Memphis Grizzlies and New York Knicks 2023-2024 Fan Engagement NLP Data Analysis

For this project I chose to focus on the Memphis Grizzlies and New York Knicks, specifically scraping data from their two respective subreddits: r/memphisgrizzlies and r/NYKnicks. I originally chose these two teams because of their differing market values, as I had intended to draw conclusions related to this market difference. However, I ended up taking a different approach to this project.

Data Collection and Cleaning

The first step was data collection and cleaning. The first dataset I downloaded was the 2023-2024 NBA schedule. The full schedule included games from every team in the league, so I had to filter the schedule to ensure I only had game data regarding the Grizzlies and Knicks. From there I created a new 'game_date' column in my dataframe which was type datetime. For this project it was necessary to have clean and consistent date data. Next I initialized my access to the Reddit API and began to scrape data from the two subreddits, collecting as many post URLs as possible. These URLs were all originally in a list, but I added them to a dataframe with a new datetime column for when the URL's corresponding post was made. Next I filtered the URLs into four groups. Group 1: Memphis Grizzlies Game Day Posts. Group 2: Memphis Grizzlies Not Gameday Posts. Group 3: New York Knicks Game Day Posts. Group 4: New York Knicks Not Gameday Posts. Again, these groups were all still URLs, so the next step entailed getting the desired text from these URLs. Each link was converted to a JSON file, and the post's title, body text, and comments were saved into a new column in the dataframe. The four post groups were then consolidated into the final two dataframes that I used throughout further processing. These two corresponded to the teams, and a new column was added that specified if that respective post was made on the day of a game or not.

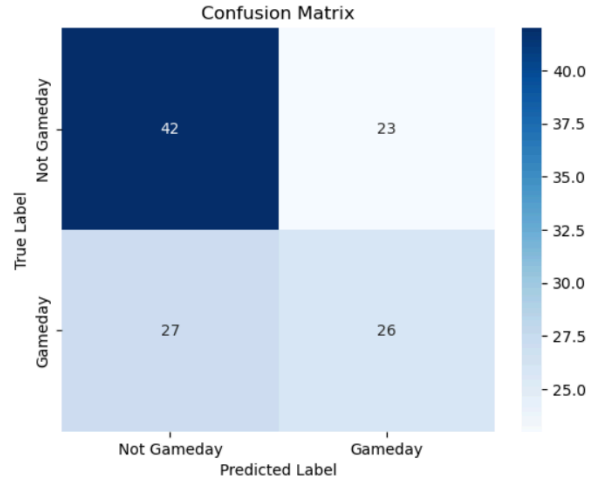
Data Modeling and Exploration

Now with the two processed data frames I could begin to mess around with the data and choose a topic to explore for the project. I chose to explore the correlation between the words used and the date of the post(do the words used indicate if the post was made on the day of a game or not) and the correlation between the words and the post's sentiment, as well as implementing a machine learning model to try and classify posts based on the two topics. In doing this I first had to tokenize the text, as well as create a Bag of Words related to said tokenized text. Now with both data frames further filtered I could begin to explore my topics. I split the topics up between the two datasets, with the Grizzlies for the first topic and the Knicks for the second. Because the second topic related to sentiment I had a little more data processing to do. Using two lists of positive and negative words, I ran through each text block and assigned a sentiment based on which type of word there was more of. From there, after some brief ML

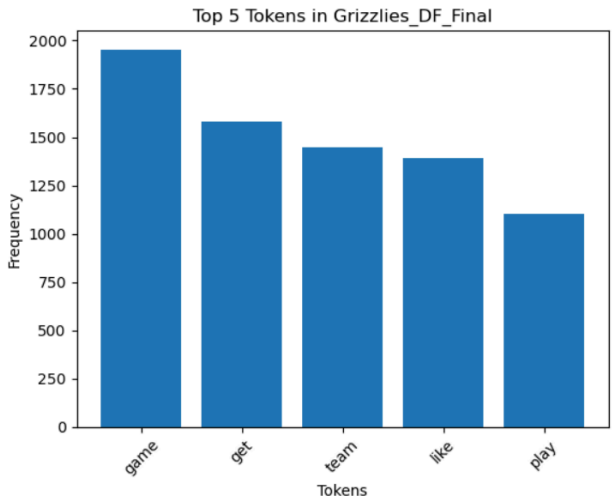
prep to the dataframes, I was ready to implement the Naive Bayes model and draw some conclusions.

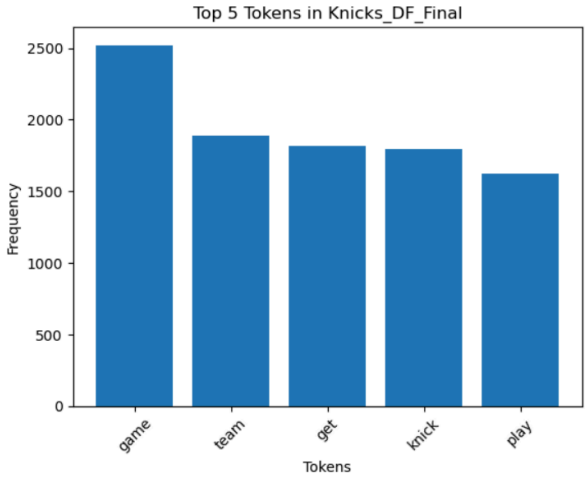
Findings

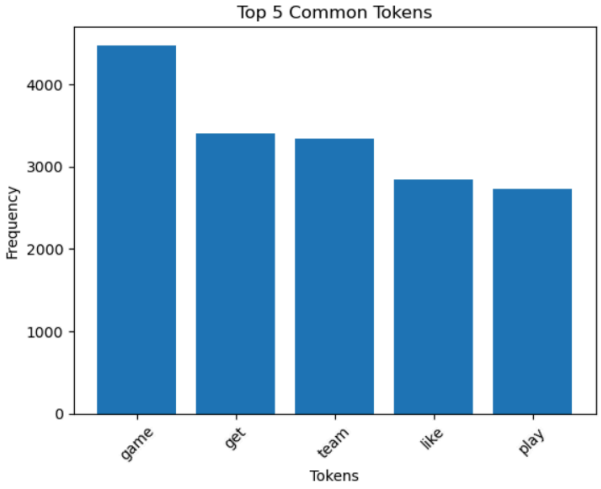
Again my first topic was attempting to classify posts into posts made on the day of a game and posts not made on the day of a game. In exploring this topic I trained two different Naive Bayes models, the difference in the two being the way my training features were extracted. After training and testing I looked into the accuracy of both models. What I noticed about both is that when changing the split of the training and test data the accuracies of both models varied wildly, sometimes by even 10 or 15 percent. Below is a Confusion Matrix for one of those splits:



At first I was confused and worried at these results, but looking closer at the data this made sense. One issue about scraping text from reddit posts is that there is going to be a lot of noise and disruptive data points. The thing about humans is that we like to type in ways that are not always conducive to the English Lexicon. Weird misspellings and slang caused a lot of noise in my data making it hard for the model to produce consistent results. The same could be said for my analysis of sentiment in the Knicks Dataframe. Oftentimes posts that were labeled as positive felt somewhat negative in reality and vice versa. Again noisy data and weird comments made it hard to accurately classify and train the ML model for classifying sentiment. Below are some more visuals that give more insight into the data. With these cleaned graphs that highlight the top five most frequent words it becomes clear that there is some consistency throughout the data. The issue, however, is that there seems to be a lot more inconsistency stemming from human inconsistency on a platform like Reddit.


Top 5 Tokens in Grizzlies_DF_Final
Frequency
2000
1750
1500
1250
1000
750
500
250
0
game
get
team
like
play
Tokens


Top 5 Tokens in Knicks_DF_Final
Frequency
2500
2000
1500
1000
500
0
game
team
get
knick
play
Tokens


Top 5 Common Tokens
Frequency
4000
3000
2000
1000
0
game
get
team
like
play
Tokens