Natural Language Processing

Language analysis between users on the r/NFL & r/fantasyfootball subreddits

Problem Statement

- I will use data collected from posts on the NFL and Fantasy football subreddits on the popular website reddit.com to build a logistic regression, classification model to predict which subreddit a specific post belongs to.
- By analyzing the language used in both subreddits, I seek to understand how to improve fan engagement for the NFL.

About the NFL:

- The NFL is the professional league that hosts the most popular sport in the United States, American Football.
- It is a multi-Billion Dollar Industry that is currently in its 100th year of operation.
- It hosts one of the most viewed shows annually, its Championship game, "The Super Bowl"

Problem:

- Despite these very impressive facts about the NFL, their engagement (Game attendance, viewership, etc.) have been declining in recent years... WHY??
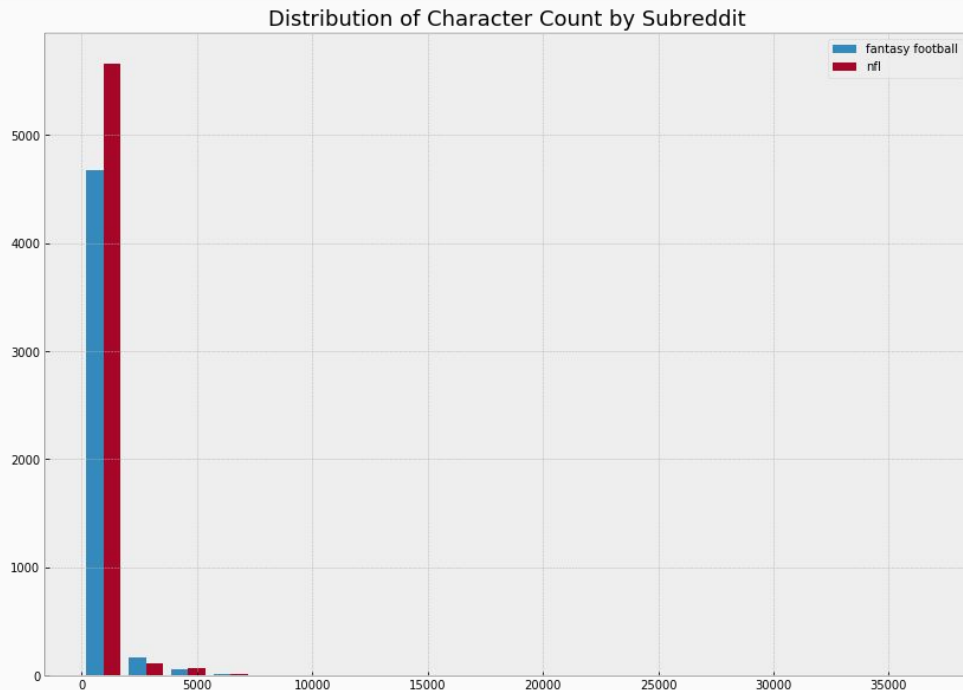
About Fantasy Football:

- Fantasy football is a virtual game where managers select their favorite players from any team in the NFL and points are collected based on each players performance in the real game.
- This is a game that have been played for many decades, but have been increasing in popularity in recent years with the help of the internet and social media.

Question:

- How does this game's popularity correlate with the real game of football and how its content is absorbed?
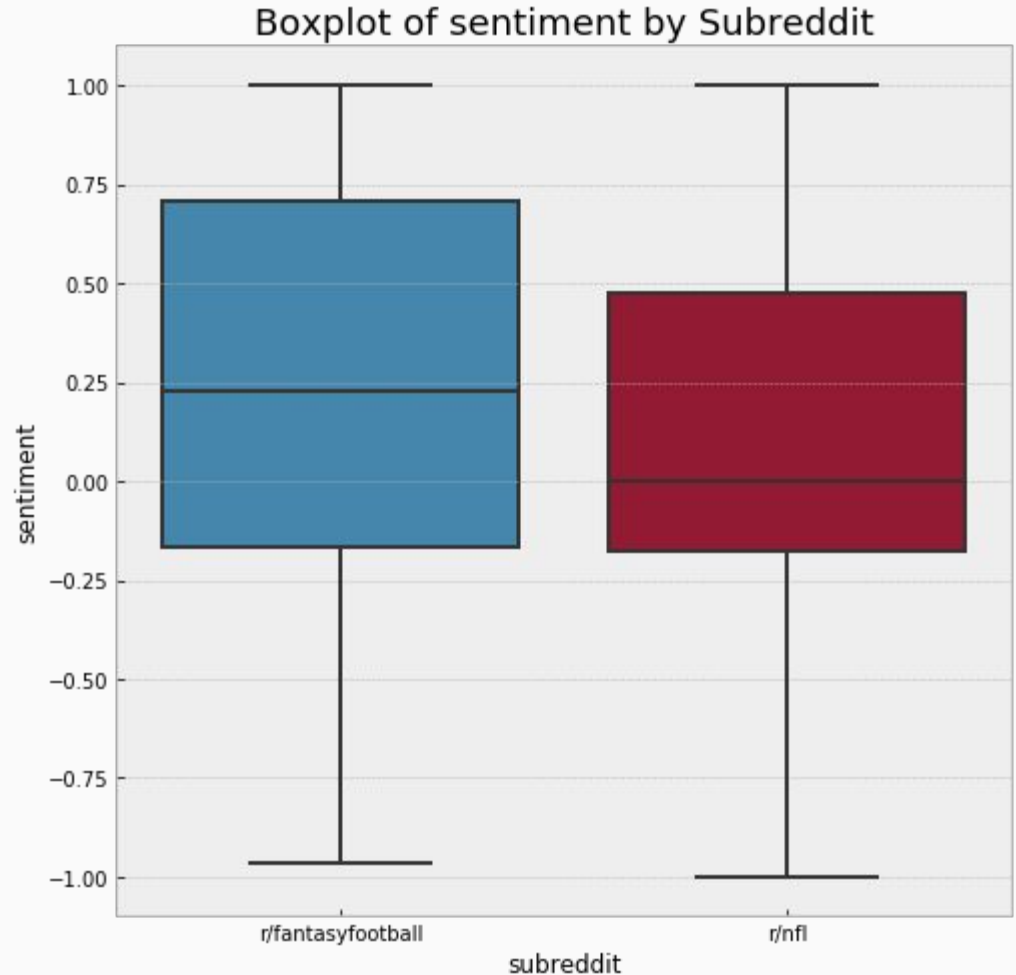
Information about Data Collected

- There are 10,819 rows of data collected, or scraped from Reddit.com
  - Of those rows 5,867 (54%) are from the NFL subreddit & 4,952 (46%) are from the fantasy football subreddit.
- There are 12 scraped or engineered predictor columns.

Distribution of Character Count by Subreddit

Sentiment Analysis
How do users typically feel?

- The text in each post was analyzed and rated, on a scale from -1 to 1, on the emotion of their language.
- -1: very negative, 0: Neutral, 1: very positive.
- Fantasy's posts are on average slightly more positive than NFL's.
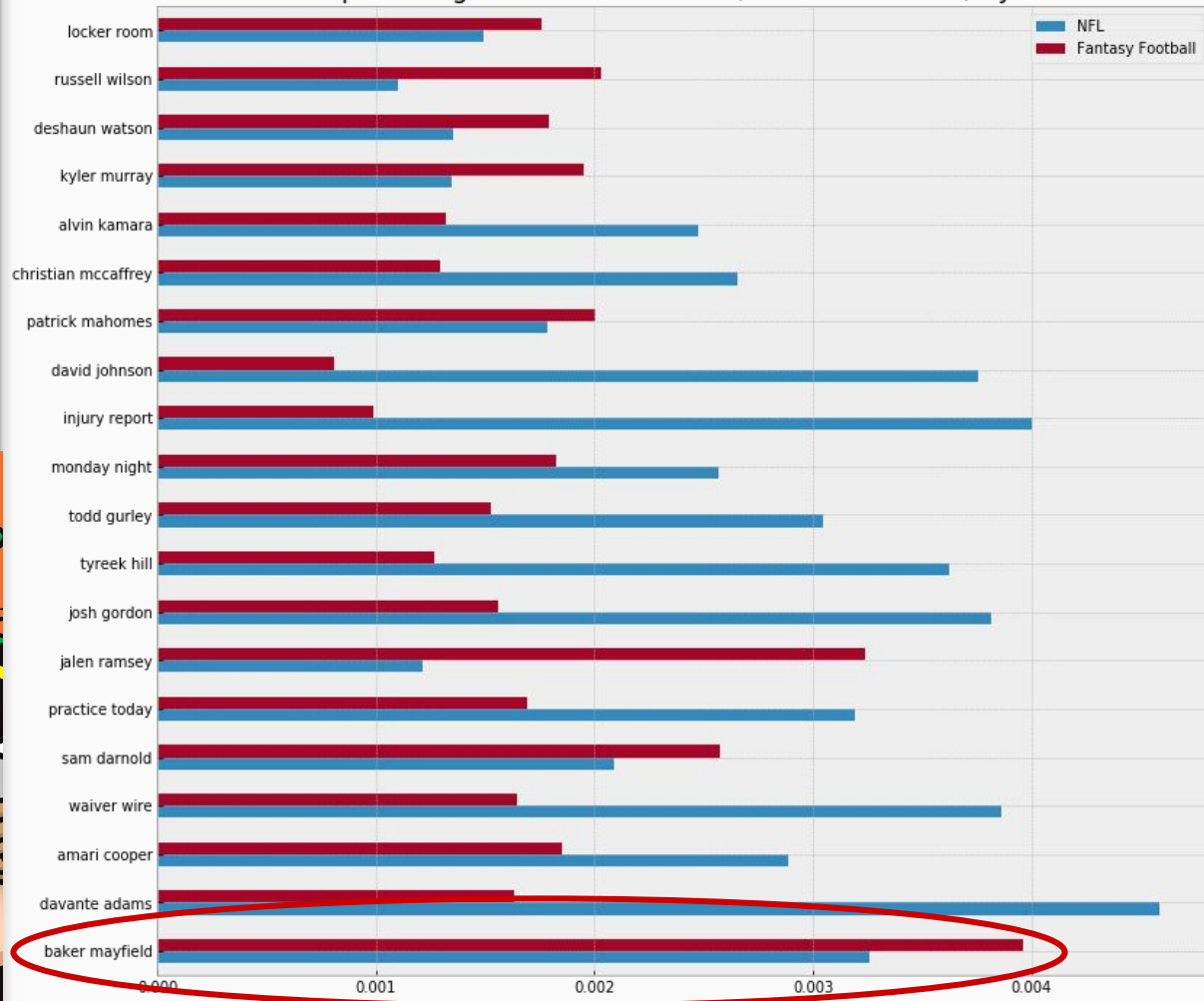

Boxplot of sentiment by Subreddit

Negative Bi-grams

- Fantasy subreddit is mentioning Baker Mayfield more often than the NFL.

Most Frequent Negative Double Words (TFIDF vectorized) by Subreddit

Legend: NFL, Fantasy Football

Categories (top to bottom):
locker room, russell wilson, deshaun watson, kyler murray, alvin kamara, christian mccaffrey, patrick mahomes, david johnson, injury report, monday night, todd gurley, tyreek hill, josh gordon, jalen ramsey, practice today, sam darnold, waiver wire, amari cooper, davante adams, baker mayfield

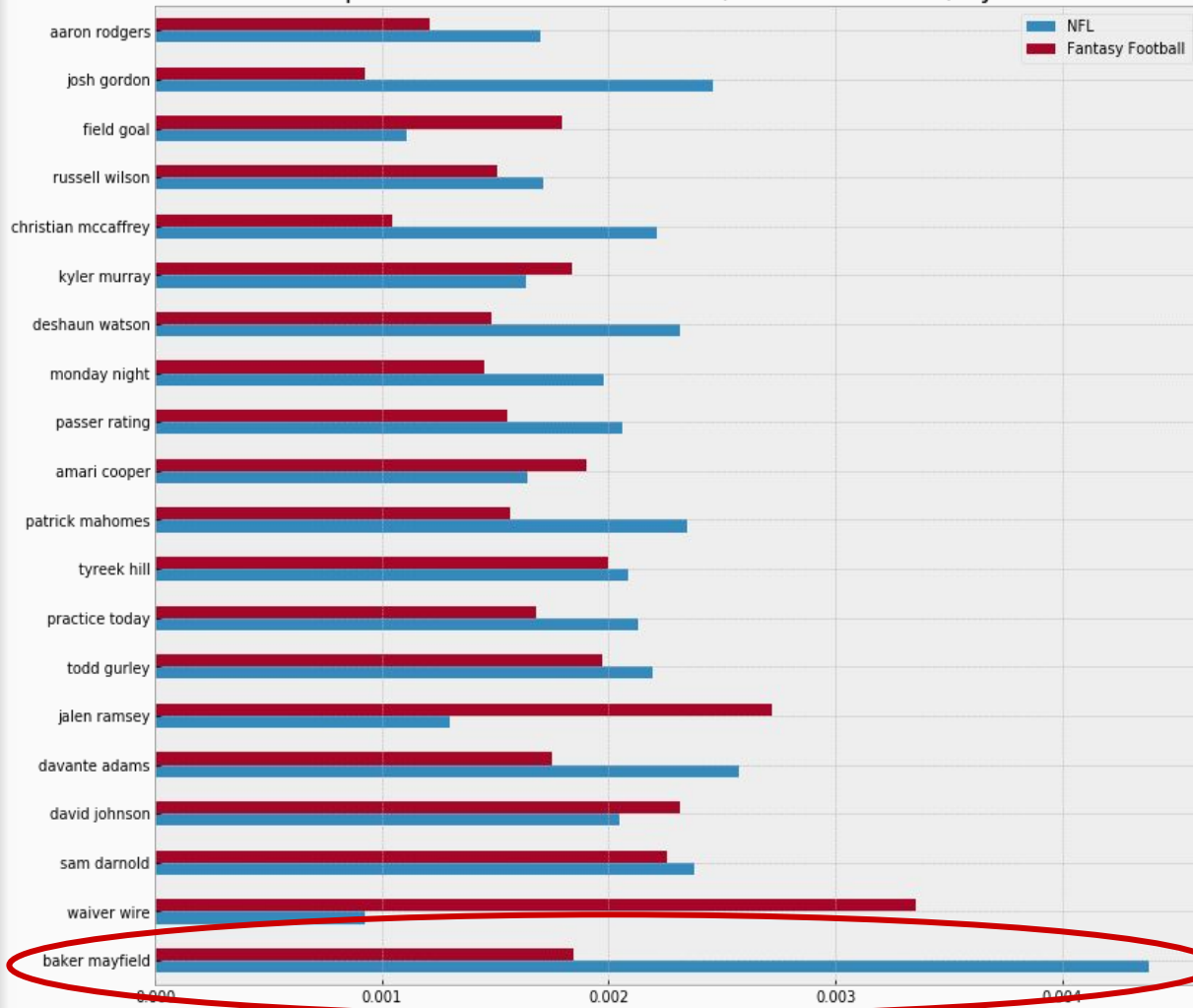X-axis: 0.000, 0.001, 0.002, 0.003, 0.004

Positive Bi-grams

- NFL subreddit is mentioning Baker Mayfield significantly more often than Fantasy.

Most Frequent Positive Double Words (TFIDF vectorized) by Subreddit

Legend: NFL (blue), Fantasy Football (red)

Categories (top to bottom): aaron rodgers, josh gordon, field goal, russell wilson, christian mccaffrey, kyler murray, deshaun watson, monday night, passer rating, amari cooper, patrick mahomes, tyreek hill, practice today, todd gurley, jalen ramsey, davante adams, david johnson, sam darnold, waiver wire, baker mayfield

X-axis: 0.000, 0.001, 0.002, 0.003, 0.004

Logistic Regression Model
Details and Performance

Input Details

- TFIDF Vectorizer used.
  - N-grams: 1 to 3 words
  - Max features: 100,000
  - Max Document Freq: 30%
- Log Reg
  - Ridge Regularization

Confusion Matrix

	Predicted NFL	Predicted Fantasy
Actual NFL	5,624	243
Actual Fantasy	366	4,586

- **Accuracy Score: 94%**

Conclusion

- Based off of the sentiment scores of both reddits it appears that users of the Fantasy Football subreddit are generally happier.

- Players are the center of the conversation (not the actual teams). The NFL should look into integrating this fact into how the game is played and Marketed.