Comparing Tinder vs Bumble Subreddits

Rebecca Patterson

July 30, 2021

Agenda

- Goal
- Overview of the Methods
- Modeling Approach
- Results of the Models

Goal

Create a model to classify posts as either belonging to the Tinder vs Bumble subreddits

Overall Approach

- In this study, I compared text from the subreddits for Tinder and Bumble
- I used the title, description (selftext) and comments fields to build a corpus to train multiple models
- After scraping and cleaning the data I ran multiple experiments to determine which model could best predict between the two subreddits

Modeling Approach

1. Data Collection

2. Data Cleaning

3. Model Tuning

Cleaning Steps

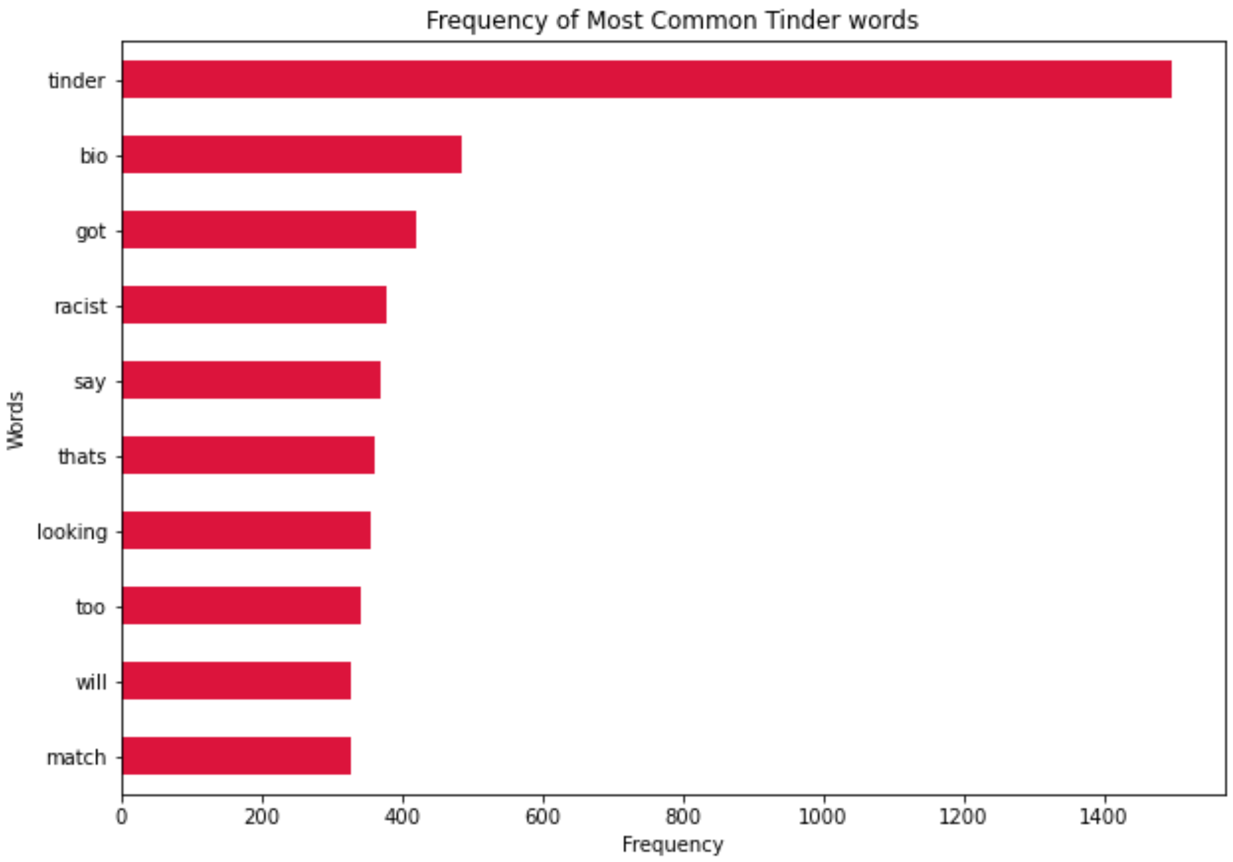
1. I removed any blank posts, or posts that had been deleted or removed

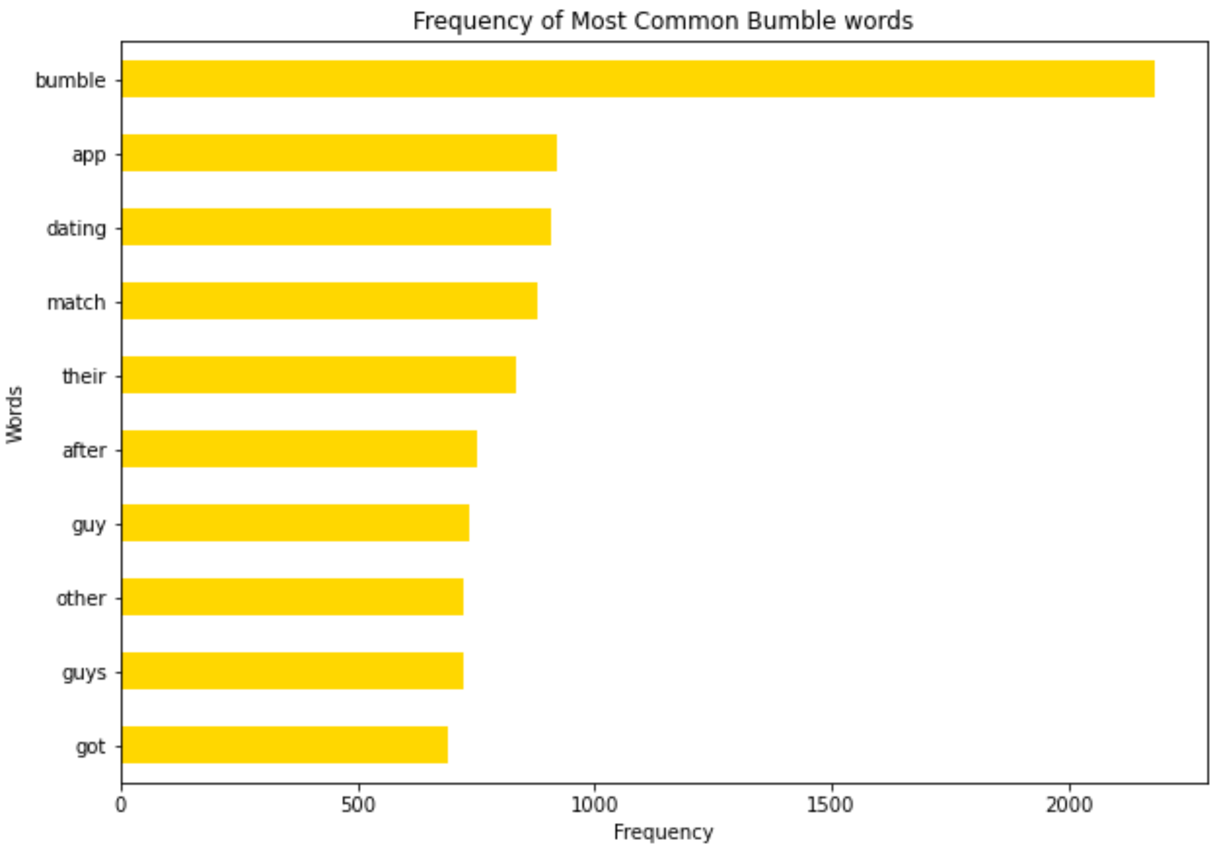
2. I removed all single character or single emoji posts and punctuation

3. I combined all text fields into one ‘combined’ text field

4. I reviewed the most common shared words to create a custom stop words list

Most Frequent Words per Subreddit



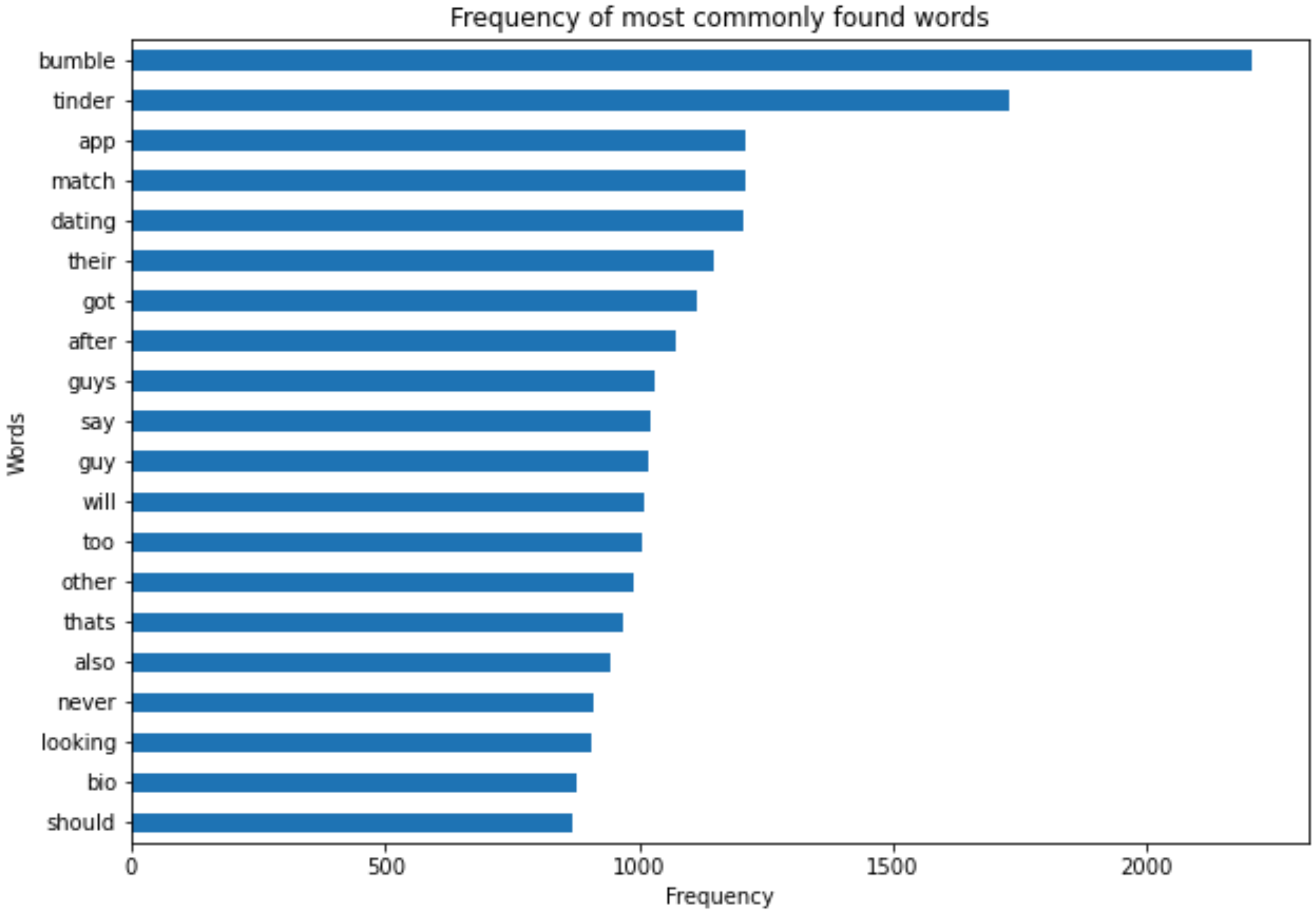


Interesting top words

Tinder : bio, racist, say, looking

Bumble: app, dating, guy, other

Most Frequent Words Combined



Modeling

Models Evaluated:

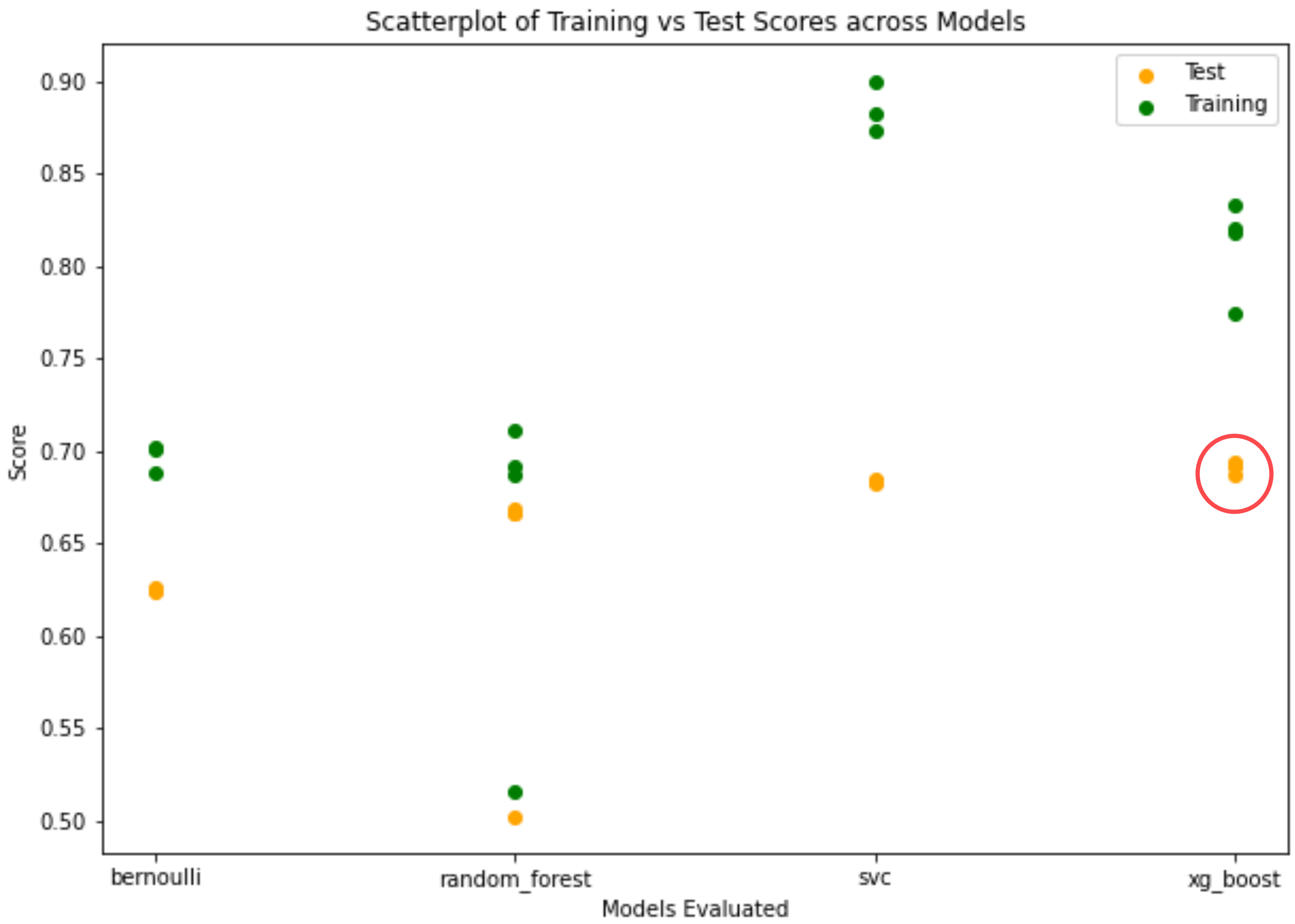
- SVC
- Bernoulli
- Random Forest
- XG Boost

Modeling Strategy

Ran BayesSearchCV across the different transformers/estimators to tune hyperparameters adjusting the following:

- **Count Vectorizer:** Max Features, Min Document Frequency, Max Document Frequency
- **SVC:** C, Coefficient, Kernel, Gamma, Degree, Shrinking
- **Bernoulli:** Alpha
- **Random Forest:** Number of Estimators, Max Depth
- **XG Boost:** Number of Estimators, Max Depth

Model Results



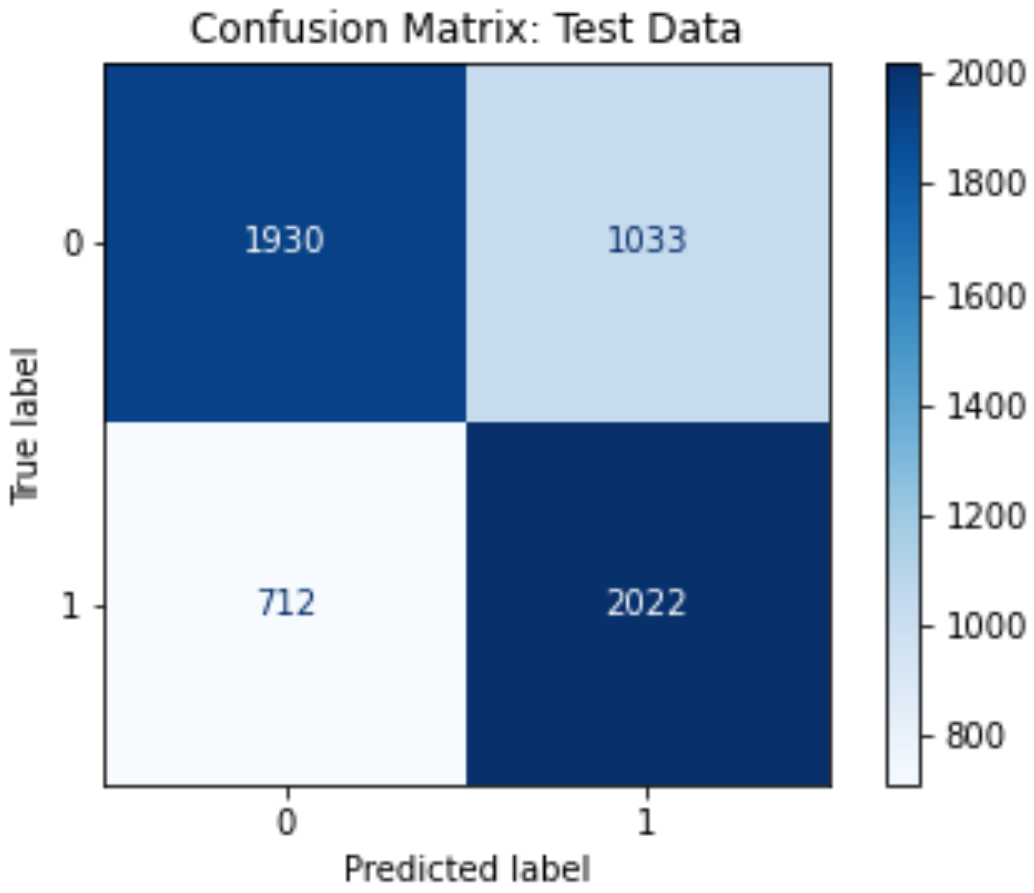
Best Model: XG Boost

XG Boost was the highest model evaluated with the following results:

- Training score: 0.83
- **Test score: 0.69**

- Best hyperparameters:
    - Cvec max_df: 0.9
    - Cvec max_features: 18,000
    - Cvec min_df: 2
    - XG max_depth: 15
    - XG n_estimators: 279

Confusion Matrix of Results



Results

- I found that these subreddits ended up being quite similar
- The model predicted accurately 70% of the time which was better than the baseline model of 50% but not as high as I would have liked
- Future work:
  - Additional analysis on text cleaning: e.g. stemming or lemmatizing, more EDA on minimum post length
  - Additional tuning: e.g. exploring additional hyperparameters for each model

Questions?