Subreddit Classification

r/harrypotter
r/lotr

Boss Sarawongsuth

PROBLEM STATEMENT

Reddit is a massive global online community forum. These communities are divided into subreddits with each being specific to a certain subject. Each subreddit usually has many moderators and admins moderating the posts to ensure that they meet the subreddit rules. With an ever-increasing userbase, the process of moderating these posts are becoming more cumbersome requiring greater amount of manpower to ensure the posts are relevant.

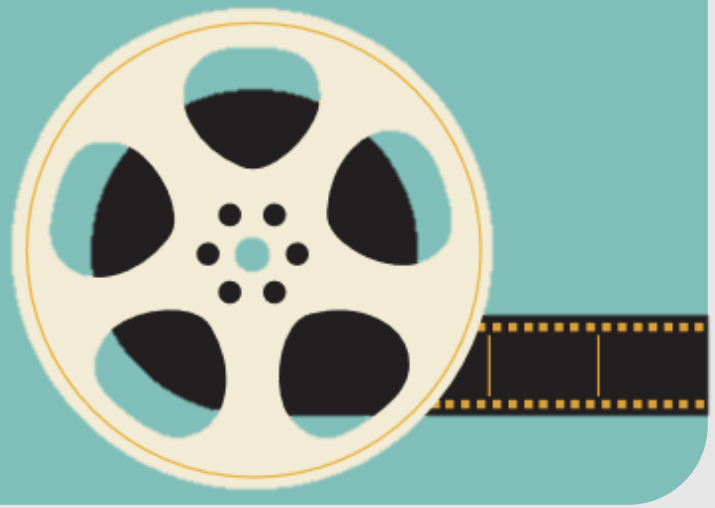
This project aims to show that the possibility of using classification models to classify whether a post belongs to a certain subreddit by primarily looking at the title and the body content.

- r/harrypotter
- r/lotr

SUBREDDITS

HARRY POTTER
r/harrypotter
993k subscribers
57%

LORD OF THE RINGS
r/lotr
523k subscribers
43%



DATASETS

The datasets for r/harrypotter and r/lotr were obtained from Reddit API.

1723 rows
&
19 columns

DATA TYPES

6
STRING
Features

---

title
selftext
subreddit
...

3
BOOL
Features

---

spoiler
author_premium
is_video

3
FLOAT
Features

---

upvote_ratio
num_reports
...

7
INT
Features

---

ups
downs
num_comments
...

EDA

Preprocessing

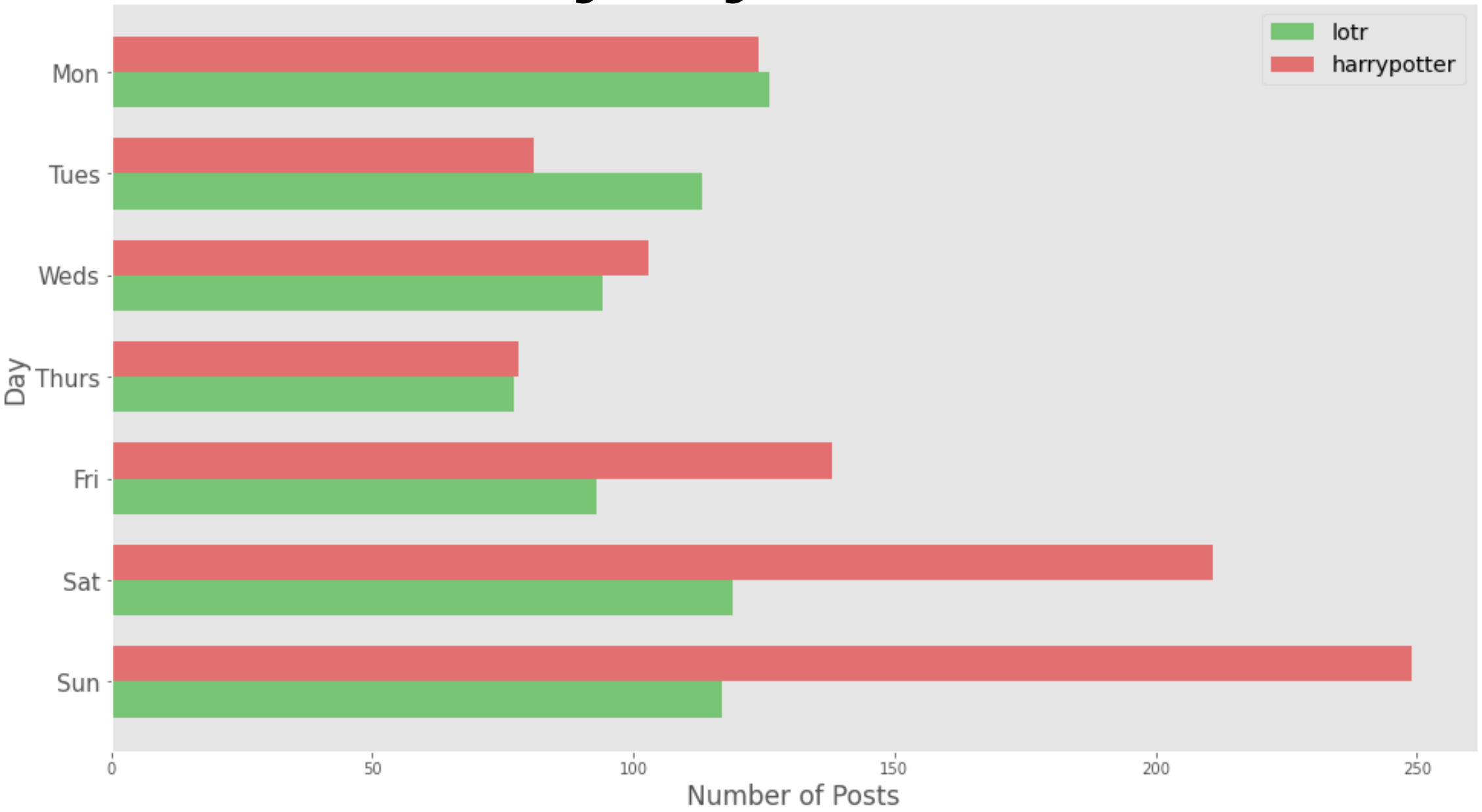
1.
FILL MISSING
VALUES

2.
Concat
Title
&
Selftext

3.
Remove
Stopwords,
Non-alphabets,
subreddit names

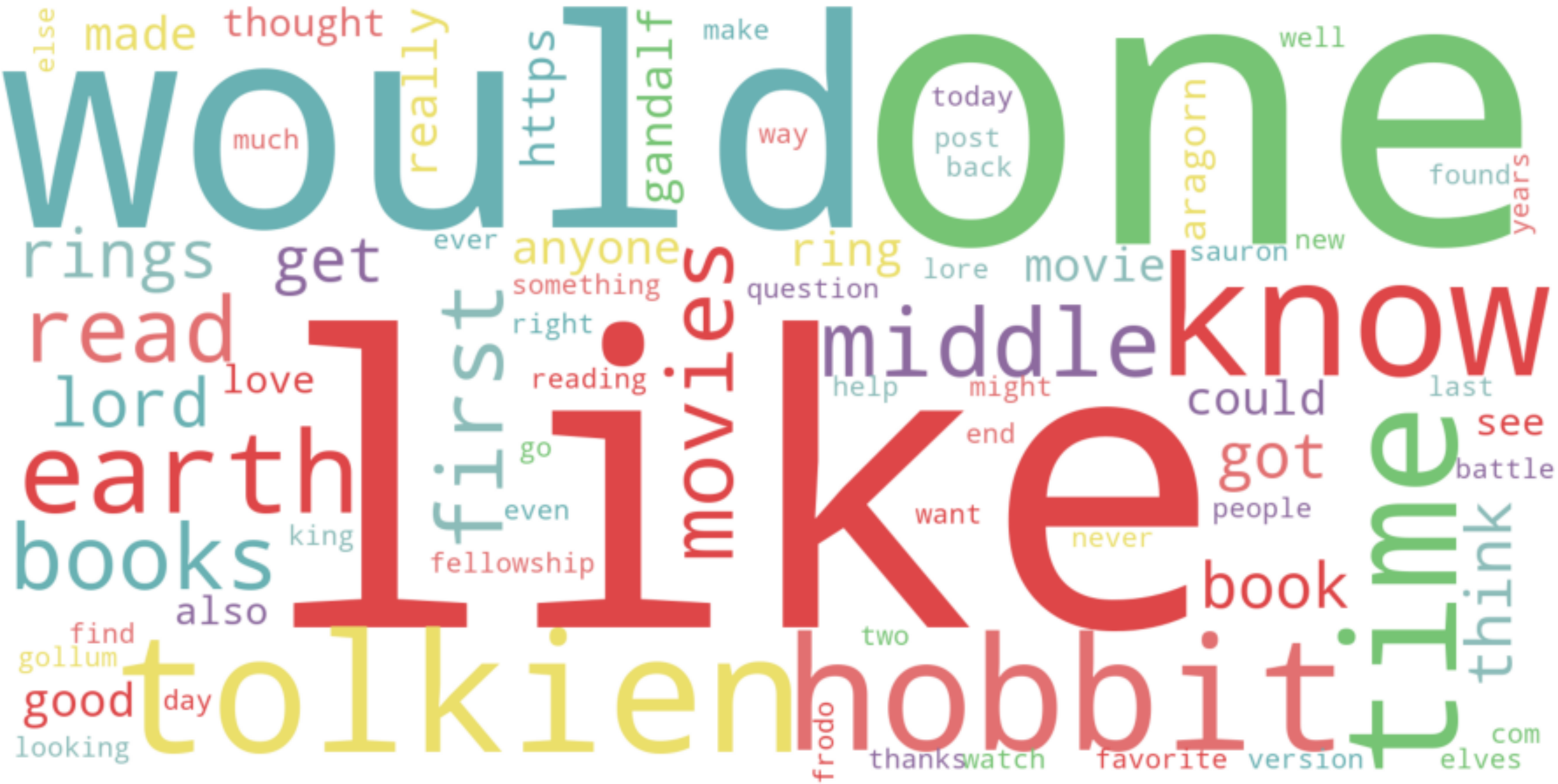
4.
Tokenizer
Stemming
TFIDF

Posts by Day & Subreddit

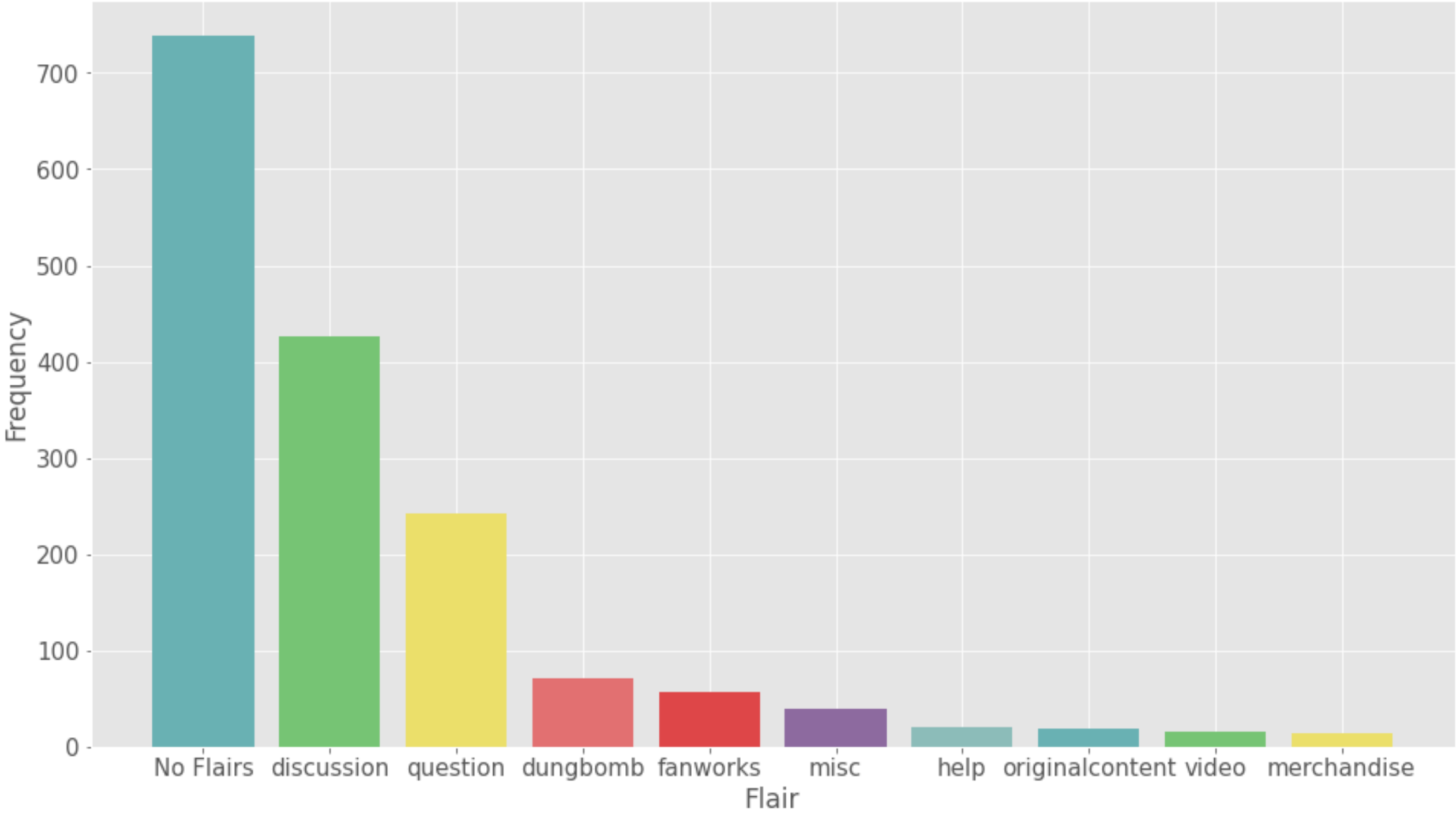


HARRY POTTER

LORD OF THE RINGS



Reddit Flairs



CLASSIFICATION MODELS

CLASSIFICATION MODELS

Logistic Regression

- 'logreg__C': 50
- 'logreg__penalty': 'l2'
- 'tvec__max_df': 0.9
- 'tvec__max_features': None
- 'tvec__min_df': 1
- 'tvec__ngram_range': (1, 1)

kNN

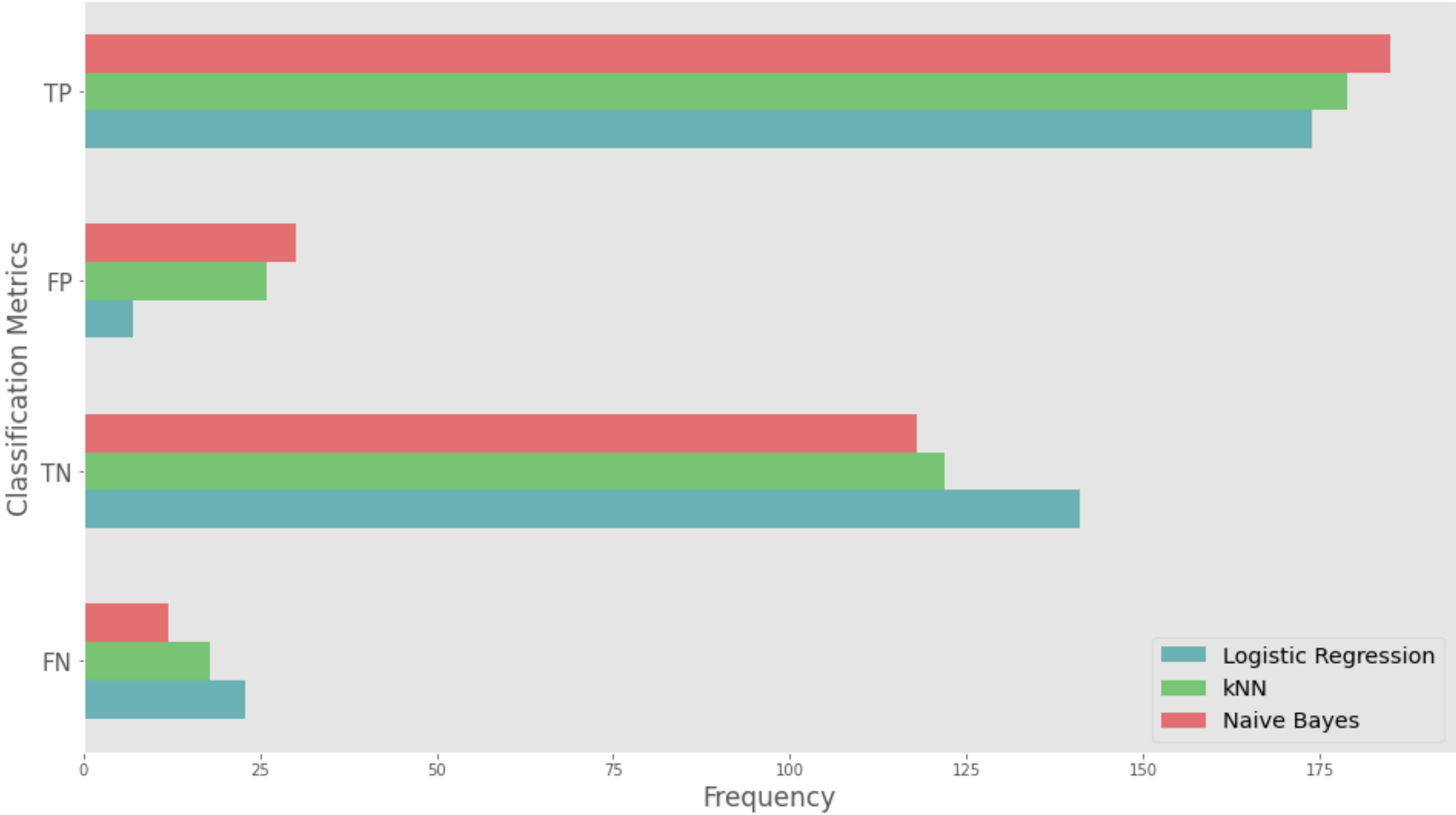
- 'knn__metric': 'minkowski'
- 'knn__n_neighbors': 15
- 'tvec__max_df': 0.9
- 'tvec__max_features': None
- 'tvec__min_df': 1
- 'tvec__ngram_range': (1, 2)

Naïve Bayes

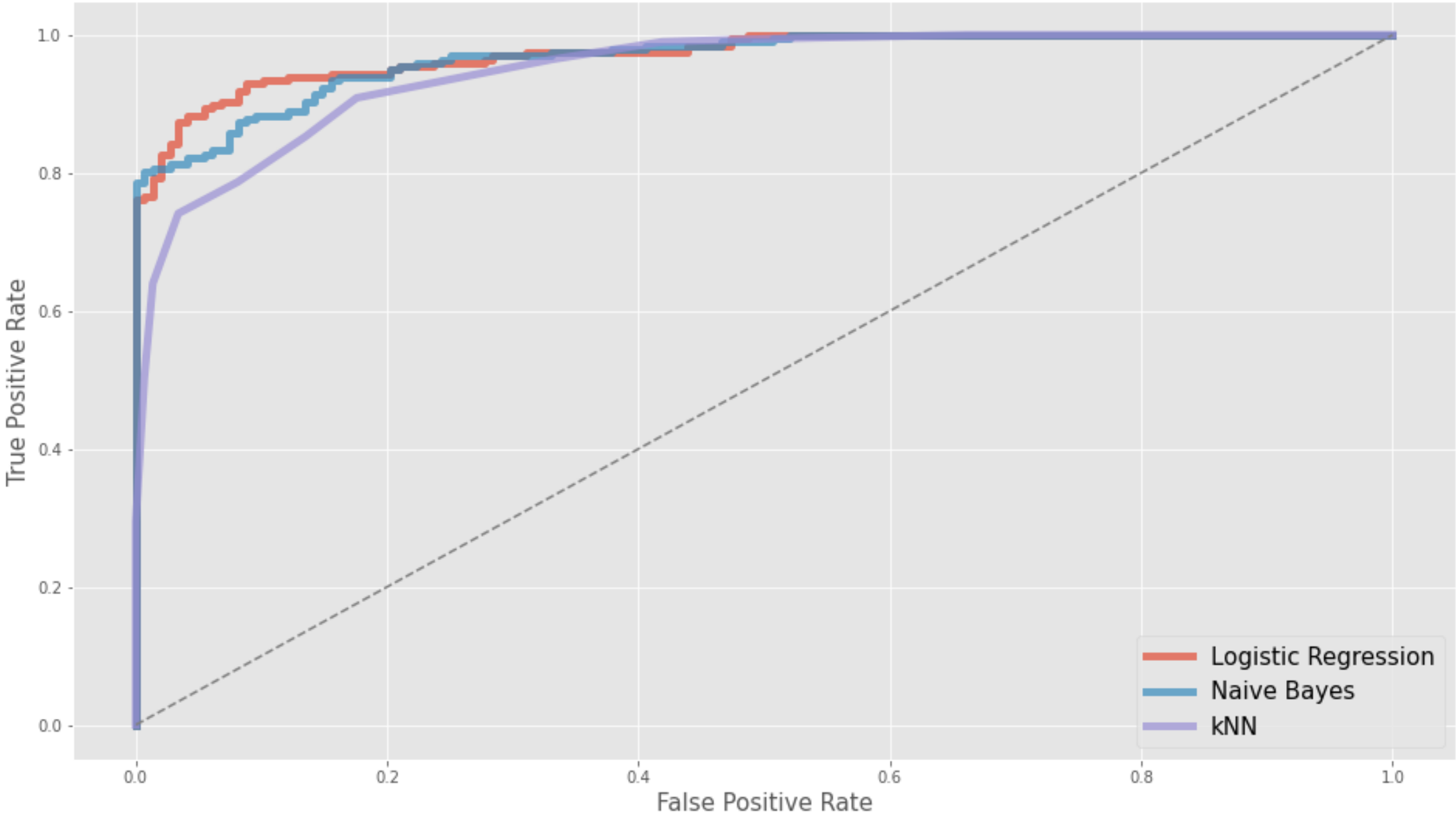
- 'tvec__max_df': 0.9,
- 'tvec__max_features': None,
- 'tvec__min_df': 3,
- 'tvec__ngram_range': (1, 2)


EVALUATION

Model Comparison



ROC Curve


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression
Naive Bayes
kNN

Evaluation Metrics

Metrics	Logistic Regression	kNN	Naive Bayes
Accuracy (Train)	1	0.89	0.95
Accuracy (Test)	0.91	0.87	0.88
Misclassification Rate (Train)	0	0.11	0.05
Misclassification Rate (Test)	0.09	0.13	0.12
Sensitivity (Train)	1	0.95	0.98
Sensitivity (Test)	0.88	0.91	0.94
Specificity (Train)	1	0.80	0.91
Specificity (Test)	0.95	0.82	0.80
Precision (Train)	1	0.87	0.94
Precision (Test)	0.96	0.87	0.86
F1 (Train)	1	0.91	0.96
F1 (Test)	0.92	0.89	0.90
AUC	0.972	0.948	0.967


INFERENCE

WORDS & SUBREDDIT LIKELIHOOD

r/harrypotter

‘harry’ : 4230000x

‘potter’ : 27000x

‘wizard’ : 5200x

‘hogwart’ : 2000x

‘weasley’ : 1100x

r/lotr

‘today’ : 0.00271x

‘gandalf’ : 0.000546x

‘tolkien’ : 0.000360x

‘hobbit’ : 0.000153x

‘ring’ : 0.000122x

*with regards to the positive class (r/harrypotter)

CONCLUSION

CONCLUSION

The project has demonstrated that it is possible to classify whether or not a post belongs in certain subreddit like r/harrypotter and r/lotr. The best performing model (Logistic) uses the texts within a post to do the classification and was shown to be quite effective.

However, if it were to be implemented in production build, it would need to use the one-vs-all approach to classify if the post belongs in the subreddit or not instead of classifying whether or not a post belongs in r/harrypotter or r/lotr.

The main limitation of this model is that it only uses word frequency to determine the subreddits. Improvements could be achieved through the use of more advanced NTL techniques that can better understand the context of the texts within the post.