Binary Classification Models with NLP

Michael Riedeman

Overview

Problem Statement

Data Acquisition

Baseline Model

Data Cleaning/EDA

Modeling Process

Conclusion/Recommendations

Questions

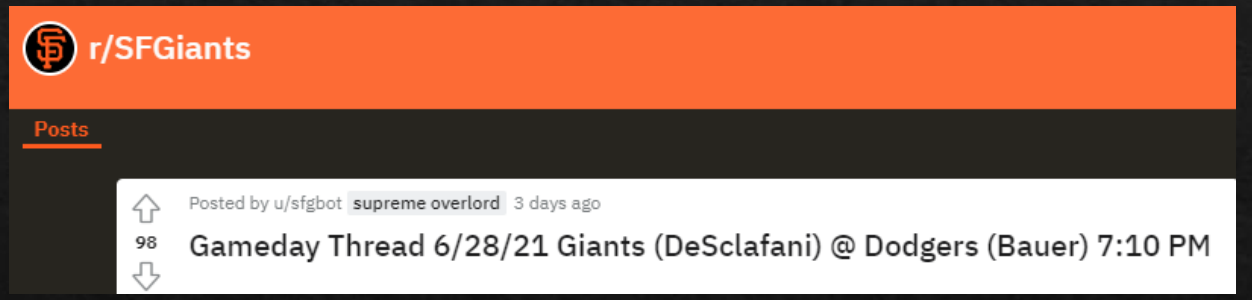
Problem Statement

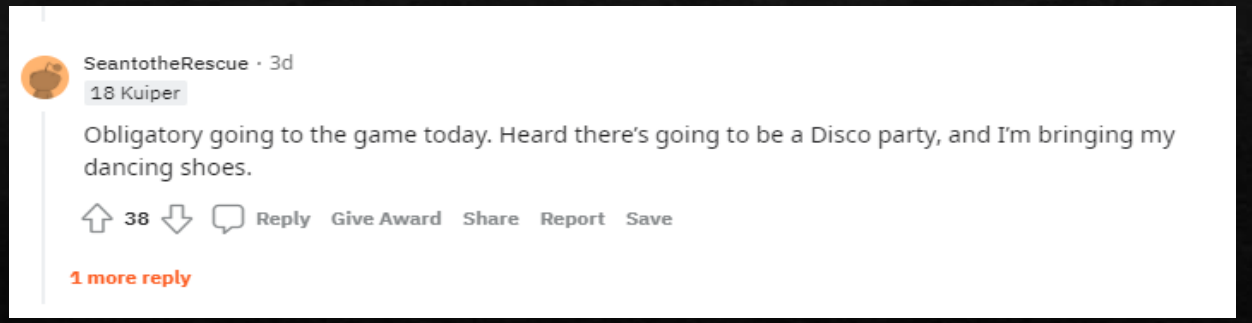
- Identify whether a comment was written in the Dodger's or Giant's subreddit by maximizing the accuracy of a binary classification model.

- Challenges:
  - Same sport
  - Rivals
  - Context of comments

Data Acquisition

- Subreddits: r/Dodgers and r/SFGiants
- Reddit's Pushshift API
- Praw Reddit API Wrapper





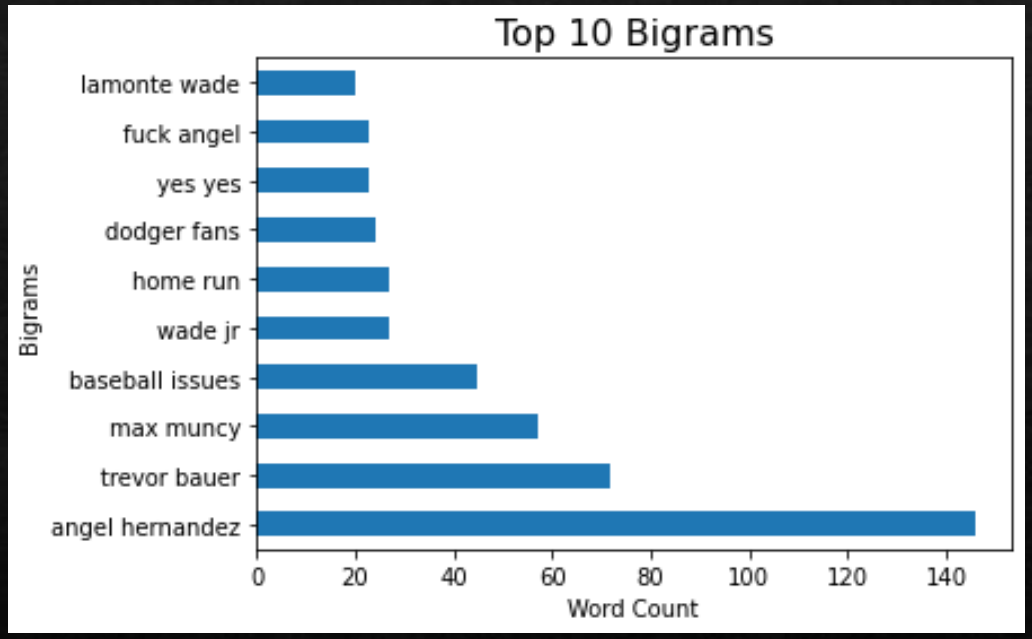
Baseline Model

- Corpus
  - Giants Game Thread: 3966 comments
  - Dodgers Game Thread: 2353 comments

- Baseline Model: 62.8%
- Optimize for Accuracy

Data Cleaning/EDA

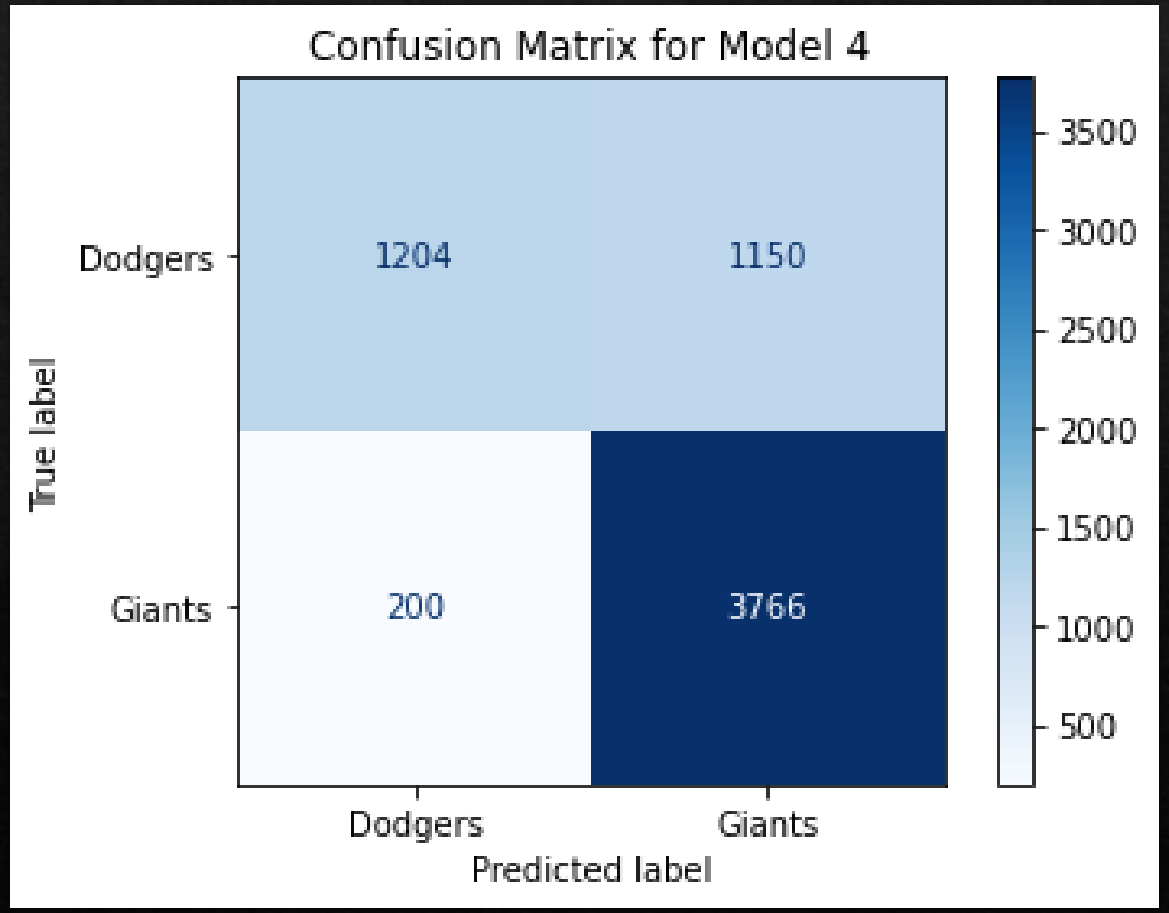
- Remove Stop Words
  - English Stop Words in Count Vectorizer
  - Urls, Spammed Bot Words
- Inclusion of Bigrams/Trigrams



Modeling Process

◈ Model 1			Accuracy
◈ Count Vectorizer	→	Random Forest Classifier	65.7%
◈ Model 2			
◈ Count Vectorizer	→	Logistic Regression	69.5%
◈ Model 3			
◈ Count Vectorizer	→	Naïve Bayes	69.8%
◈ Model 4			
◈ Tfidf Vectorizer	→	Logistic Regression	69.5%

Modeling Process

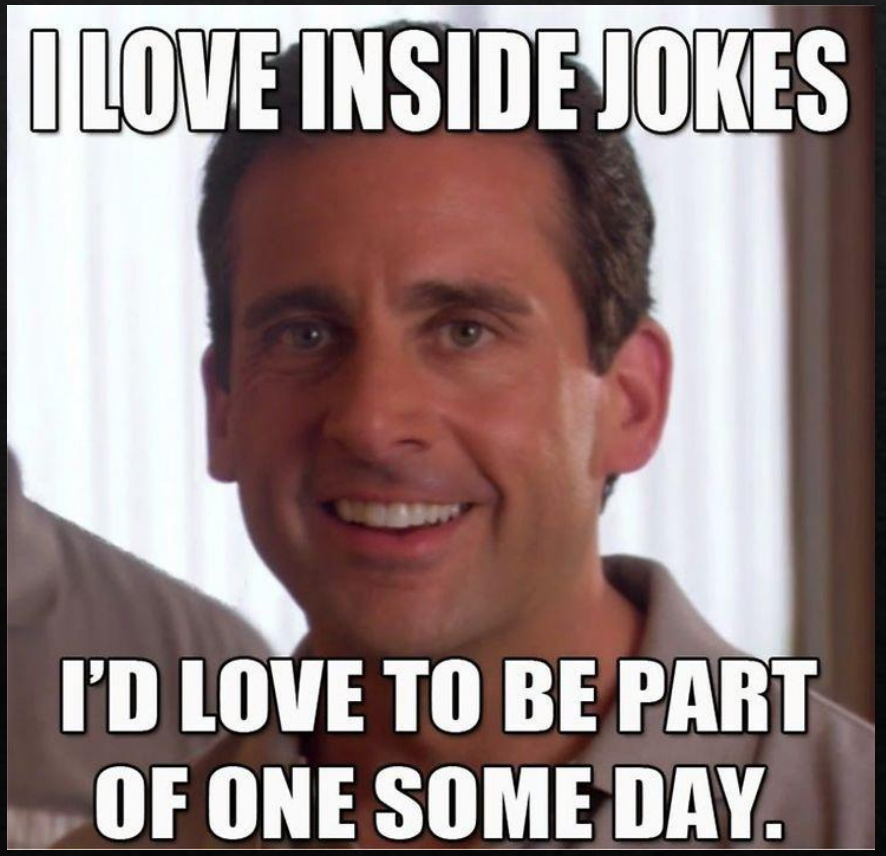


Conclusions

- Best Model:
  - Naïve Bayes with 69.8% Accuracy
- Only 7.0 % increase over baseline model (62.8%)
- Hyperparameters:
  - stop words, bigrams, trigrams

Recommendations

- Utilize SVM and tune additional hyperparameters
- Additional data cleaning
- Sentiment Analysis



Questions?