

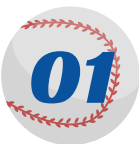


StubHub *Marketing*

Peter Murphy, Data Scientist

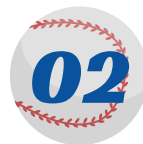


Table of contents



01

***Subreddit
Background***



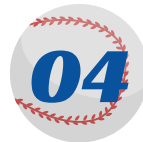
02

***Data
Processing***



03

***Model
Evaluation***



04

Recommendations



Subreddits



247k
members

r/mlb

Where to watch MLB Games?

Watch local MLB games on your local regional sports networks or regional sports apps. For out-of-market games, stream on MLB.TV or on ESPN+.

With ESPN+, please keep note that not all games will be available.

Select MLB games will be on FOX, FS1, ESPN, TBS, Apple TV+, Peacock, YouTube, and on MLB Network.

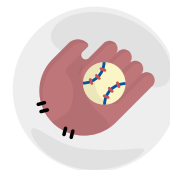
Sunday Morning: Peacock

Sunday Night: ESPN

Tuesday Night: TBS

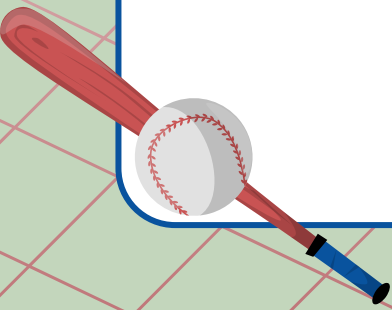
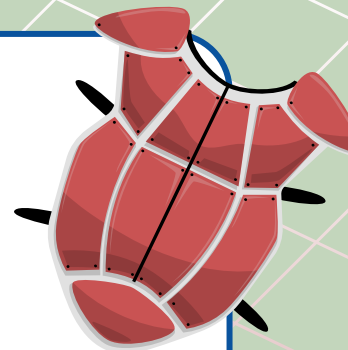
Friday Night: Apple TV+

Saturday Night: FOX



r/redsox

145k
members

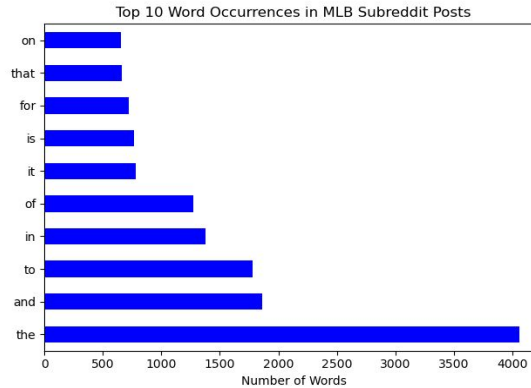


Data Cleaning

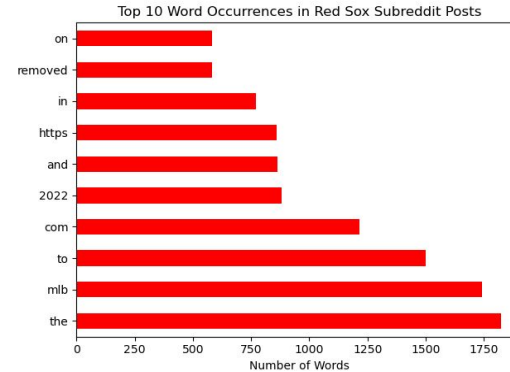
3,000 posts:

June 22, 2022 - September 17, 2021

866 unique authors



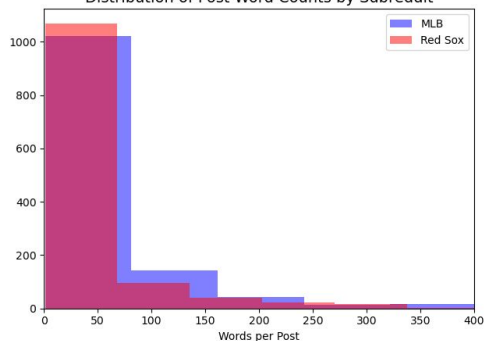
757 unique authors



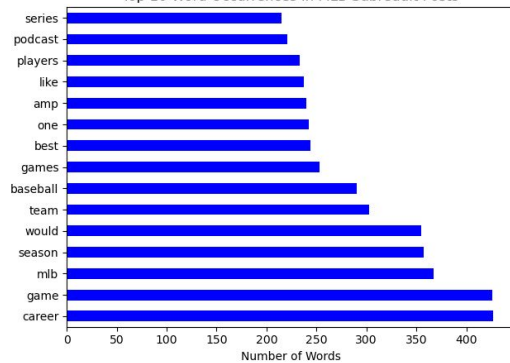
1. **“removed” / “deleted” posts**
2. **Outliers?**
3. **Removed video posts**
4. **stop words (“https”, “com”, “www”)**

EDA

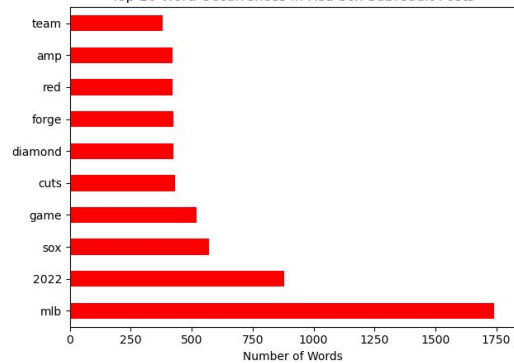
Distribution of Post Word Counts by Subreddit



Top 10 Word Occurrences in MLB Subreddit Posts



Top 10 Word Occurrences in Red Sox Subreddit Posts

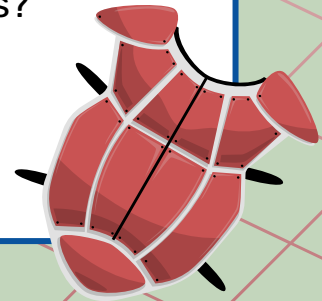




Problem Statement

To inform our marketing campaign, this project's purpose is two-fold:

- 1) Can we predict with 70% accuracy whether a post came from the Red Sox or MLB subreddit?
- 2) Are there discernible differences in sentiment between users on the MLB and Boston Red Sox subreddits?







Model Performance

Baseline Model Accuracy

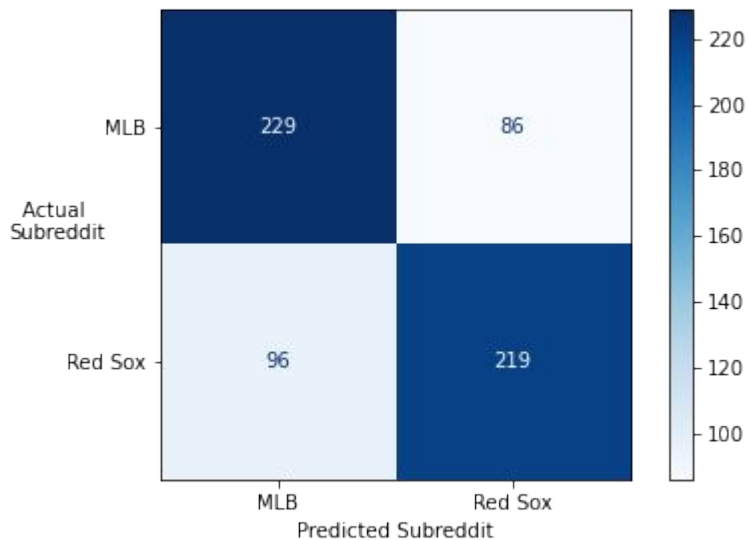
50.02%

Attempt	Model Description	Train Accuracy	Test Accuracy	F1 Score
1	Multinomial Naive Bayes (Count Vectorizer)	0.875	0.700	0.703
2	Multinomial Naive Bayes (Tfidf Vectorizer)	0.88	0.713	0.718
3	Logistic Regression (Count Vectorizer)	0.894	0.727	0.749
4	Logistic Regression (Tfidf Vectorizer)	0.824	0.730	0.737
5	Random Forest (Count Vectorizer)	0.750	0.713	0.706
6	Random Forest (Tfidf Vectorizer)	0.751	0.660	0.670
7	Extremely Randomized Trees (Count Vectorizer)	0.713	0.656	0.624
8	Ada Boost Classifier (Count Vectorizer)	0.800	0.727	0.743
9	Ada Boost Classifier (Tfidf Vectorizer)	0.802	0.694	0.717
10	Gradient Boosting Classifier (Count Vectorizer)	0.901	0.735	0.746
11	Gradient Boosting Classifier (Tfidf Vectorizer)	0.903	0.716	0.75



Best Model

Random Forest (Count Vectorized)



Tuning

- **English stop words**
- **200 estimators**
- **5 levels to split on**
- **Unigrams**

Recommendations

r/mlb

r/redsox

Sentiment: 0.25

0.14

Limitations:

- **Time of year**
- **Team Performance**
- **High-volume posters**

Next Steps

- **Additional fan bases**
- **Expand Samples**
- **Additional Sports**

r/redsox Author	Posts (out of 1,260 Total)
RedSoxGameday	127
trixster110497	69
Tal_Venada	28
Keith_H25	21
lordbloodstar	18
3ric8223	16
TalkinMetsWithRob	15
retroanduwu24	15
canitryto	15

Thanks!



Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics and images by **Freepik**

