By Makena Jones
January 18, 2022

Harry vs. Frodo?:
When Fandoms Collide

● ● ●

Distinguishing posts from the subreddits of two of the largest fantasy fandoms, Harry Potter and Lord of the Rings

A General Assemby - DSI 1128 Project

What classification model best distinguishes if a post is from r/lotr or r/harrypotter?

What about when the post contains no character names or words associated with the respective lores?
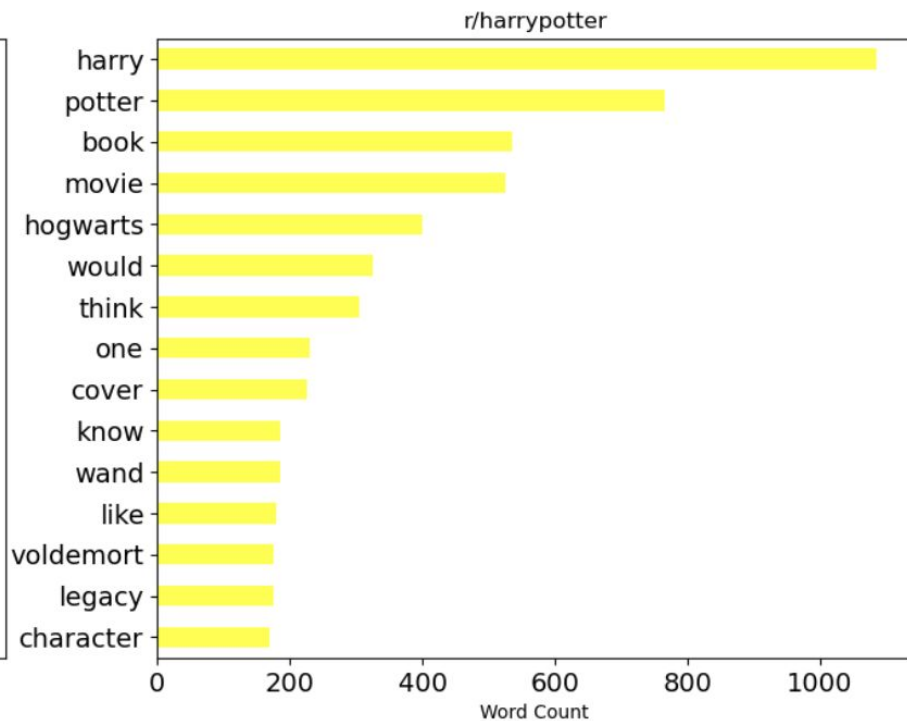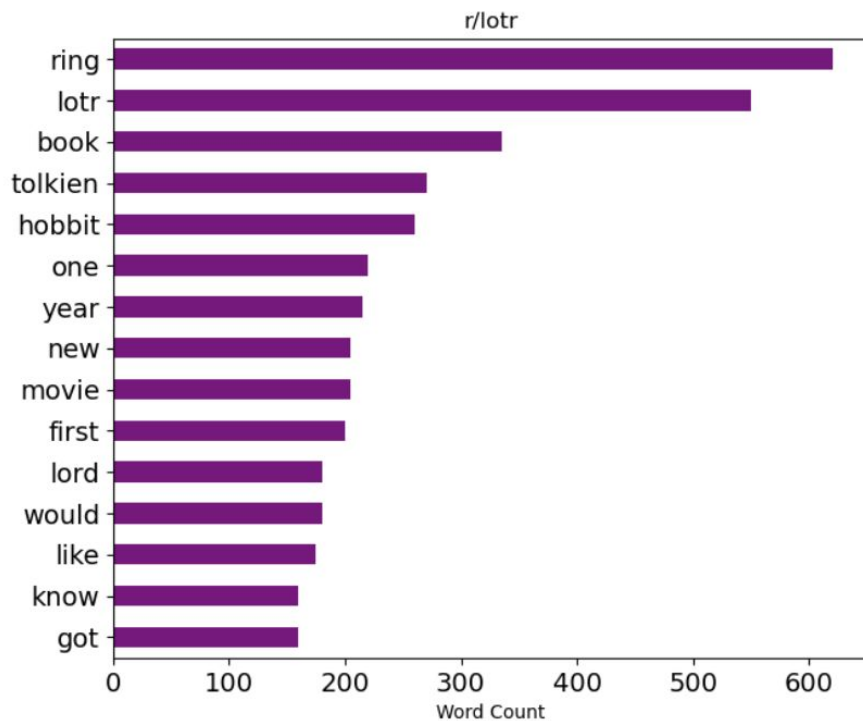

Source:
link

OVERVIEW

- Why would we want to create this model?

- Methods for testing and evaluating different models.

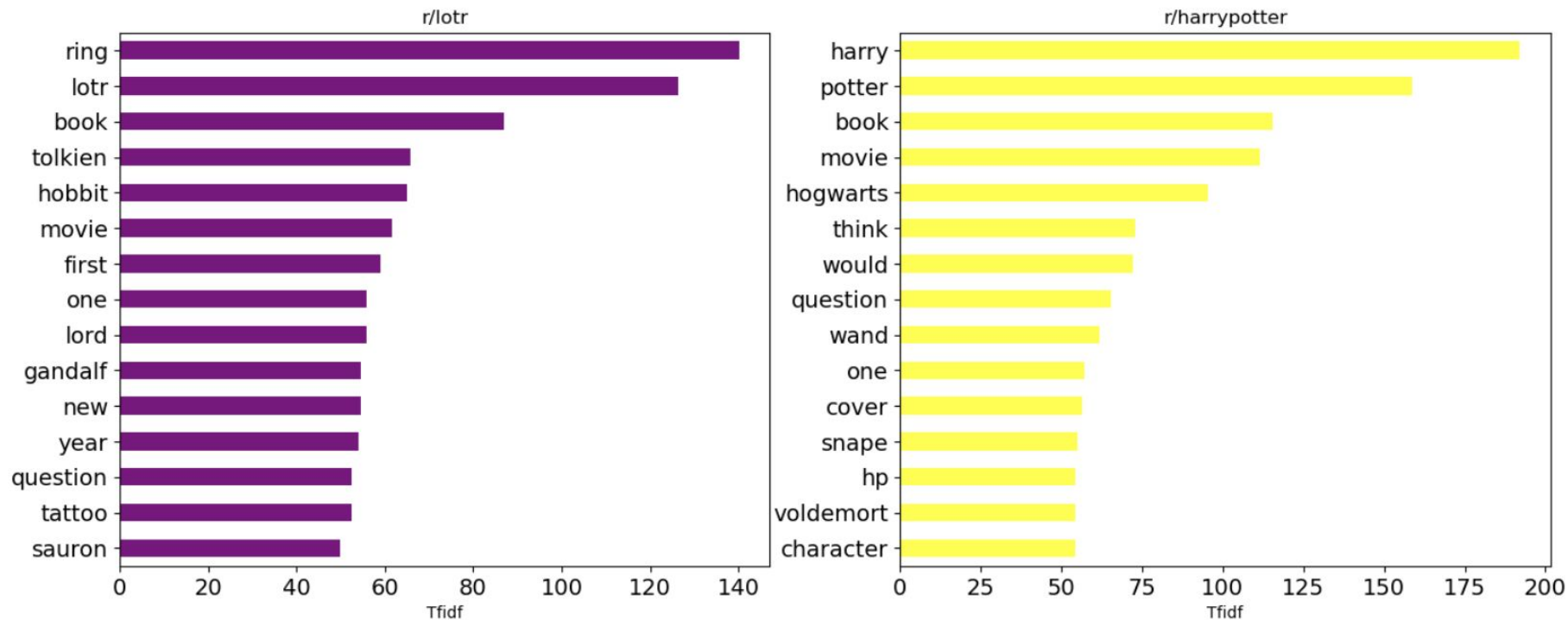- Best performing model.

- Going forward.

Data Exploration

LOTR and HP Subreddit Word Count in "Title" of Posts

r/lotr

Word	Word Count
ring	~620
lotr	~550
book	~340
tolkien	~270
hobbit	~260
one	~220
year	~215
new	~205
movie	~205
first	~200
lord	~180
would	~180
like	~175
know	~160
got	~160

r/harrypotter

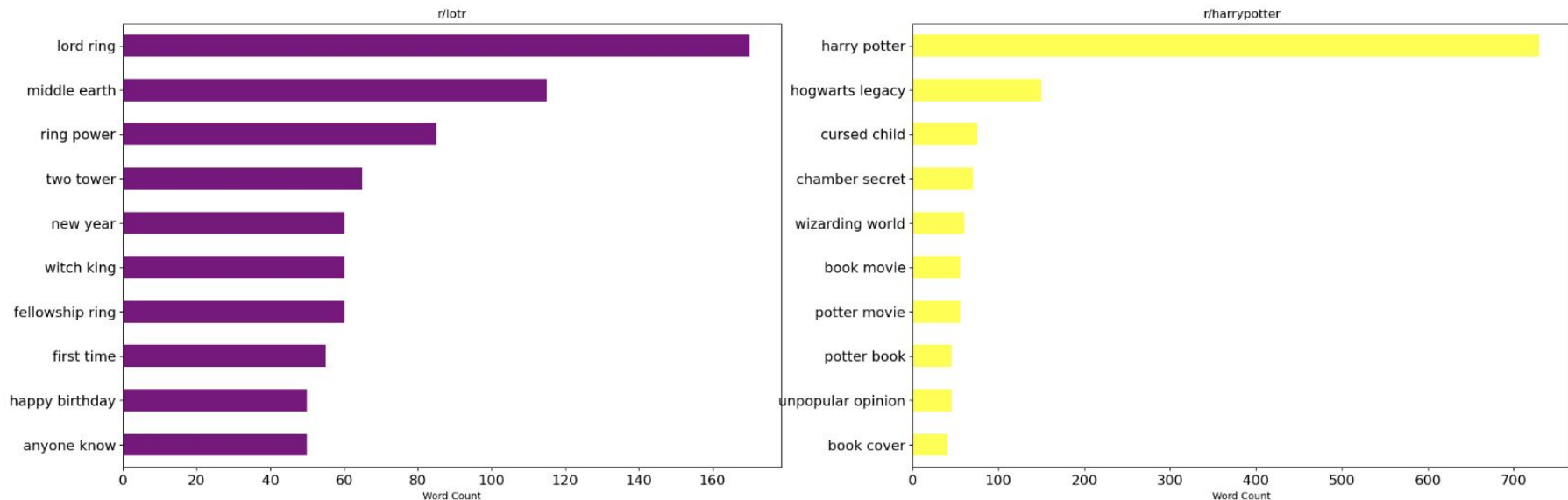| Word	Word Count
harry	~1090
potter	~770
book	~540
movie	~525
hogwarts	~400
would	~320
think	~305
one	~230
cover	~225
know	~185
wand	~185
like	~180
voldemort	~180
legacy	~180
character	~175

Data Exploration

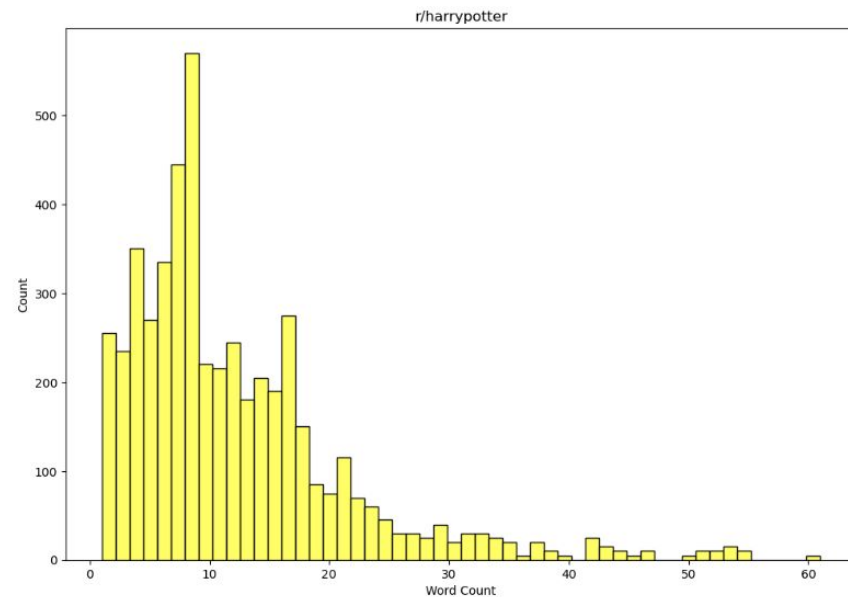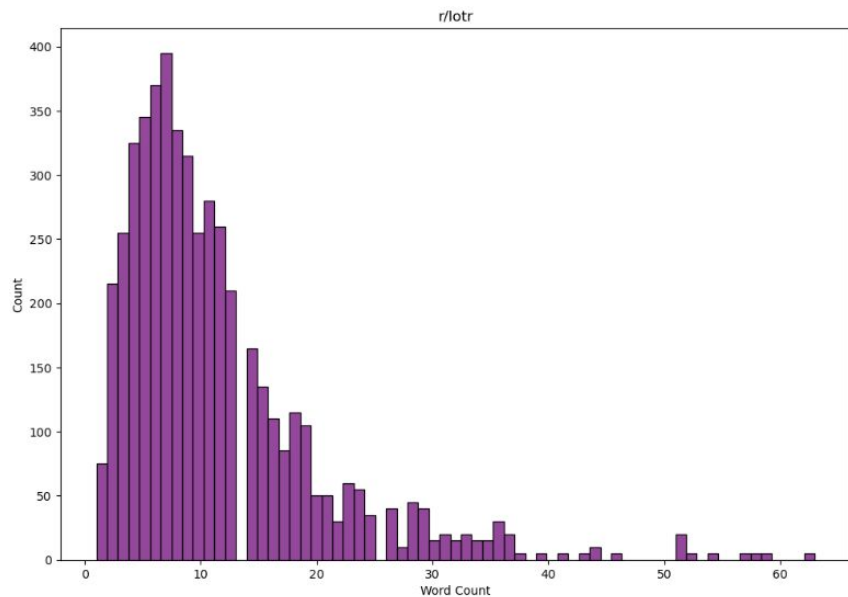LOTR and HP Subreddit Tfidf in "Title" of Posts

Data Exploration

LOTR and HP Subreddit Bigram Count in "Title" of Posts
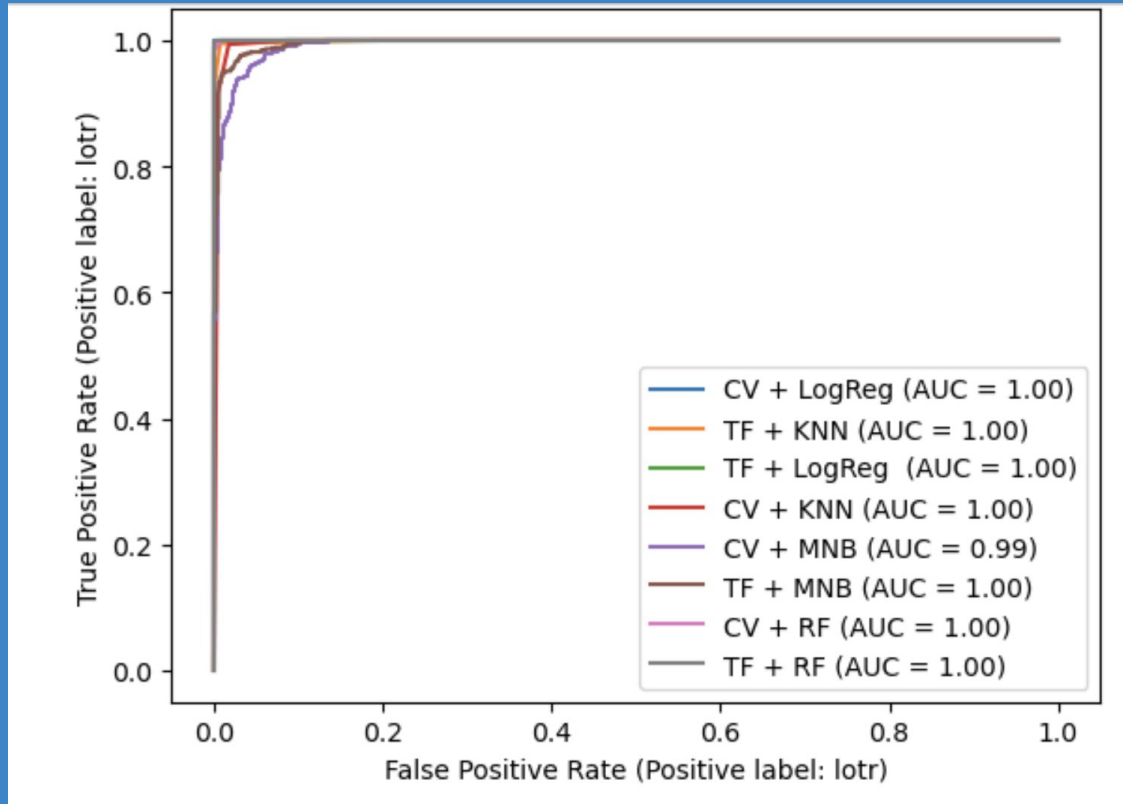
Data Exploration

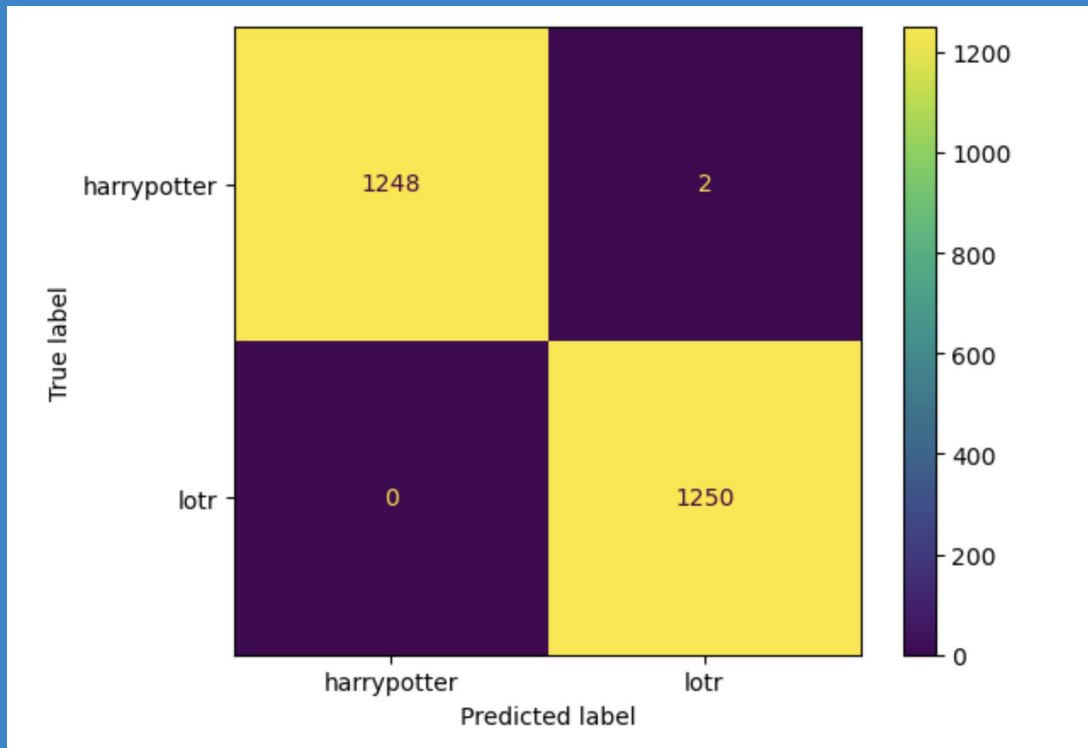Word Count Distributions of Subreddit Posts

Model Testing

L

Legend:
- CV + LogReg (AUC = 1.00)
- TF + KNN (AUC = 1.00)
- TF + LogReg (AUC = 1.00)
- CV + KNN (AUC = 1.00)
- CV + MNB (AUC = 0.99)
- TF + MNB (AUC = 1.00)
- CV + RF (AUC = 1.00)
- TF + RF (AUC = 1.00)

Model Testing

- Best performing model = Count Vectorizer + Logistic Regression

Findings and Conclusions

- Very easy to distinguish the two subreddits using NLP.

- Going forward with second part of problem statement.

- Applications for use.

Thank you :)