Spotify vs Netflix

Yunus Herman

The purposes :

1. Understand current trends using Reddit
2. Better understand customer’s preferences
3. Which model is better to predict subreddit

Audience :

Spotify or Netflix Data scientists

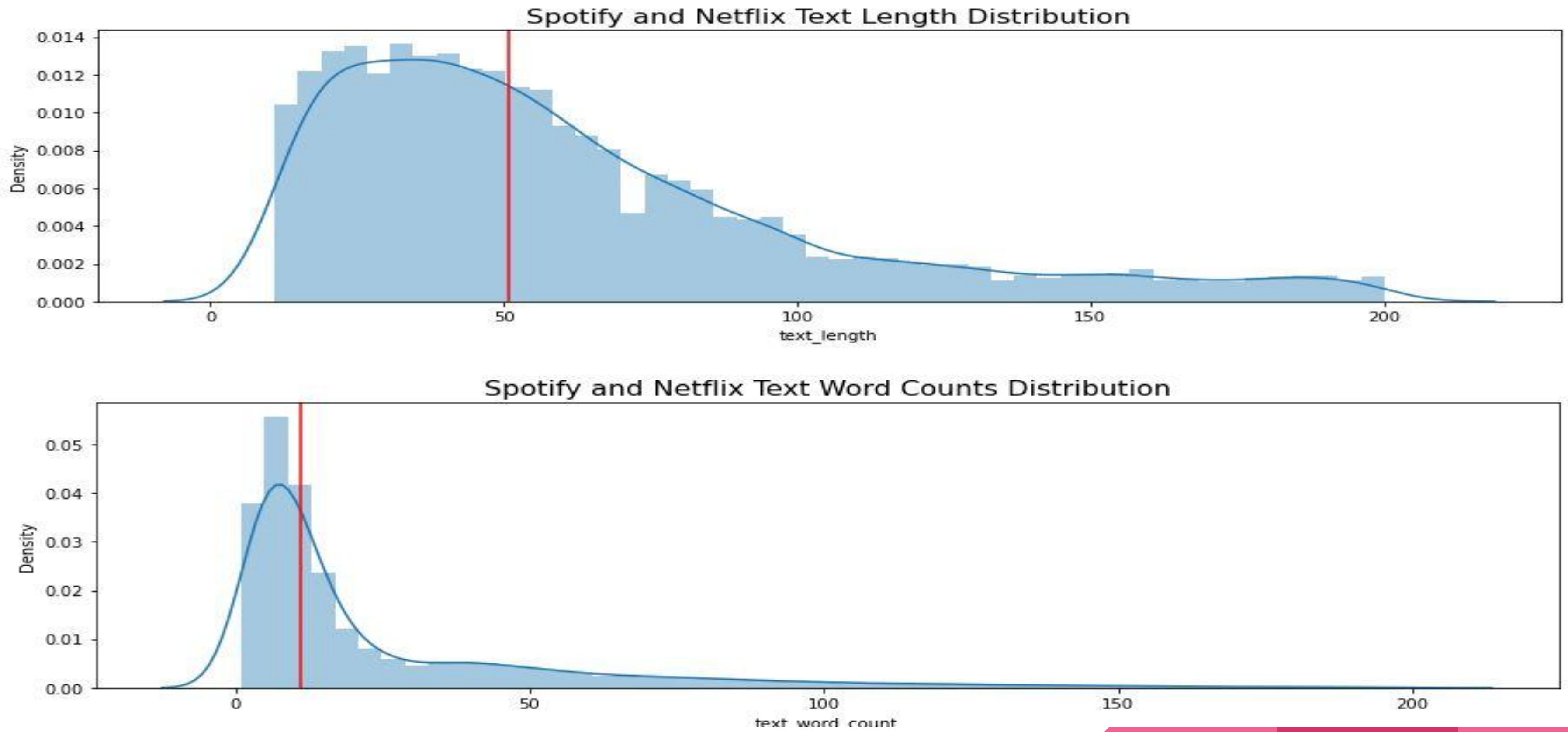
Webscraping

1. Using pushshift API (https://api.pushshift.io)

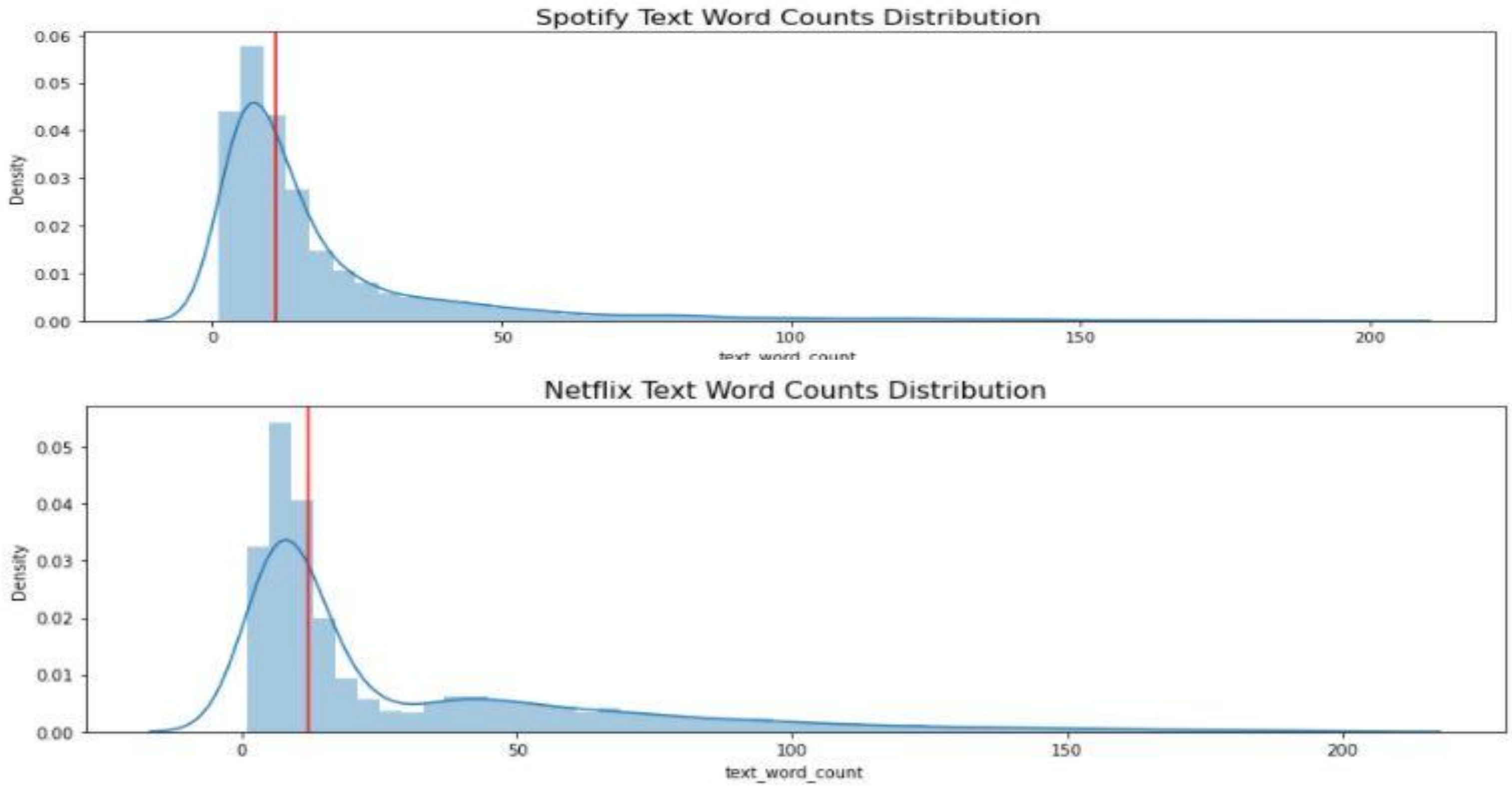
2. Subreddits : Spotify and Netflix

3. 10,000 rows each Subreddit, before Jan 31,2021

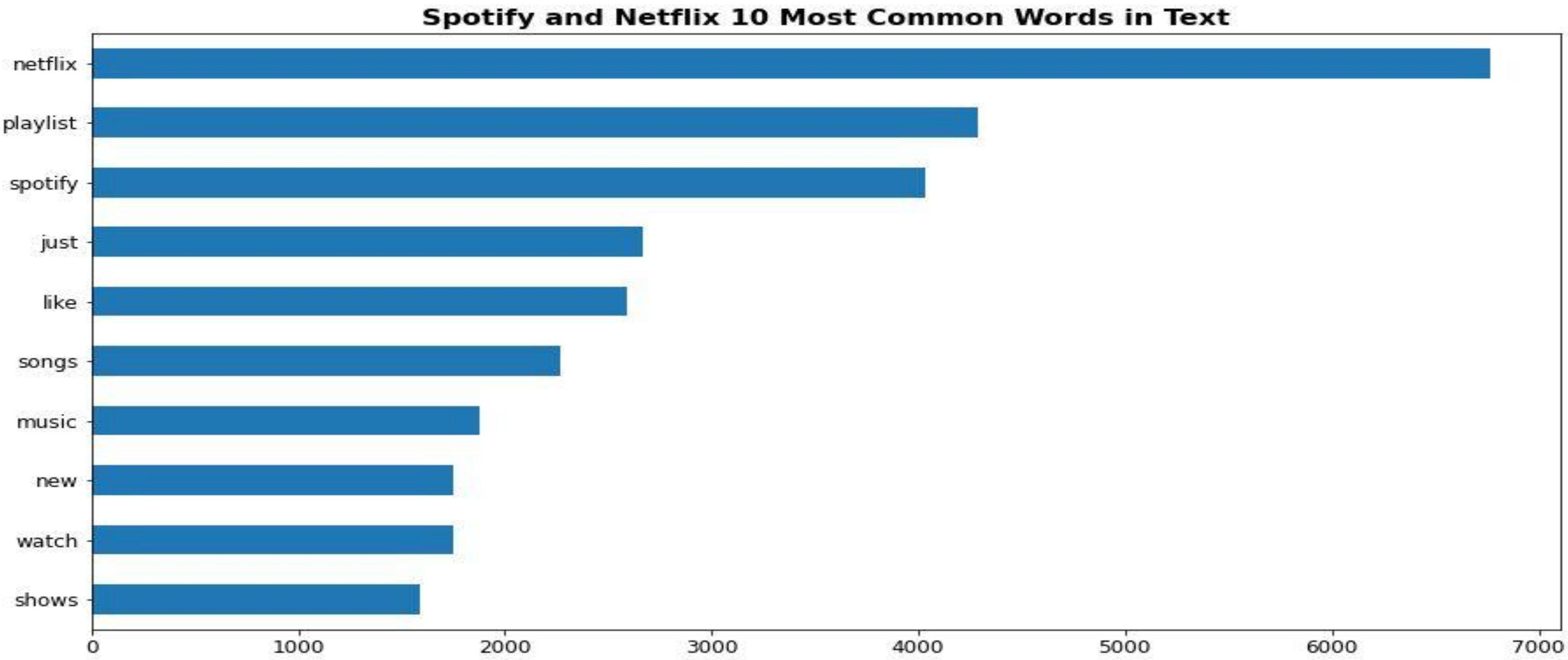
Distribution:



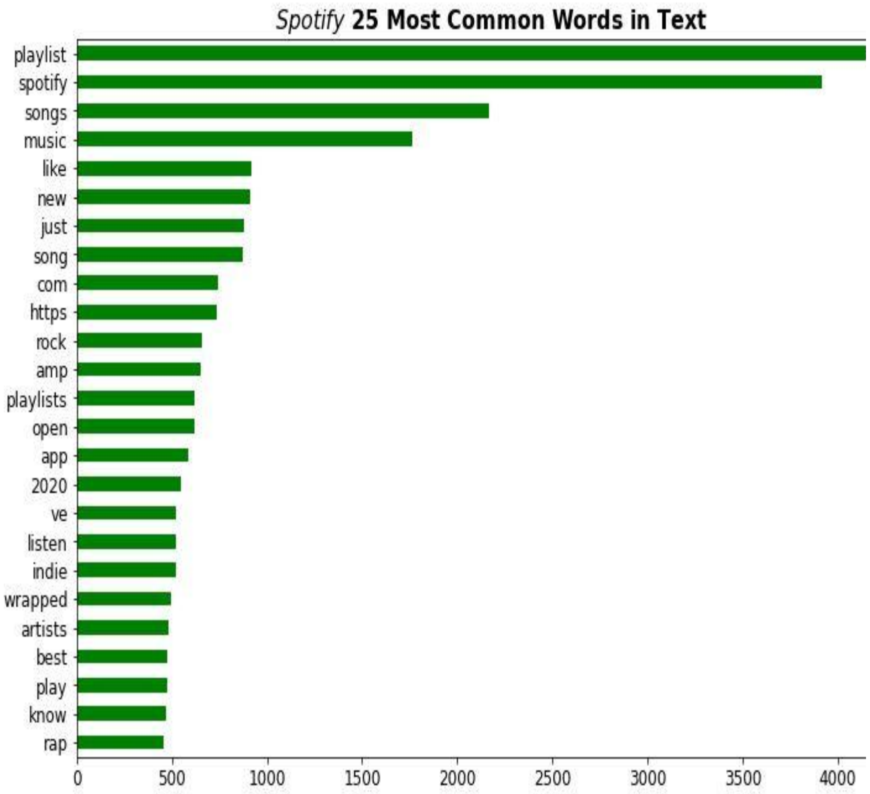
More words for Netflix

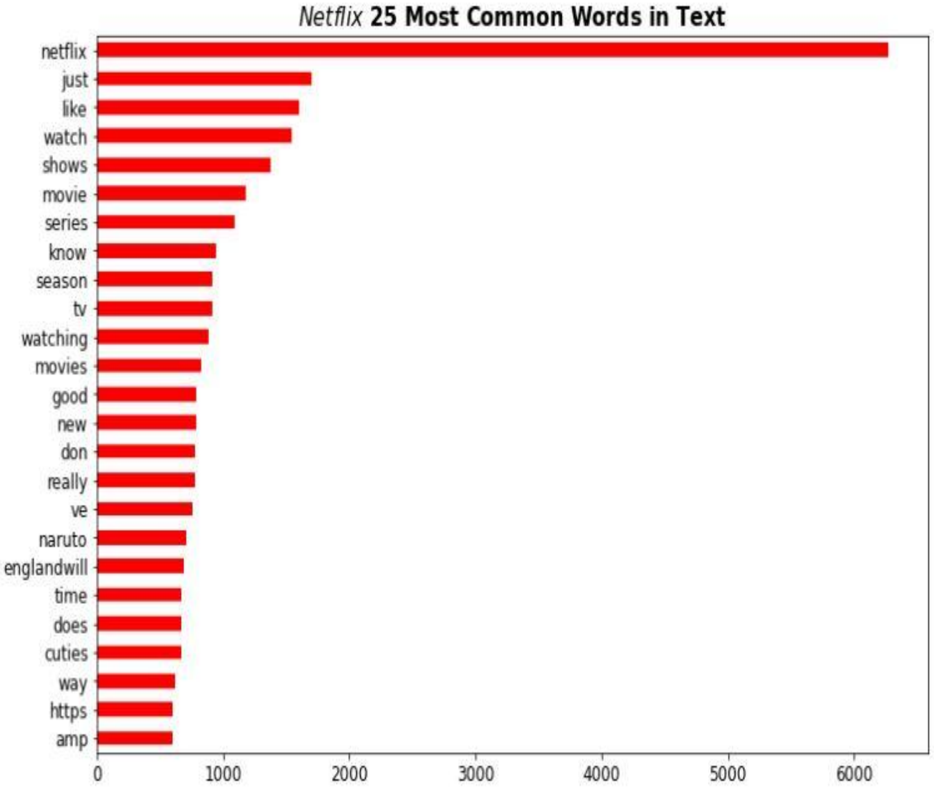


Top 10 common words



Spotify vs Netflix





Modeling

Using Count Vectorizer and TF-IDF with english stop words:

Baseline : 51 %

1. Bayes

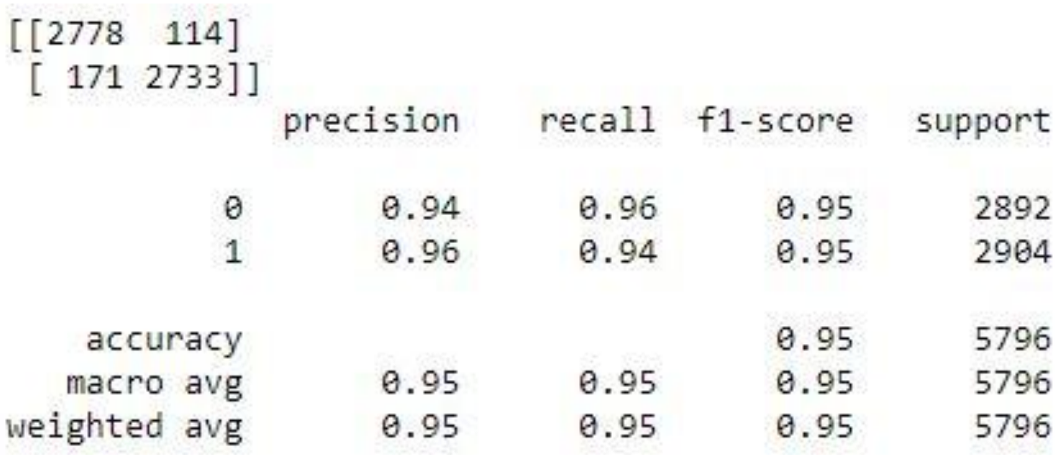
2. KNN

3. Random Forest Classifier

Bayes

Train score : 97.5 %

Test score : 95 %

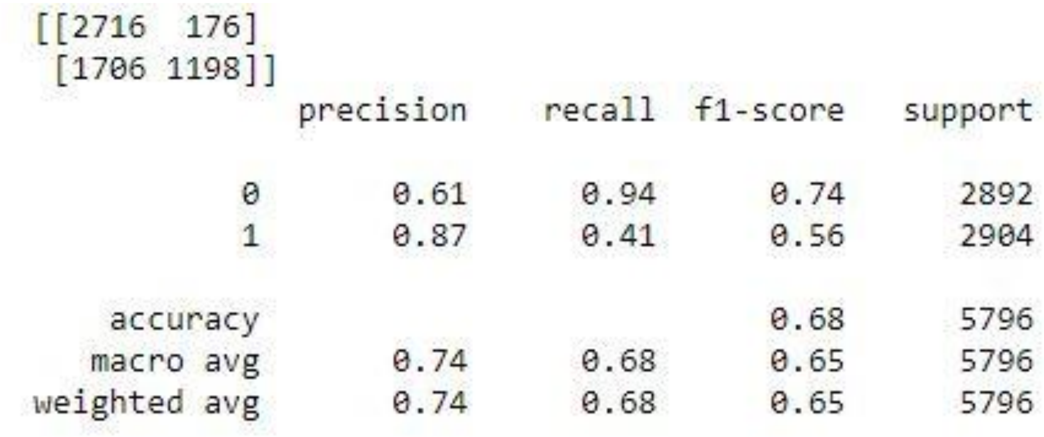


```
[[2778  114]
 [ 171 2733]]
              precision    recall  f1-score   support

           0       0.94      0.96      0.95      2892
           1       0.96      0.94      0.95      2904

    accuracy                           0.95      5796
   macro avg       0.95      0.95      0.95      5796
weighted avg       0.95      0.95      0.95      5796
```

KNN

Train score: 75%

Test score : 67.5%

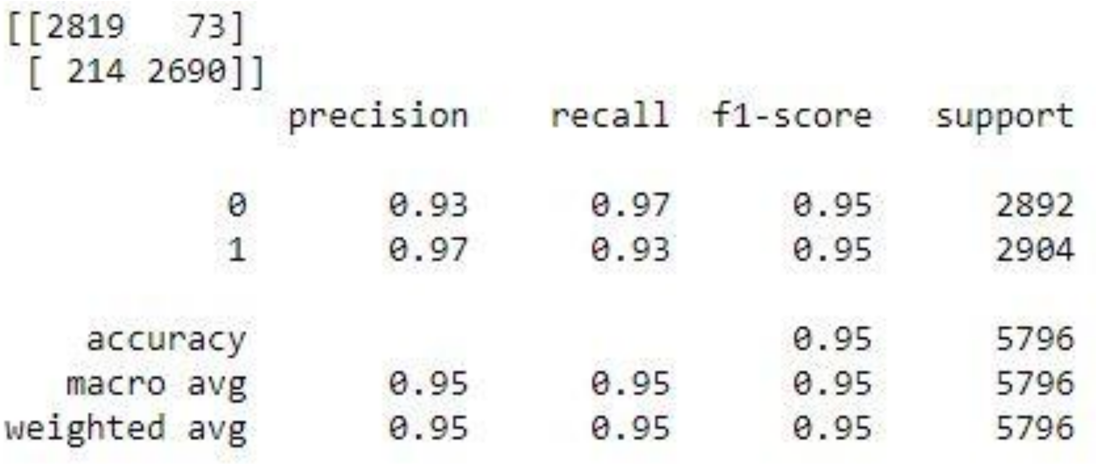


```
[[2716  176]
 [1706 1198]]
              precision    recall  f1-score   support

           0       0.61      0.94      0.74      2892
           1       0.87      0.41      0.56      2904

    accuracy                           0.68      5796
   macro avg       0.74      0.68      0.65      5796
weighted avg       0.74      0.68      0.65      5796
```

Random forest Classifier

Train score : 99.9%

Test score : 95%



```
[[2819   73]
 [ 214 2690]]
              precision    recall  f1-score   support

           0       0.93      0.97      0.95      2892
           1       0.97      0.93      0.95      2904

    accuracy                           0.95      5796
   macro avg       0.95      0.95      0.95      5796
weighted avg       0.95      0.95      0.95      5796
```

Conclutions:

Bayes and Random forest classifier model is the best to predict classification data

In this project or similar project