Project 3: Predicting the source of NFL and NBA related reddit posts

•••

Luka Kelly

Problem Statement

I work as a data scientist for the Fox Sports 1 (FS1) sports talk show "Undisputed with Skip (Bayless) and Shannon (Sharpe)". My producers have tasked me with using reddit as a source of information to build three models that can take in unseen data from both the r/nfl and r/nba subreddits, and to understand what topics are being discussed the most in the NFL and NBA communities currently. If our models can appropriately predict whether unseen data is from the r/nfl or r/nba subreddits with above 95% accuracy in both the training and testing data, the model will be considered a success. An additional success will be if I can pull at least 5 clear topics from each subreddit dataframe using bi-grams, in order to help understand what topics might be good for upcoming segments on the show.

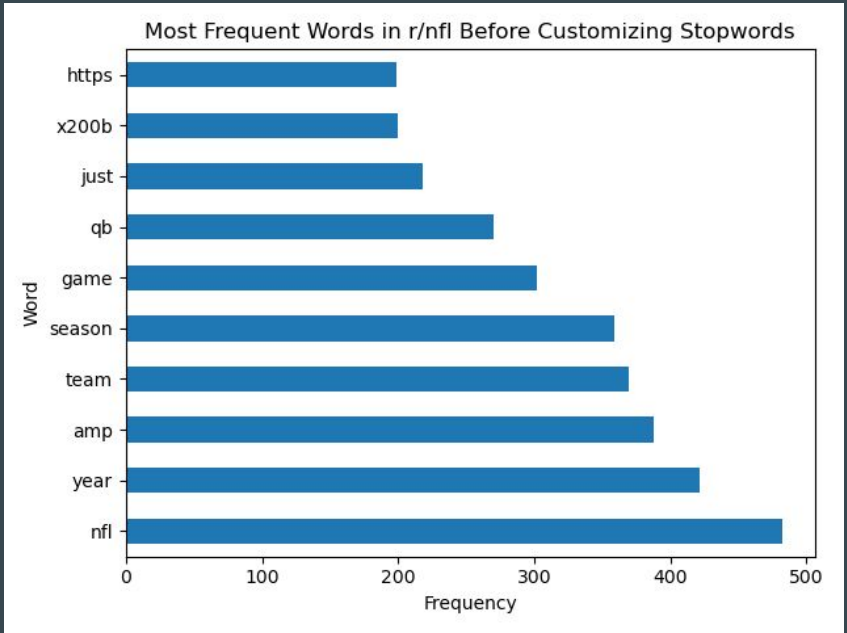
From this model, some tasks that I will be charged with in the future (outside of this project) are examining which league (NFL or NBA) is being discussed more currently by fans.

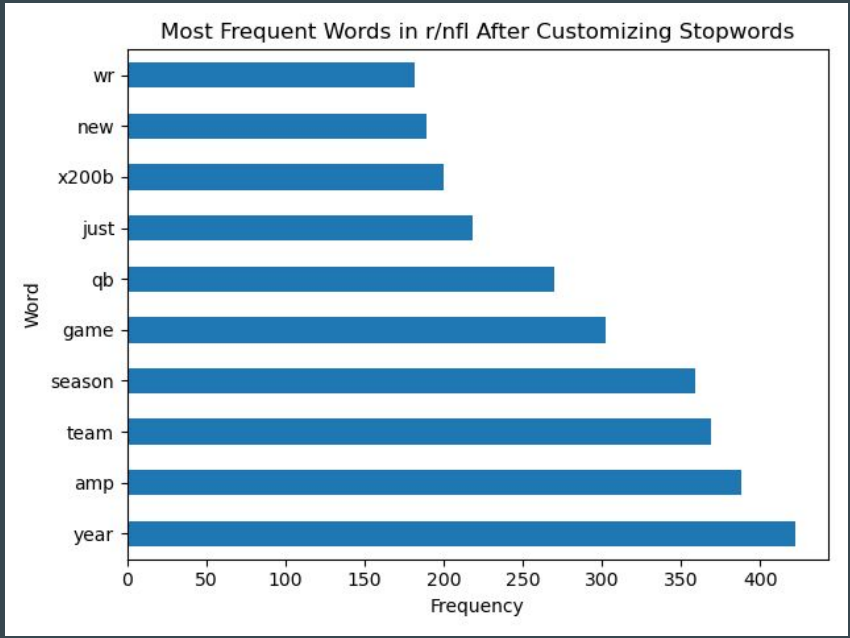
Problem Statement Context

For those who don't know, Shannon Sharpe is a now Hall of Fame inductee who played tight end in the NFL for many years, won three super bowls, and is considered one of the greatest tight ends of all time. Skip Bayless is a former sportswriter who is now one of the biggest names in sports talk shows. They both have big personalities and often get into heated debates regarding whatever topic is at hand, which almost exclusively seems to be related to events happening in the NFL or NBA - which is why I have been tasked with using data from the r/nfl and r/nba subreddits in my model.

EDA and Preprocessing

r/nfl High Frequency Words - Before and After Customizing Stopwords

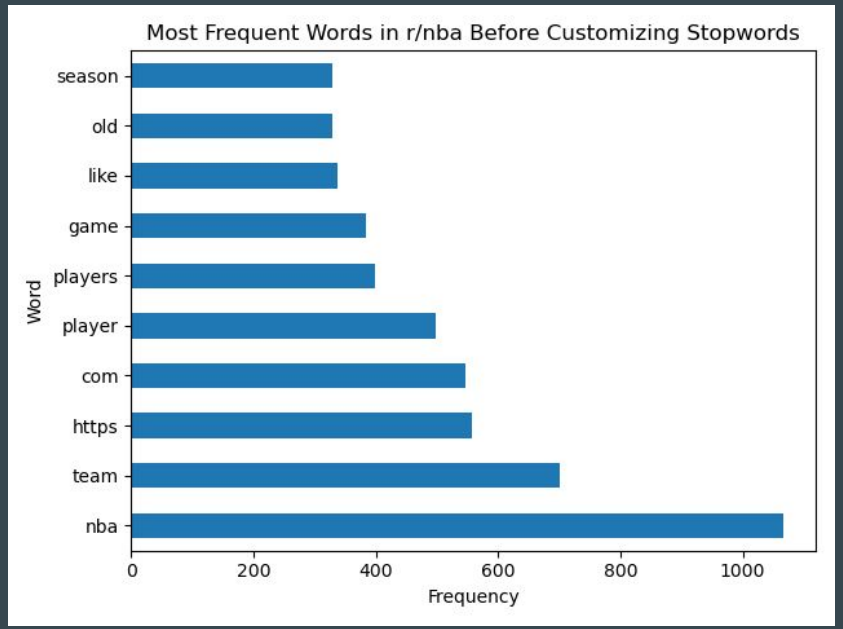


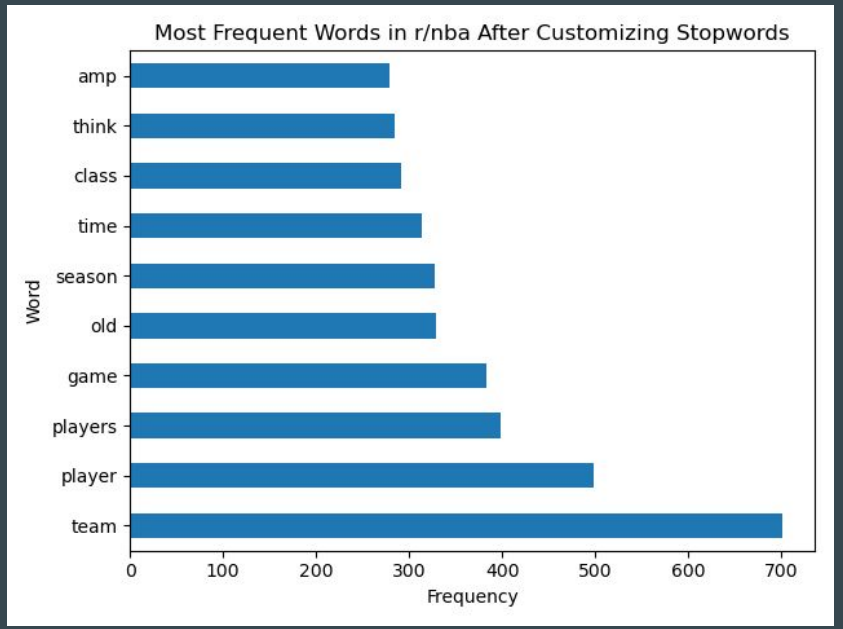


Stopwords: 'english'

Stopwords: 'english', 'like', 'nfl', 'https'

r/nba High Frequency Words - Before and After Customizing Stopwords

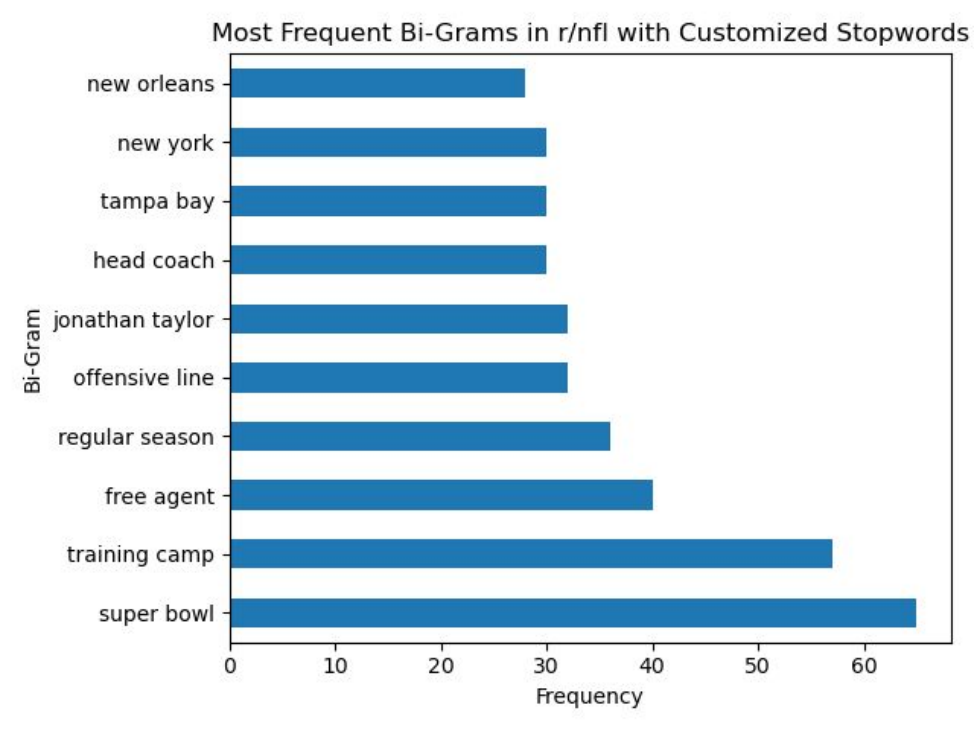




Stopwords: 'english'

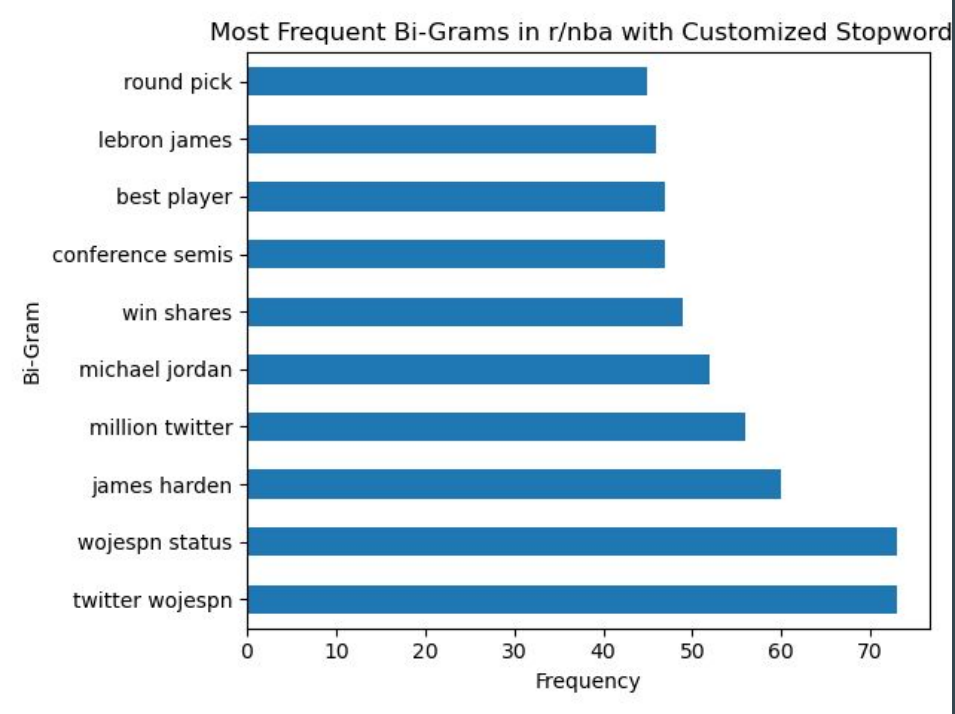
Stopwords: 'english', 'nba', 'com', 'https', 'like'

r/nfl High Frequency Bi-Grams - After Customizing Stopwords



Stopwords: 'english', 'like', 'nfl', 'https', 'amp', 'www', 'reddit', 'com', 'utm_source', 'utm_medium', 'web2x'

r/nba High Frequency Bi-Grams - After Customizing Stopwords



Stopwords: 'english', 'nba', 'com', 'https', 'like', 'reddit', 'amp', 'comments'

Modeling

Modeling Info and Methodology

- Null Baseline Accuracy: 53%
- Stopwords tested in all models:
  - 'english'
  - Custom Stopwords: 'nba', 'com', 'https', 'like', 'reddit', 'amp', 'comments', 'like', 'nfl', 'www', 'utm_source', 'utm_medium', 'web2x'
- Created pipelines for all models
  - ‘TfidfVectorizer’ used in pipelines of all models
- Set parameters to gridsearch over for each model
  - Bernoulli Naive Bayes: Max Features: [2000, 3000, 4000, 5000], Stop Words: (listed above), N-gram range: [(1, 1), (1, 2)]
  - Logistic Regression: Max Features: [2000, 3000, 4000], Stop Words: (listed above), Min Document Frequency: [2, 4], Max Document Frequency: [1.0, 0.8, 0.5]
  - Random Forests: Number of Decision Trees: [100, 150, 200], Stop Words: (listed above), Max Depth: [None, 1, 2, 3, 4, 5]
- Grid-searched to find best parameters for each model

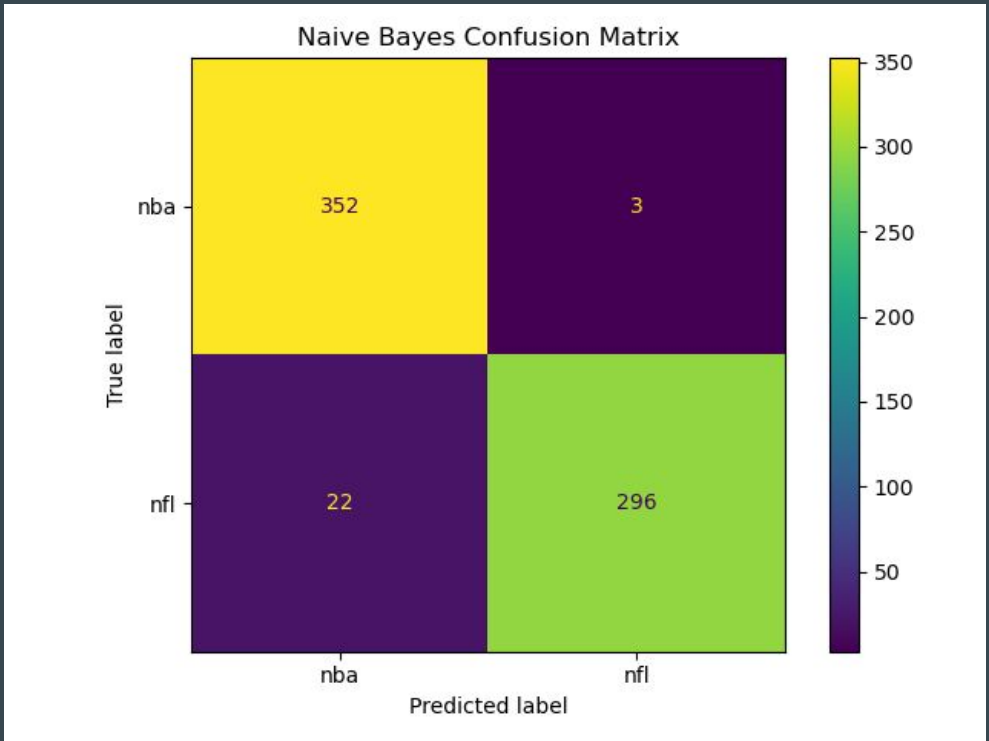
Bernoulli Naive Bayes

Best Parameters:

- Maximum Features: 5000
- N-gram range: (1, 2)
- Stop Words: 'english'

Training Accuracy Score	0.98
Testing Accuracy Score	0.96

Bernoulli Naive Bayes



Accuracy = (True Positives + True Negatives (648)) / total (673)

Accuracy = 96%

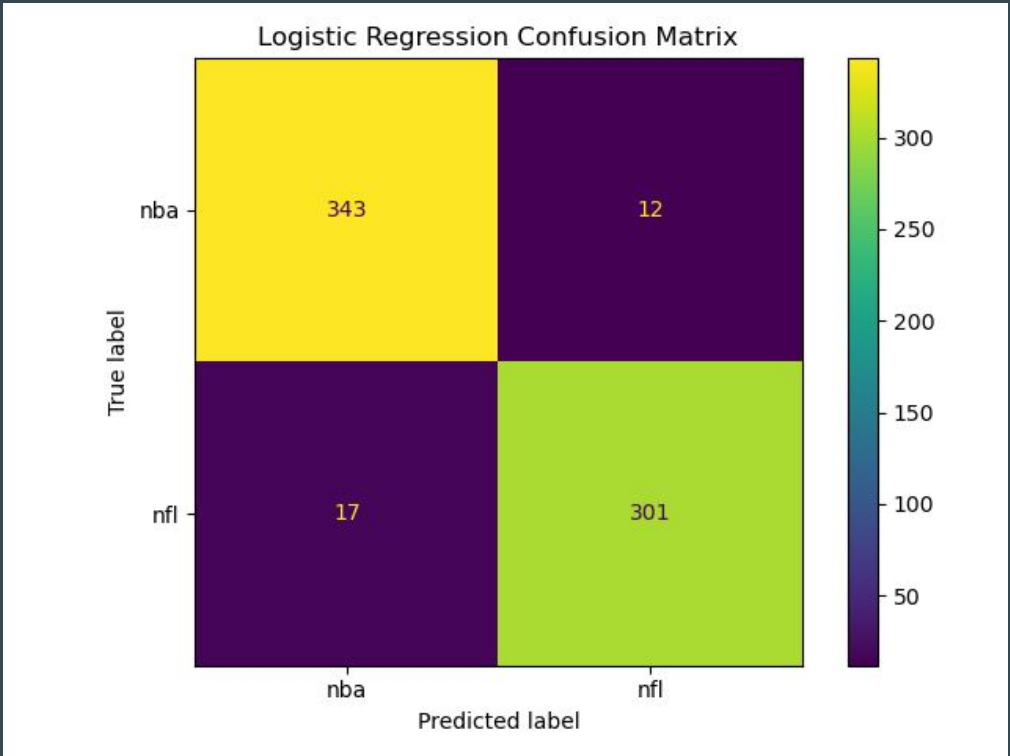
Logistic Regression

Best Parameters:

- Max Features: 4000
- Max Document Frequency: 1.0
- Min Document Frequency: 2
- Stop Words: english

Training Accuracy Score	0.99
Testing Accuracy Score	0.96

Logistic Regression



Accuracy = (True Positives + True Negatives (648)) / total (673)

Accuracy = 96%

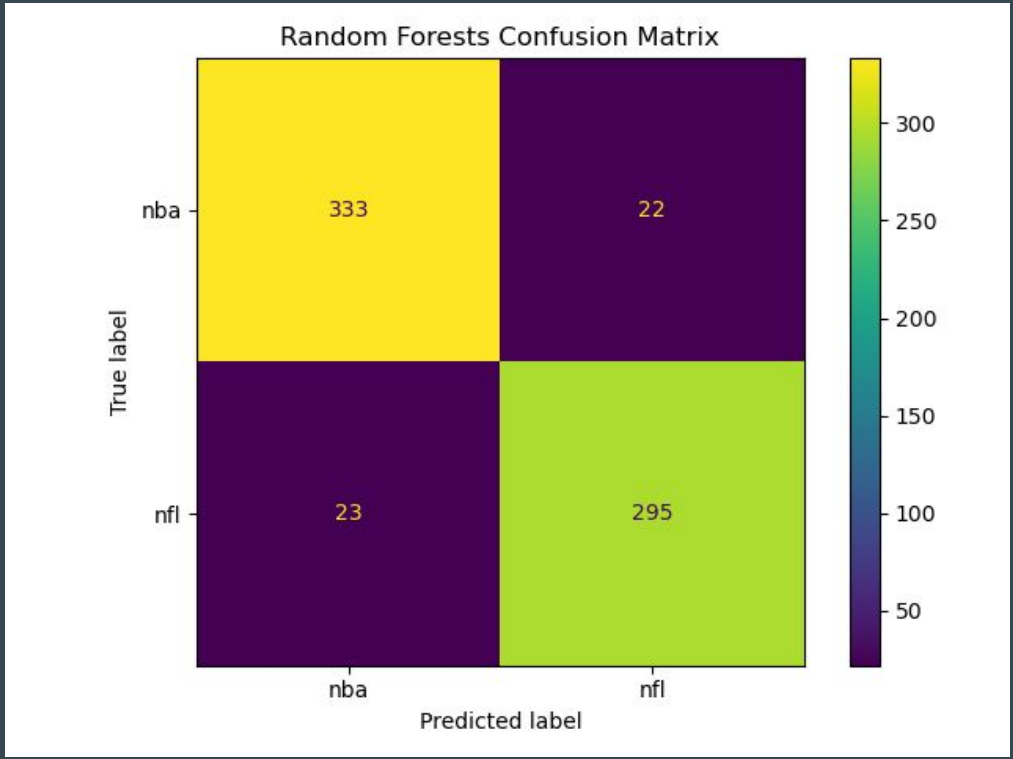
Random Forests

Best Parameters:

- Max Depth: None
- Number of Decision Trees: 200
- Stop Words: english

Training Accuracy Score	1.0
Testing Accuracy Score	0.93

Random Forests



Accuracy = (True Positives + True Negatives (625)) / total (673)

Accuracy = 93%

Evaluation and Recommendations:

In consideration of our baseline of 0.53 (rounded), all three of our models significantly outperformed. All three of our models also scored either extremely high (Naive Bayes: 0.98, Logistic Regression: 0.99) or perfect (Random Forests: 1.0) on our testing data. All of our models were not underfit(high bias) and only Random Forests was overfit(high variance), as displayed in the table below:

Model	Train Accuracy	Test Accuracy
Naive Bayes	0.98	0.96
Logistic Regression	0.99	0.96
Random Forests	1.0	.93

The metric most relevant to our first problem objective is accuracy - we want to be as correct as possible when interpreting if new data falls under the NFL or NBA subreddits. Our first measure of success was if our model(s) would exceed 95% accuracy, which all of them did on both the training and testing score, with exception to Random Forest on our training score, as mentioned above.

Evaluation and Recommendations Continued:

After cleaning the data, conducting EDA, preprocessing, and filtering noise out of the top most-occuring bigrams we were able to meet our second measure of success: finding at least 5 clear topics from each subreddit dataframe using bi-grams, in order to help understand what topics might be favorable for upcoming segments on the show. As a data scientist in the sports media industry, my domain knowledge tells me that the clearest of those topics are as follows, along with a description of why these topics might be trending on reddit currently, and recommendations of how they can be discussed by Skip and Shannon:

- NFL:
  - Super Bowl (maybe the topic is around predicting which teams will make it to the super bowl this year?)
  - Training Camp (report on the events and happenings of training camps currently being held around the NFL?)
  - Free Agent (reporting on big news in regards to free agency recently, which is typically a big topic during this point in the NFL off-season?)
  - Regular Season (predicting how the NFL regular season will pan out for certain teams?)
  - Jonathan Taylor (reporting on any news regarding the Indianapolis Colts running back who recently requested a trade from the team?)
- NBA:
  - Twitter Wojespn (Adrian Wojnarowski is the 'hottest' reporter in the NBA when it comes to top news, especially regarding trades and free agency - is there anything big he's reported on recently?)
  - James Harden (James Harden recently requested a trade from the Philadelphia 76ers - maybe report on any new updates from that situation?)
  - Michael Jordan (Jordan recently finalized a deal to sell the Charlotte Hornets, any updates on that topic?)
  - Lebron James (After losing in the Western Conference Finals this past season, Lebron is reportedly considering retirement - any new information on that?)
  - Round Pick (The NBA draft was held in the past couple months, with one of the most anticipated prospects since Lebron James being drafted first overall - any news on how that prospect (Victor Wembanyama), or any of the other draft picks, are performing leading up to the start of the NBA season?)

Conclusion:

From our analysis and modeling, we can conclude that using either Naive Bayes or Logistic Regression as our models will give us high accuracy when it comes to predicting if new data is from the r/nfl or r/nba subreddit. We can expect 96% accuracy from both models when taking in new data.

We can also conclude that the bi-grams above are the top, most-occuring, clear topics being mentioned in the r/nfl and r/nba subreddits currently.

Future Steps:

One future step that can be taken with our data is to continue our search for n-grams within our data, expanding to exploring tri-grams and potentially even 4-grams. If we are able to pull legible tri- and 4-grams, the extra text may give us further insight to each topic. For example, Lebron James creates a lot of media around him - if his name is being mentioned on the r/nba subreddit, it could be for a number of reasons. Finding tri-grams and 4-grams could potentially help us better understand the context of his name being mentioned.

Thank You!