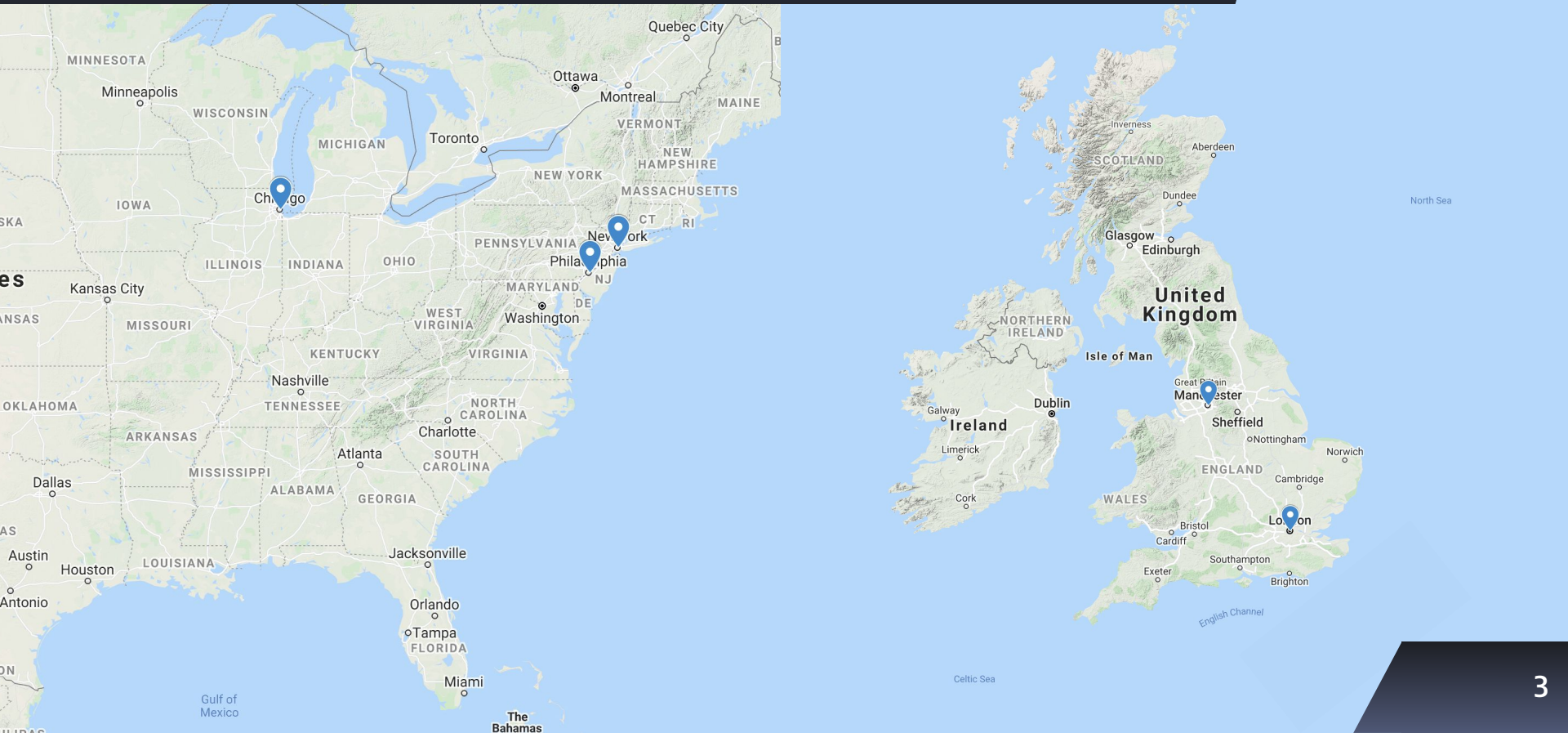
Market Research: Sports Fans in the U.S. and England

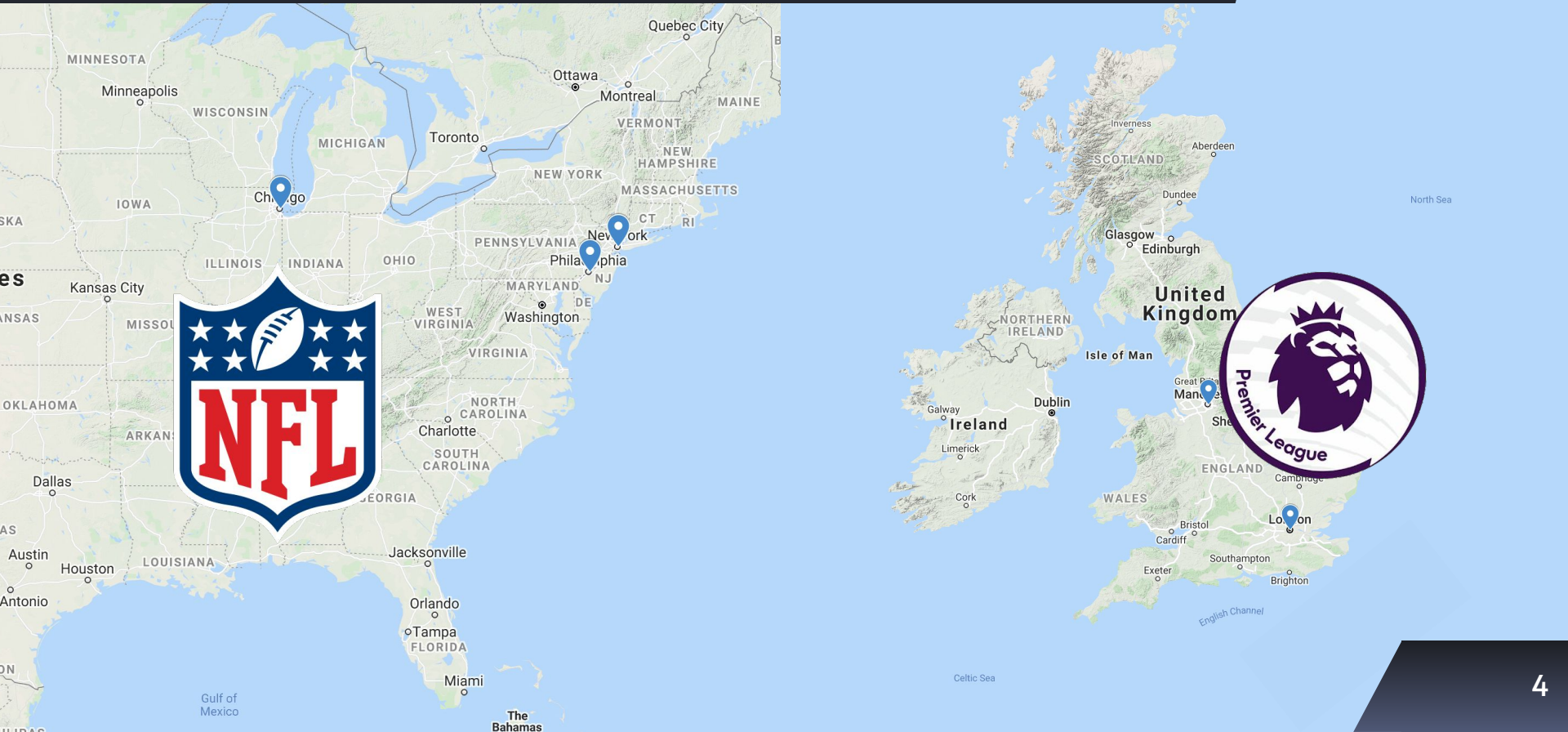
Riley Robertson

The Client

The Client


Quebec City
Ottawa
Montreal
Minneapolis
MINNESOTA
WISCONSIN
MICHIGAN
Toronto
MAINE
VERMONT
NEW HAMPSHIRE
NEW YORK
MASSACHUSETTS
IOWA
CT
RI
PENNSYLVANIA
ILLINOIS
INDIANA
OHIO
NJ
MARYLAND
DE
Kansas City
MISSOURI
WEST VIRGINIA
Washington
KENTUCKY
VIRGINIA
Nashville
TENNESSEE
NORTH CAROLINA
Charlotte
ARKANSAS
Atlanta
SOUTH CAROLINA
MISSISSIPPI
ALABAMA
GEORGIA
Dallas
Austin
Houston
LOUISIANA
Jacksonville
Orlando
Tampa
FLORIDA
Miami
Gulf of Mexico
The Bahamas
Inverness
Aberdeen
SCOTLAND
Dundee
North Sea
Glasgow
Edinburgh
United Kingdom
NORTHERN IRELAND
Isle of Man
Dublin
Galway
Ireland
Limerick
Cork
Sheffield
Nottingham
Norwich
ENGLAND
Cambridge
WALES
Bristol
Cardiff
Southampton
Exeter
Brighton
English Channel
Celtic Sea

The Client


Quebec City
MINNESOTA
Minneapolis
Ottawa
Montreal
WISCONSIN
MAINE
VERMONT
MICHIGAN
Toronto
NEW HAMPSHIRE
NEW YORK
MASSACHUSETTS
IOWA
CT
RI
PENNSYLVANIA
ILLINOIS
INDIANA
OHIO
NJ
Kansas City
MARYLAND
DE
WEST VIRGINIA
Washington
VIRGINIA
NORTH CAROLINA
Charlotte
SOUTH CAROLINA
Dallas
GEORGIA
Austin
Houston
LOUISIANA
Jacksonville
Orlando
Tampa
FLORIDA
Miami
Gulf of Mexico
The Bahamas
Celtic Sea
Inverness
Aberdeen
SCOTLAND
Dundee
North Sea
Glasgow
Edinburgh
United Kingdom
NORTHERN IRELAND
Isle of Man
Dublin
Galway
Ireland
Limerick
Cork
ENGLAND
WALES
Bristol
Cardiff
Southampton
Exeter
Brighton
English Channel
NFL
Premier League

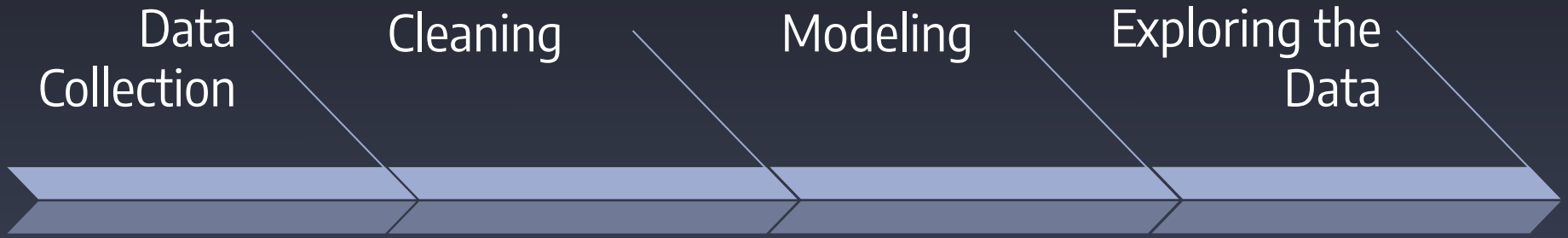
OverArmor's Requests:

1. Collect data from each league's respective Reddit community to build a clean database for investigating the colloquial language used by each fan-base

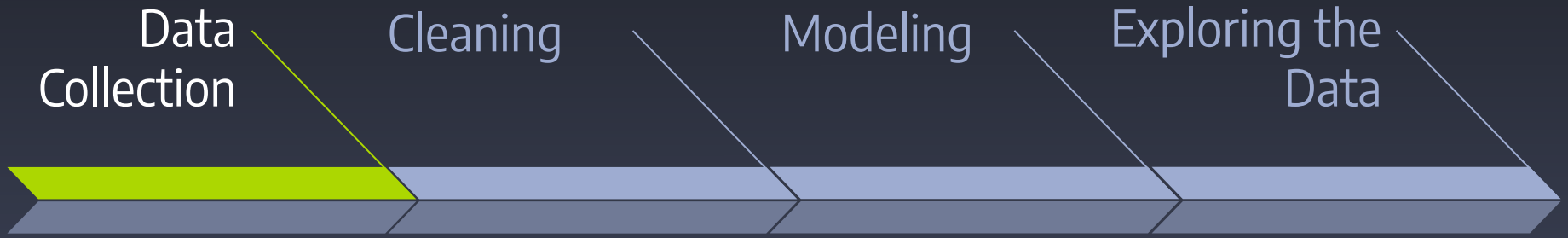
2. Based on that data, create a machine learning model that can accurately classify a post if the source is unknown

3. Provide insight into the vernacular of each fanbase by creating lists of the most used and most significant words to each community

Workflow Overview

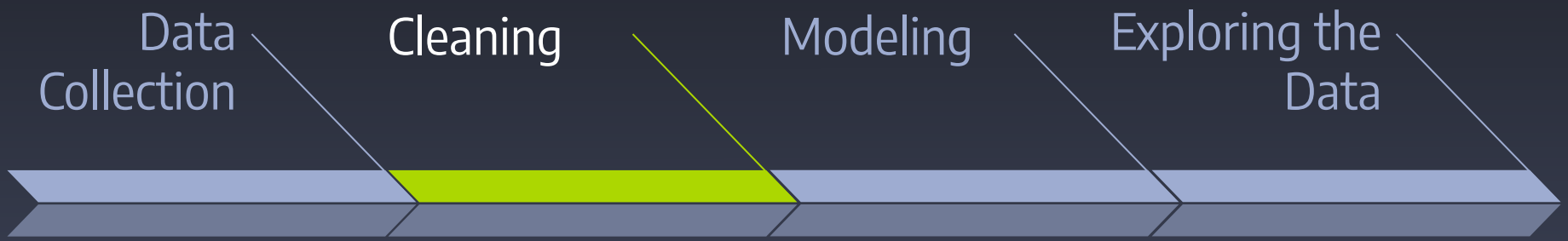


Workflow Overview



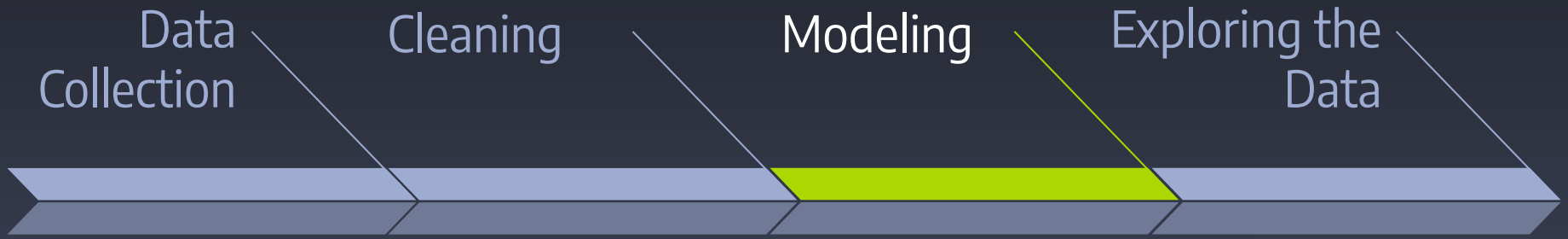
- 100k posts per subreddit
- Lots of cleaning required, but room for refinement and balancing the two subs

Workflow Overview



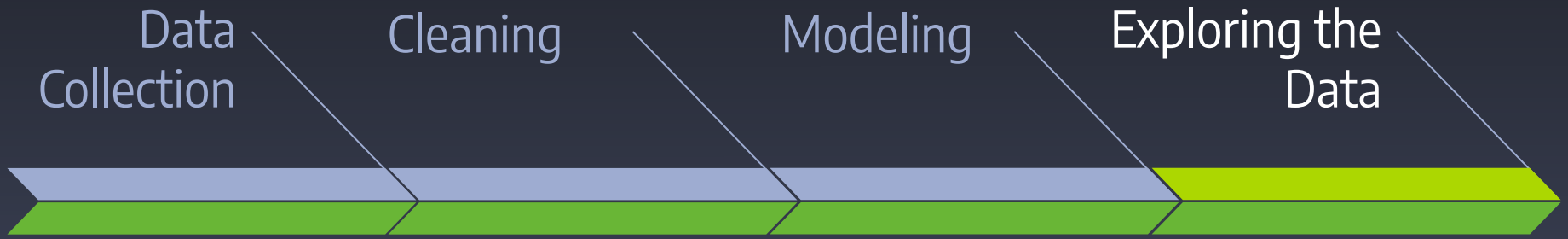
- Decreased to 12k total posts with a 56% / 44% split
- That number serves as my baseline - the number to beat

Workflow Overview



- Logistic Regression and Random Forest for interpretability
- Both returned excellent results

Workflow Overview

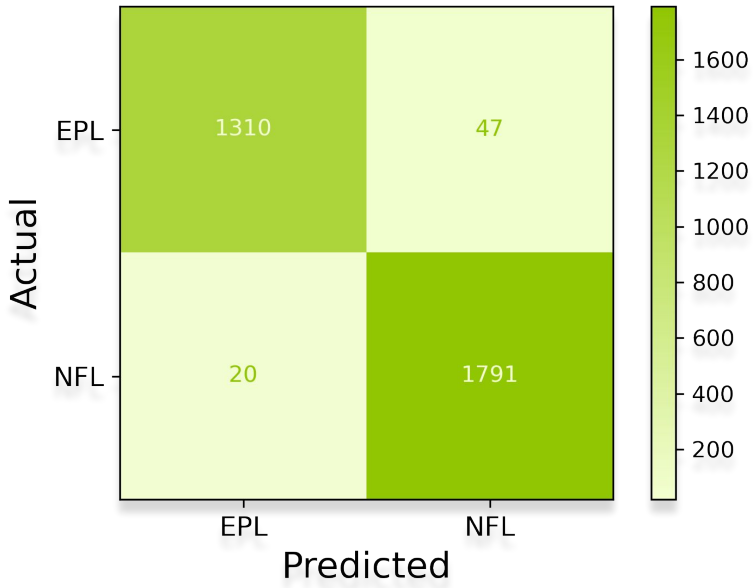


- EDA produced several lists that I can provide to OverArmor to help them craft their messaging

Model Comparison

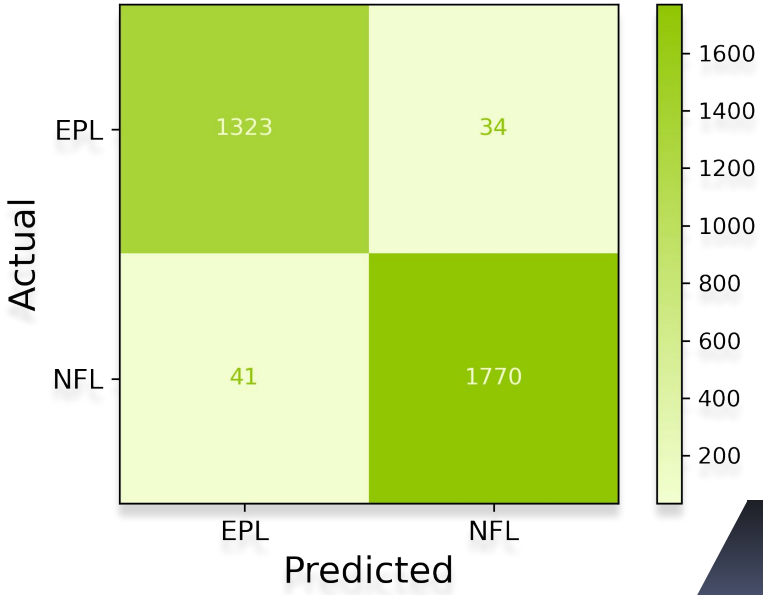
Logistic Regression

97.88% Test Accuracy

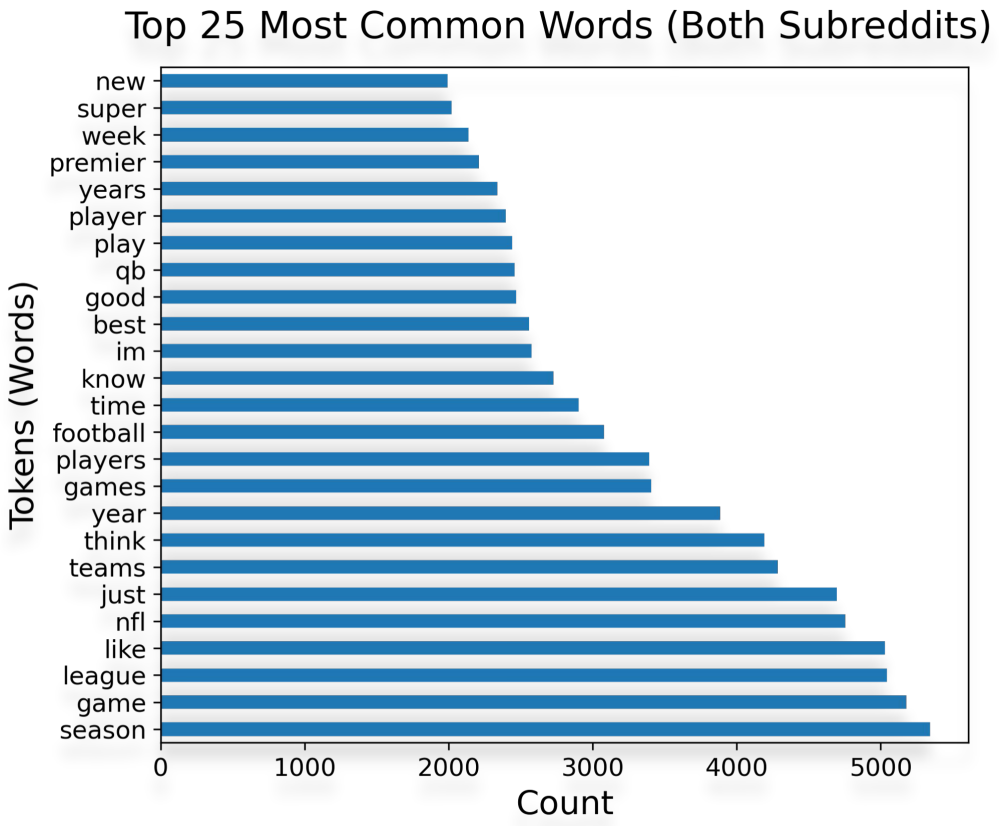


Random Forest

97.41% Test Accuracy



The Final Request - Count/Freq



The Final Request - Count/Freq

	count_all	freq%_all	count_nfl	freq%_nfl	count_epl	freq%_epl
team	7688.0	0.6214031684448755	5510.0	0.7892852026930239	2178.0	0.4040066777963272
season	5566.0	0.4498868412544455	3628.0	0.5196963185790001	1938.0	0.35948803561491377
game	5391.0	0.4357419980601358	4063.0	0.5820083082652915	1328.0	0.24633648673715453
league	5150.0	0.4162625282896864	1362.0	0.1951009883970778	3788.0	0.7026525690966425
like	5049.0	0.40809893307468476	3428.0	0.4910471279186363	1621.0	0.30068632906696346
nfl	4945.0	0.399692854833495	4898.0	0.7016186792723106		
just	4720.0	0.3815066278693825	3175.0	0.454805901733276	1545.0	0.2865887590428492
teams	4360.0	0.35240866472680243	3100.0	0.4440624552356396	1260.0	0.2337228714524207
think	4195.0	0.3390720982864533	2805.0	0.40180489901160293	1390.0	0.25783713596735297
year	4026.0	0.3254122211445199	3450.0	0.4941985388912763	576.0	0.10684474123539232
players	3430.0	0.2772389266084707	2304.0	0.3300386764073915	1126.0	0.2088666295677982
games	3389.0	0.27392499191723246	2401.0	0.34393353387766795	988.0	0.18326841031348545

The Final Request - Count/Freq

interested					193.0	0.036346516007532956
jaguars			440.0	0.060748308711859725		
joe			441.0	0.06088637304984123		
jones			468.0	0.06461411017534172		
kane					165.0	0.031073446327683617
leagues					220.0	0.04143126177024482
leeds					237.0	0.04463276836158192
leicester					405.0	0.07627118644067797
lions			514.0	0.07096506972249068		
list			452.0	0.06240508076763772		
live					236.0	0.044444444444444446
london					204.0	0.0384180790960452

Conclusions

1. Collected and cleaned data
2. Created 2 models that prioritized accuracy and succeeded in classifying posts at 97+% rates
3. Gathered various lists of language used in each subreddit

Other:

- Focus on particular teams and players who are mentioned more frequently in the subreddits

With more time...

1. Dive much deeper into the language
2. Look at frequency of 2 and 3 word strings (bi-grams/tri-grams)
3. Investigate misidentified posts
4. Run further testing on posts my models haven't seen

The Final Request - Model Results

r/PremierLeague

Feature Name	Odds
premier	0.14511658195545027
pl	0.32218635485068015
liverpool	0.346444771065564
league	0.4051798927874466
arsenal	0.4234677526175186
epl	0.43721412092479994
club	0.4381214188759073
chelsea	0.44666040454104927
united	0.4907167042099343

r/nfl

Feature Name	Odds
nfl	12.774481362420794
qb	2.334324906682998
bowl	2.229995708321256
yards	1.8885643568614652
draft	1.8811256326337626
rnfl	1.736748422376564
offense	1.60905265437232
field	1.6027441511167724
eagles	1.572497816145432