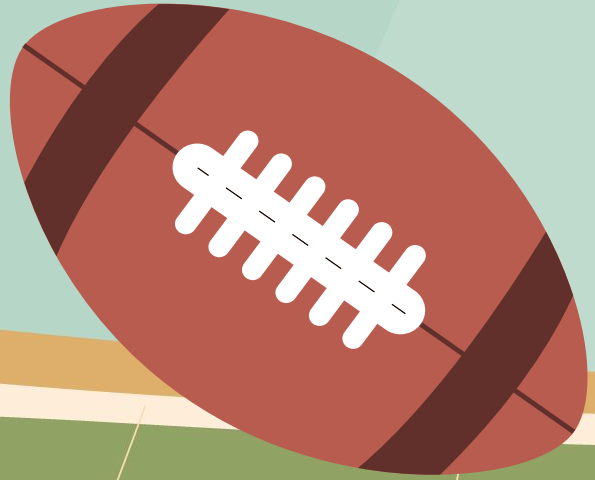
Subreddit Classification

Project 3

Football v Football

r/NFL



r/soccer

The Problem

- Do you want to know which football people are talking about?
- Do you want to know if you're in America or Europe?
- This project involves the creation of multiple models that classify reddit posts as belonging to either r/NFL or r/soccer
- Using data taken from the first two weeks of October 2022

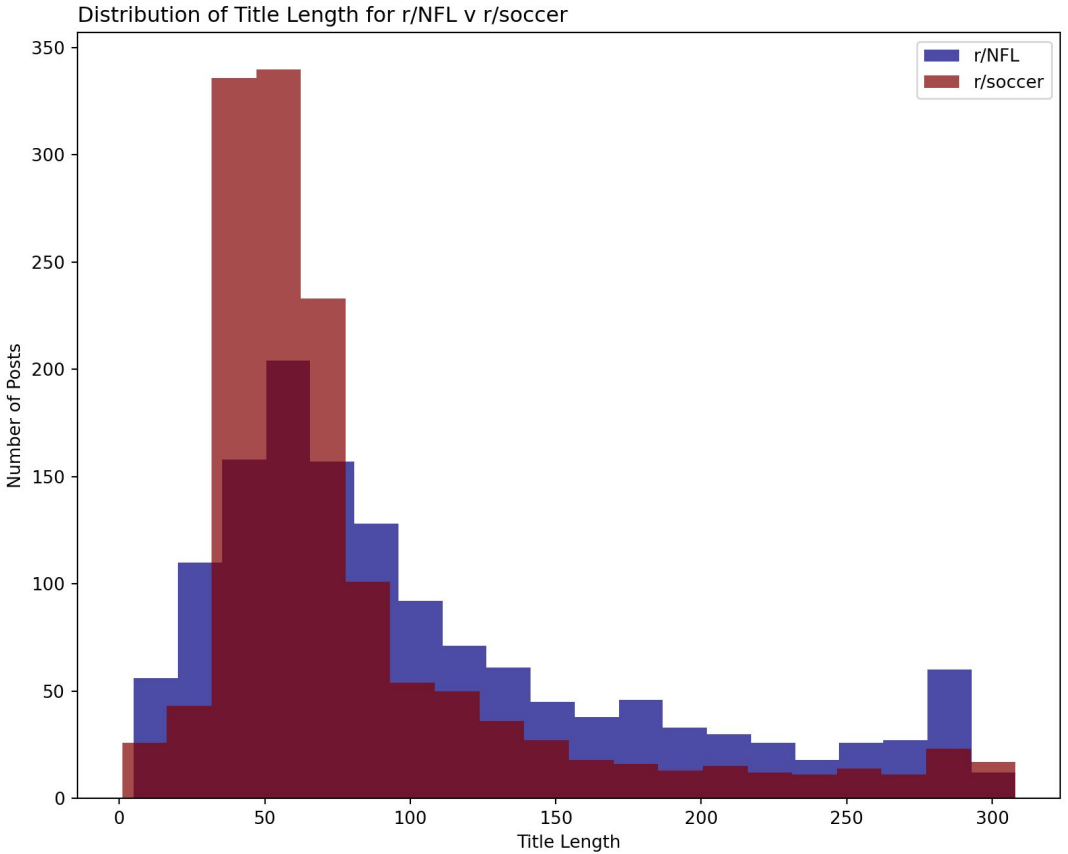
Subreddit Comparison

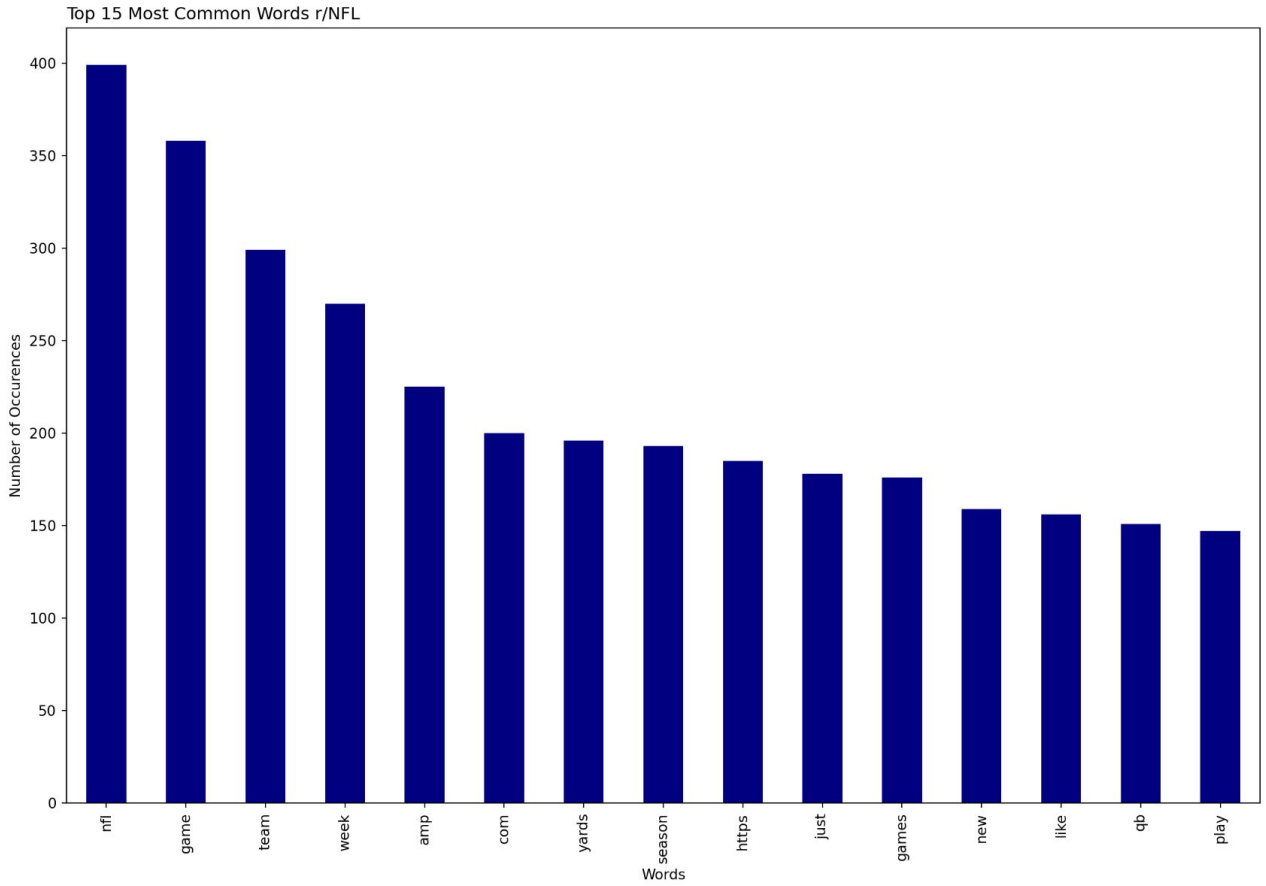
r/NFL

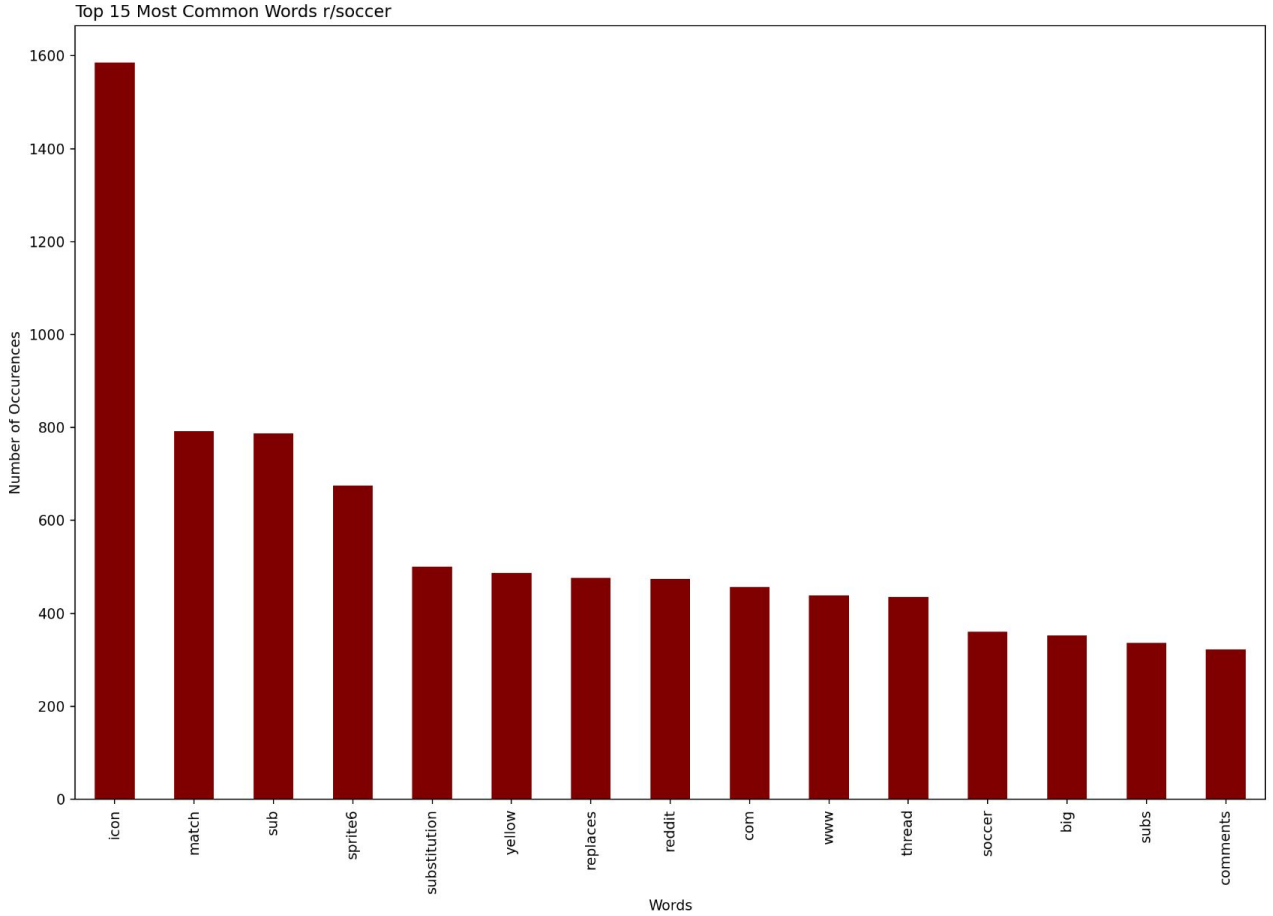
- Subreddit for following the NFL specifically
- 3 million members
- Majority of posts are links to Twitter

r/soccer

- Subreddit for all soccer leagues
- 3.5 million members
- More match threads and highlights


Distribution of Title Length for r/NFL v r/soccer
r/NFL
r/soccer
Number of Posts
Title Length
0
50
100
150
200
250
300
350


Top 15 Most Common Words r/NFL
Number of Occurences
0
50
100
150
200
250
300
350
400
nfl
game
team
week
amp
com
yards
season
https
just
games
new
like
qb
play
Words


Top 15 Most Common Words r/soccer
Number of Occurences
0
200
400
600
800
1000
1200
1400
1600
icon
match
sub
sprite6
substitution
yellow
replaces
reddit
com
www
thread
soccer
big
subs
comments
Words

Icon?

- The most common word by far is ‘icon’
- There are other confusing words like ‘sprite6’

From Sports Illustrated

The Culprit Revealed

MATCH EVENTS | *via ESPN*[2]

45' Substitution, PSV Eindhoven. Luuk de Jong replaces Anwar El Ghazi.

55' Goal! PSV Eindhoven 1, Arsenal 0. Joey Veerman (PSV Eindhoven) left footed shot from the centre of the box to the top right corner. Assisted by Luuk de Jong.

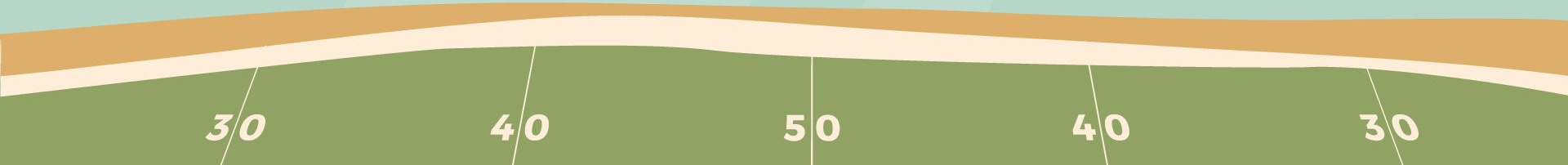
56' Substitution, Arsenal. Thomas Partey replaces Albert Sambi Lokonga.

57' Substitution, Arsenal. Bukayo Saka replaces Martin Ødegaard.

58' Kieran Tierney (Arsenal) is shown the yellow card for a bad foul.

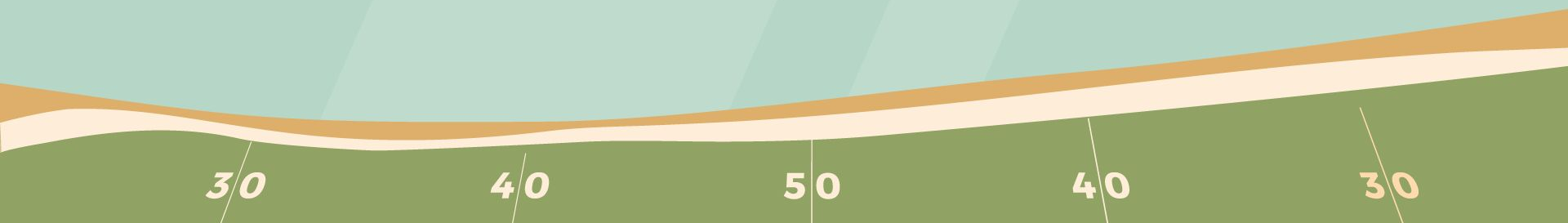
63' Goal! PSV Eindhoven 2, Arsenal 0. Luuk de Jong (PSV Eindhoven) header from the centre of the box to the bottom right corner. Assisted by Cody Gakpo with a cross following a corner.

From reddit.com/r/soccer



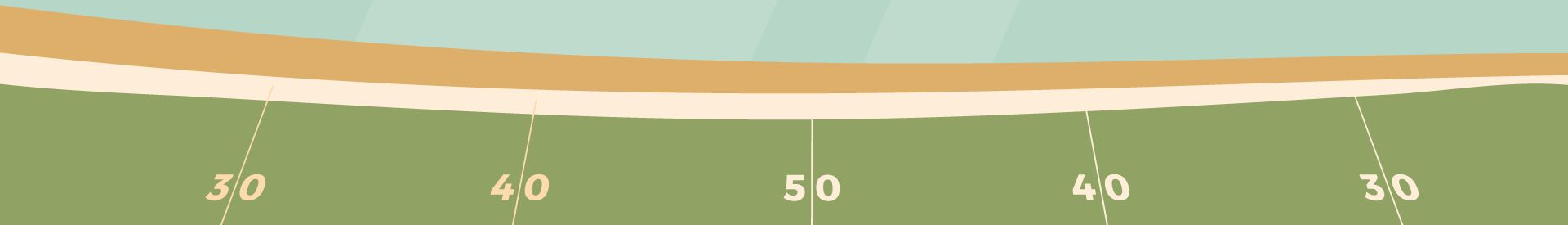
Random Forest

- A simple model that still produces good results
- The random forest should have lower variance than other tree based models
- This model is 93.2% accurate on test data



Support Vector Machine

- Good fast model for generating predictions
- Works well with NLP data
- This model’s accuracy is 91,7%



- Both models vastly outperform the baseline of 50/50
- Overall, the random forest model works the best

THANKS!

Do you have any questions?