Predicting Subreddits

March 4, 2022

NBA or Chess?

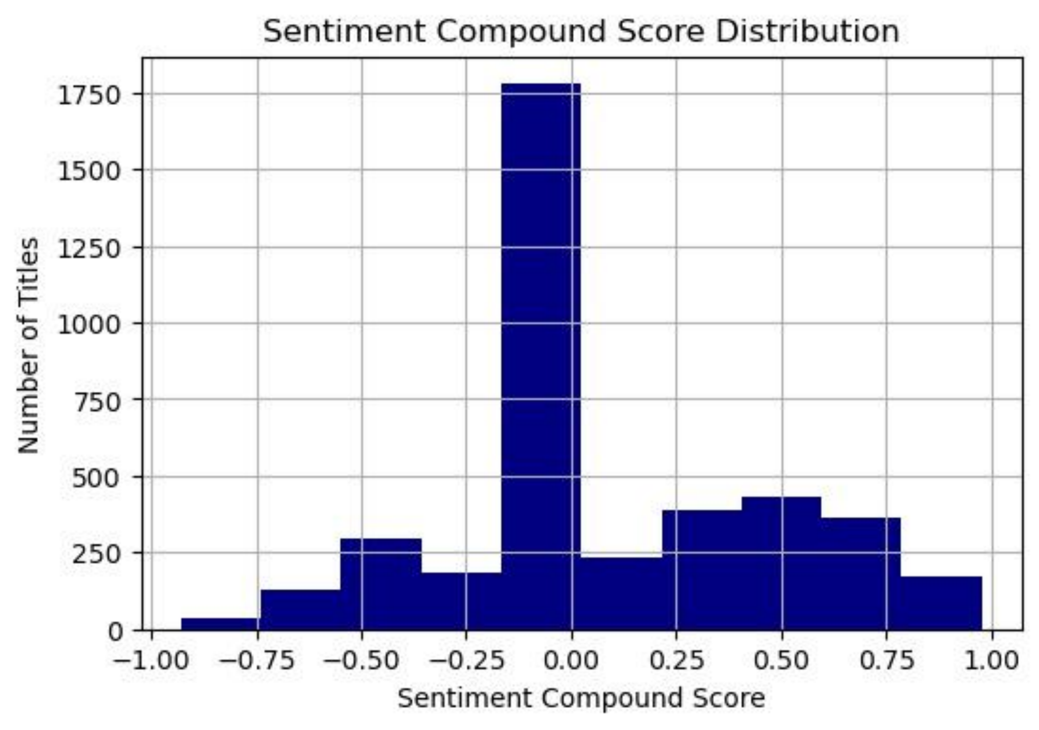
Overview

1. EDA
2. Modeling
3. Conclusions
4. Next Steps

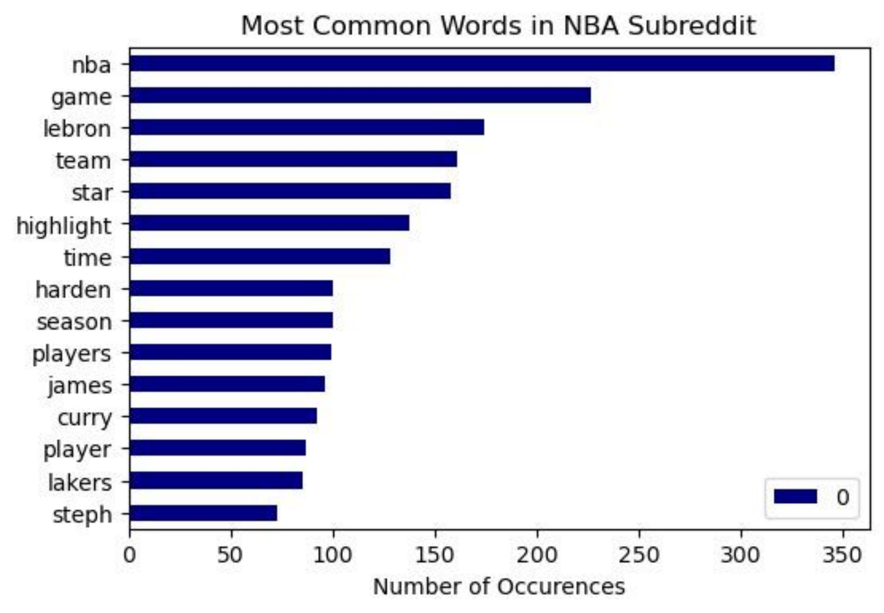
EDA

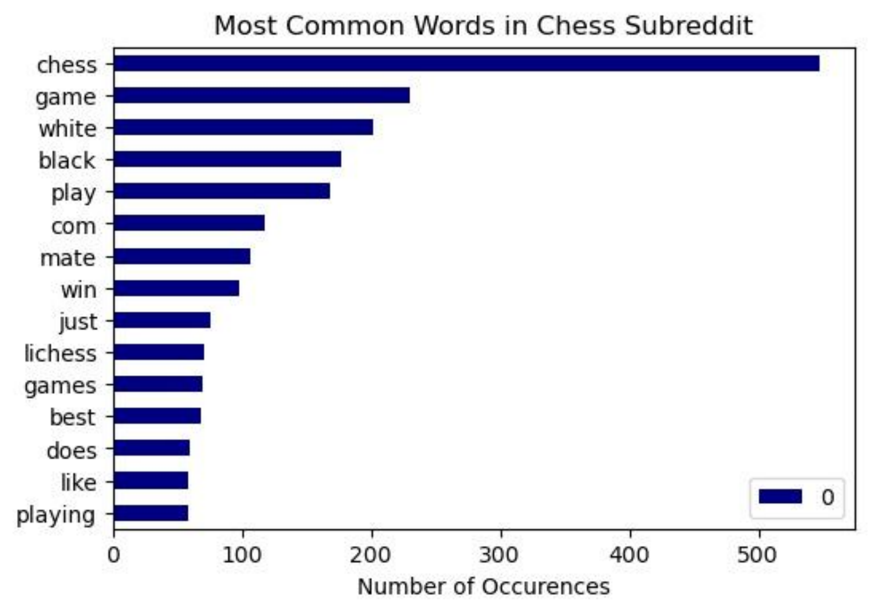
Engineered Features

- Title + Self-text
- Word Count
- Length (in characters)
- Sentiment Analysis



EDA

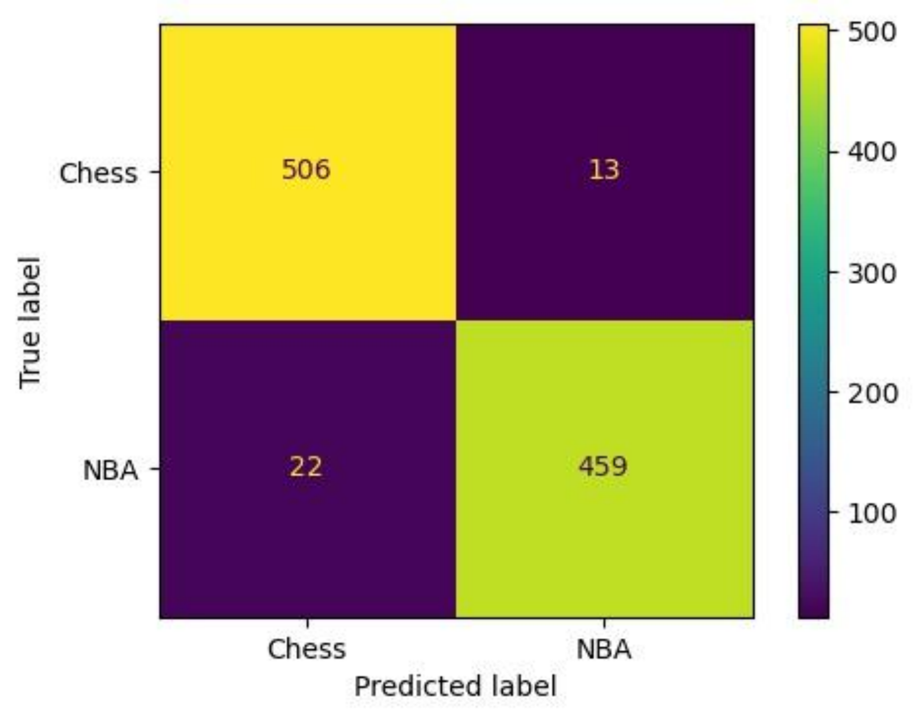

Most Common Words in NBA Subreddit
nba
game
lebron
team
star
highlight
time
harden
season
players
james
curry
player
lakers
steph
0
0
50
100
150
200
250
300
350
Number of Occurences

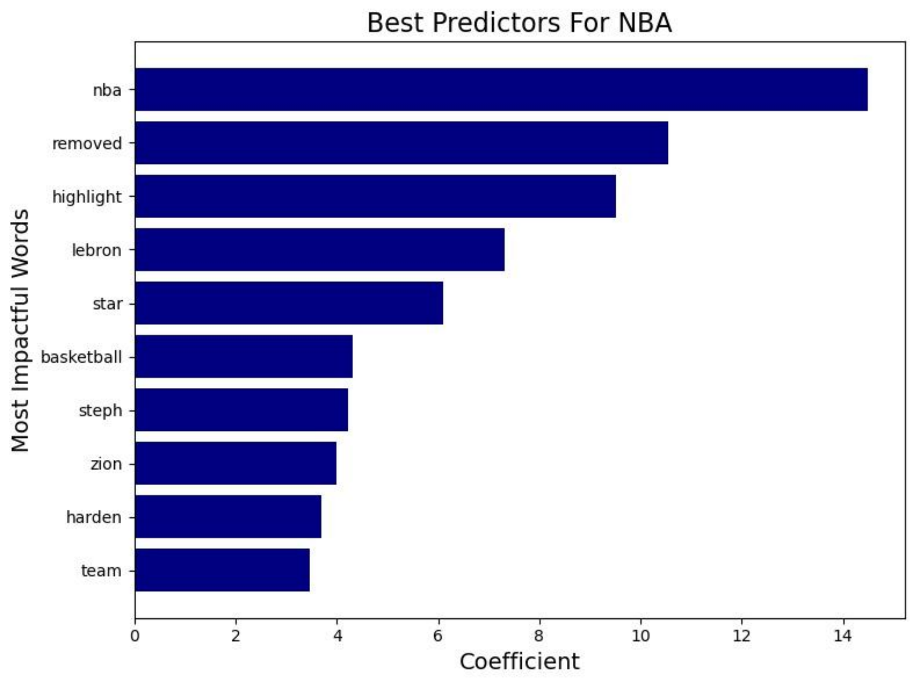

Most Common Words in Chess Subreddit
chess
game
white
black
play
com
mate
win
just
lichess
games
best
does
like
playing
0
0
100
200
300
400
500
Number of Occurences

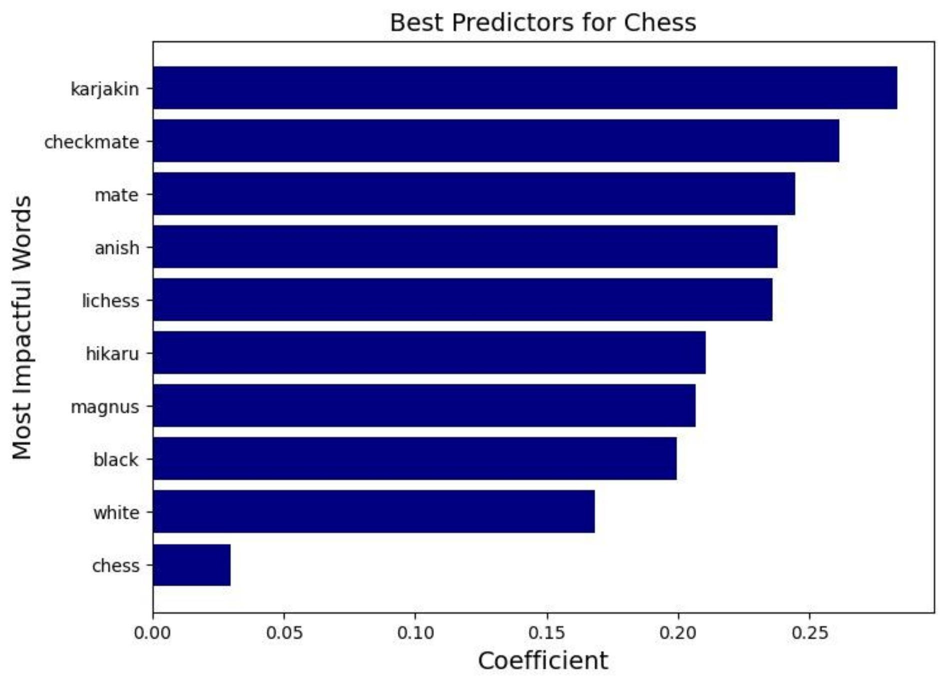
Modeling

	Model 1	Model 2
Model Type	Logistic Regression	Random Forest
Text Analysis	CountVectorizer	TFIDF
Other Methodologies	Pipeline, FeatureUnion, GridSearchCV	Pipeline, GridSearchCV
Parameters	stop words, max features, min df, max df	stop words, max features, min df
Accuracy Score	95.9%	95.5%

Confusion Matrix




Best Predictors For NBA
Most Impactful Words
nba
removed
highlight
lebron
star
basketball
steph
zion
harden
team
0
2
4
6
8
10
12
14
Coefficient


Best Predictors for Chess
Most Impactful Words
karjakin
checkmate
mate
anish
lichess
hikaru
magnus
black
white
chess
0.00
0.05
0.10
0.15
0.20
0.25
Coefficient

Conclusions & Next Steps

Conclusions

- Yes, it was possible to create a model that beats the baseline model of 50% Accuracy.
- Consider the relationship between the two subreddits.

Next Steps

- Additional CountVectorize tuning (different token patterns)
- More feature engineering
- Look into time patterns.