**Analyzing NCAA CFB Playoff Contenders and Oklahoma: A Reddit Sentiment Analysis**

Group 5

Braum Russell, Jacob Conger, Noah Hunter, Kayla Nguyen, Nathaniel Pumphrey,

Lucas West, Sam Harkins

Dr. Heshan Sun

MIT 5702.100 Analytics with Python

October 9th, 2024

**Problem Statement**

This project's objective is to enhance our understanding of how to use Python and relevant packages to scrape and analyze Reddit data, specifically focusing on the current landscape of college football. Ultimately, we aim to explore the following questions:

1. Identify which teams are most mentioned in Reddit discussions about attending the playoffs
   1. More specifically, are the Oklahoma Sooners favored to be in the playoffs?
2. What do Reddit users think about the Oklahoma Sooners?
   1. Is there a correlation between average compound score and in-game performance?

**Description of Data**

We sourced our data from Reddit by using PRAW to make read calls to Reddit’s API. For the first part of our analysis, we extracted comments from the r/CFB thread called “What is your 12-Team Playoff Prediction?” and created a Pandas data frame containing the number of mentions each team had.

For our second analysis, we used PRAW to pull comments from specific post-game threads in r/Sooners. These comments included fans’ post-game analysis and their opinions of how Oklahoma played. We analyzed the sentiment scores for each comment and created another Pandas data frame holding score values. We used this process for game threads against Temple, Houston, Tulane, Auburn, and Tennessee.

**Python Package Used and Code**

This project utilized several python packages. The packages that we imported are as follows:

|  |  |
| --- | --- |
| PRAW | This package was used to access Reddit’s API. |
| NLTK | This package was used to perform our sentiment analysis. |
| Re | This package was used for searching and manipulating strings based on specific patterns. |
| Pandas | This package was used for data frames. |
| Matplotlib | This package was used to visualize the data. |

To obtain each thread of comments for both question 1 and 2, we used PRAW to authenticate to Reddit’s API and collect data from our chosen subreddits. Once imported and connected successfully, we then pasted a URL to a specific subreddit thread we wanted to scrape data from. Since Reddit uses hierarchical comments, we used the “replace\_more” function to ensure we load every nested comment. The general process/code is shown in Figure 1 below:

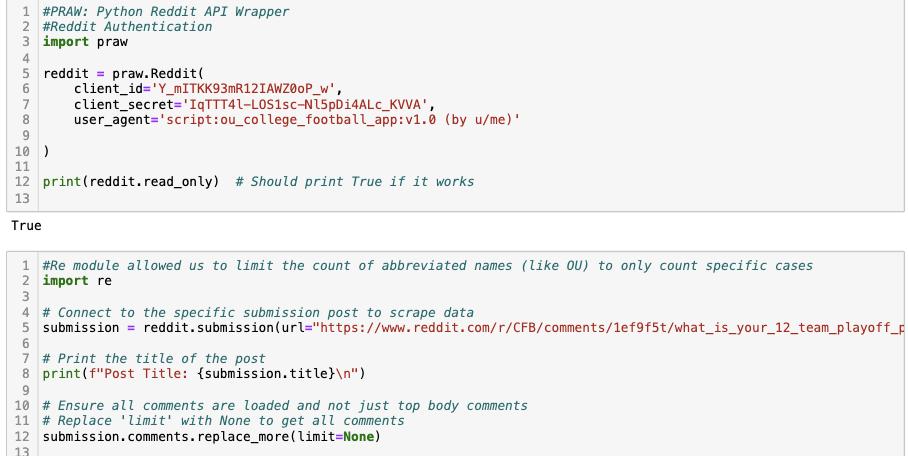


Fig 1. General PRAW Installation & Data Collection

*Question I Code*

To determine what the top twelve teams mentioned in the comments were, we first imported the Re module, which allowed us to limit the count of abbreviated names (like OU) to only count specific cases using boundaries (‘\b’). In the body of the code, we created two dictionaries — one that stores keys of team names and values of a list of all team aliases, and one that stores keys of team names and values of the count of their mentions. Using a loop, we went through each comment and counted how many times a team was mentioned, using the re package to ensure specific matches. Figure 2 below outlines this process:

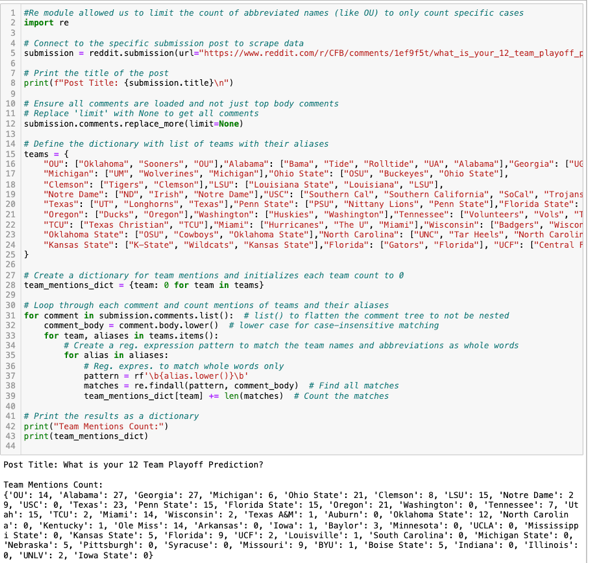


Fig 2. Sample Code from Top Teams Analysis

To organize our data, we imported the Pandas package so we could create data frames that display our results into a two-dimensional table (Fig. 3). We then exported this table into a CSV file so that we could visualize the results in Tableau. Prior to creating our final visualization, we imported MatPlotLib to create a bar graph and see the preliminary results of our analysis (Fig. 4).

A screenshot of a computer

Description automatically generated

Fig 3. Top Teams Data Frames (Pandas)

A screenshot of a computer screen

Description automatically generated

Fig 4. Top Teams Preliminary Visual (Matplotlib)

*Question II Code*

After the authentication to Reddit’s API listed earlier, we created a list including every comment from the game thread URL we selected by creating a loop to append every ‘comment.body’ to the empty ‘comments\_list’ list. See Figure 5 below.



Fig 5. Code Creating List of Game Thread Comments

Using our comment data for each OU game, we performed a sentiment analysis using the NLTK package. We used a loop to calculate sentiment scores (positive, negative, neutral, and compound) for each comment in the list and stored them in a dictionary. This process is captured below in Figure 6.

A screenshot of a computer

Description automatically generated

Fig 6. Code to Perform Sentiment Analysis

After the sentiment analysis, we followed a similar procedure as the previous problem where we imported the Pandas package to create data frames from the “team\_mentions\_dict” dictionary to organize the comment number and their respective sentiment scores into a table. This data frame was also saved as a CSV file. Finally, we computed the mean compound scores for each game to see what the overall sentiment was.

A screenshot of a computer code

Description automatically generated

Fig 7. Average Compound Score Code

**Major Results & Findings**

*Question I Analysis*

The following team are what users believed would make the twelve-team playoff according to /CFB:

A graph of different colored bars

Description automatically generated

Fig 8. Top 12 Reddit Mentions for Playoff Consideration

As shown above, Notre Dame led the way with 29 mentions by Reddit users. Just behind, with 27 mentions each, were Georgia and Alabama, respectively. By far, the most interesting team mentioned by Reddit users was Florida State, who is currently 1-3 on the season. Additionally, Notre Dame, who was mentioned the most in the comments, has already lost one game on the season thus far. The Oklahoma Sooners were mentioned 14 times, putting them at a tie with Ole Miss and Miami in the bottom of the rankings.

*Question II Analysis*

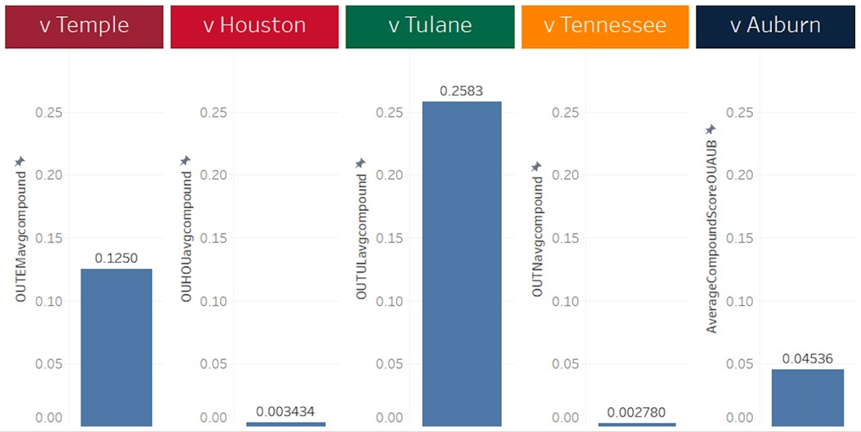


Fig 9. OU Game Outcomes & Compound Scores

Our second question looked to determine the sentiment scores for each game, determining if the comments on the game thread were positive, neutral, or negative. Our analysis found that every Oklahoma Football game this year resulted in sentiment that was slightly above neutral overall. Even in games where the OU offense and defense managed to perform well, i.e., Temple, the overall sentiment was still barely above neutral– sitting at 0.1250. This year, the highest sentiment seen in one of the football games was versus Tulane at home, where our program showcased a sentiment of 0.2583. As for the lowest, both our Houston and Tennesse matchups delivered more negative discussions amongst Reddit users. These figures give basic insight into how fans reacted when looking at the game as a whole. While there might be more positive comments during one portion of the game when OU scores or more negative comments when our QB throws an interception, this program takes all comments into account and provides one sentiment number to determine the tone of the entire thread.

**Conclusion**

In conclusion, our analysis suggests that Reddit users believe Notre Dame, Alabama, and Georgia are the leading contenders for the playoffs. The Oklahoma Sooners, while ranked among the top twelve teams in this subreddit (tied with two others), were positioned toward the lower end. Sentiment regarding the Sooners on the r/sooners subreddit has been neutral overall, with discussions surrounding games against Temple, Tulane, and Auburn, generating slightly more positive sentiment compared to the Houston and Tennessee games. This sentiment appears to align with the outcomes of those games, with less favorable reactions potentially reflecting disappointment in the Sooners’ performance during certain matchups.

Overall, we believe the model we developed offers valuable insights and can be adapted to analyze sentiment across sports beyond college football. The code can be modified for various sports, such as softball, enabling users to gauge fan reactions to wins and losses for any team. Looking ahead, we are interested in the sentiment trends for the remainder of the season.

References

Discussion thread - Houston Cougars (0-1) at (15) Oklahoma Sooners (1-0). (2024a, September). https://www.reddit.com/r/sooners/comments/1fbjax8/discussion\_thread\_houston\_cougars\_01\_at\_15/

Game discussion - Temple vs (16) Oklahoma. Friday 09/30. (2024b, August). https://www.reddit.com/r/sooners/comments/1f597c1/game\_discussion\_temple\_vs\_16\_oklahoma\_friday\_0930/

Game thread: (21) Oklahoma vs Auburn. (2024d, September). https://www.reddit.com/r/sooners/comments/1frmkt3/game\_thread\_21\_oklahoma\_vs\_auburn/

Game Thread: (6) Tennessee Volunteers - (15) Oklahoma Sooners. 09/21/2024. (2024c, September). https://www.reddit.com/r/sooners/comments/1fmgmn2/game\_thread\_6\_tennessee\_volunteers\_15\_oklahoma/

Pregame Discussion - Tulane Green Wave (1-1) @ (15) Oklahoma Sooners (2-0). (2024e, September). https://www.reddit.com/r/sooners/comments/1fgnj97/pregame\_discussion\_tulane\_green\_wave\_11\_15/

What is your 12 Team Playoff Prediction? (2024h, August). https://www.reddit.com/r/CFB/comments/1ef9f5t/what\_is\_your\_12\_team\_playoff\_prediction/?sort=new"