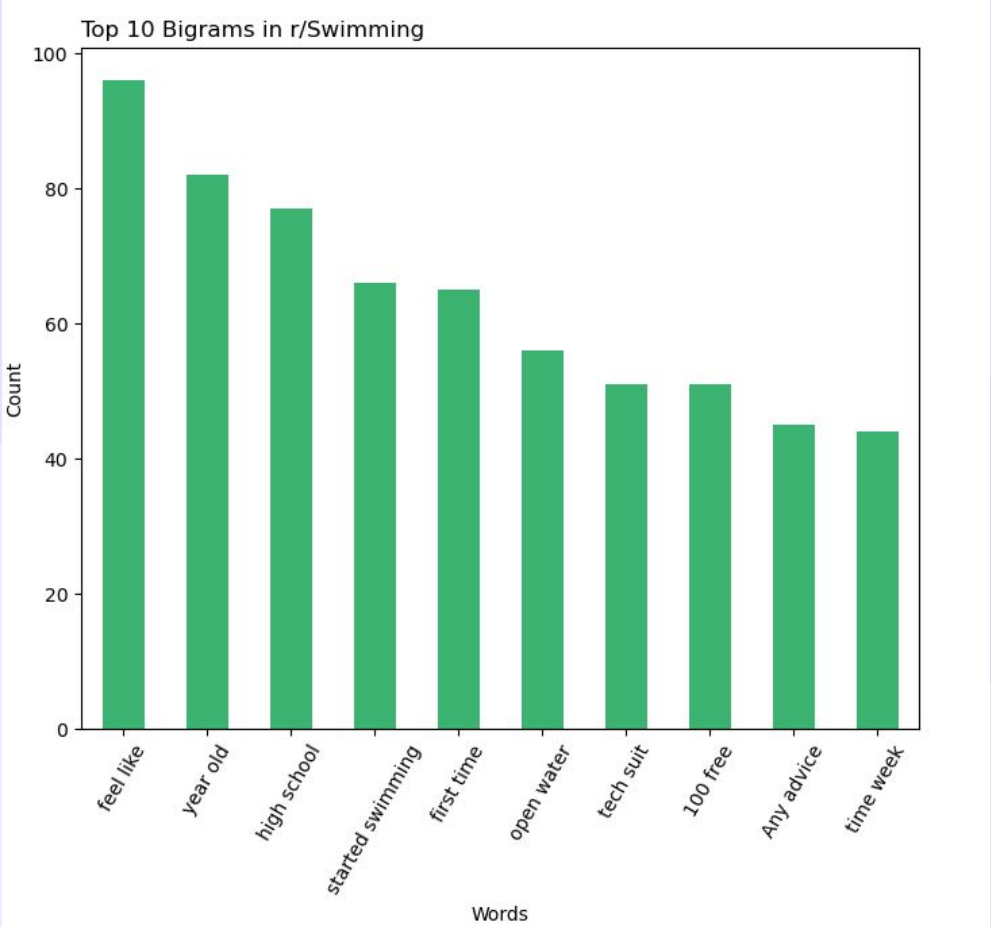
r/Swimming


Top 10 Words in r/Swimming
Count
0
200
400
600
800
1000
1200
swimming
swim
wa
time
like
get
pool
year
water
back
Words

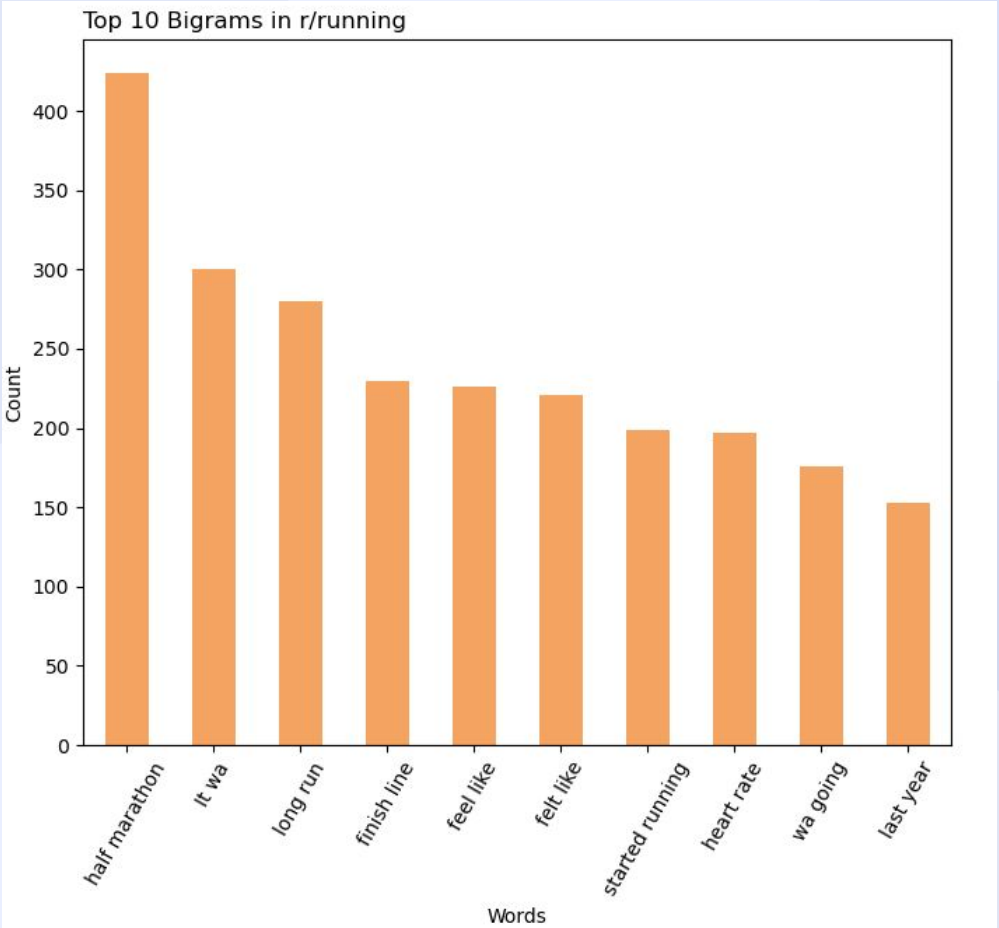
r/running


Top 10 Words in r/running
Count
0
1000
2000
3000
4000
5000
6000
wa
run
running
mile
time
race
like
day
marathon
The
Words

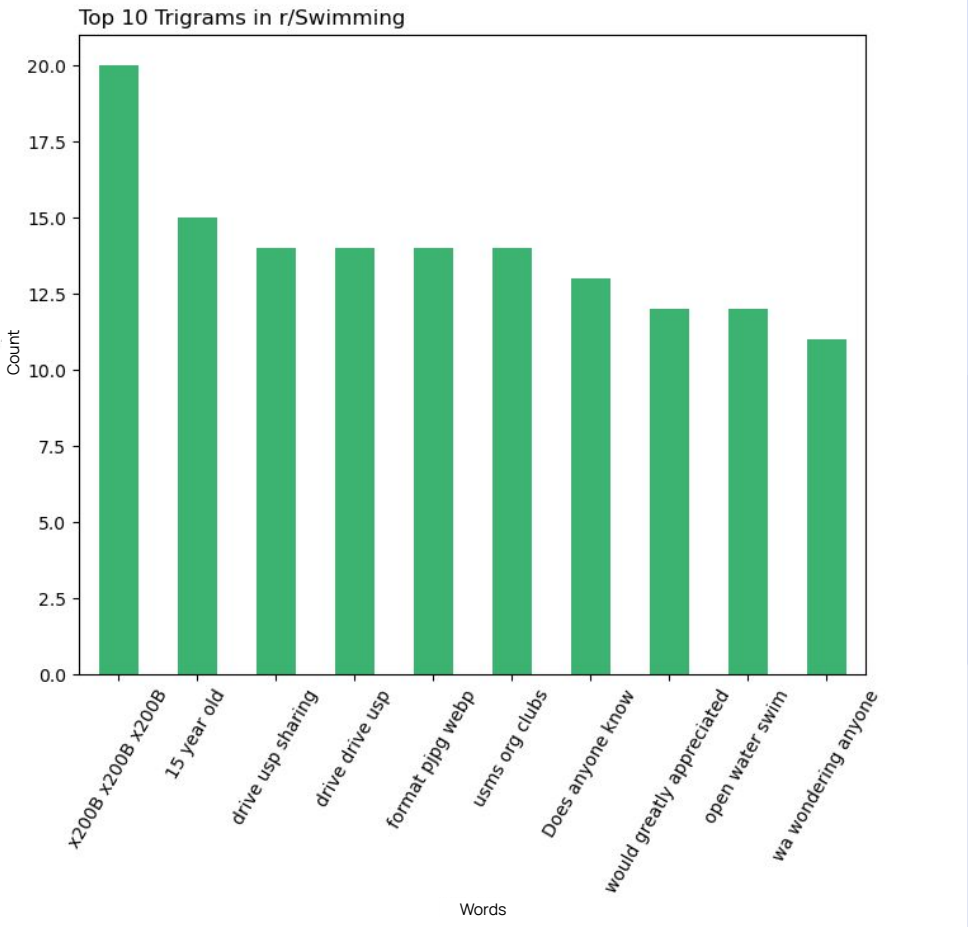
r/Swimming: BIGRAMS



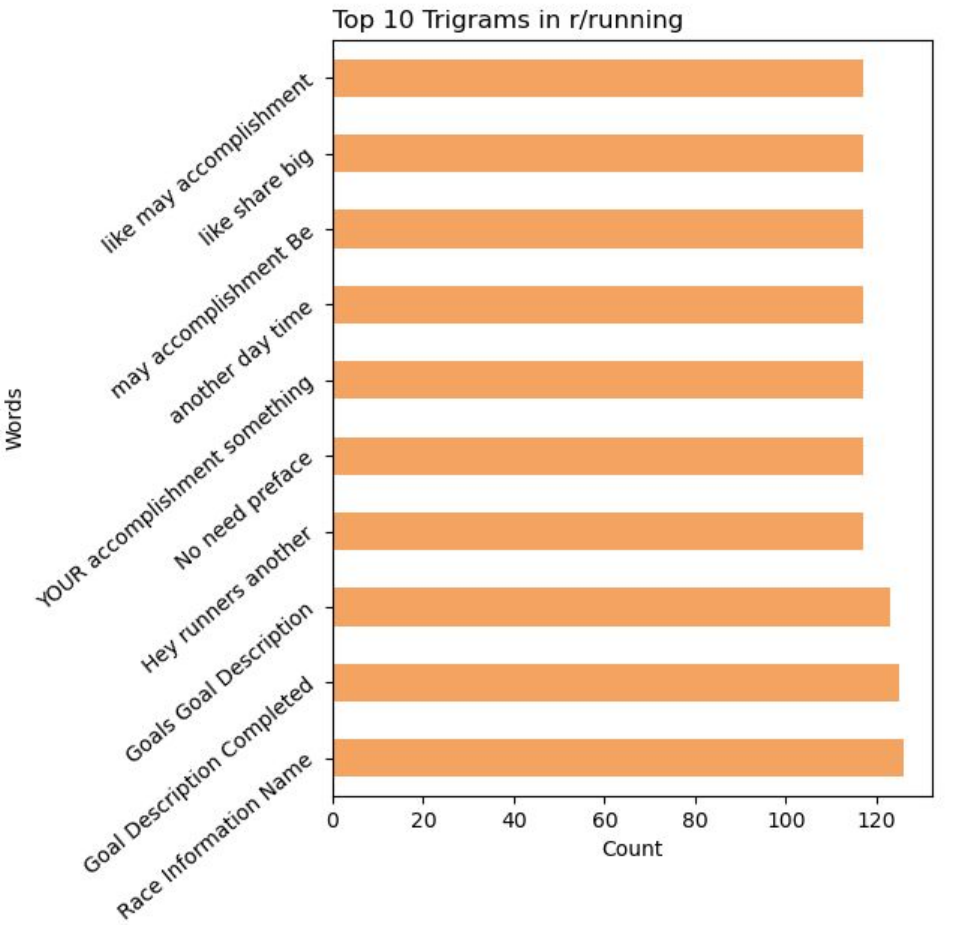
r/running: BIGRAMS



r/Swimming: TRIGRAMS



r/running: TRIGRAMS



WHO'S HAPPIER?

Guess!

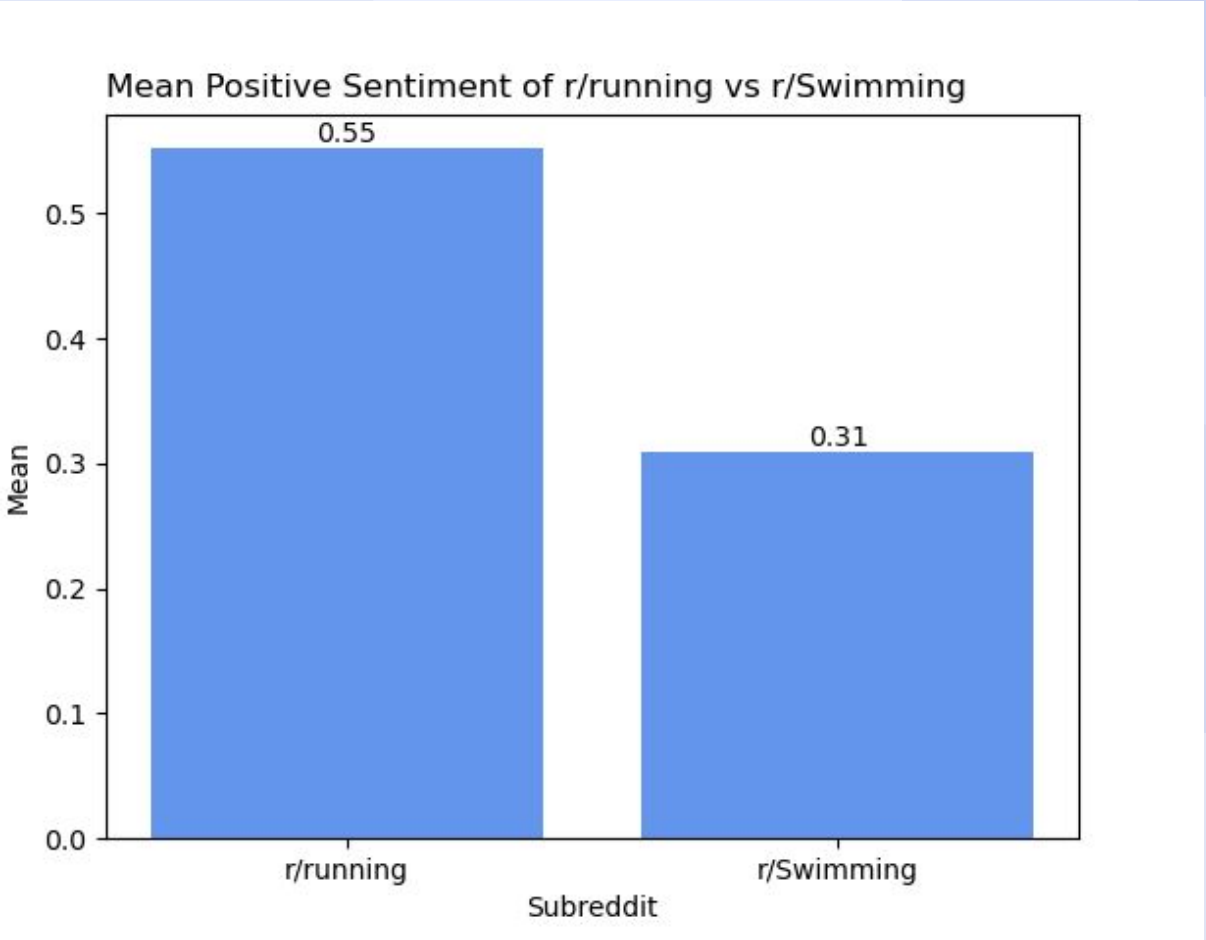
WHO'S HAPPIER?

r/running Post Sentiment:

0.55

r/Swimming Post Sentiment:

0.31



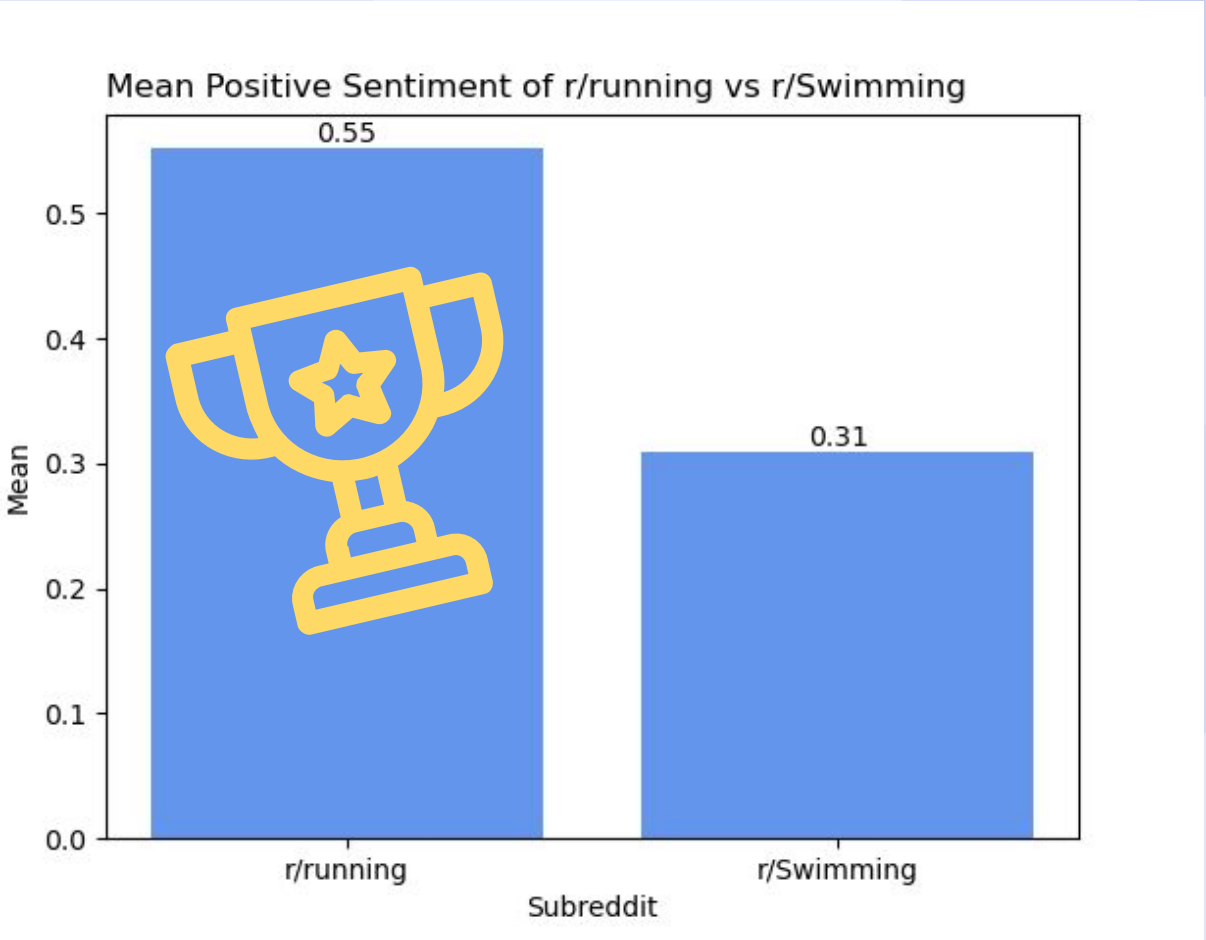
WHO'S HAPPIER?

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

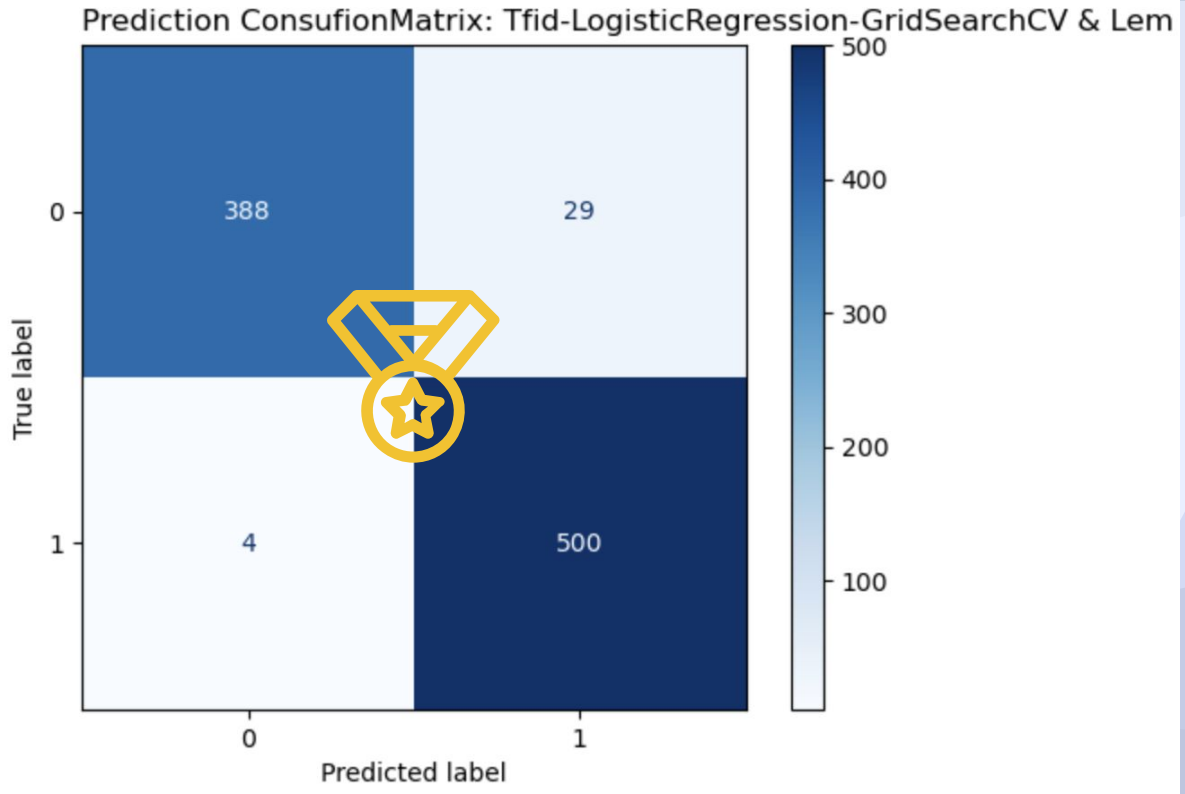
PREDICTIVE MODELS

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

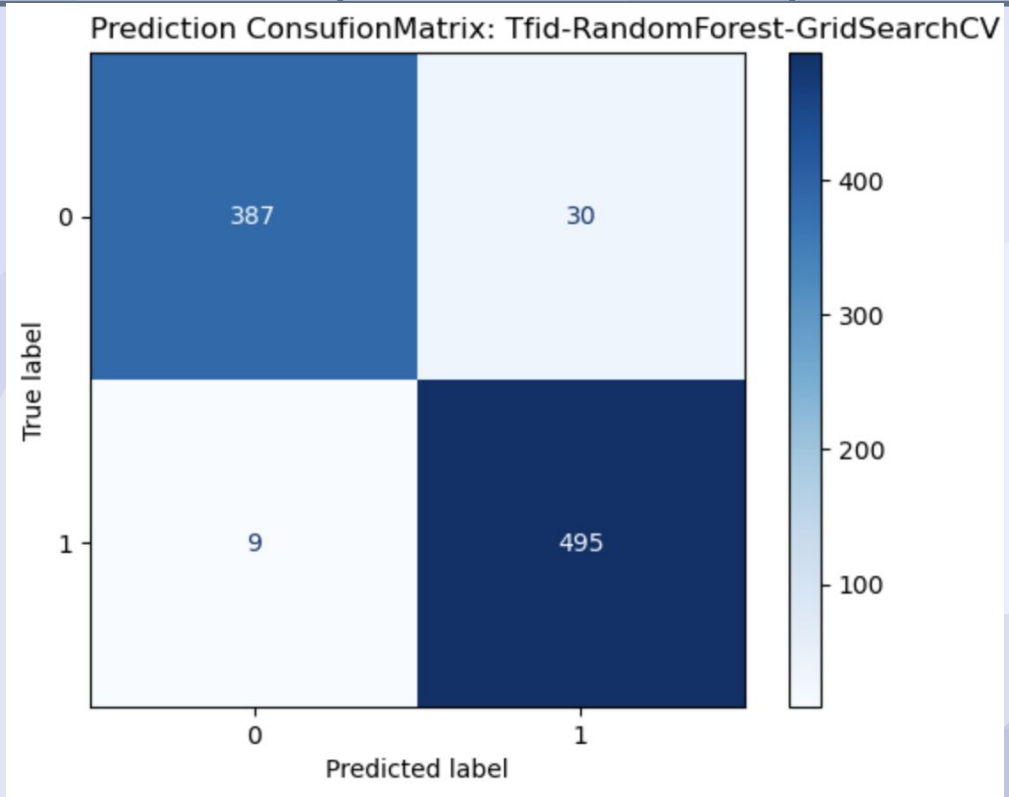
PREDICTIVE MODELS

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

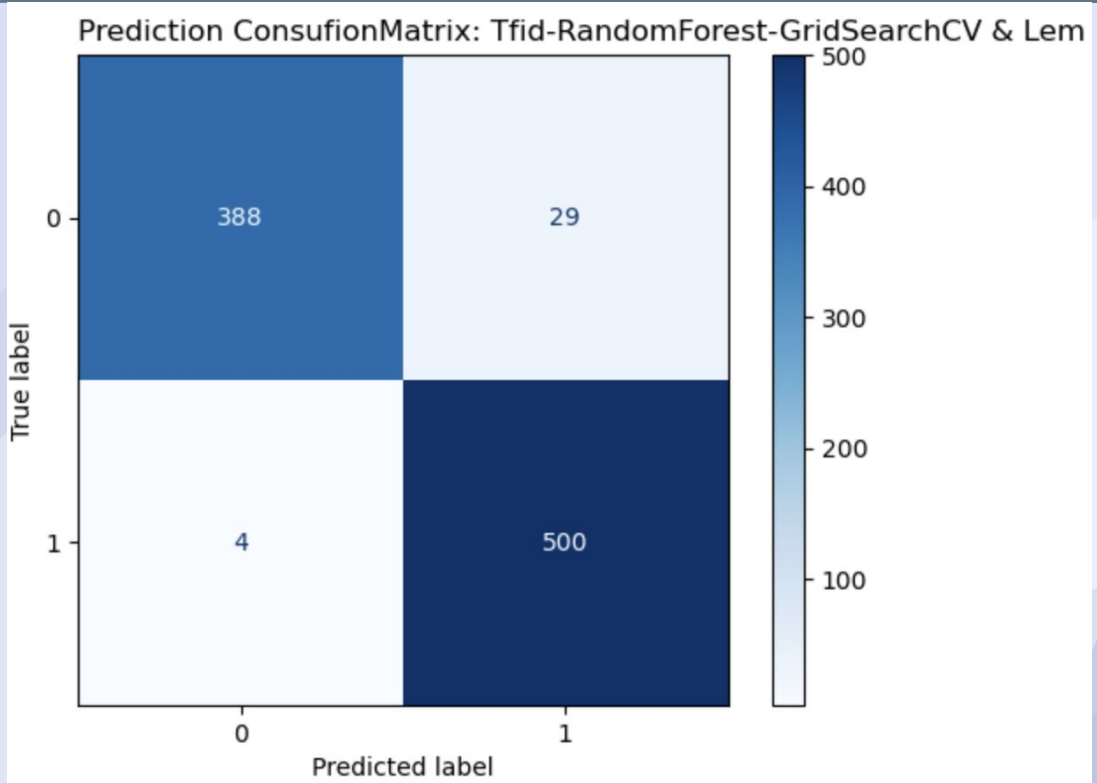
MODEL	TRANSFORMER	ESTIMATOR	TRAIN	TEST
1	TfidVectorizer w/ Lemmatizer	Logistic Regression w/ GridSearchCV	0.98	0.9685



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



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



04

RECS & STEPS

RECS

RECS

96.8%

For now, use this model

w/ Lemmatized Data, TfidVectorizer, RandomForest, GridSearchCV

RECS

96.8%

For now, use this model

w/ Lemmatized Data, TfidVectorizer, RandomForest, GridSearchCV

r/Swimming

Advertise to this subreddit

Is the less happy subreddit (but maybe not sad)

NEXT STEPS

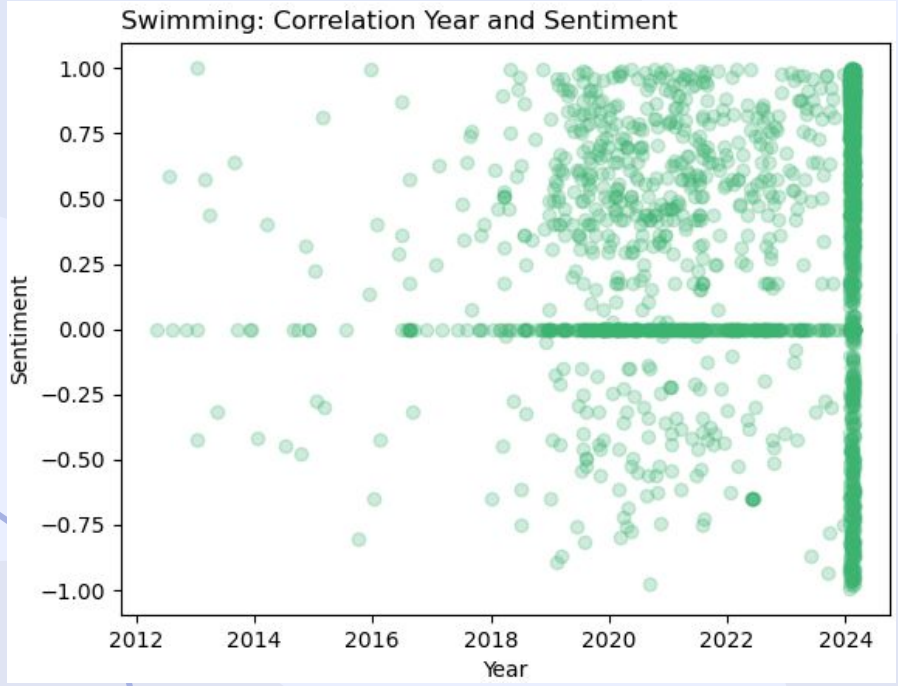
SENTIMENT
& DATETIME

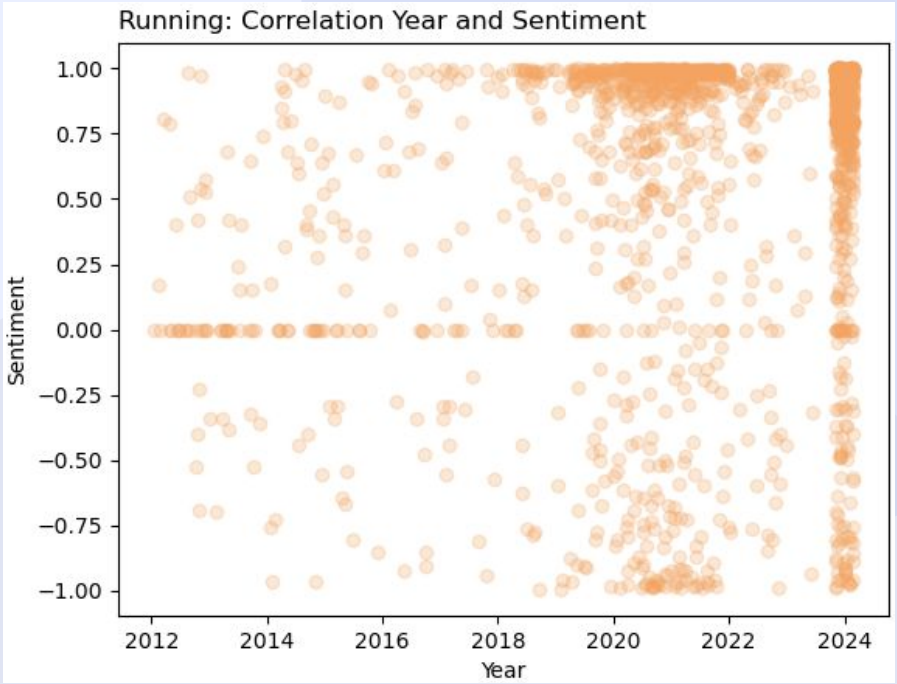
STEMMING
& FILTER

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