CLASSIFICATION MODELING AND NATURAL LANGUAGE PROCESSING OF SPORTS SUBREDDITS

ANALYSIS AND VISUALIZATION BY:
TEMPLE MOORE
DATA SCIENTIST

Sports Subreddits?

- Reddit is a news aggregation, rating, and discussion website, allows users to create communities
- “There is a subreddit for everything”
- r/nba and r/nfl are the two of the most popular subreddits

Classification

- My task was to create classification models to predict whether or not a post originated from a specific subreddit
- Objective was to find the most important terms in classifying the posts. From there, use domain knowledge of the subreddits to analyze why the terms were deemed important by the model and to infer the nature of discussion in the subreddits

Process

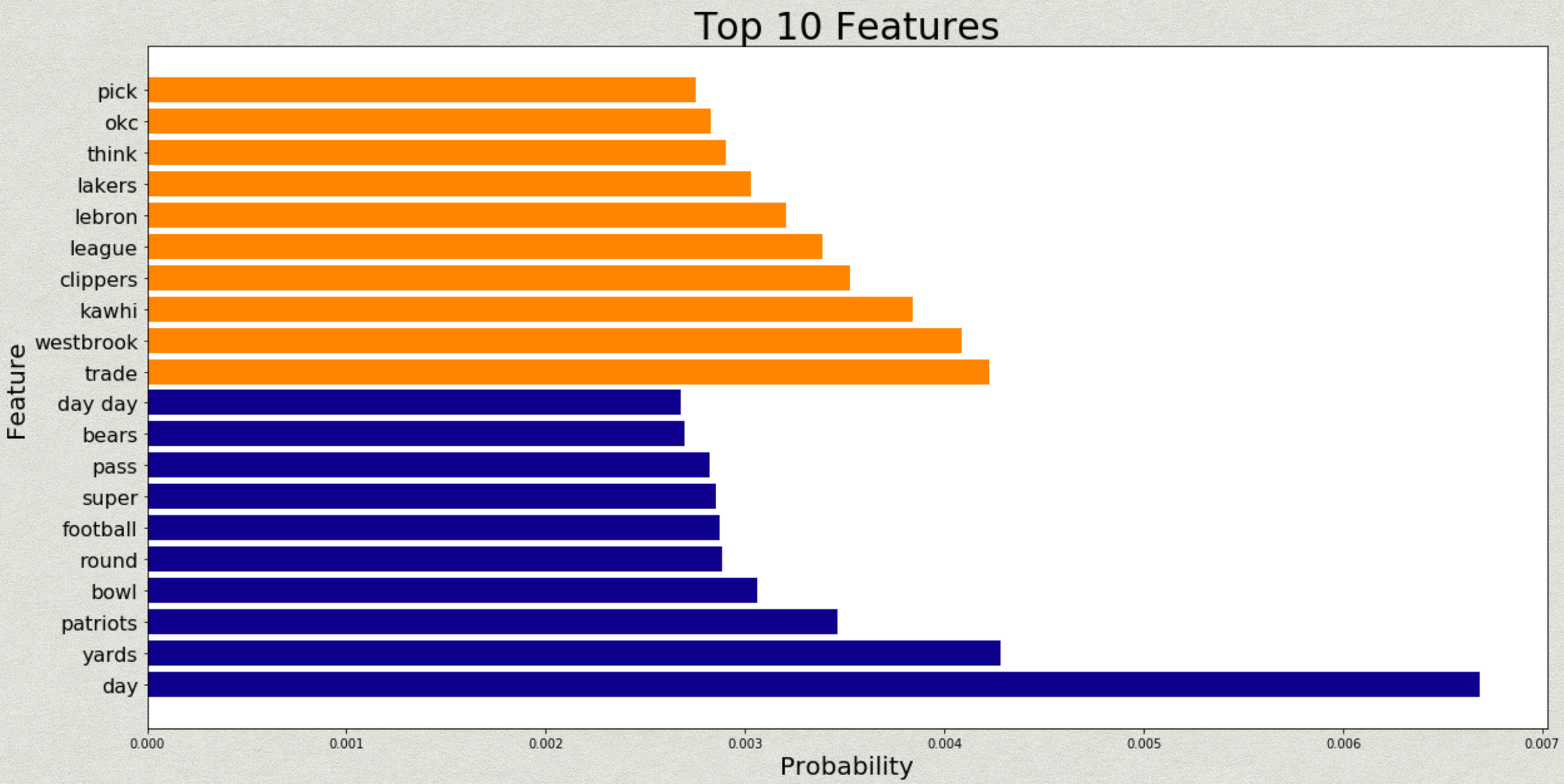
- Used Reddit's built in API to gather data
- Combined the title of posts and the "Self text" for analysis
- Used regular expressions to clean the text and remove words frequently classified in both subreddits
- Used two vectorization methods: Count Vectorization and TF-IDF

MULTINOMIAL NAIVE BAYES CLASSIFICATION

NB with Count Vectorized Features

- Accuracy score of 95.37% with a precision of 95%, this was the best performing model
- Predicted correctly from r/nfl 97% of the time, and r/nba 91%
- Most important words for r/nba referenced free agency moves like "westbrook", "kawhi", "lebron", but weighted more generally related terms like "trade" and "pick" as important.
- Important words in classifying r/nfl: “day”, “yards”, “patriots”, and “round”

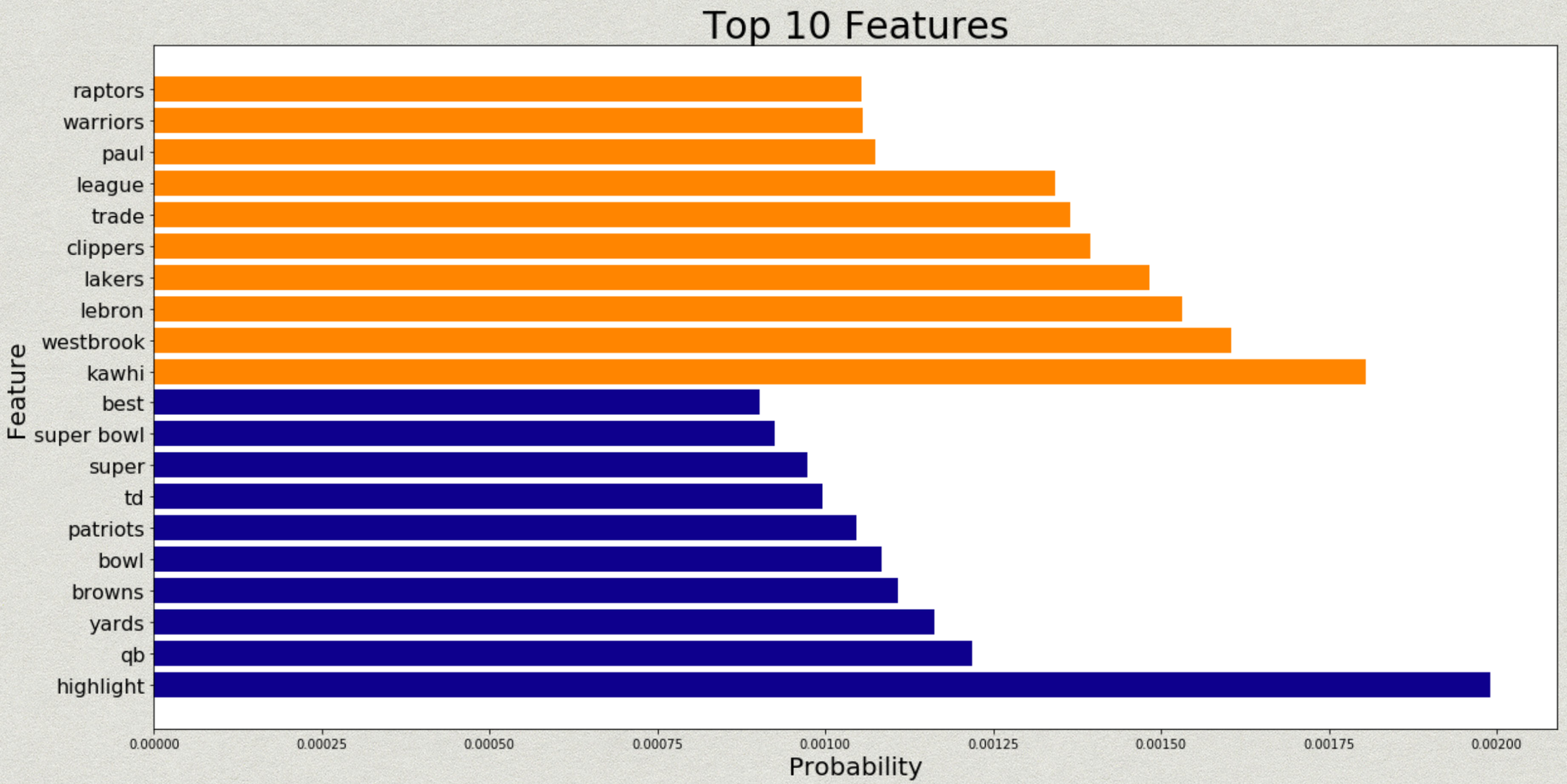
Naive Bayes with Count Vectorized Features



NB with TF-IDF Features

- Accuracy score of 93.39%, weighted precision of 94%
- Correctly predicted r/nfl 98% of the time, r/nba 91%
- Words found in discussions about free agency in r/nba were wighted as important
- Important words for r/nfl changed: “highlight”, “qb”, “yards”, and “browns”

Naive Bayes with TF-IDF Vectorized Features



SUPPORT VECTOR MACHINE CLASSIFICATION

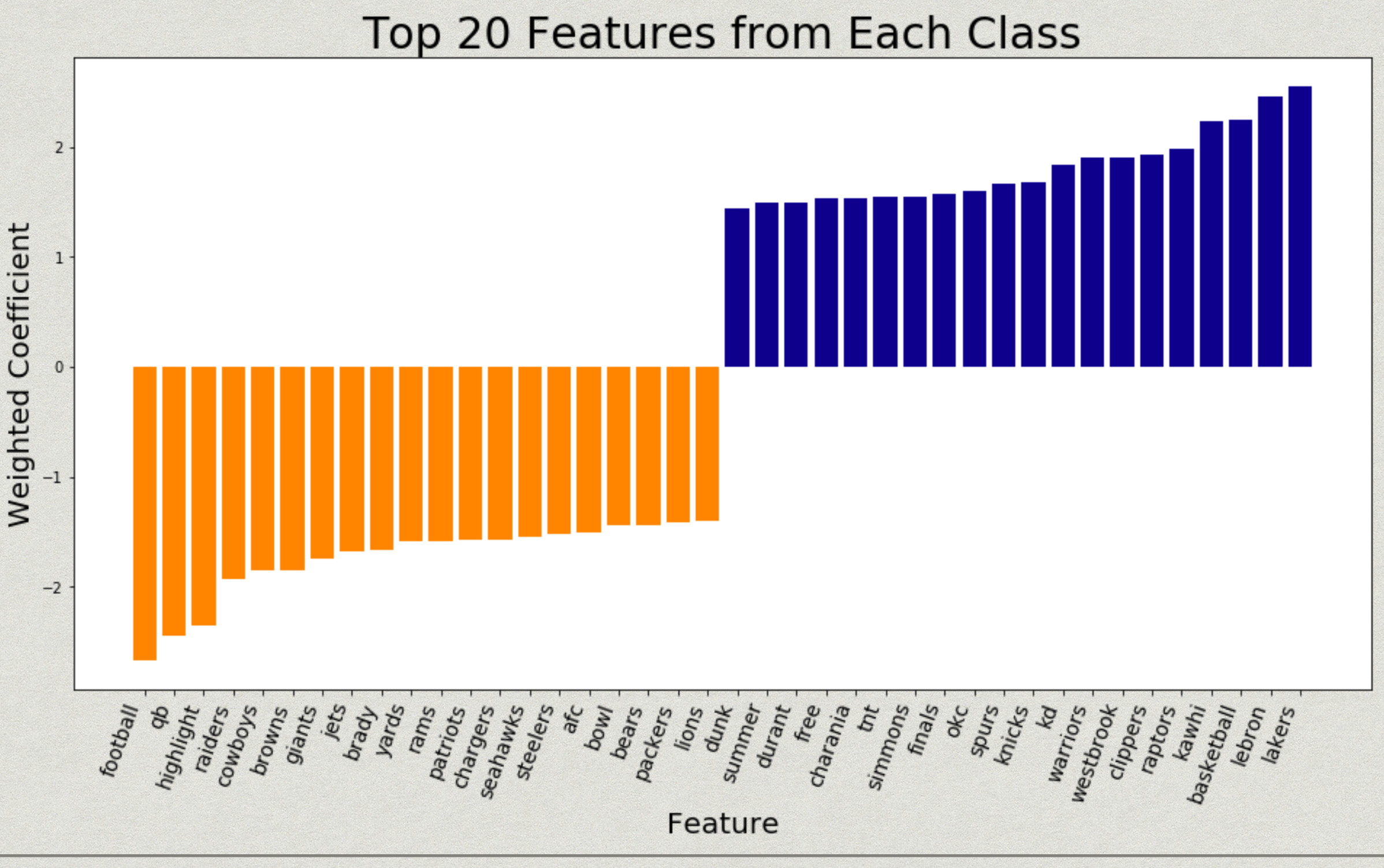
SVM Classifier with TF-IDF Features

- Accuracy score of 93.14% and a weighted precision of 93%

- Important words for r/nba: “lakers”, “lebron”, “basketball”, “kawhi”

- Important words for r/nfl: “football”, “qb”, “highlight”, “raiders”

SVC with TF-IDF Vectorized Features

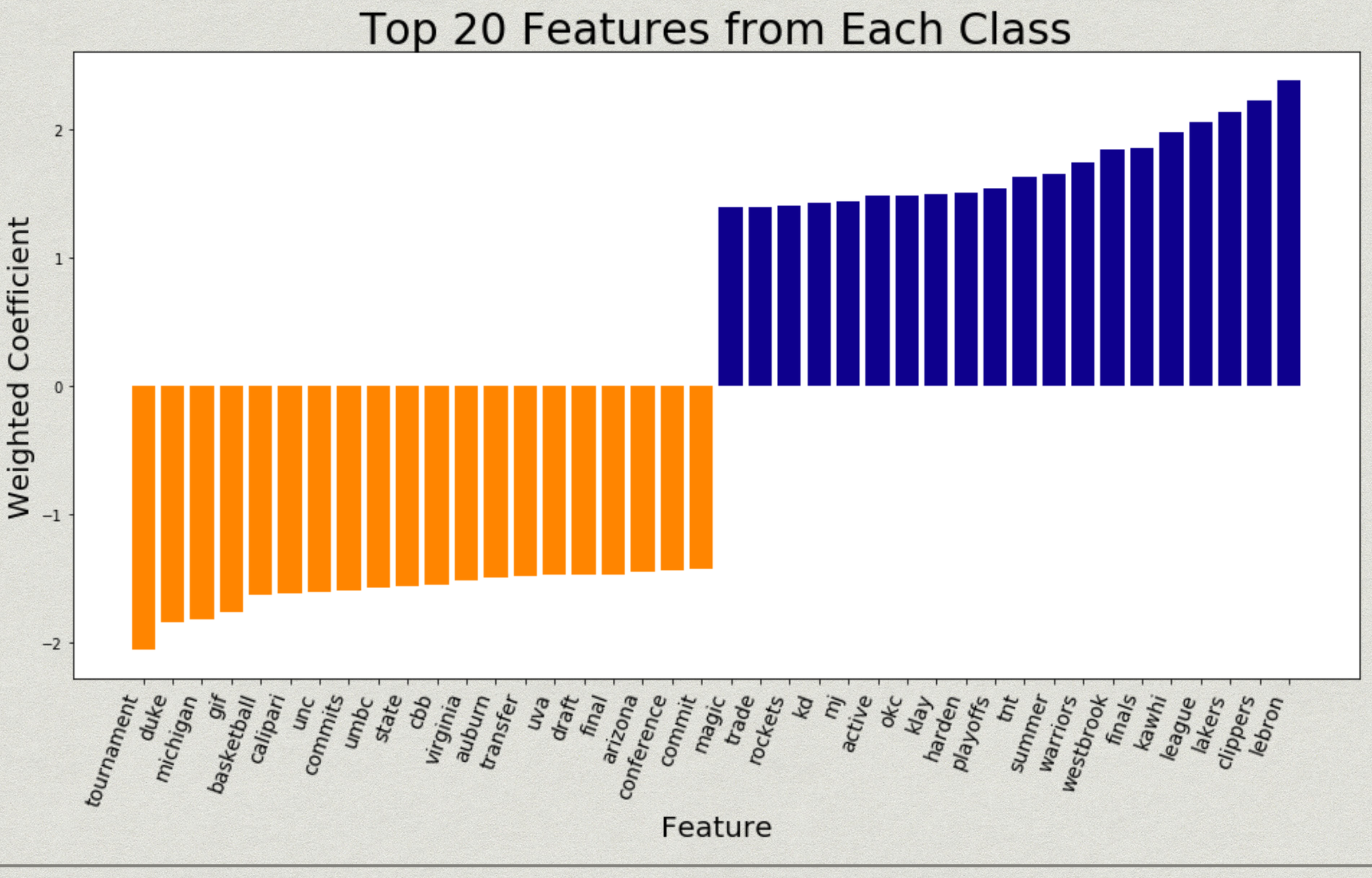


LETS ADD TWO MORE SUBREDDITS INTO THE EQUATION!

r/nba and r/CollegeBasketball SVM Classifier with TF-IDF

- Performed similarly: accuracy score 91.39%, weighted precision of 91%
- r/CollegeBasketball: “tournament”, “duke”, “michigan”, and surprisingly, “basketball”

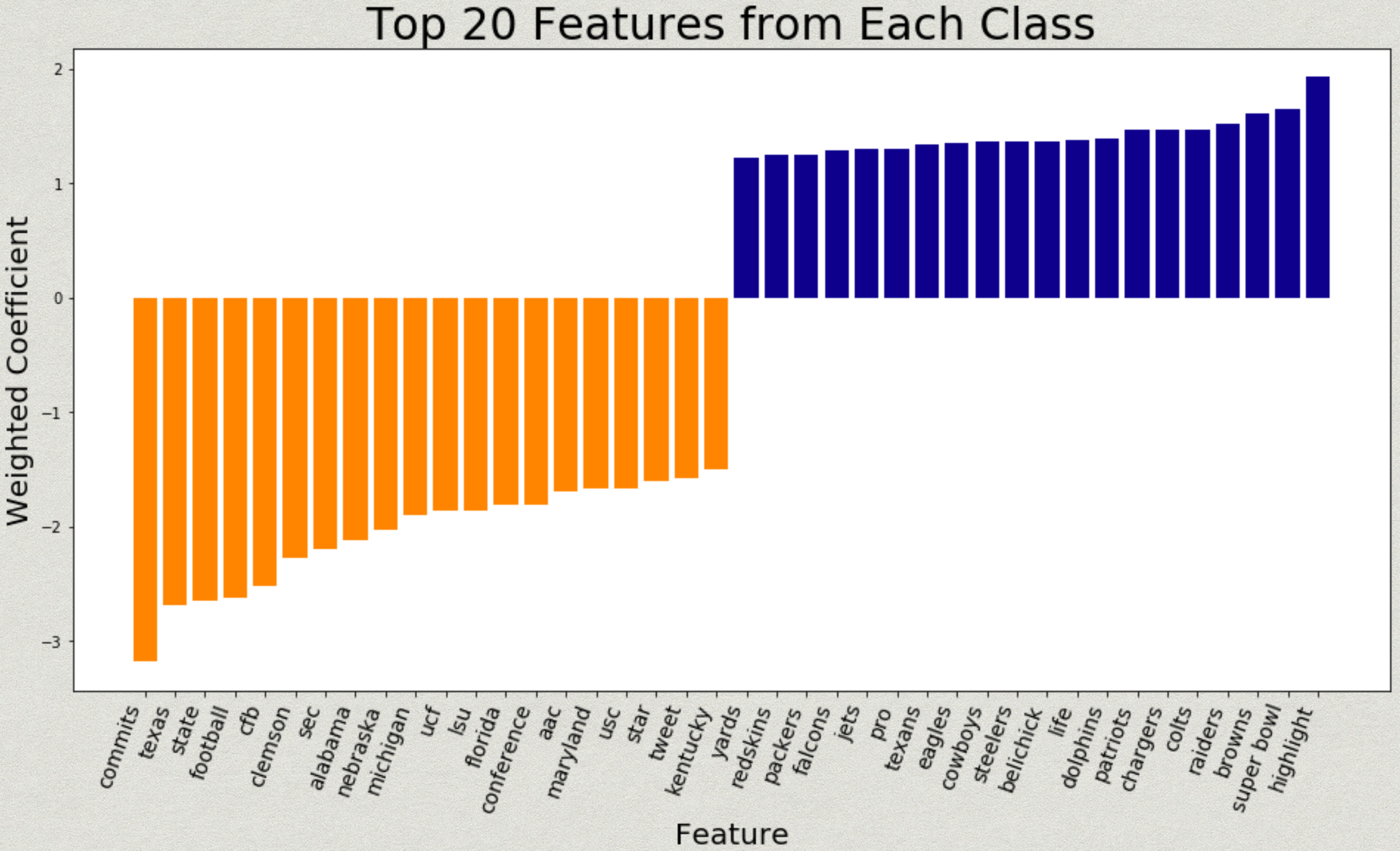
r/CollegeBasketball vs. r/nba



r/cfb and r/nfl
SVM Classifier with TF-IDF features

- Accuracy score of 92.43% and precision of 93%
- Important terms for r/cfb: “commits”, “texas”, “state”
- Terms that would be frequently used in both were classified under r/cfb. i.e. “tweet”, “conference”, “football”

r/cfb vs. r/nfl


Top 20 Features from Each Class
Weighted Coefficient
2
1
0
−1
−2
−3
commits
texas
state
football
cfb
clemson
sec
alabama
nebraska
michigan
ucf
lsu
florida
conference
aac
maryland
usc
star
tweet
kentucky
yards
redskins
packers
falcons
jets
pro
texans
eagles
cowboys
steelers
belichick
life
dolphins
patriots
chargers
colts
raiders
browns
super bowl
highlight
Feature

Conclusions

- For r/nba, player names and transactions were most important in classification
- For r/nfl, team oriented, game specific terms like "td" and "yard" were important across all models
- Collegiate variants: school pride is important, coaches and recruiting come up more

THANK YOU