API & NLP
r/CFB & r/CollegeBasketball

Anthony Zurke
August 23, 2021

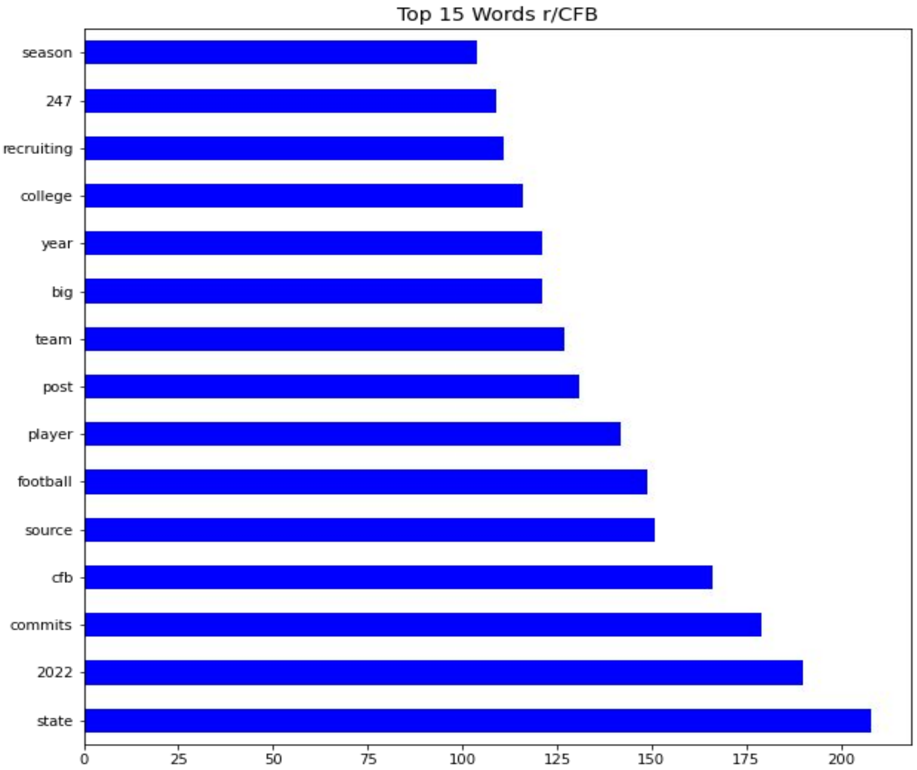
Problem Statement

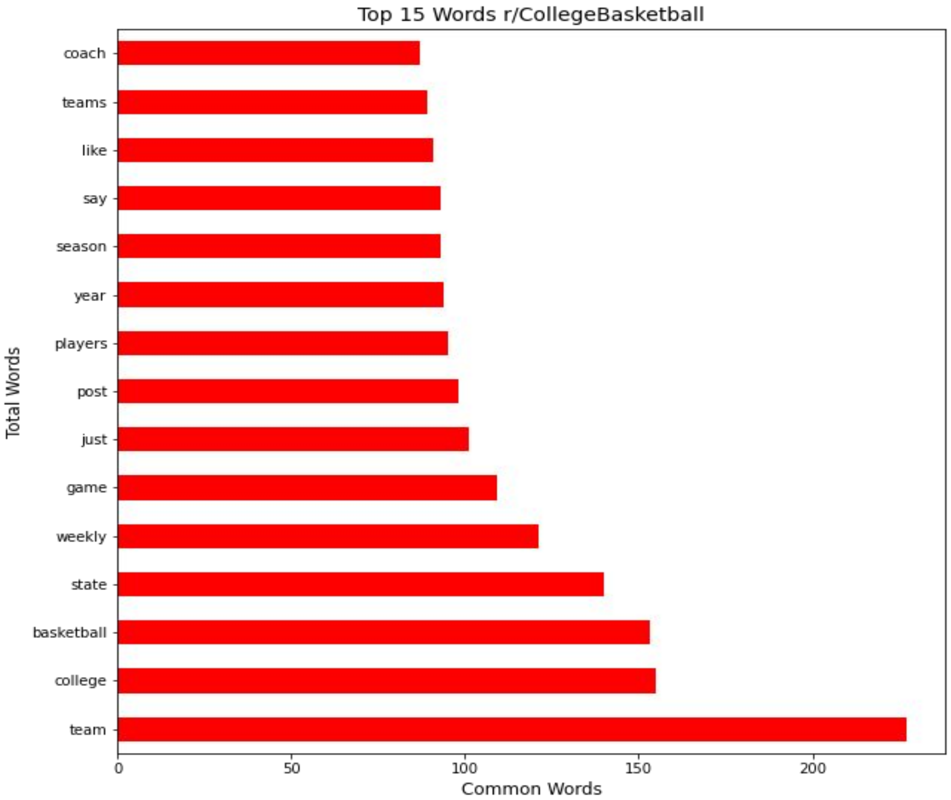
I have partnered with a sports marketing agency to expand their market into college athletics. I have generated 2 models, a Logistic Regression Model, and Random Forest Classifier, using posts from r/CFB and r/CollegeBasketball on Reddit. These will be used to gather data from these models to distinguish 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



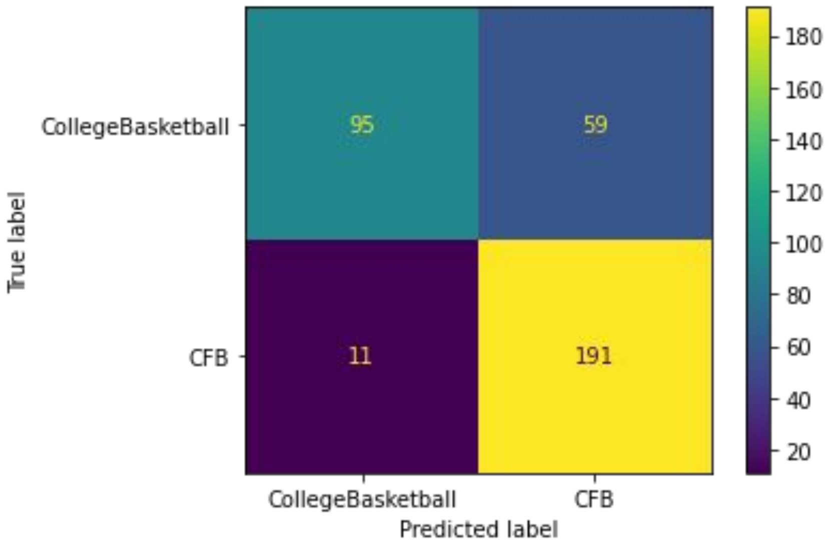


Data Exploration

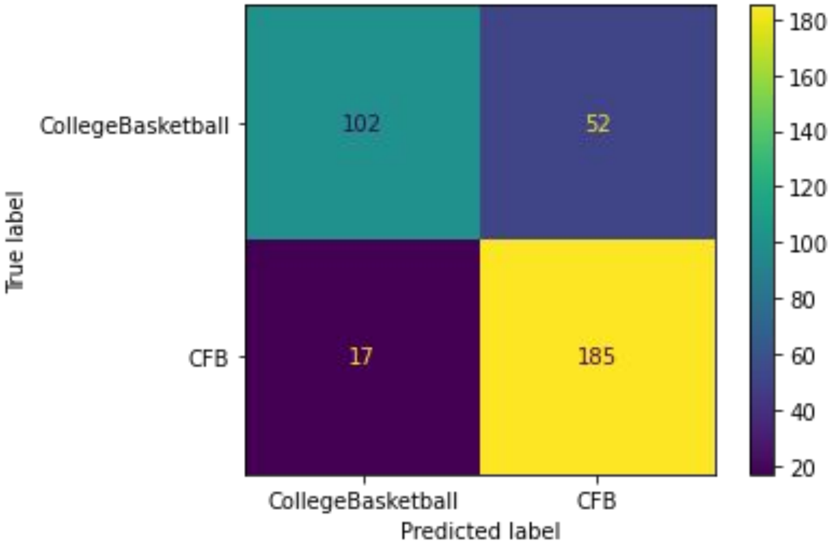
Subreddit	Polarity	Subjectivity
r/CFB	0.0731	0.4245
r/CollegeBasketball	0.1175	0.4825

Model Performance

Logistic Regression
Accuracy: 0.8033707865168539
Sensitivity: 0.9455445544554455
Precision: 0.764
F1: 0.8451327433628318
Cross val score: 0.8225352112676056



Random Forest
Accuracy: 0.8061797752808989
Sensitivity: 0.9158415841584159
Precision: 0.7805907172995781
F1: 0.8428246013667428
Cross val score: 0.8403755868544602



Conclusions & Recommendations

Logistic Regression was able to predict posts with 80.3% accuracy and the Random Forest Classifier had an accuracy score 81%. the Random Forest Classifier also had a slightly higher cross val score of .84 compared to the Logistic Regression cross val score of .82.

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.