Baseball or Basketball?

By Brandon Hoskins

Problem Statement

Gathering text data from Reddit posts from the baseball and basketball subreddits using Pushshift's API, we are comparing the accuracy scores of logistic regression and random forest models.

What is Reddit and What are Subreddits?

Reddit is a social news and discussion website where users can upload content such as links, text posts, and images. These posts are voted up or down by other users.

These posts are organized by subject into user created boards called “subreddits”, which covers a variety of topics.

Data Collection

Using Pushshift's API, I was able to collect 1000 posts from the baseball and basketball subreddits.

The posts from the baseball subreddit were mostly from the last week, as there were many posts during the MLB playoffs that were being played last week.

The posts from the basketball subreddit were from the past month, which included the NBA playoffs.

Data Cleaning Process

After going through the data, I decided to only use the title of the posts in my models for prediction. I chose not to use the descriptions under the posts because there were many posts that did not have a description.

Pre-Processing & EDA

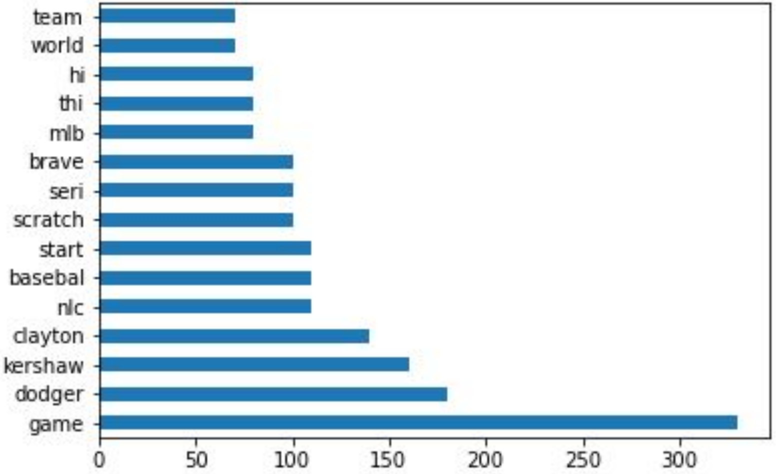
When dealing with text data such as reddit posts, we have to do pre-processing steps to make the text data easier for the computer to work with. With both the baseball and basketball subreddit posts, I had to complete the following steps:

- Removing special characters and punctuation marks
- Stemming
- Tokenizing
- Removing stop words

The Top 15 Most Common Words in the Baseball Subreddit

The most common word in the baseball subreddit is game.

Most of these posts were from last week during the NLCS playoff series between the Los Angeles Dodgers and Atlanta Braves. Which would explain why the words Dodgers and Braves were so common. Along with Dodgers' ace Clayton Kershaw.

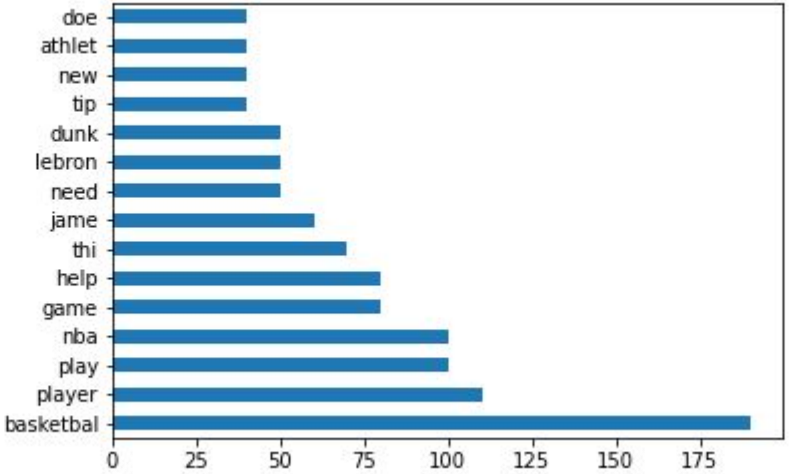


Top 15 Most Common Words in the Basketball Subreddit

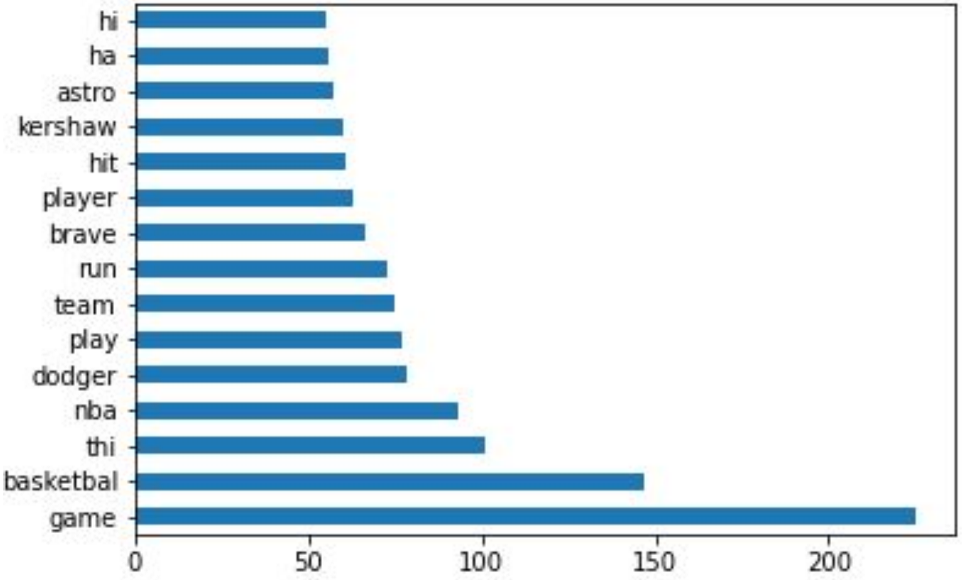
The most common word in the basketball subreddit is...basketball.

LeBron James' name is also among the common words in this subreddit after he just won his 4th NBA Championship with the Lakers during the time of these posts.

I also found interesting that words like tip, need, and help were among the most common words.



Top 15 Most Common Words in the Combined Dataset Used in Our Model



Baseline Accuracy Score for Our Models

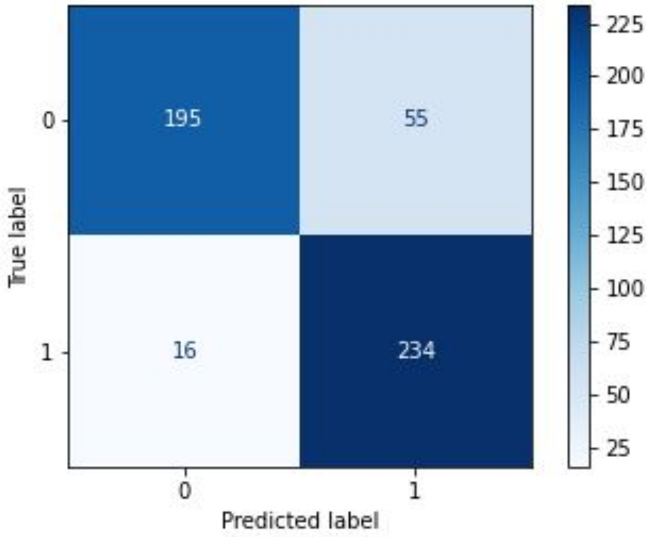
The baseline accuracy score for both our training and test data is .50 or 50%.

Our data has exactly 1000 posts from both the baseball and basketball subreddits.

Logistic Regression Model Evaluation

- Accuracy Score on Training Data: 0.906
- Accuracy Score on Testing Data: 0.858

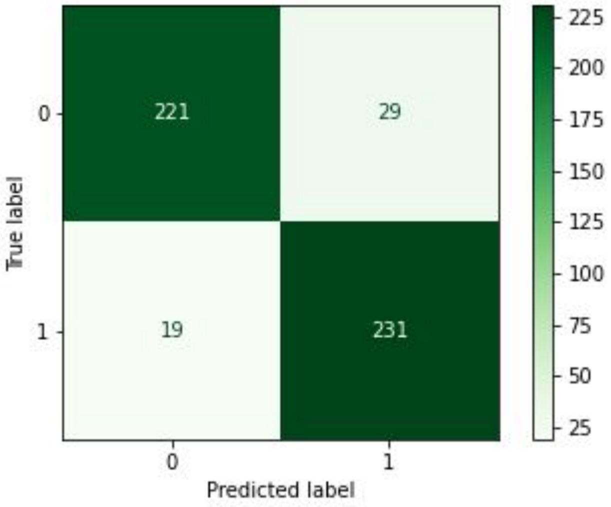
- True Positives = 234
- False Positives = 55
- True Negatives = 195
- False Negatives = 16
- Specificity = 0.78
- Sensitivity = 0.936



Random Forest Model Evaluation

- Accuracy Score on Training Data: 0.998
- Accuracy Score on Testing Data: 0.904

- True Positives = 231
- False Positives = 29
- True Negatives = 221
- False Negatives = 19
- Specificity = 0.884
- Sensitivity = 0.924



Conclusion

When comparing both of these models, it is clear that the random forest model performed significantly better than the logistic regression model in terms of accuracy.

In the future, I would like to use more posts from the past year to possibly better our models. I would also like to try other models and compare them to random forest.