Group 3 Project 3

Jance Ng
Jin Min Wood
Adrian Kong

Problem Statement

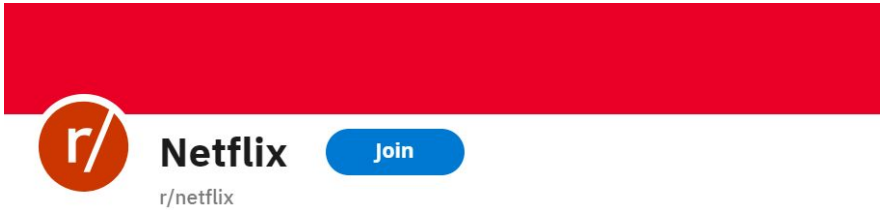
We are part of the marketing team for a tv/movie production company. Our company is interested to sign a contract with either Netflix or Disney+ to stream our shows on their platform. Before we make our decisions, we would like to see which platform suits our tv/movie genre the best.

To get the sentiments of the users of the platform, we dive into their subreddit community to extract what the community have to say about these platforms.

We will make use of natural language processing and classification models to classify reddit posts into either netflix or disneyplus. This will allow us to make use of the review of our tv/movies and run it through the classification model to see which platform is a better fit for our tv/movie.

Scope of project

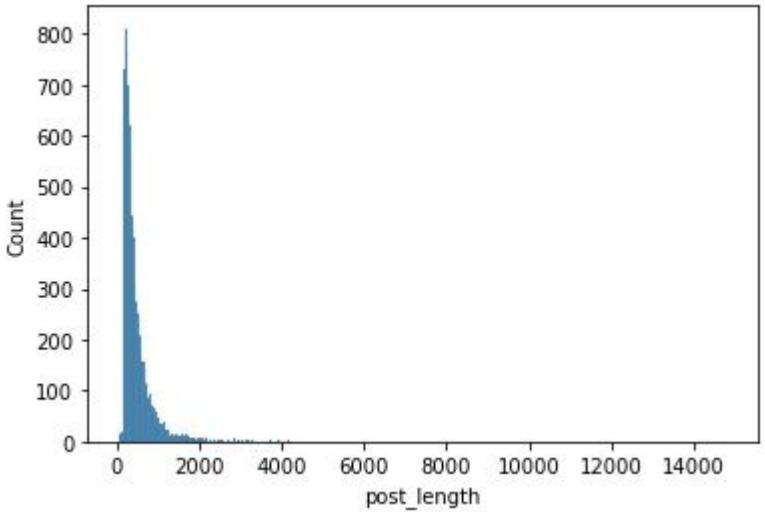
Subreddits:



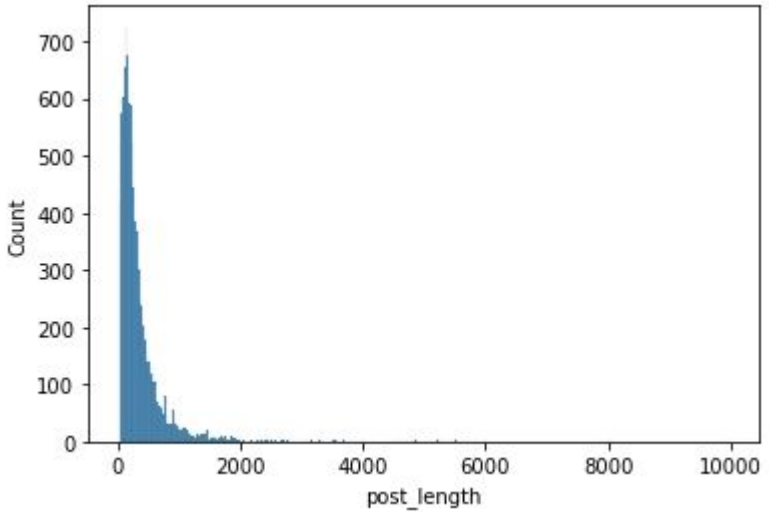
Length of post: at least 30 words

No. of unique posts: 9,800 each

Exploratory Data Analysis

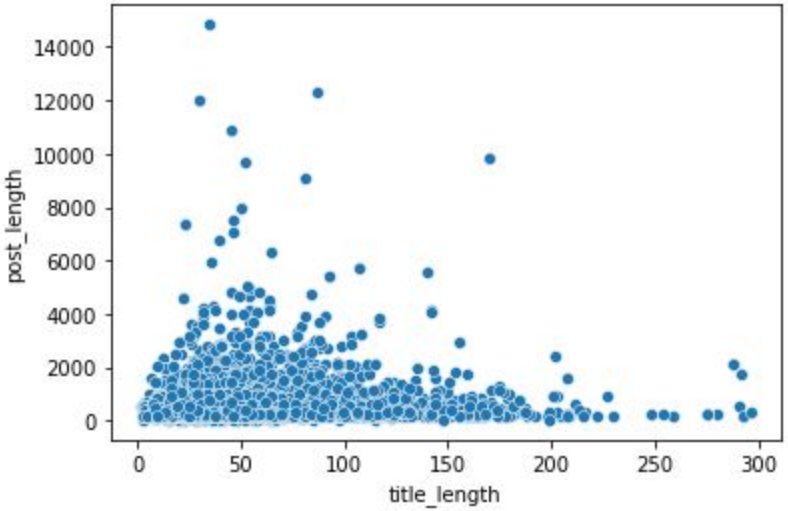


Netflix

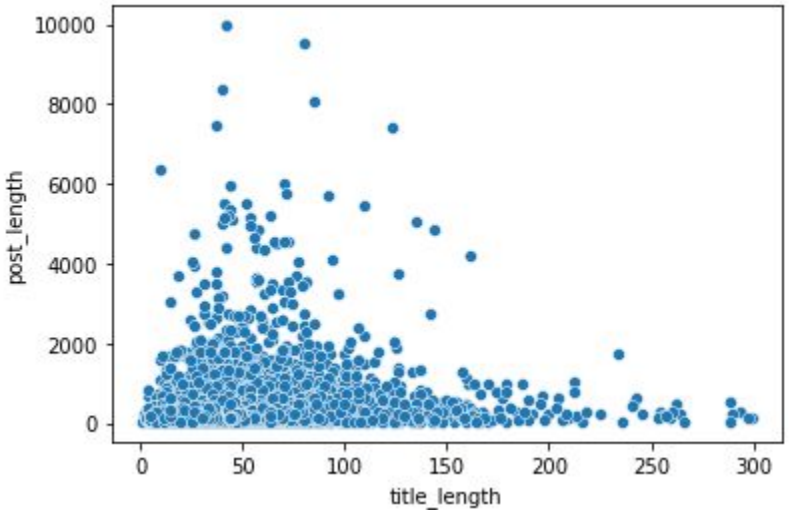


DisneyPlus

Exploratory Data Analysis



Netflix



DisneyPlus

Modelling

Baseline model:

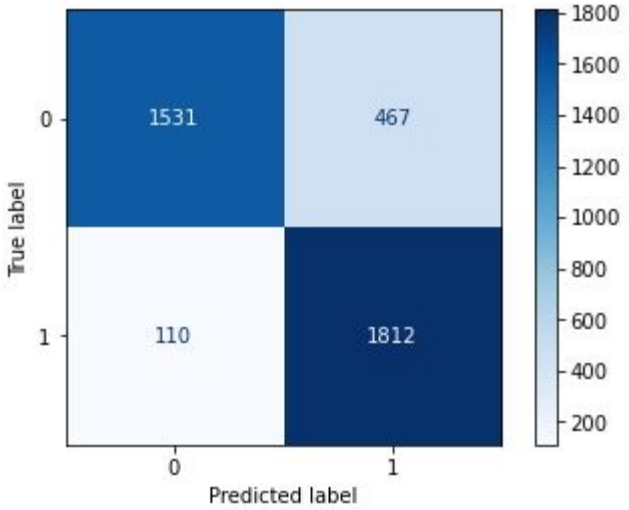
- 50% accuracy since equal number of posts from each subreddit

Models we will be using to try to beat the baseline score:

- Classification Algorithm - Random Forest Algorithm and Logistic Regression
- Vectorizer - Countvectorizer and TF-IDF vectorizer
- Word form - Tokenized, Lemmatized and Stemming

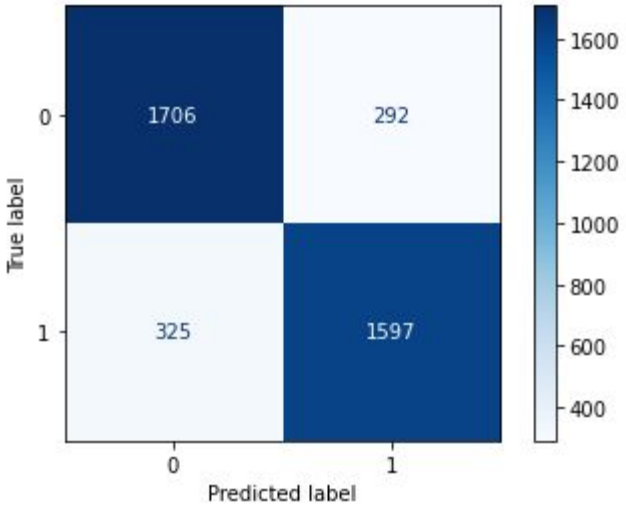
Model 1ai

Classification Algorithm	Random Forest
Vectorizer	TF-IDF
Word Form	Tokenized
Train set accuracy score	0.86096
Test set accuracy score	0.85280
F1 score	0.86265



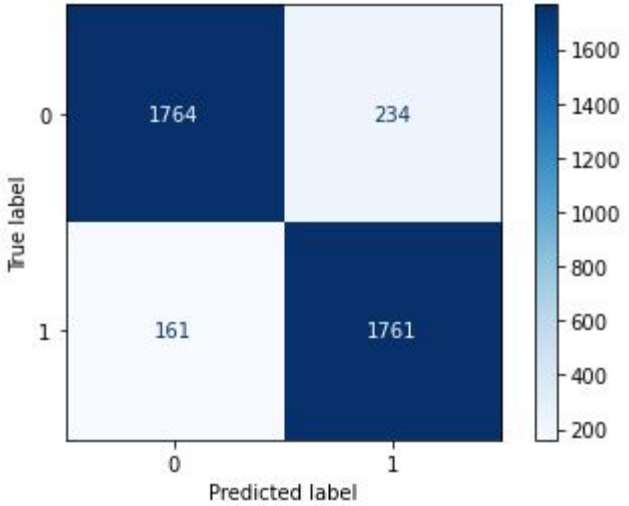
Model 1aii

Classification Algorithm	Random Forest
Vectorizer	Countvectorizer
Word Form	Tokenized
Train set accuracy score	0.85771
Test set accuracy score	0.84260
F1 score	0.83810



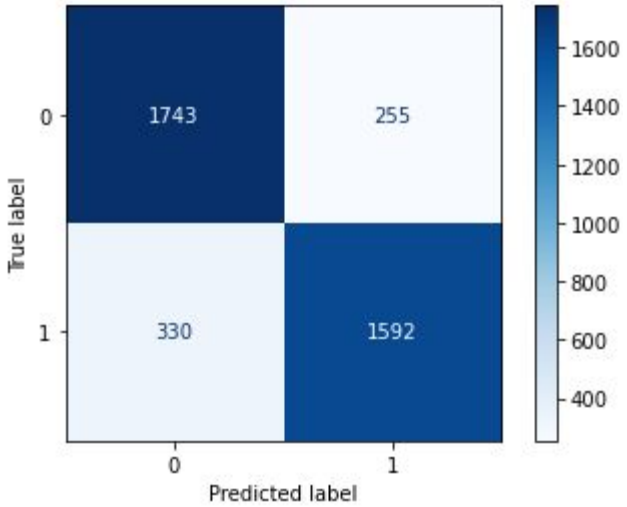
Model 1bi

Classification Algorithm	Logistic Regression
Vectorizer	TF-IDF
Word Form	Tokenized
Train set accuracy score	0.94304
Test set accuracy score	0.89923
F1 score	0.89915



Model 1bii

Classification Algorithm	Logistic Regression
Vectorizer	Countvectorizer
Word Form	Tokenized
Train set accuracy score	0.86128
Test set accuracy score	0.85076
F1 score	0.84478



Evaluation

In addition to the training and testing accuracy score (to look for overfitted models), we will be looking at the F1 score as well

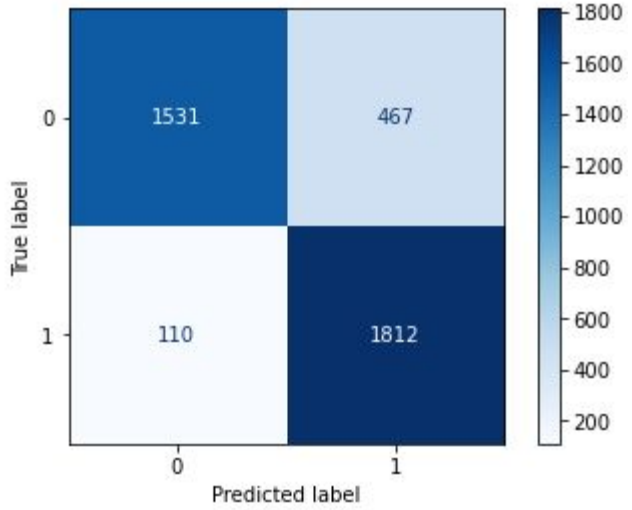
- We are a neutral company, and we seek to minimise both
  - false negatives (classified as disneyplus even though they should be classified as netflix) and
  - false positives (classified as netflix even though they should be classified as disneyplus)

We seek to optimise the F1 score, which will give us the harmonic mean of precision and recall scores.

Evaluation

Model	Training Score	Testing Score	F1 Score
1bi	0.943048	0.899235	0.899158
2bi	0.943048	0.899235	0.899158
3bi	0.943048	0.899235	0.899158
1ai	0.860969	0.852806	0.862652
2ai	0.867666	0.854082	0.861902
3ai	0.867347	0.848214	0.846926
1bii	0.861288	0.850765	0.844786
2bii	0.861288	0.850765	0.844786
3bii	0.861288	0.850765	0.844786
2aii	0.856122	0.843367	0.841262
3aii	0.858801	0.840816	0.840164
1aii	0.857717	0.842602	0.838100

Classification Algorithm	Random Forest
Vectorizer	TF-IDF
Word Form	Tokenized
Train set accuracy score	0.86096
Test set accuracy score	0.85280
F1 score	0.86265



Evaluation

Top 1 - 10 words

	importance
disney	0.052049
netflix	0.047973
plus	0.014836
and	0.007692
the	0.007016
show	0.006789
of	0.006607
to	0.005878
that	0.005848
it	0.005758

Top 11 - 20 words

shows	0.005208
this	0.004976
on	0.004863
but	0.004771
is	0.004445
for	0.004361
watched	0.004151
anyone	0.004115
in	0.004023
was	0.003997

From the top 20 words, other than the top 3 words: 'disney', 'netflix' and 'plus', which probably relates to disneyplus, netflix and disneyplus respectively, the remaining words seem to be quite generic.

The top 2 words is also ~4-5x more important than the 3rd/4th words.

This tells us that if our movie reviews do not have 'disney' or 'netflix' in them (which is highly likely since our shows are not on either platforms yet), our model may possibly have a hard time classifying the reviews. This would possibly impact the accuracy of our model.

Recommendations

Improvement to the model could be made by removing words with high correlation to the subreddits such as 'disney', 'netflix', 'plus'.