SUBREDDITS CLASSIFICATION

FERAS ALTWAL, DSIR-824

PROBLEM STATEMENT

To Classify posts from two subreddits:

NBA and Soccer

SUBREDDITS

- **NBA:**
  - 50K post titles
  - July 26, 2020 – October 8, 2020

- **Soccer:**
  - 50K post titles
  - July 13, 2020 – October 8, 2020

WORKFLOW

Data Collection

Pushshift API

Data Cleaning

Merge data
Remove duplicates

Visualization

Word frequency
Title length

preprocessing

CountVectorizer
TfidfVectorizer

Modeling

Logistic Regression
Naïve Bayes
Random Forest

Validation

Precision
Accuracy

DATA COLLECTION

- **Pushshift API**

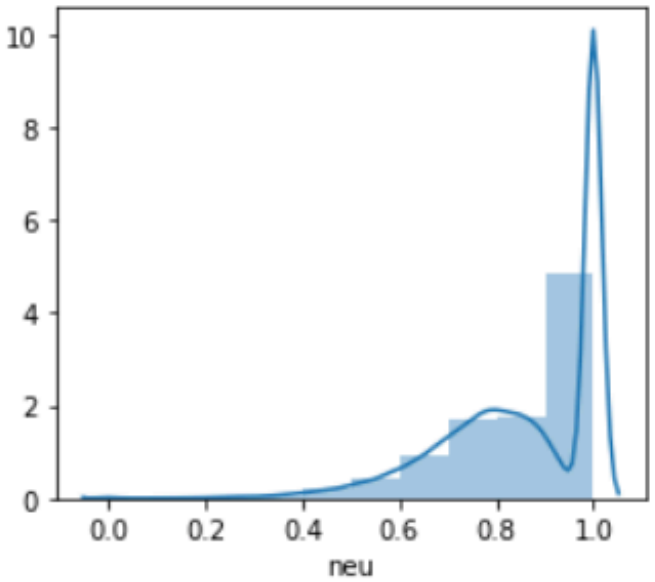
- **Collected 50K posts for each subreddit:**
  - title, created utc, selftext, subreddit, author, permalink
  - Target: NBA: 0, Soccer: 1

DATA CLEANING

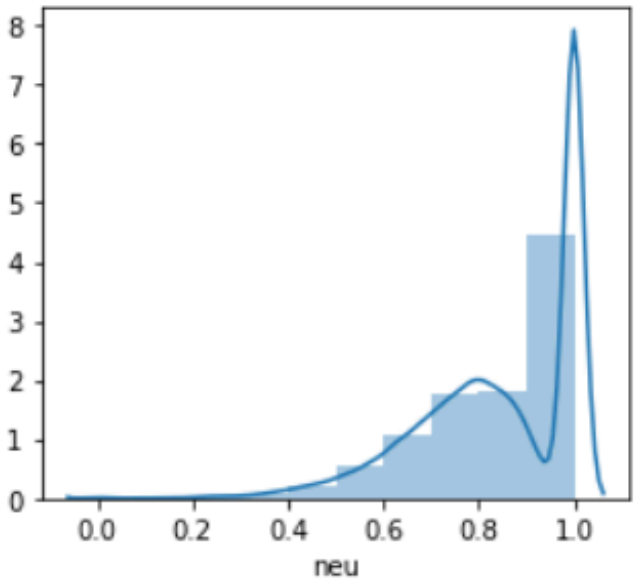
- Combined NBA and Soccer subreddits
- Removed duplicates (5261 duplicate posts)

SENTIMENT ANALYSIS

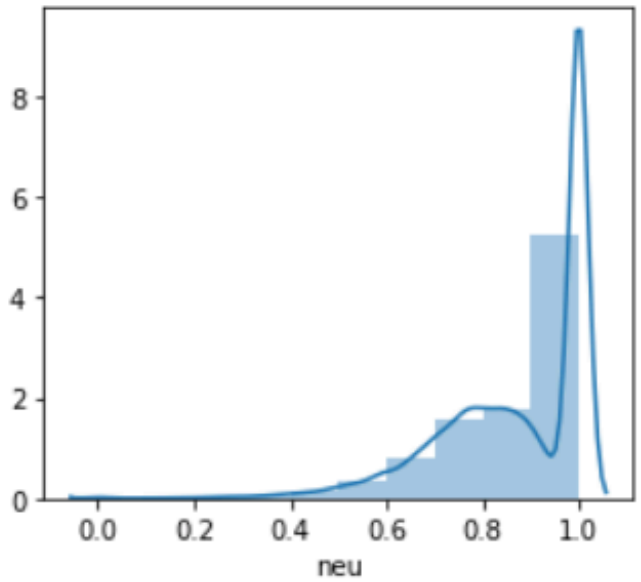
Both



NBA

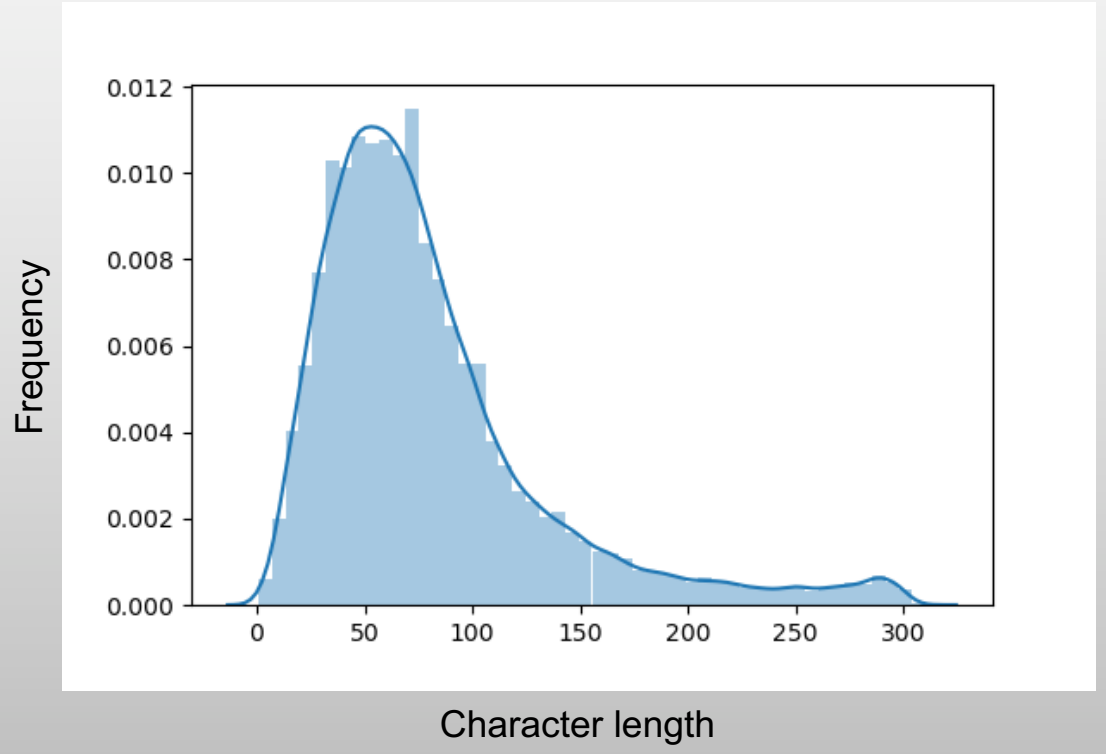


Soccer

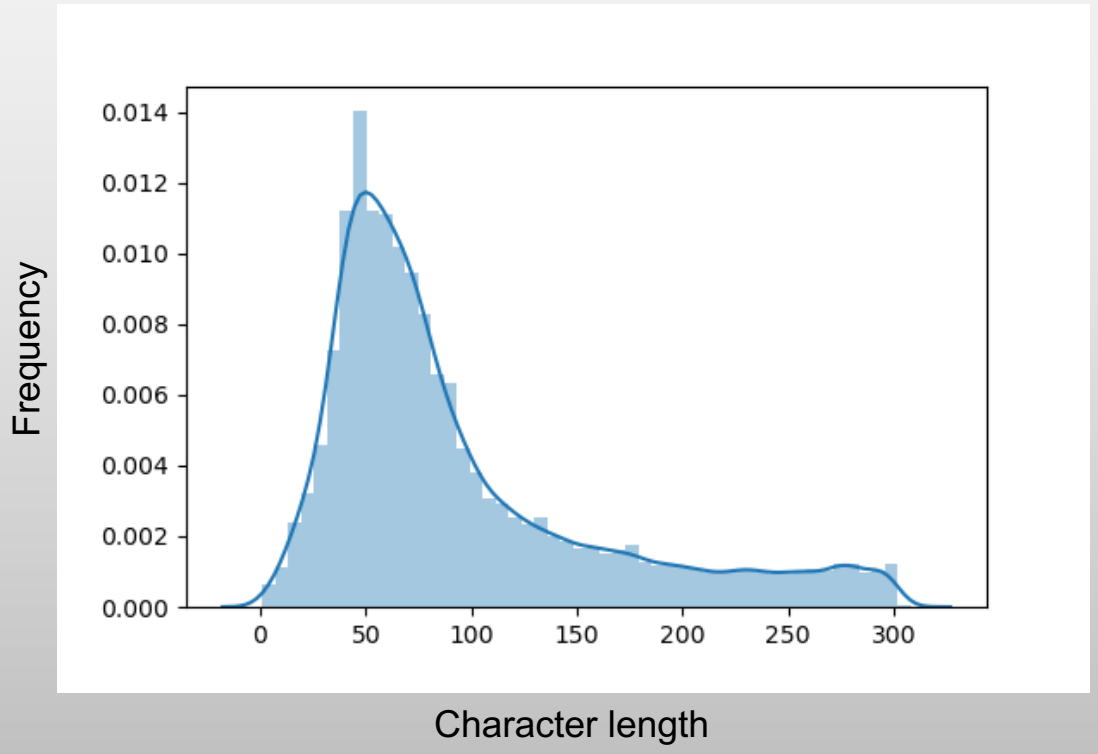


TITLE LENGTH

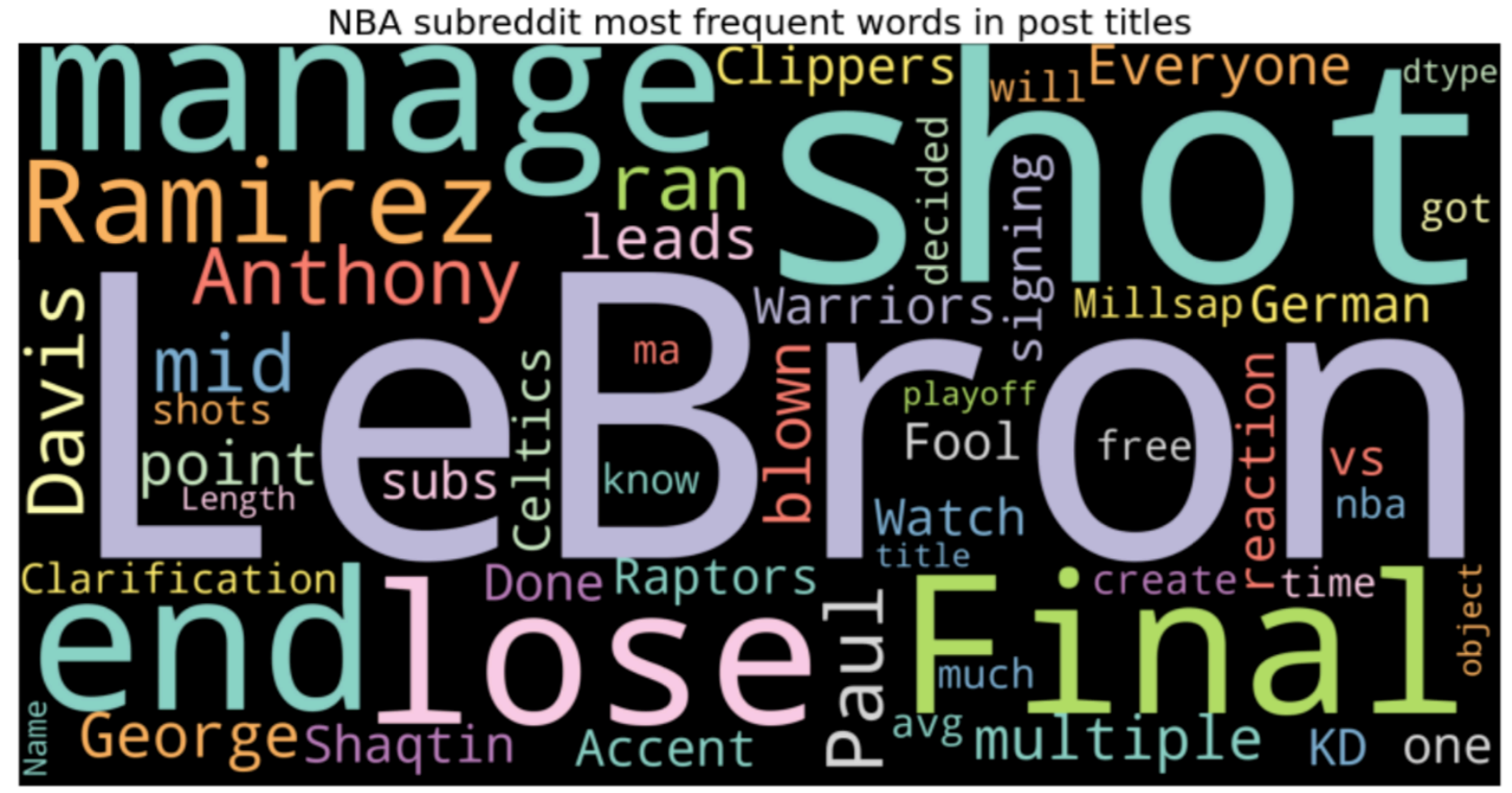
NBA



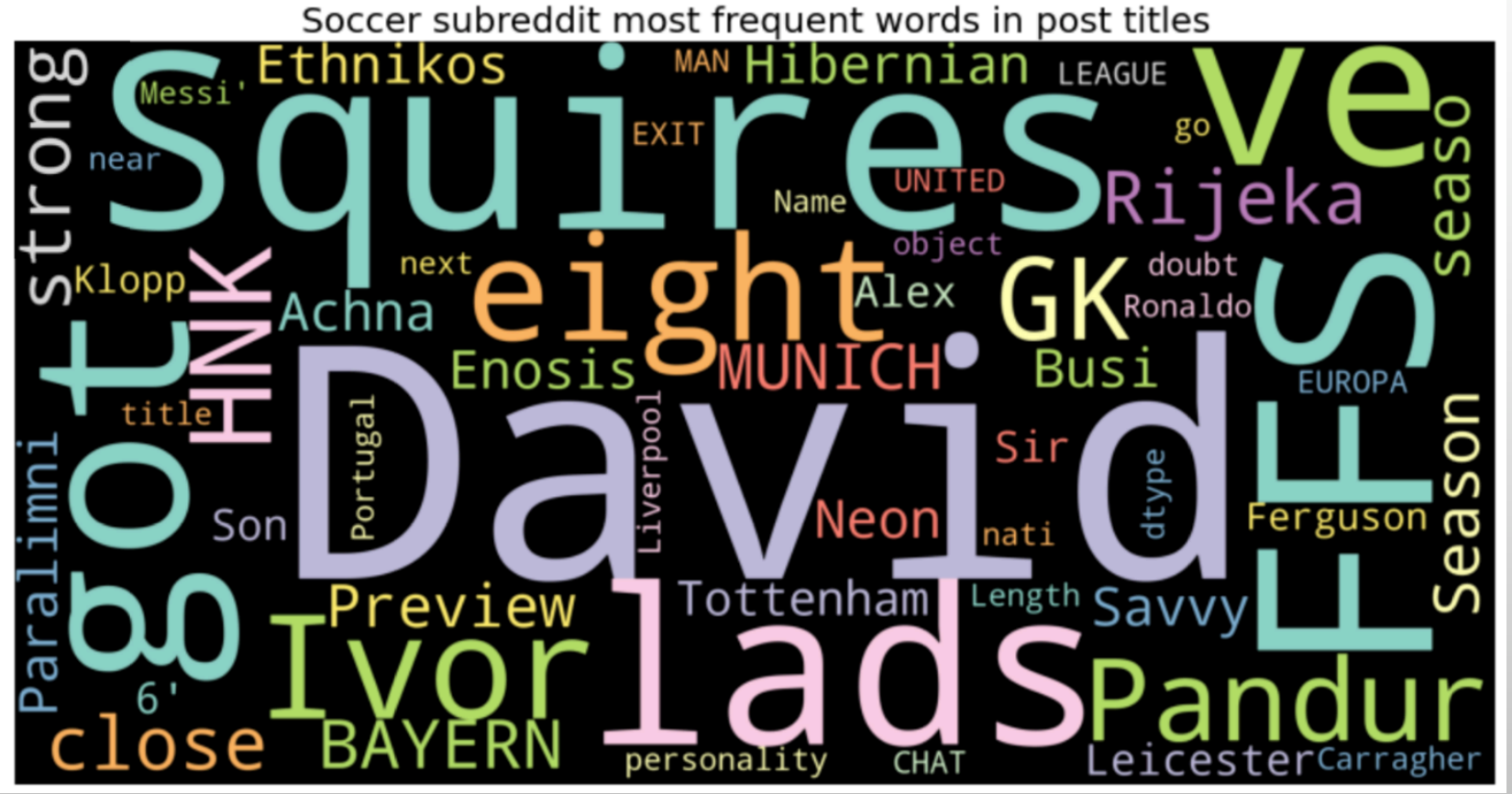
Soccer



MOST USED WORDS



MOST USED WORDS



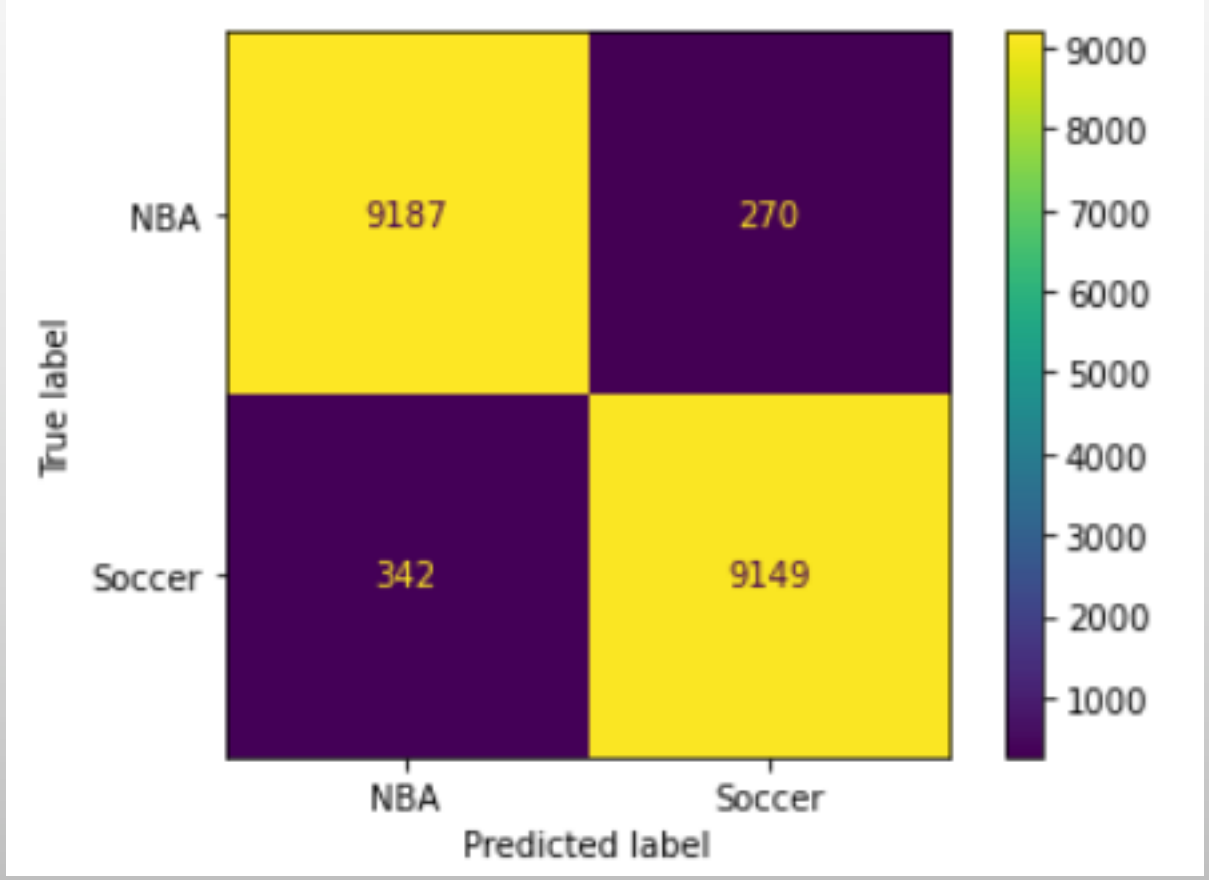
MODELING
LOGISTIC REGRESSION

Training score: 0.998
Testing score: 0.968

Precision: 0.971

Accuracy: 0.968

Logistic Regression



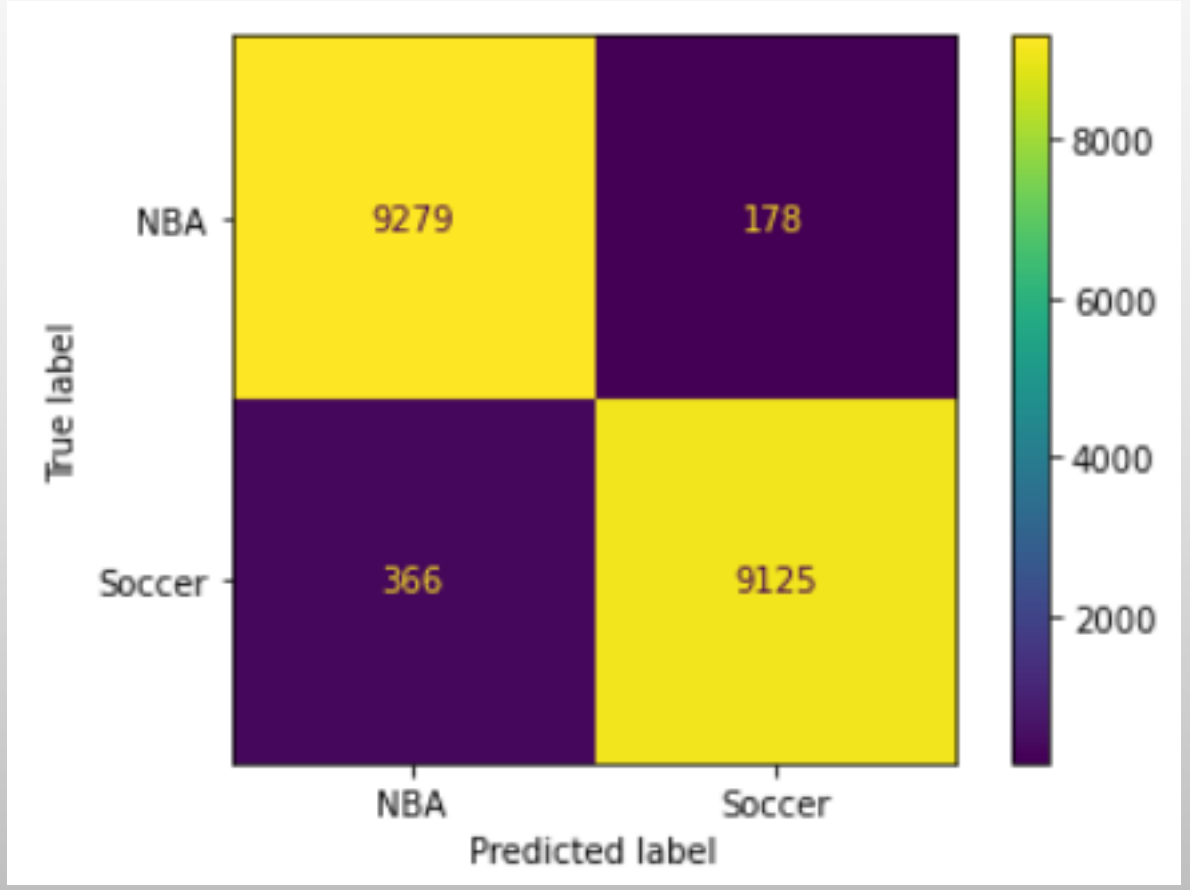
MODELING
NAÏVE BAYES

Training score: 0.989
Testing score: 0.971

Precision: 0.980

Accuracy: 0.971

Naïve Bayes



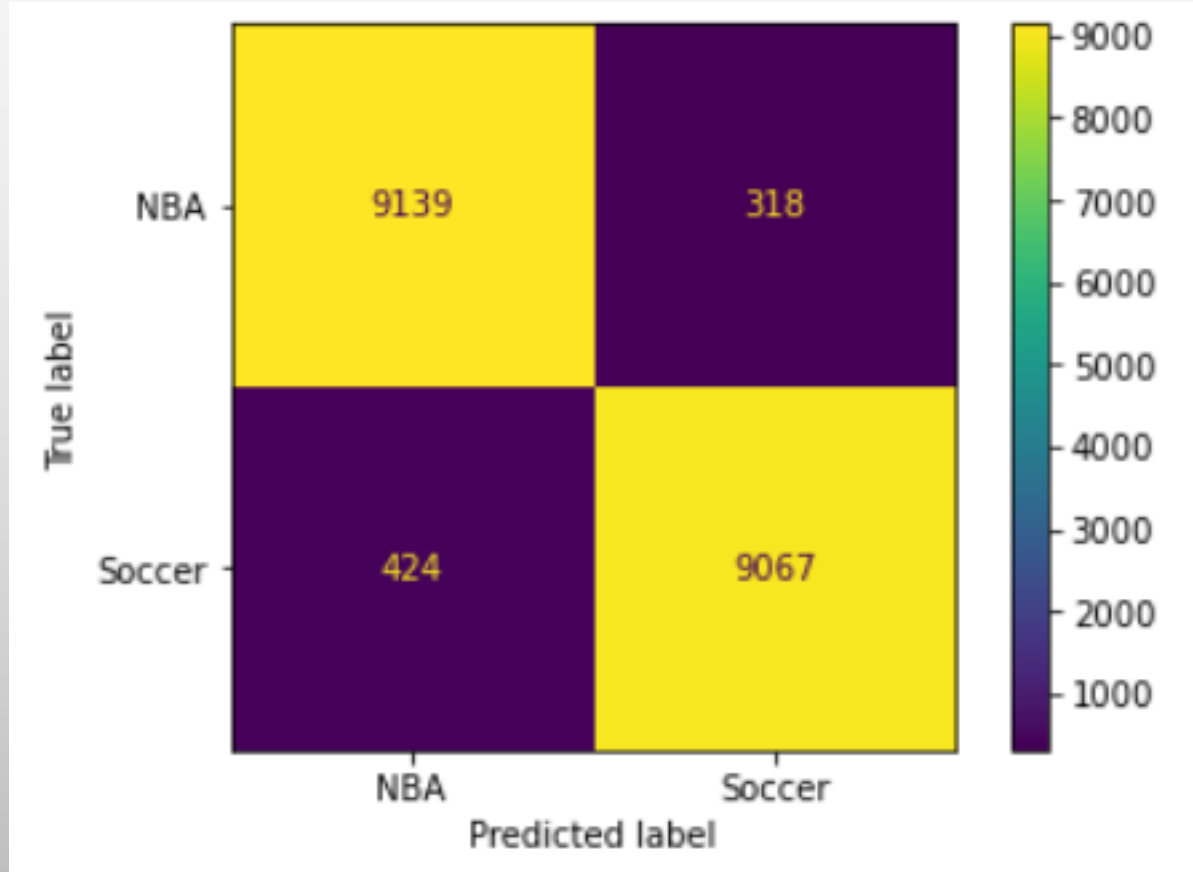
MODELING
RANDOM FOREST

Training score: 0.999
Testing score: 0.960

Precision: 0.966

Accuracy: 0.960

Random Forest



VALIDATION

	Training score	Testing score	Precision	Accuracy
Logistic Regression	0.998	0.968	0.971	0.968
Naïve Bayes	0.989	0.971	0.980	0.971
Random Forest	0.999	0.960	0.966	0.960

CONCLUSIONS & FUTURE WORK

- CountVectorizer with Naïve Bayes performed the best
- Logistic regression is also a good model (needs tuning)
- Future work:
  - Try other classification models (e.g. SVC)
  - More tuning with hyperparameters to improve models
  - Streamlit, to make the model available for use