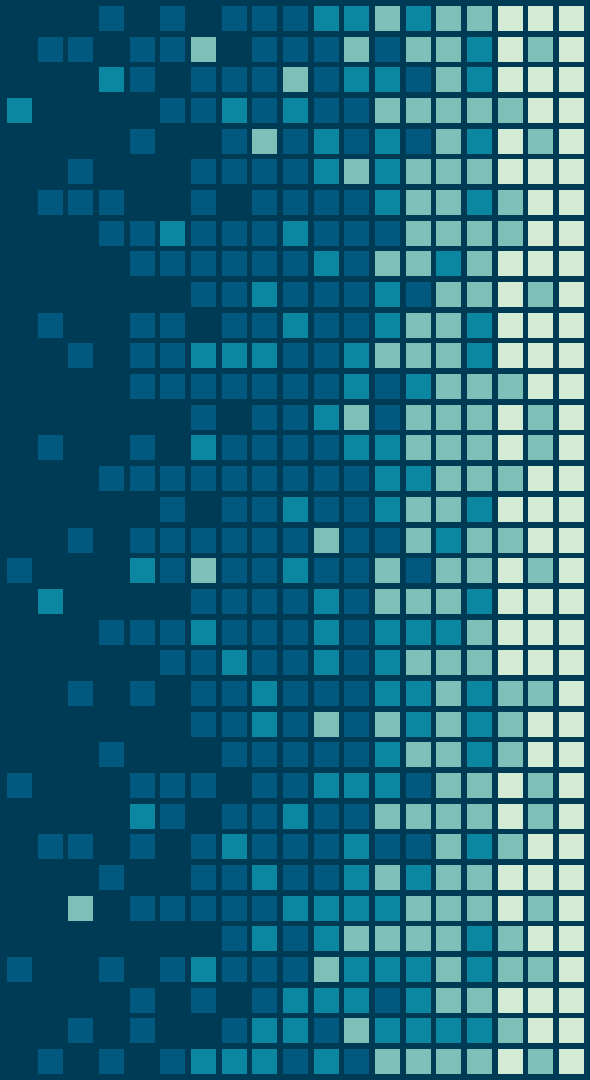
WHAT SPORT ARE WE TALKING?

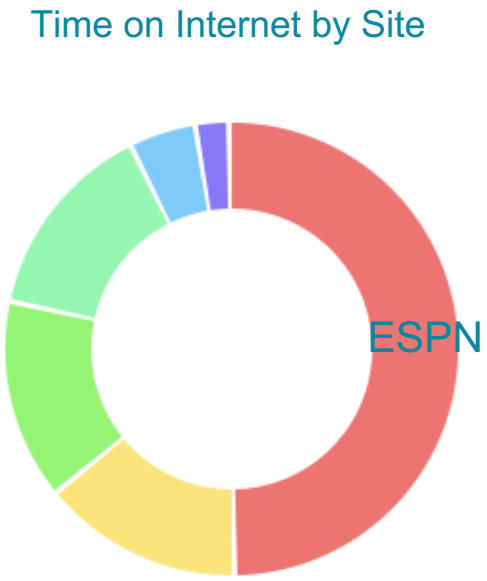
USING NLP & CLASSIFICATION TO DIFFERENTIATE BETWEEN R/NFL & R/NBA

COLLEEN KENNEY
GENERAL ASSEMBLY 2019



BACKGROUND

- I LOVE sports
- I spend a lot of time on ESPN
- Some people could care less



Source: Google Chrome WebTime Tracker

PROBLEM STATEMENT

How can I differentiate between posts on r/nfl and r/nba?

APPROACH

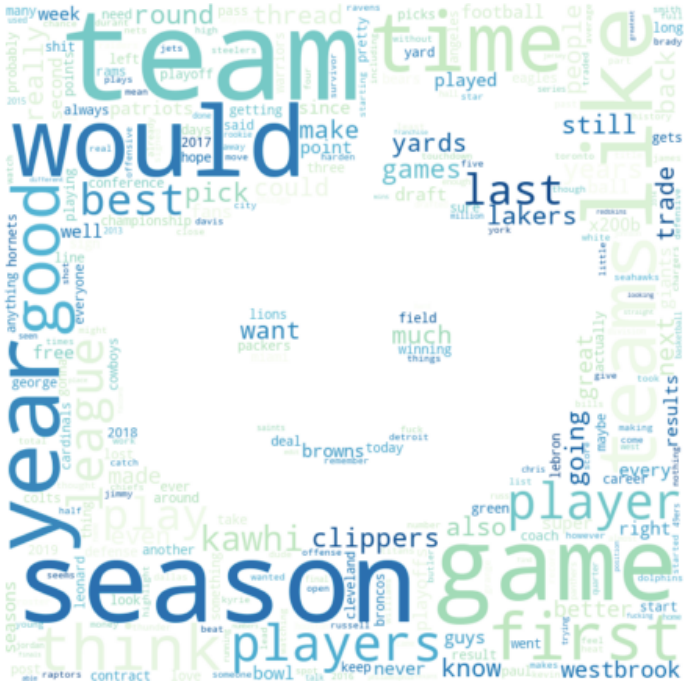
WEBSCRAPING

1,731 posts
from r/nfl & r/nba combined

5,502 comments
from r/nfl & r/nba combined

60%
from r/nba

EXPLORATORY DATA ANALYSIS



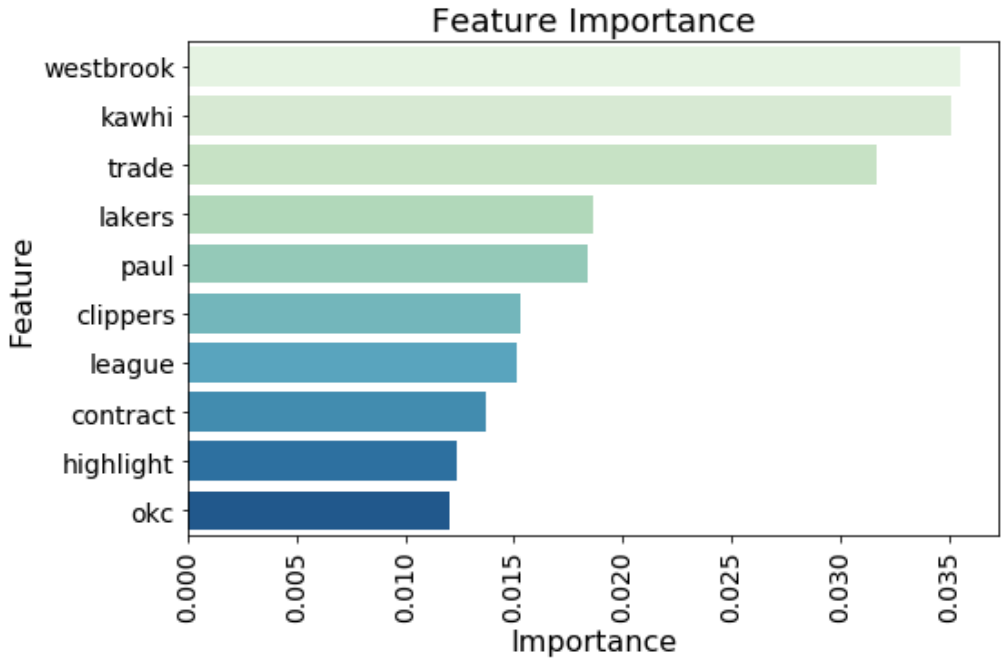
MODELING

- Use GridSearchCV to find best vectorizer parameters based on logistic regression
- Define function to show best parameters and scores for dictionary of models
- Implement function using vectorized data

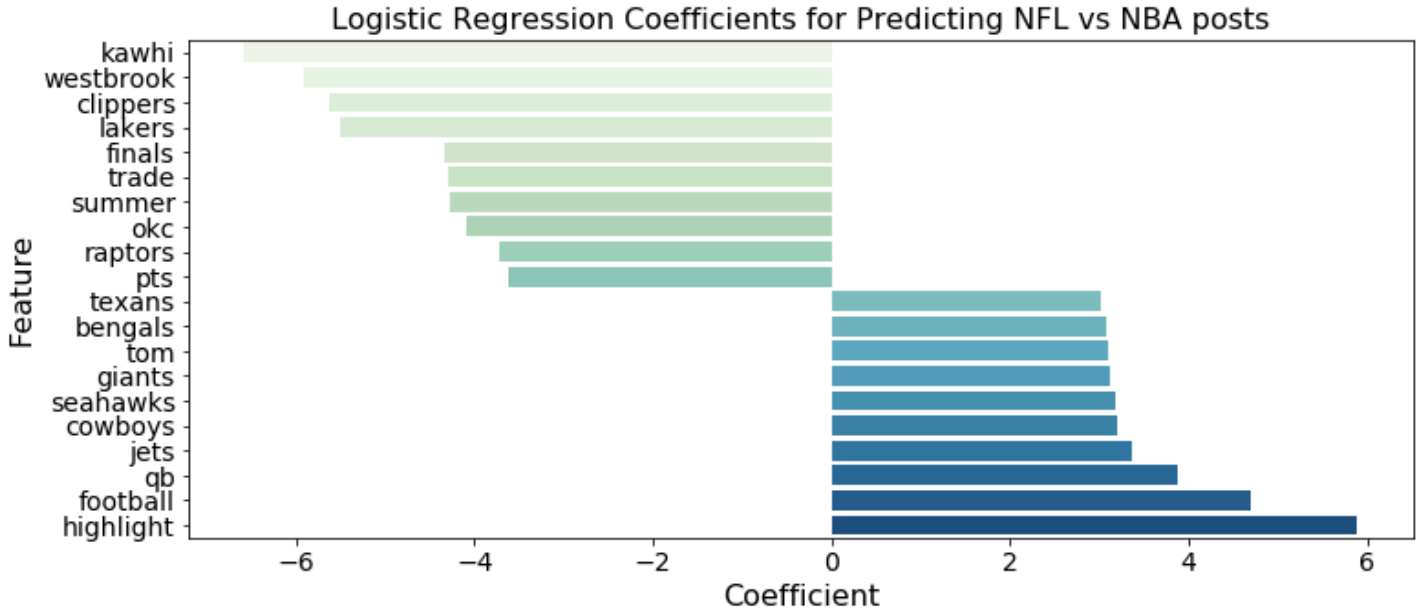
ACCURACY RATES

Iteration	Logistic Regression	Multinomial Naïve Bayes	Random Forests
Posts & Comments	**84.91%**	**85.63%**	**83.53%**
Posts	**95.61%**	**95.61%**	**91.45%**
Posts without ‘nba’ or ‘nfl’	**93.07%**	**92.61%**	**87.76%**

FEATURE IMPORTANCES



COEFFICIENT STRENGTH



CONCLUSIONS & NEXT STEPS

- Posts are easier to differentiate than posts & comments
- Logistic Regression was able to differentiate posts without the use of ‘nba’ or ‘nfl’ with a 93.07% accuracy rate.
- Player names, team names, positions, and unique verbs proved to be stronger differentiators
- Gather additional data over time to get more relevant terms overall rather than right now
- Scrape sports-related subreddits to expand into multi-class models