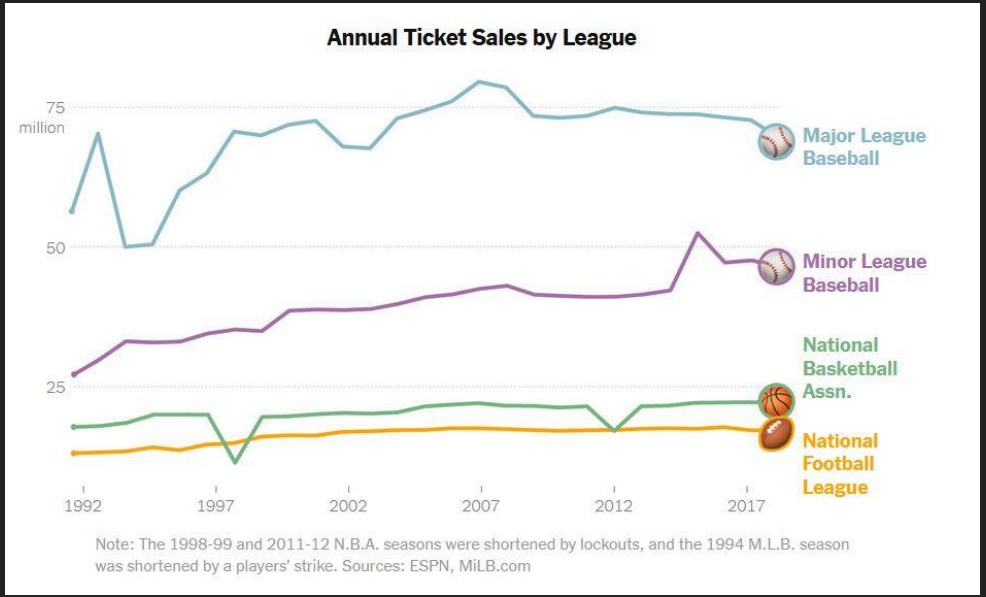
MLB Advertising

Matthew Burrell
MLB Data Scientist

Problem

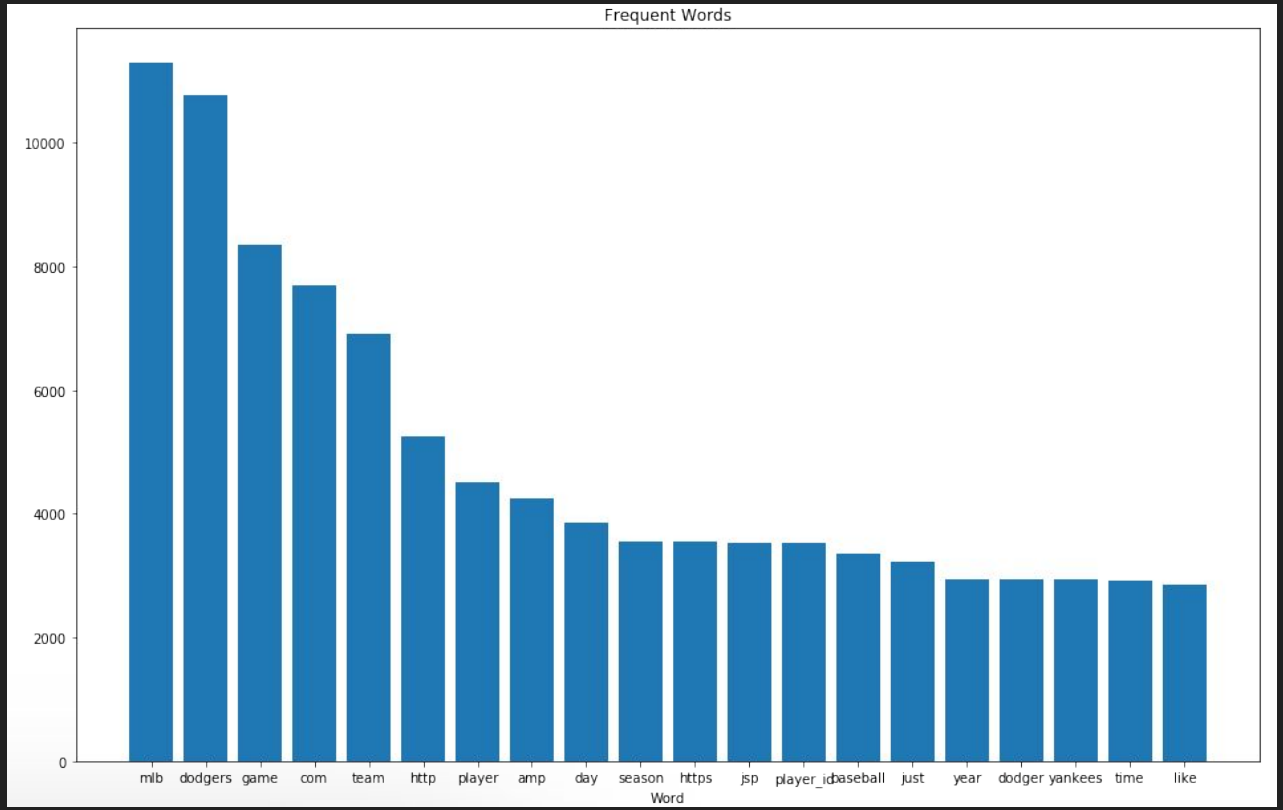
- MLB declining in popularity
  - falling attendance
  - Porr World Series ratings
  - lack of nationally recognized stars.
- Covid-19 pandemic shutting down in game attendance
- Custom ads with social media and live stream shares .
- The most accurate model possible with the goal to launch a model by next season.



Data

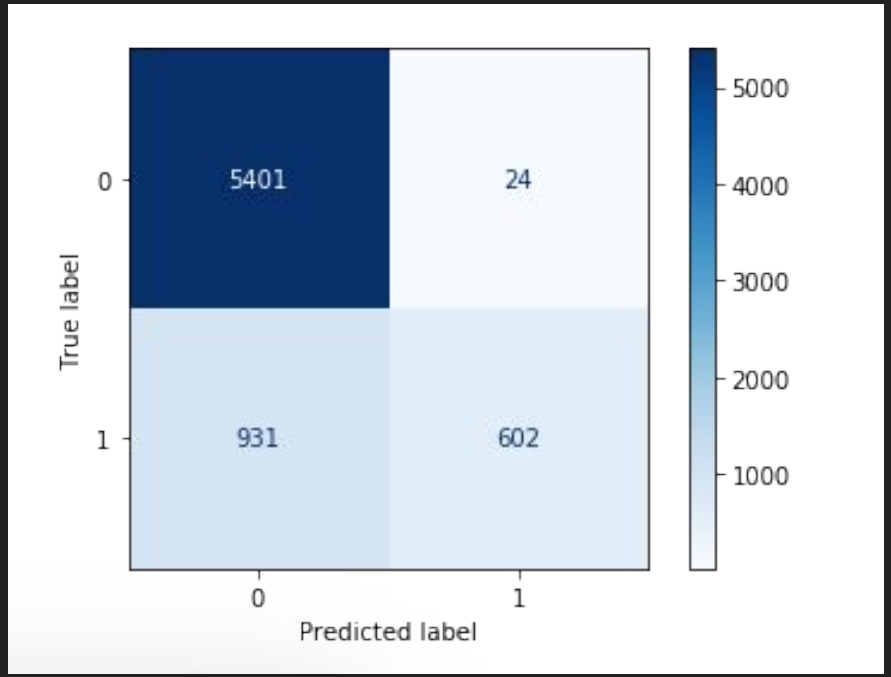
- Pulled from two subreddits Dodgers and NYYankees
  - 26k post from the Dodgers s
  - 6k post form the Yankee
- Missing body of the post are from media like pictures or videos.
- The distributions of all lengths and word counts are skewed to the right.
- 44 emoji post have been removed
- 275 single word data dropped

Frequent Words



Modeling

- Baseline model accuracy is about 77.97%
- Random forest classifier with an accuracy of 85.89%
- The misclassification rate of the model is 15.31%
- The true positive rate of the model is 39.26%
- The true negative rate of the model is 99.56%
- The precision if the model is 96.17%



Conclusion

- The model can predict at 84% accuracy what post are from Dodger fans or Yankee fans
- True positive rate is low at 39%
  - Misclassified a lot Yankee fans as dodger fans
- True negative rate is 99.56%
  - Misclassified some Dodger fans as Yankee fans

Next steps

1. Expanded current model to all 25 teams
    a. Not as accurate as but gets the next step to targeting users
2. Try Term Frequency-Inverse Document Frequency (TF-IDF) Vectorizer
    a. TF-IDF is a score that tells us which words are important to one document, relative to all other documents.
    b. words that occur often in one document but don't occur in many documents contain more predictive power.

References

How Popular Is Baseball, Really?,The New York Times, accessed 27 July 2020, <ttps://www.nytimes.com/interactive/2019/10/22/sports/baseball/baseball-popularity-world-series.html>