API & NLP r/CFB & r/CollegeBasketball

Anthony Zurke

August 23, 2021

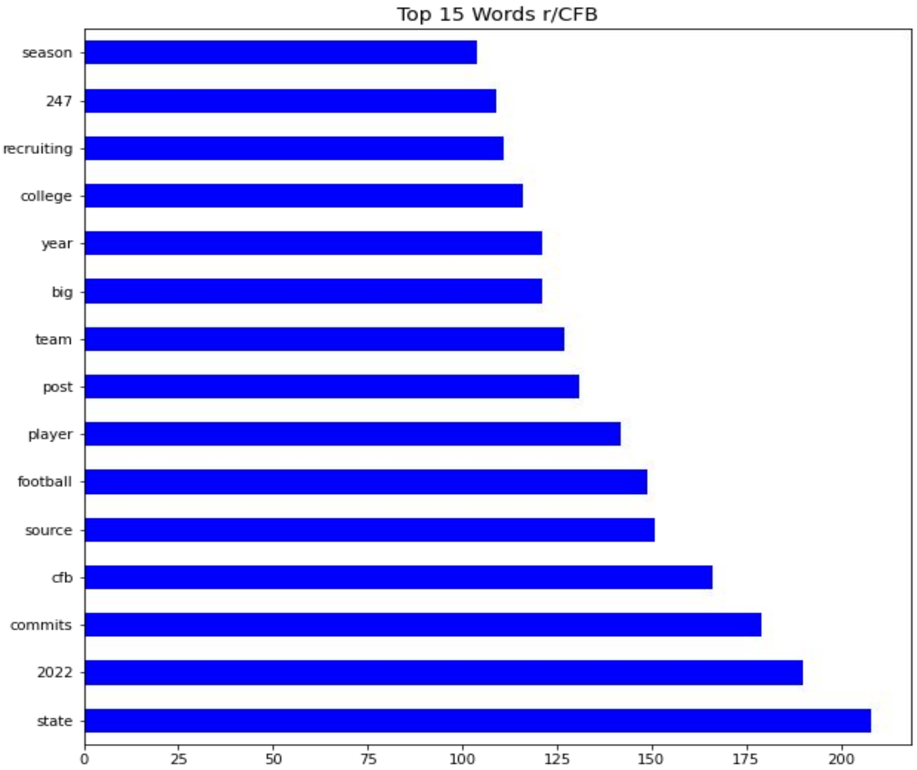
Problem Statement

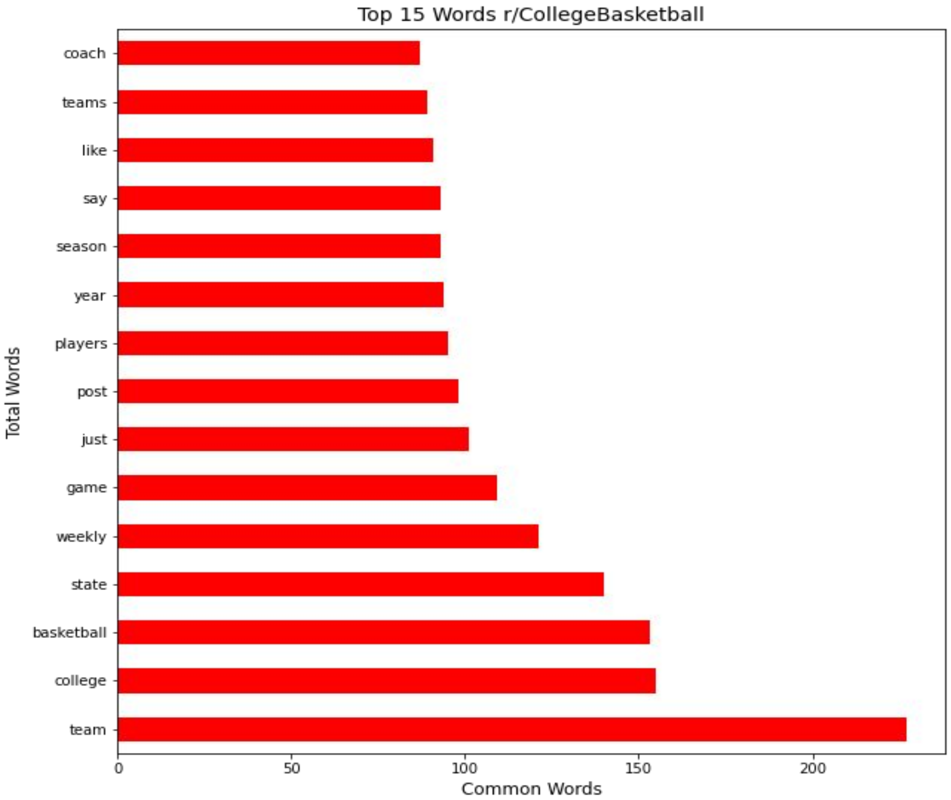
My company has partnered with a sports marketing agency to expand their market into college athletics. We have generated 2 models, a Logistic Regression Model, and Random Forest Classifier, using posts from r/CFB and r/CollegeBasketball on Reddit. We will gather the data from these models ot distinguish most commonly talked about topics and also, differentiating them to provide the most effective marketing strategies.

Data Cleaning and EDA

- 2,500 posts from both subreddits
- Dropped duplicated rows, null values, and outliers
- Remapped subreddits to 1: r/CFB 0: r/CollegeBasketball
- Added column ‘post’ combined ‘selftext’ and ‘title’
- Created two new columns ‘post_length’ and ‘post_word_count’

CountVectorizer Most Common Words





Data Exploration

CFB Polarity: 0.06879804651208767

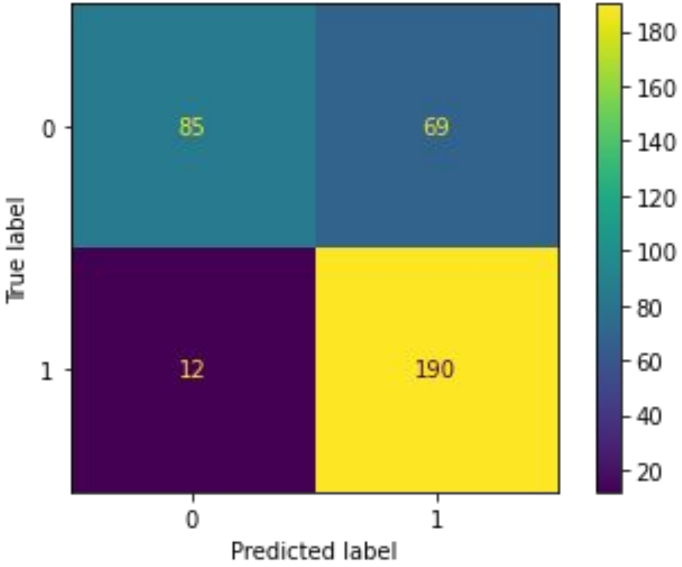
CFB Subjectivity: 0.4241496792781037

CollegeBasketball Polarity: 0.1222440153525197

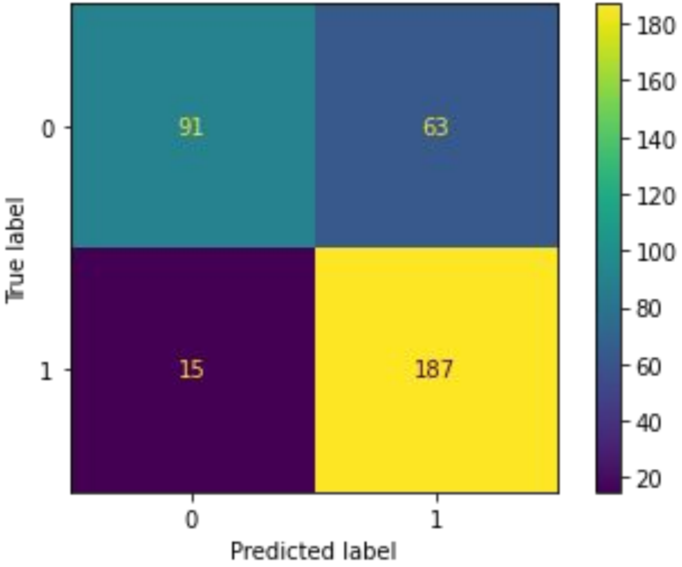
CollegeBasketball Subjectivity: 0.48171778531016124

Model Performance

Logistic Regression
Accuracy: 0.7724719101123596
Sensitivity: 0.9405940594059405
Precision: 0.7335907335907336
F1 Score: 0.824295010845987



Random Forest
Accuracy: 0.7808988764044944
Sensitivity: 0.9257425742574258
Precision: 0.748
F1 Score: 0.827433628318584



Conclusions & Recommendations

Logistic Regression was able to predict posts with 77% accuracy and the Random Forest Classifier had an accuracy score 78%. the Random Forest Classifier also had a higher cross val score of .81 compared to the Logistic Regression cross val score of .79.

I would recommend that sports marketing agency that has partnered with us to use the Random Forest Classifier to properly market towards their target market and to take advantage of the opportunities on recruiting websites.