NLP Modeling to differentiate between Soccer and Fifa Subreddits

DSI 523 Project 3 - Rick Powell

Table of Contents

Background

Problem Statement

Data Analysis

Modeling

Conclusion

https://ichef.bbci.co.uk/onesport/cps/976/cpsprodpb/5FA1/production/_126618442_gettyimages-1242992752.jpg

Background

r/soccer

Subreddit for Association Football (Soccer)

3.5M members

11 v 11 team sport

It is played by approximately 250 million players in over 200 countries and dependencies[source]

World’s Most Popular Sport

https://media.npr.org/assets/img/2022/06/06/ap22156795241469_custom-ac5673971ad7b73fb5af7e3c0605d08abc59f80e-s1200-c85.webp

Background

r/FIFA

Reddit for the video game FIFA

606,640 Members

Best selling video game franchise in the world[source]

Allows players to play as their favorite players on their favorite teams

https://en.wikipedia.org/wiki/FIFA_22

Problem Statement

As a data scientist for EA Sports, we want to examine what people are posting about in the r/soccer subreddit and compare that to what people are posting about in the r/FIFA subreddit in order to make recommendations on how to improve the game.

Data Analysis

Data Pulled 9/21/22

7978 Unique Posts

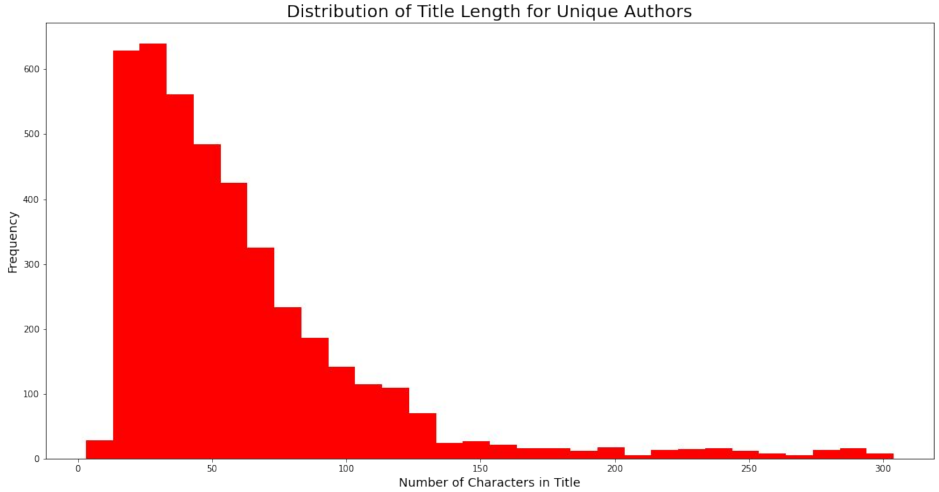
4201 Unique Authors

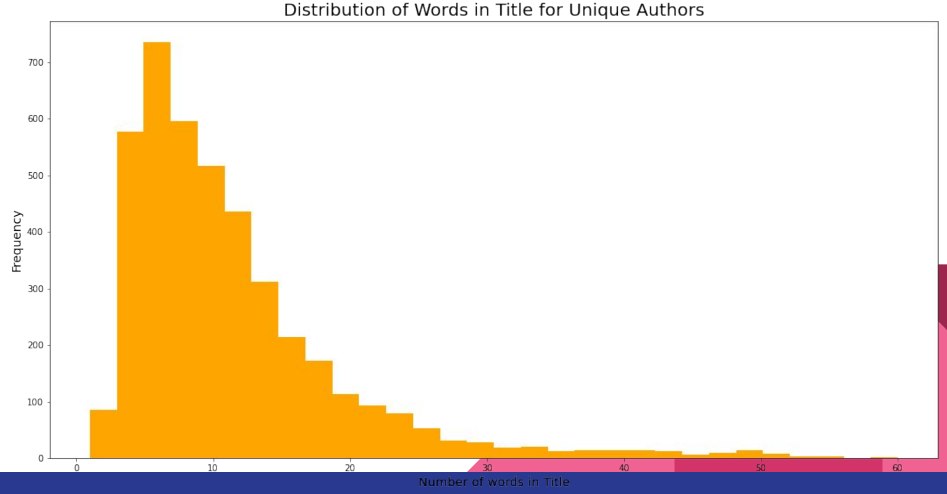
Average Title: 10.05 words

55.8 characters

Longest Post: 3427 words

19614 characters


Distribution of Title Length for Unique Authors

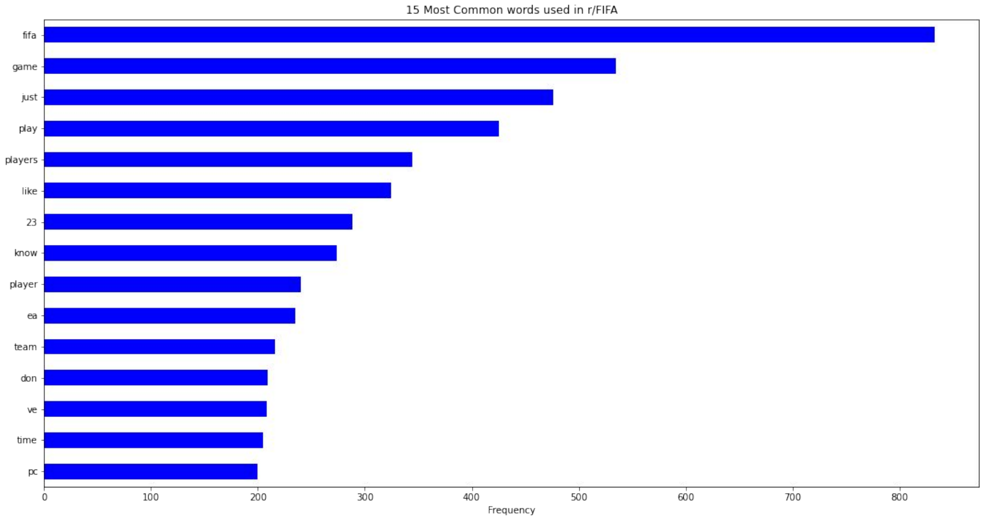

Distribution of Words in Title for Unique Authors

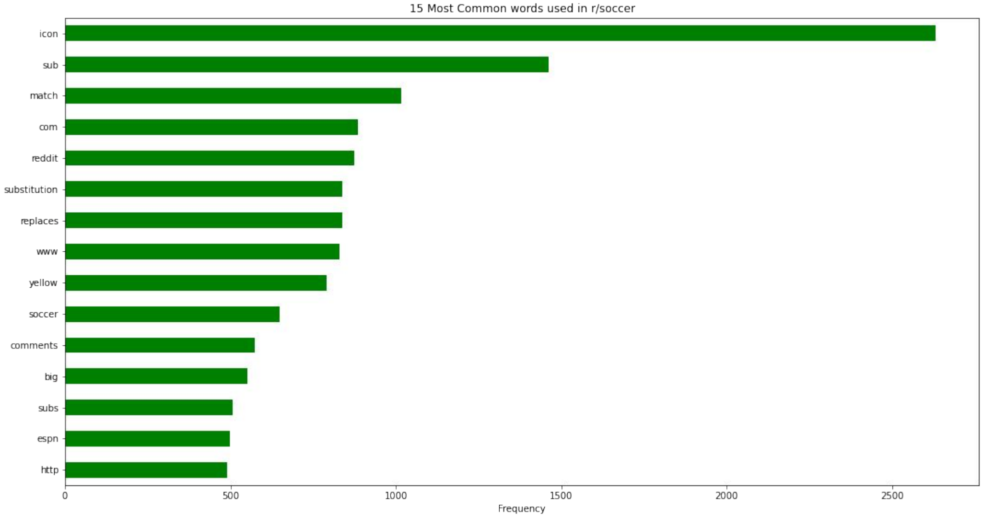
Data Analysis

Most common words used in posts on each subreddit.

Stopwords have been removed

https://www.teamusa.org/-/media/TeamUSA/Soccer/Yedlin_DeAndre/Yedlin_DeAndre_101121_1440x810_.png





Modeling

Baseline Model

r/soccer posts: 3978

r/FIFA posts: 4000

Baseline Model is 50.14%

https://media.wired.co.uk/photos/610d3e28647eec9baf1ca396/16:9/w_2560%2Cc_limit/Copy%2520of%2520F22_G4_Mbappe_30_-25102260ed97dfd39b22.73803290.png

Modeling

RandomForest

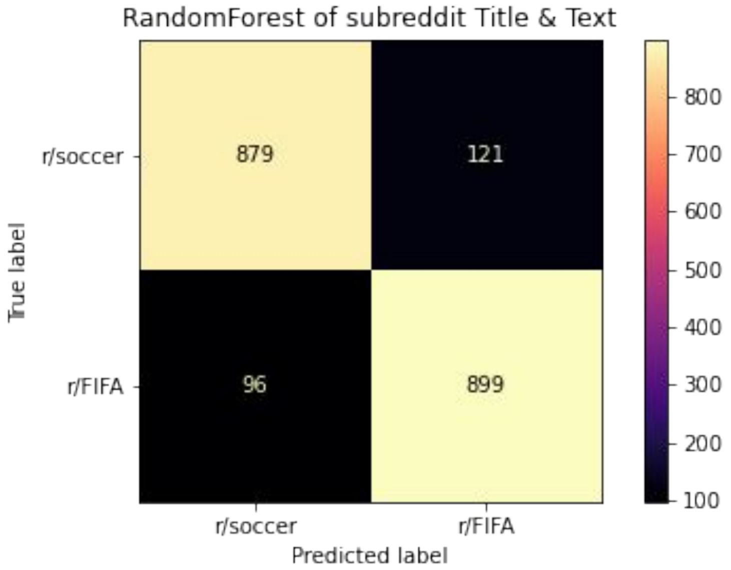
Training Score: 0.938

Testing Score: 0.891

Accuracy: 89.1%

Recall: 90.3%

Specificity: 87.9%



Modeling

Multinomial Naive Bayes

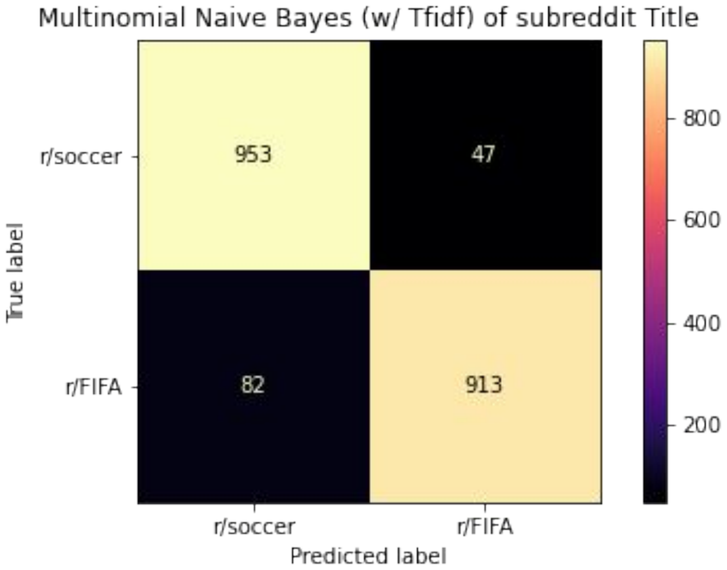
Training Score: 0.962

Testing Score: 0.935

Accuracy: 93.5%

Recall: 91.8%

Specificity: 95.3%



Conclusion

Production model is the Multinomial Naive Bayes Model on the Subreddit Title

- 93.5% Accuracy
- Not as Overfit as RandomForest
- No preference on minimizing Recall vs Specificity as False Positives and False Negatives are weighed equally

Conclusion

r/soccer focused on player stats most common words: shots, passes, goals, time

r/FIFA is focused more on the players themselves (most common words: players, player, play, team, and tots (team of the season)).

My recommendation to EA Sports is to push their Competitive FIFA leagues (ePremier League, eMLS). A lot of the stats and keywords that are commonly found in the r/soccer community would be more prevalent in r/FIFA if there was a much smaller subset of leagues that was followed by the entire community.

Questions?

https://images.pexels.com/photos/47730/the-ball-stadion-football-the-pitch-47730.jpeg