PROJECT 3:

WEB SCRAPING & NLP

SUBREDDIT MODELING

AGENDA

SECTION 1

DATA

- Problem Statement
- Assumptions
- Exploratory Data Analysis

SECTION 2

MODELING

- Baseline Models
- Hyperparameter Adjustments

SECTION 3

CONCLUSION & RECCOMENDATIONS

- Final Model
- Conclusion
- Questions

PROBLEM STATEMENT

Reddit is a social media platform that hosts discussion boards (called Subreddits) on various topics ranging from entertainment, business, politics, and self-help to name a few. Users are able to write posts that other users can interact with by either commenting or "up-voting" posts they like.

Streaming services, like Netflix and DisneyPlus, are subscription based websites that studios now offer to give viewers direct access to previous and upcoming films & TV shows. As of March 2021, only Sony does not have its own independent streaming service.

For this project, posts have been scrapped from the Netflix and DisneyPlus Subreddits in order to develop a classification model that will predict which Subreddit the post has been generated from and what is being discussed within these Subreddits.

ASSUMPTIONS

GENERAL KNOWLEDGE OF STREAMING

ALL POSTS ARE HUMAN GENERATED

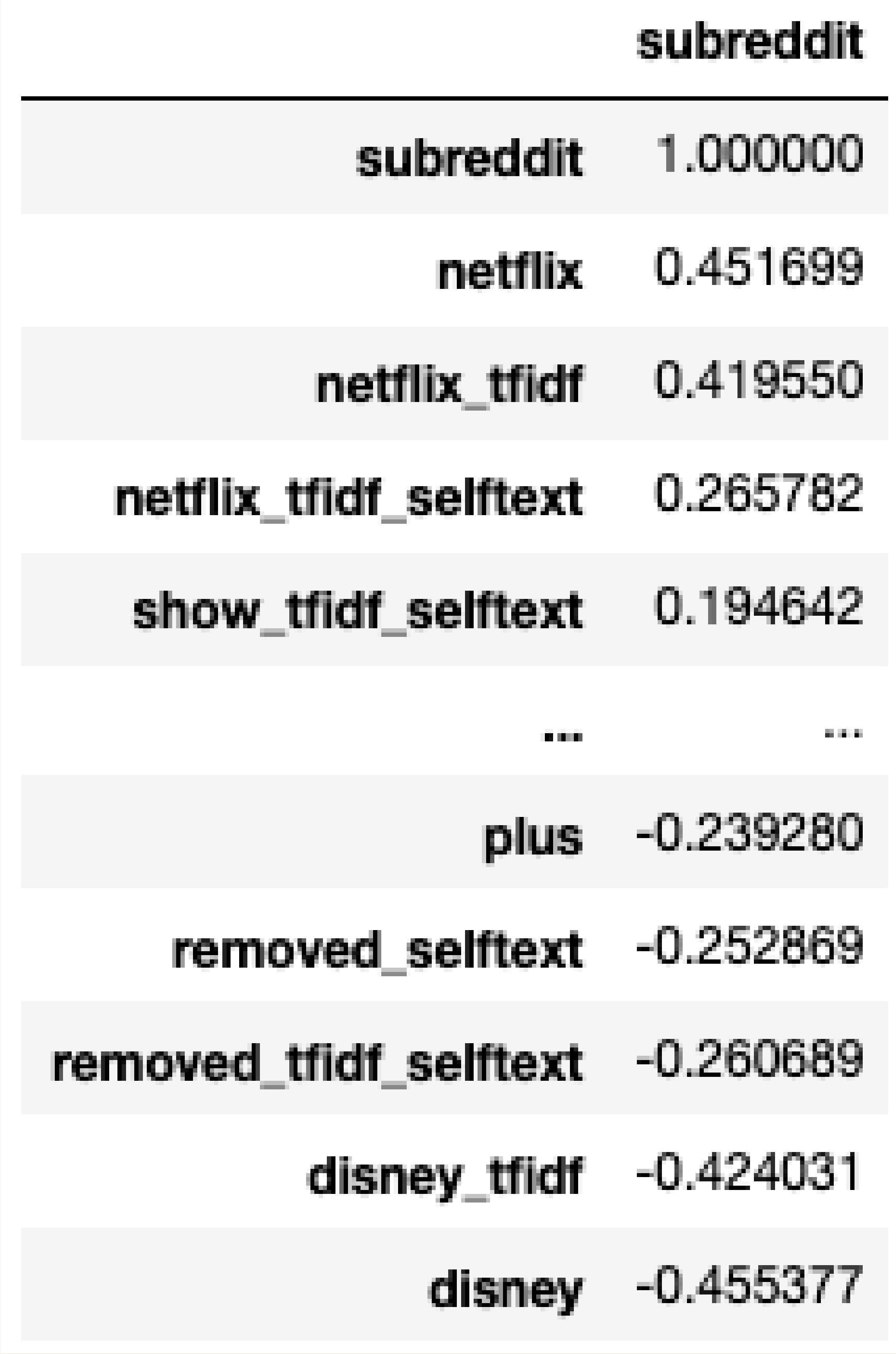
*THE WORDS 'NETFLIX' AND 'DISNEY' ARE NOT EXCLUSIVE TO THEIR RESPECTIVE SUBREDDIT.

EDA

Feature correlation: which words or other columns in the data are most helpful in predicting the appropriate Subreddit (scores closer to 1 predict Netflix, score closer to -1 predict Disney).

Word Count: what words were most prevalent and their distribution between Subreddits.

Distribution of other features by Subreddit: distribution of other data columns by Subreddit.

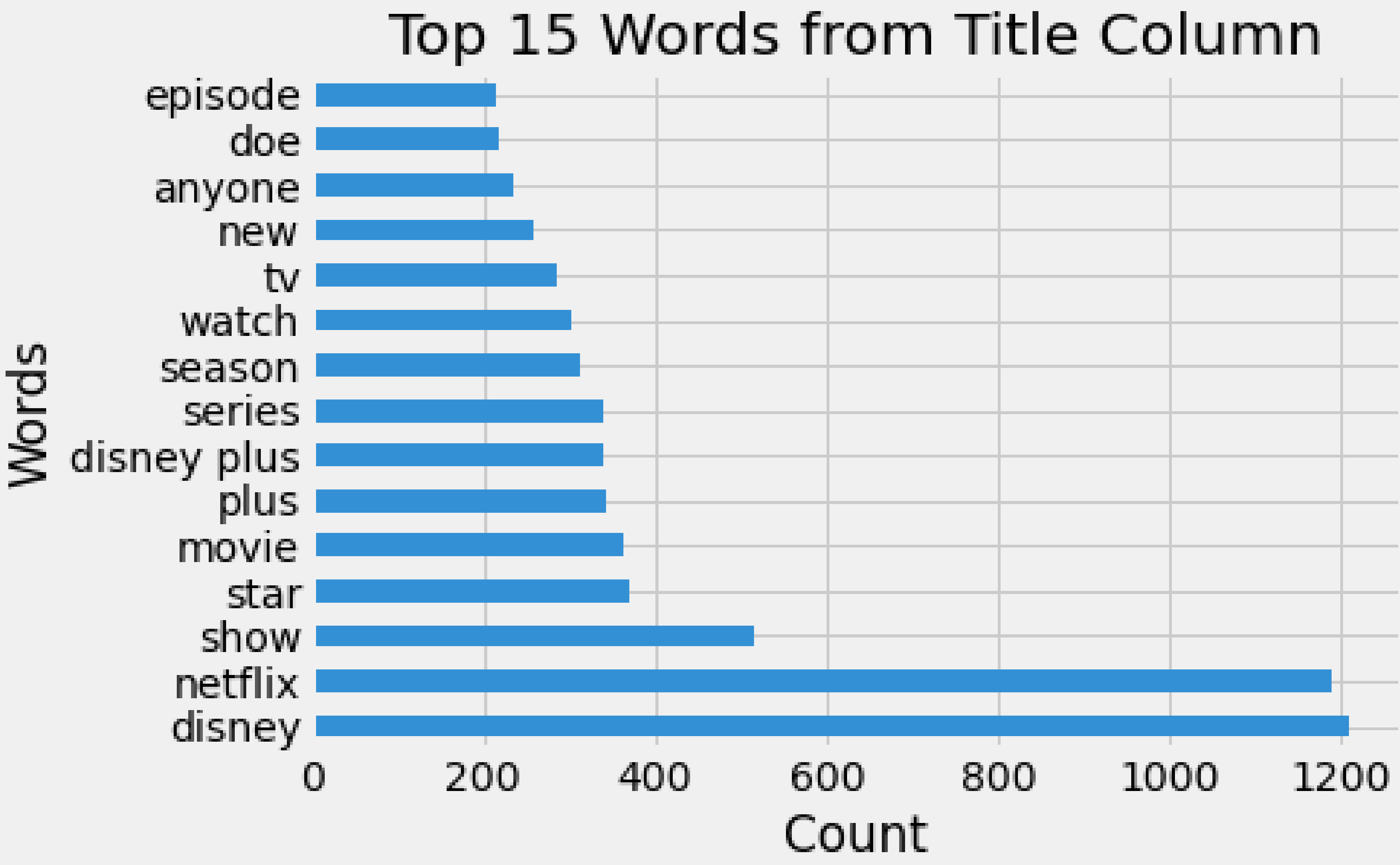


	subreddit
subreddit	1.000000
netflix	0.451699
netflix_tfidf	0.419550
netflix_tfidf_selftext	0.265782
show_tfidf_selftext	0.194642
...	...
plus	-0.239280
removed_selftext	-0.252869
removed_tfidf_selftext	-0.260689
disney_tfidf	-0.424031
disney	-0.455377

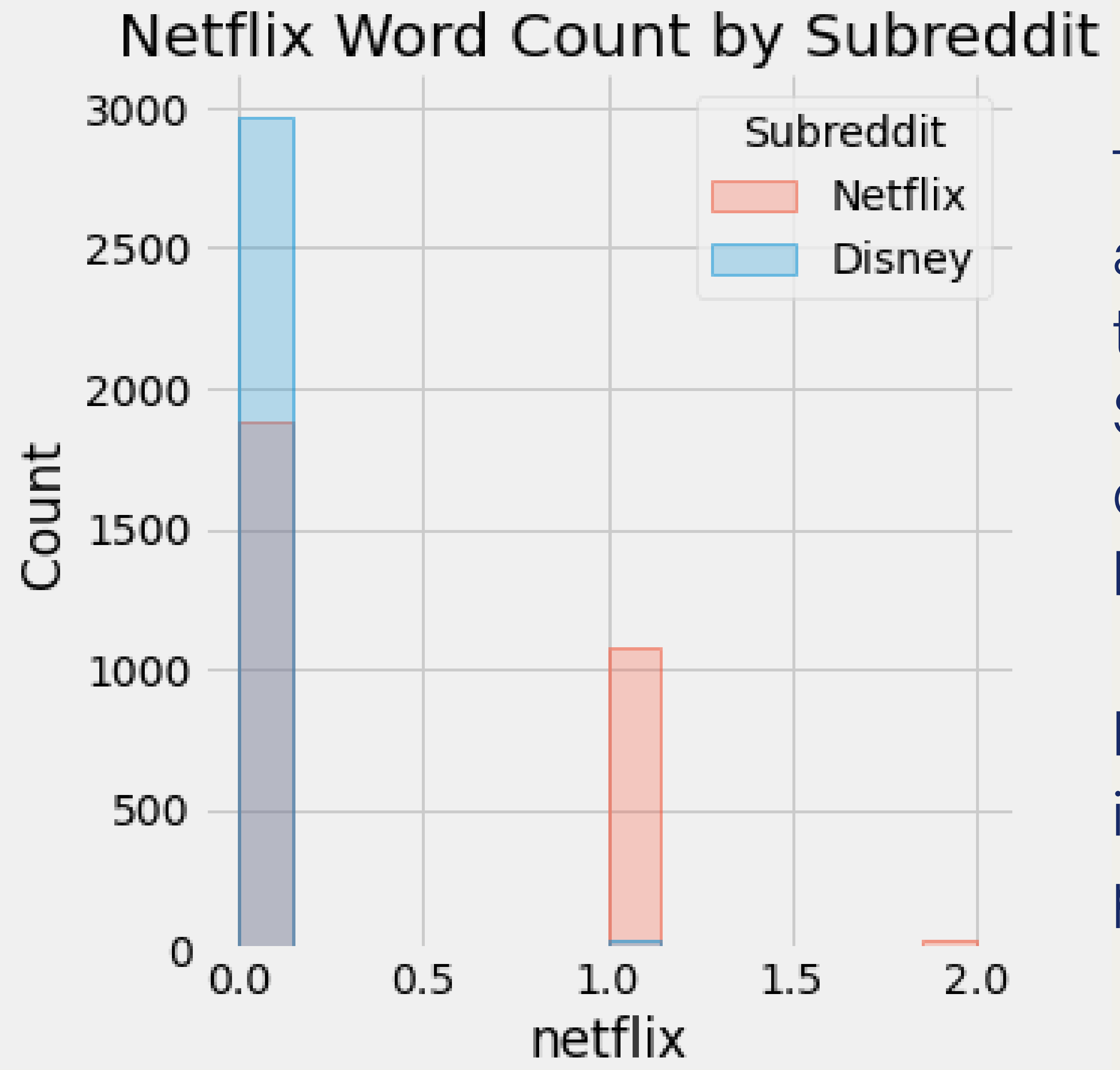
MOST FREQUENT WORDS

Disney & Netflix are the most frequent words (also had highest & lowest correlation respectively). Third assumption is disproven.

Theme: frequent discussion around watch[ing] new movies and TV series/seasons/episodes.

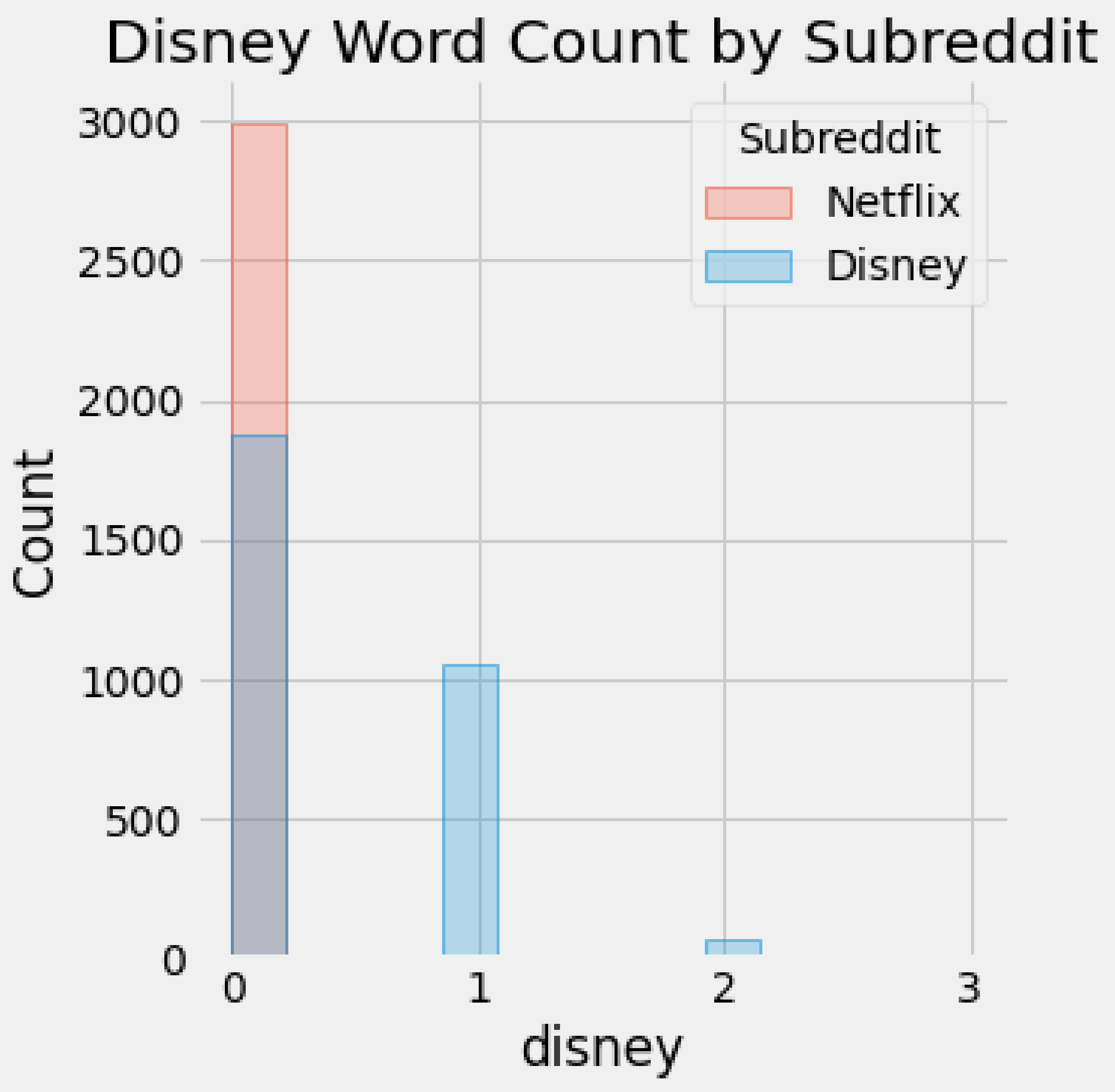


NETFLIX/DISNEY WORD COUNT BY SUBREDDIT



The word Netflix does appear in one post in the DisneyPlus Subreddit, but Disney does not appear in the Netflix Subreddit.

Neither word appears in more than half of the posts.



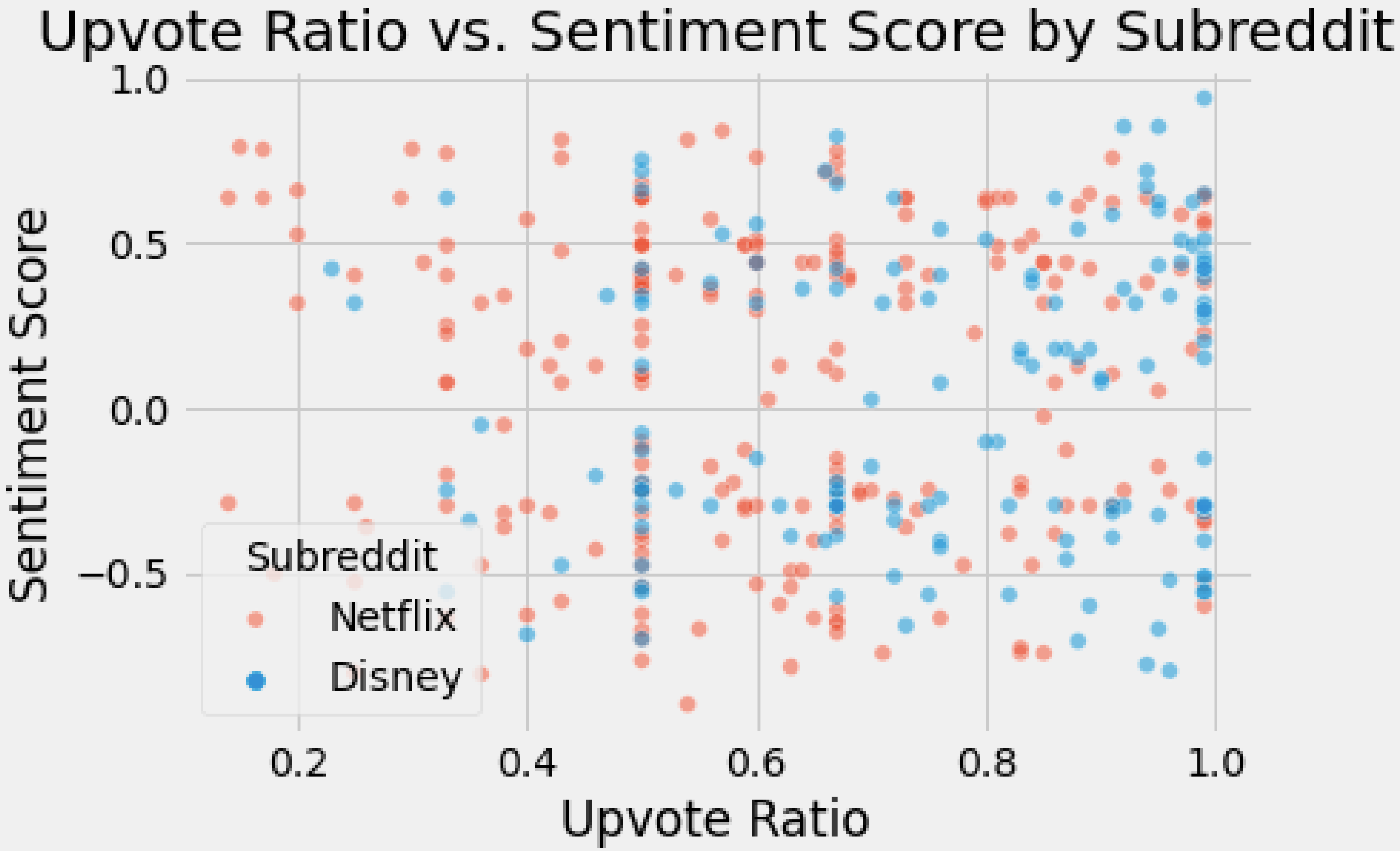
SENTIMENT SCORE / UPVOTE RATIO

The majority of sentiment scores were 0.

The majority of up-vote ratios were 1.

This graph plots them against each other, removes the majority score, and shows which Subreddit they come from.

Over 80% of the posts with a sentiment score above 0 and up-vote ratio below 0.5 were from the Netflix Subreddit

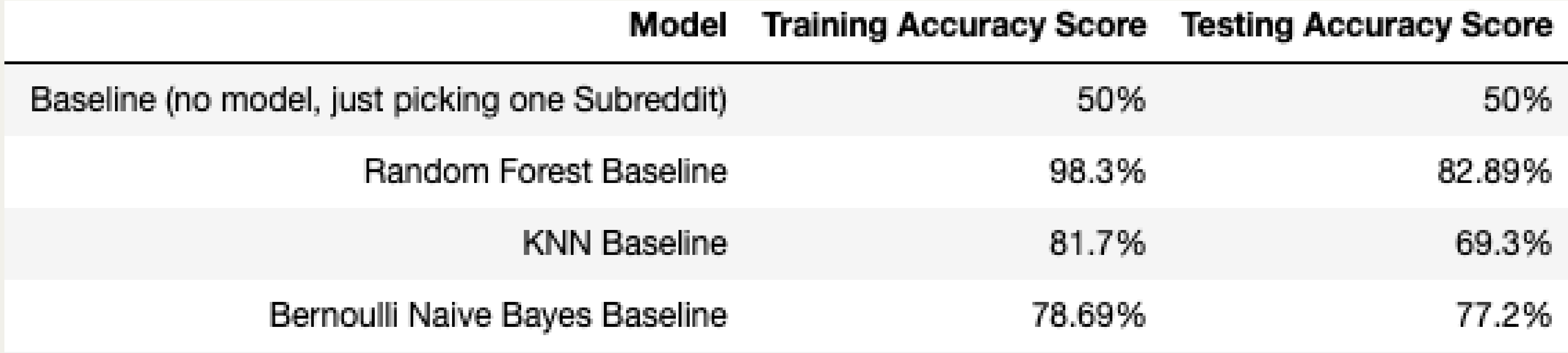


MODELS: BASELINE MODELS

Focus is on optimizing accuracy score: predict as many observations properly.

Bernoulli model showed least overfitting.

Random Forest gave the highest accuracy score on testing data. This is the model that I focused on tuning.



Model	Training Accuracy Score	Testing Accuracy Score
Baseline (no model, just picking one Subreddit)	50%	50%
Random Forest Baseline	98.3%	82.89%
KNN Baseline	81.7%	69.3%
Bernoulli Naive Bayes Baseline	78.69%	77.2%

RANDOM FOREST MODEL TUNING

Adjust Hyperparameters

Model increased accuracy score on testing data to 84.6%.

Accuracy score of training data decreased to 95.69%. Model decreased variance but is still overfit

Create New Features

Model accuracy score on testing data fell slightly to 84.1%.

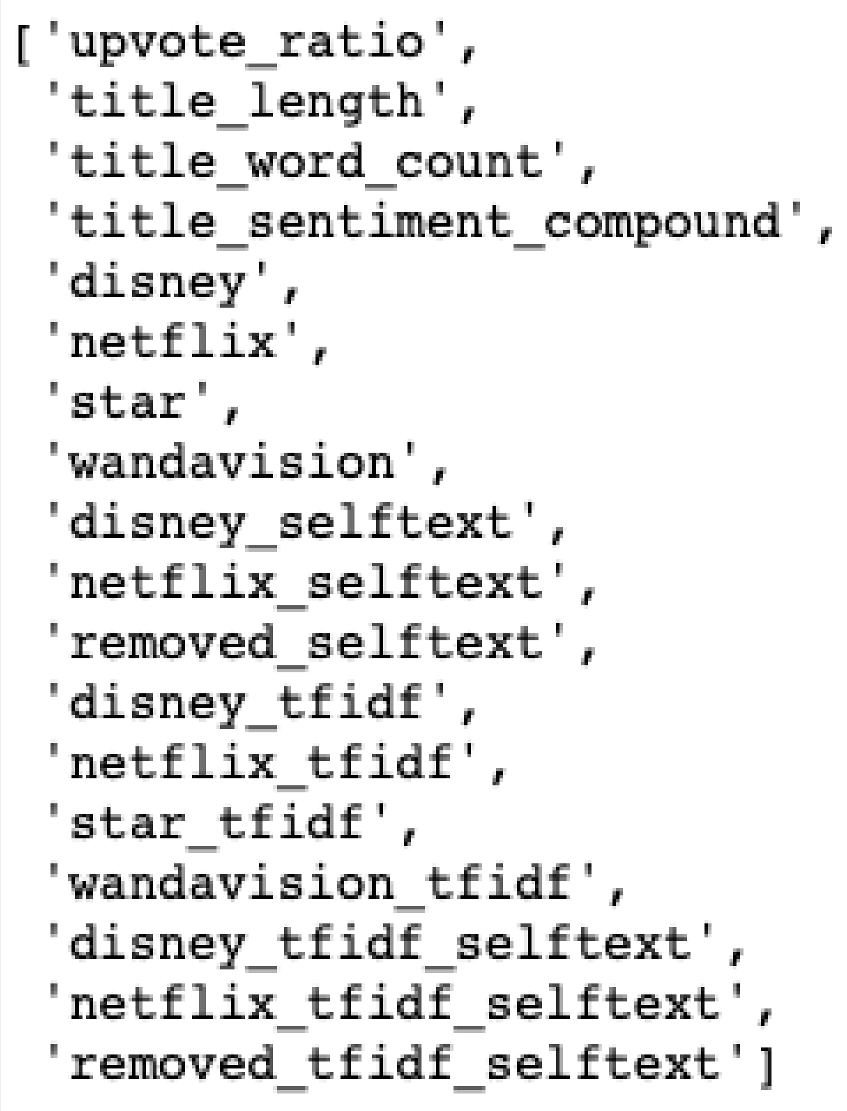
Accuracy score of training data also decline to 94.97%. Newer model decreased variance more, but is still overfit.

RANDOM FOREST MODEL TUNING

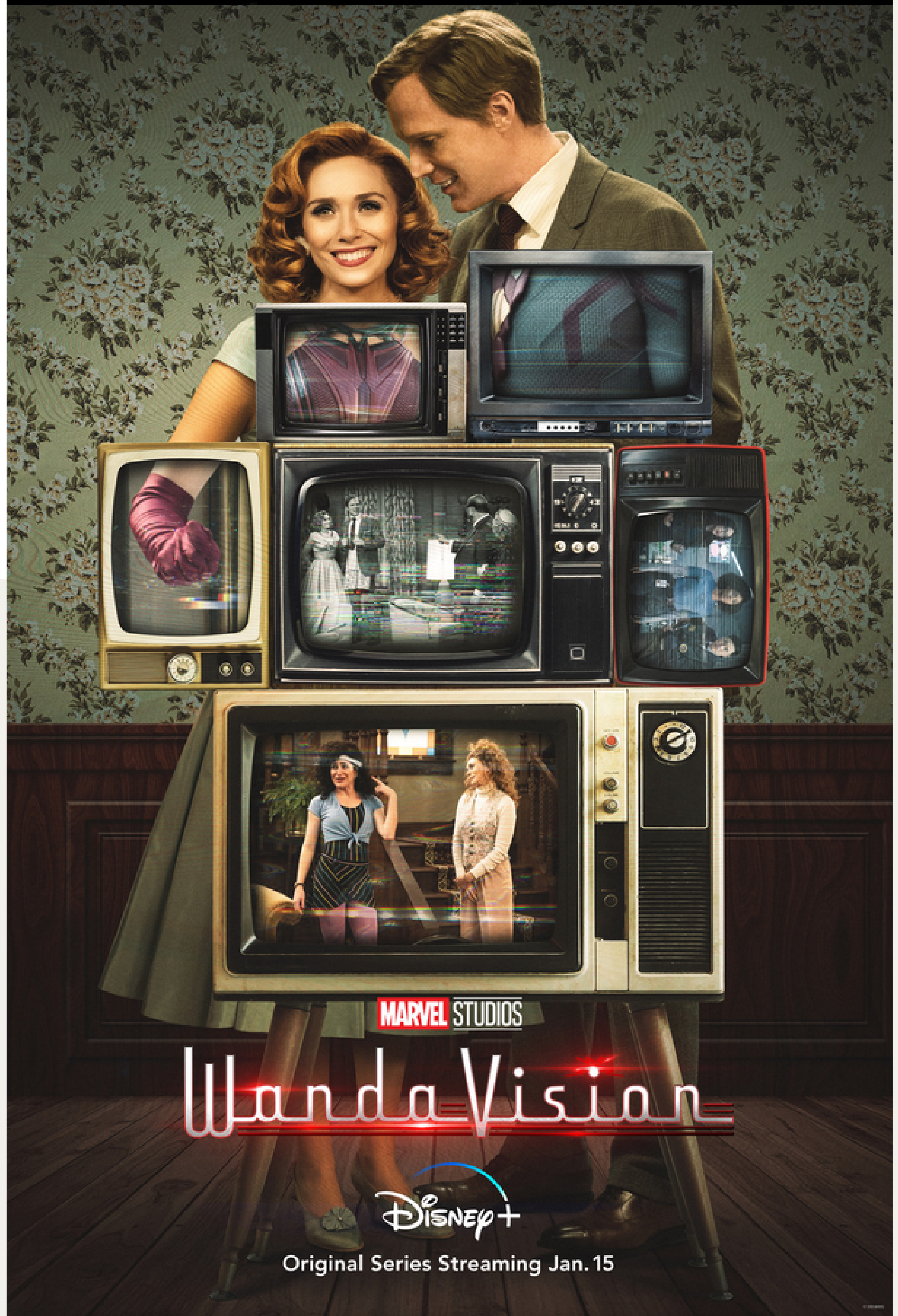
Remove Features With Low Importance

Original model had 165 features.

Used 1% as the cut-off for importance: scores closer to 100% are ideal but the feature with highest importance (title_length) had a score of 11%



```
['upvote_ratio',
 'title_length',
 'title_word_count',
 'title_sentiment_compound',
 'disney',
 'netflix',
 'star',
 'wandavision',
 'disney_selftext',
 'netflix_selftext',
 'removed_selftext',
 'disney_tfidf',
 'netflix_tfidf',
 'star_tfidf',
 'wandavision_tfidf',
 'disney_tfidf_selftext',
 'netflix_tfidf_selftext',
 'removed_tfidf_selftext']
```



RANDOM FOREST MODEL TUNING

Adjust Hyperparameters with Select Features

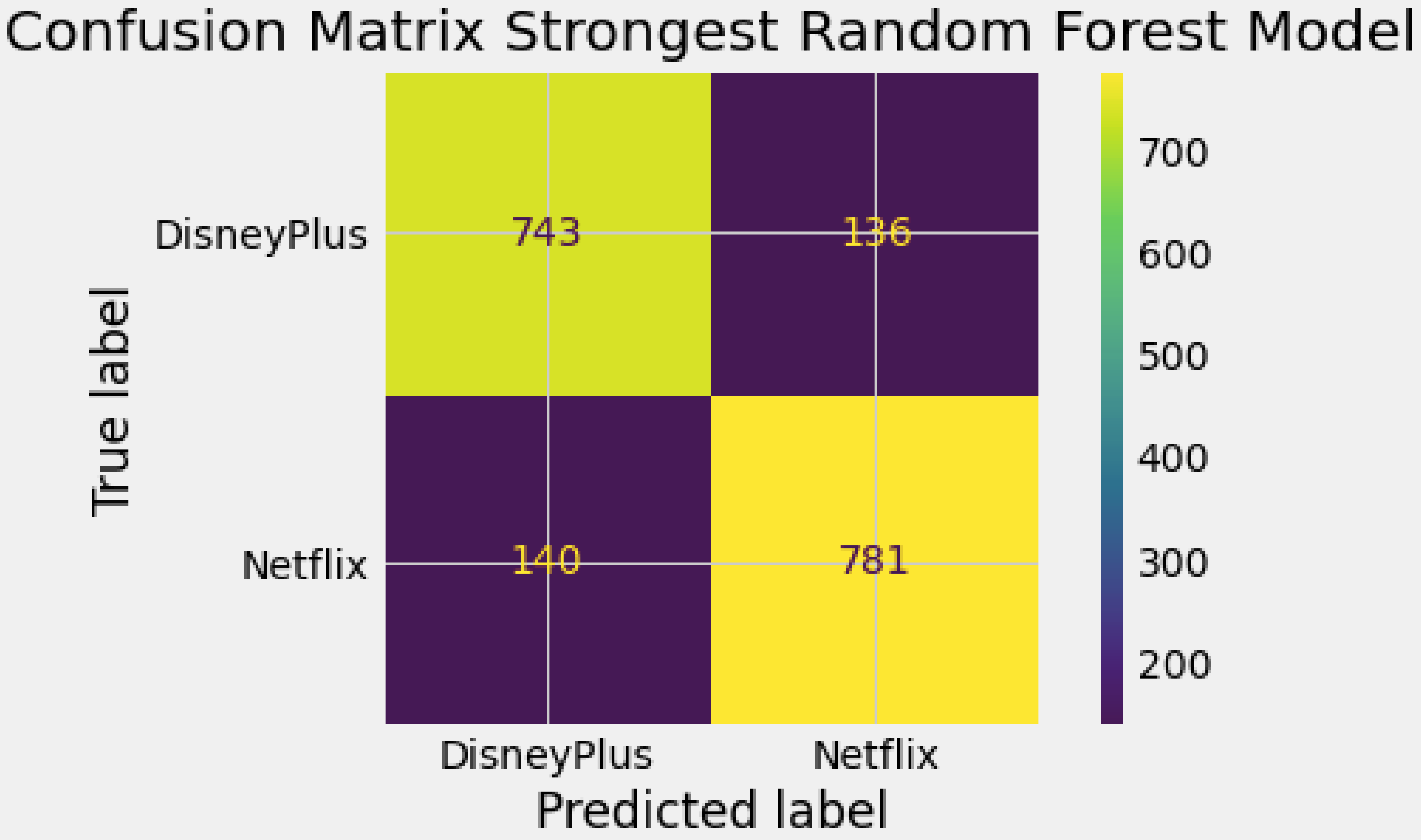
Model decreased accuracy score on testing data to 81.1%.

Accuracy score of training data decreased to 94.4%. Model decreased variance but is still overfit

Create New Features with Select Features

Model accuracy score on testing data increased slightly to 81.5%.

Accuracy score of training data increased to 94.6%. Newest model increased variance and is still overfit.



STRONGEST RANDOM FOREST MODEL

Confusion matrix to the left shows all of the predictions in the Random Forest model that adjusted hyperparameters (nothing else) and if they were correct predictions or incorrect predictions.

RECOMMENDATIONS

Random Forest Model With Hyperparameter Adjustments Was The Strongest Model

But still overfit!

Pull More Words or Use other NLP Processing Techniques

Use techniques like BERT or filter in other Sentiment Scores (only used compounded score)

Adjust Bernoulli Naive Bayes Model

Could return a higher accuracy score with less variance than the final Random Forest Model.

CONCLUSION

Themes discussed on both Netflix and DisneyPlus Subreddit revolves around watching the newest movies or TV shows.

All models created were more successful in predicting the correct Subreddit than the baseline of no model. The Random Forest Model was the strongest model in terms of accuracy score.

QUESTIONS?