SUBREDDIT CLASSIFIER: MOVIE VS MUSIC

LIONEL LWAMBA

WHY IS THIS IMPORTANT?

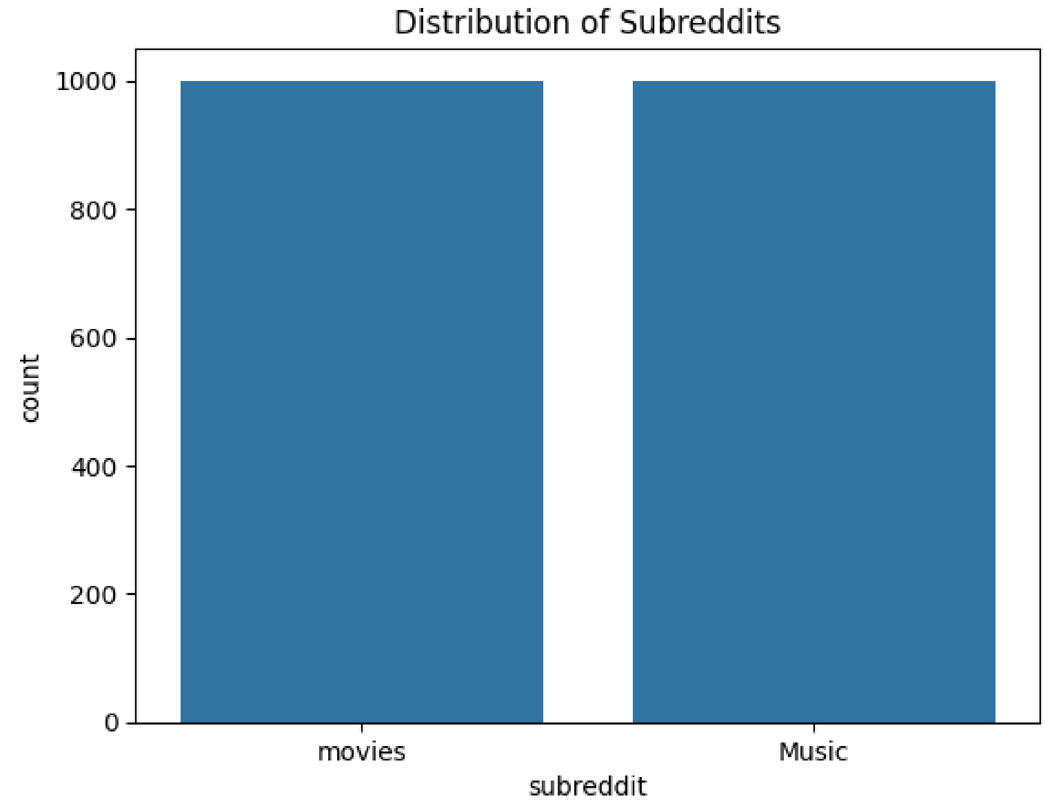
- **Problem:** Online communities are flooded with content. Effective classification is important for users to find what interests them.

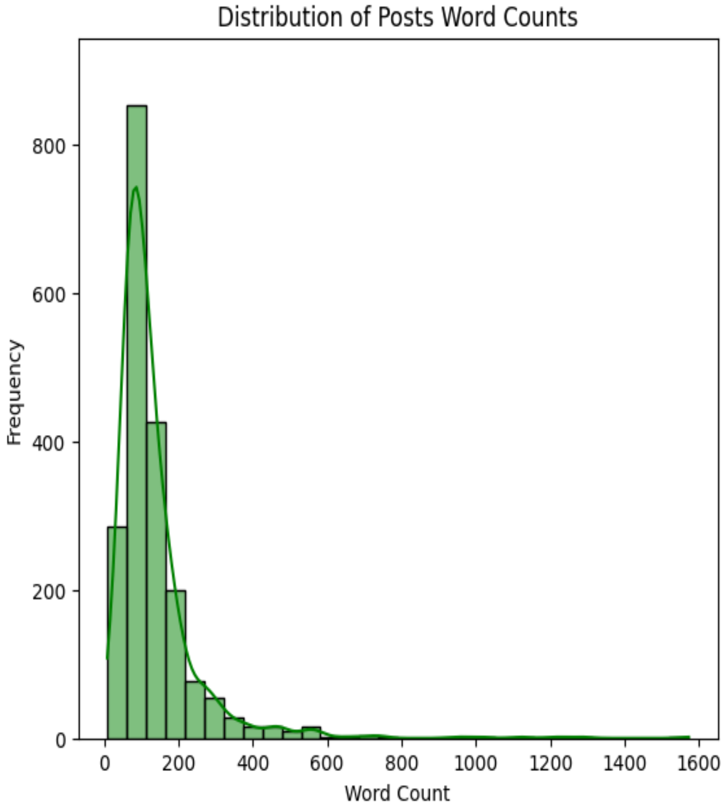
- **Goal:** Build an NLP model to automatically categorize Reddit posts as "Movies" or "Music."

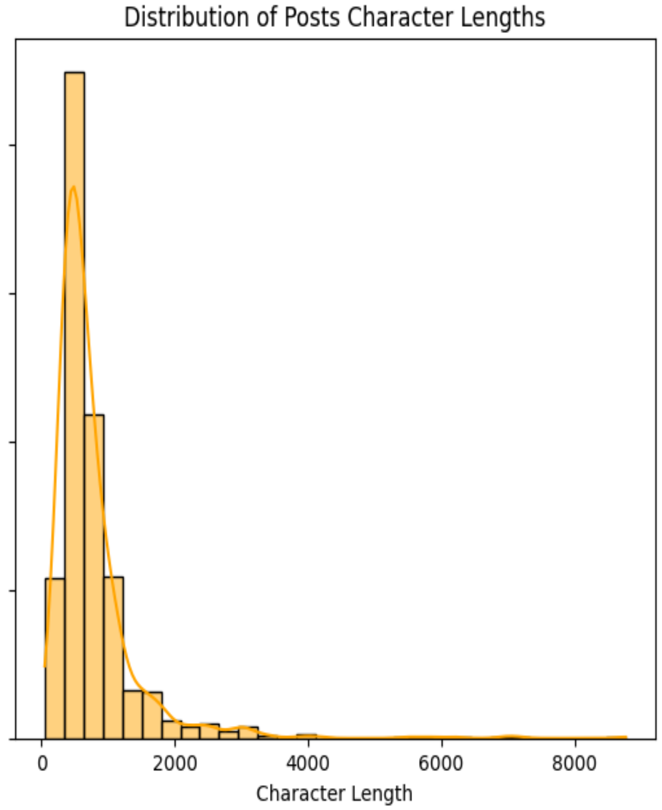
DATA EXPLORATION

- Scraped 2000 Reddit post
- Main Features: Title and Self-Text
- Remove duplicates, and collect data with Title and Self-Text not empty
- Label: Music-0, Movies-1

- Post Character Length Median: 769
- Post Word Count Standard Deviation: 708
- Post Word Count Mean: 135
- Word Count Standard Deviation: 126

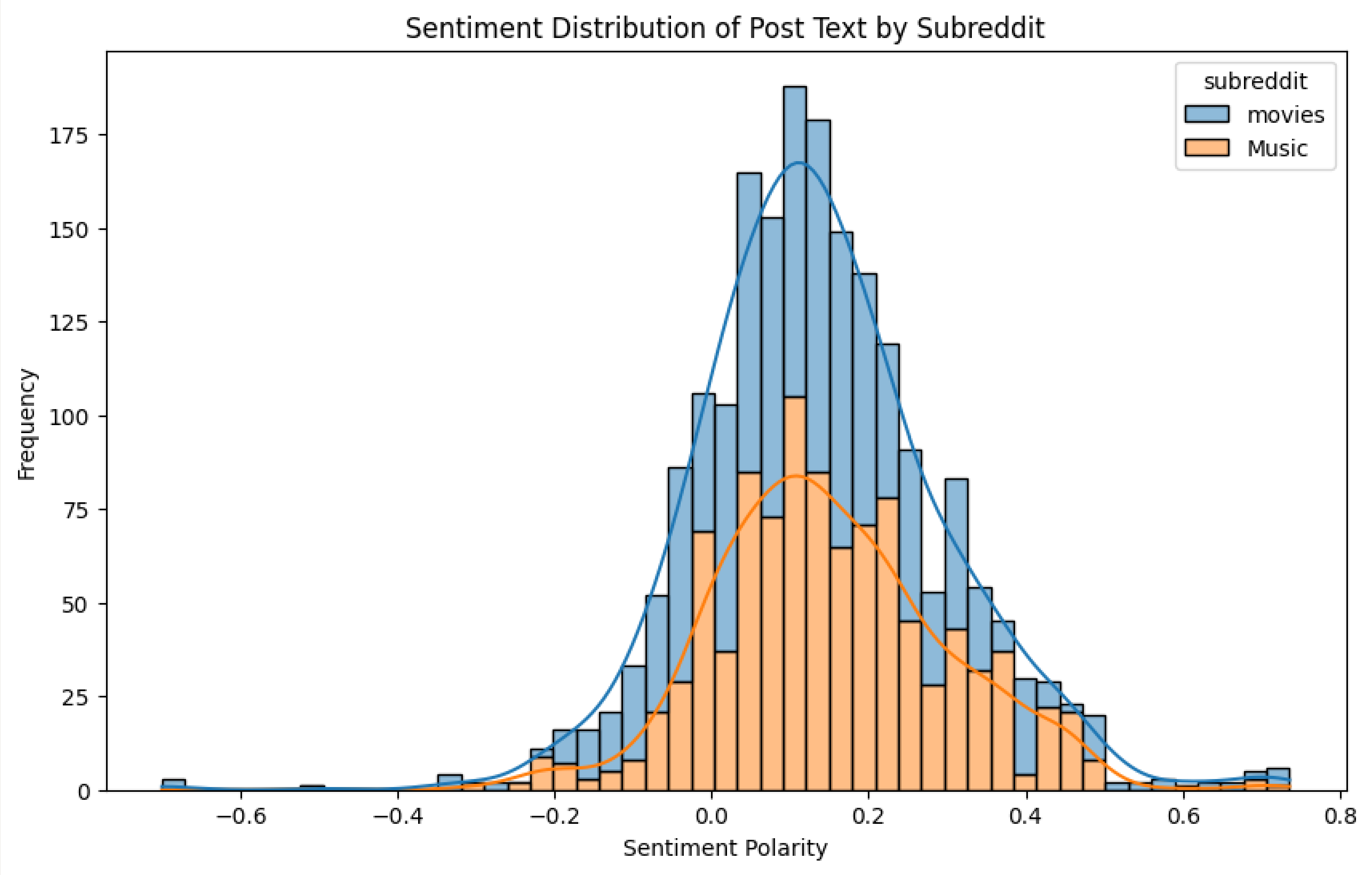






EXPLORING SENTIMENT AND TOPICS ON REDDIT

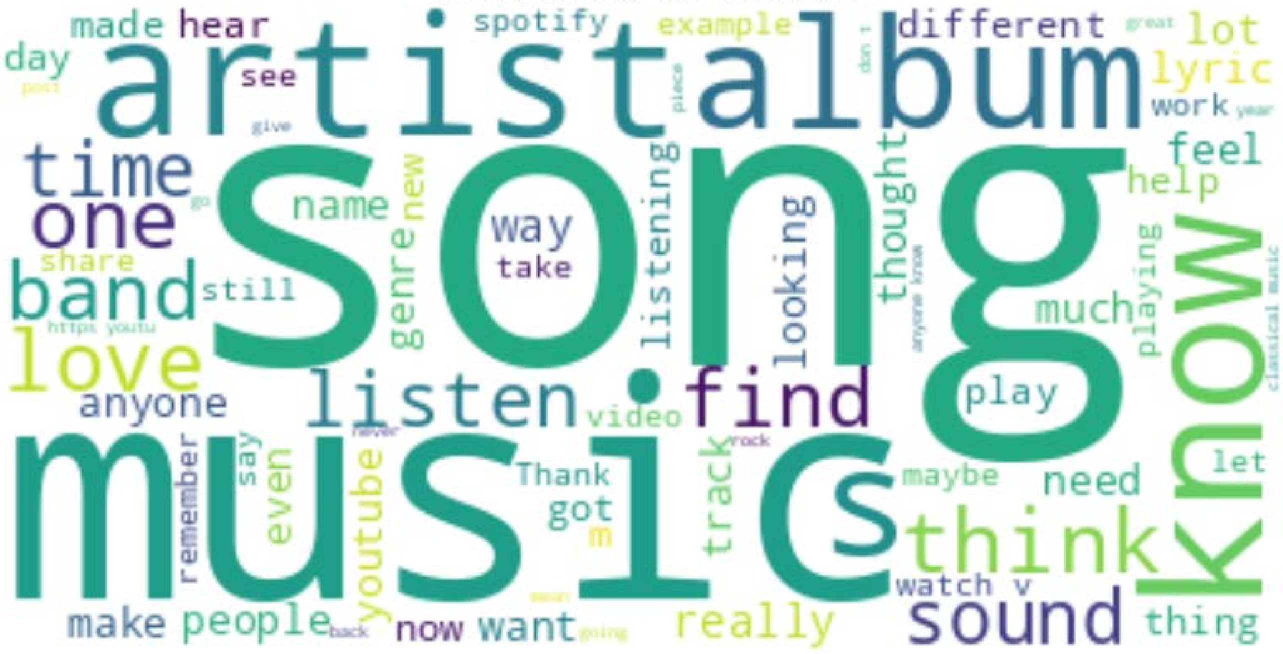
- Analyze sentiment polarity in movie and music posts on Reddit.
- Most data falls within -.2 and .5
- Neutral to slightly positive sentiment



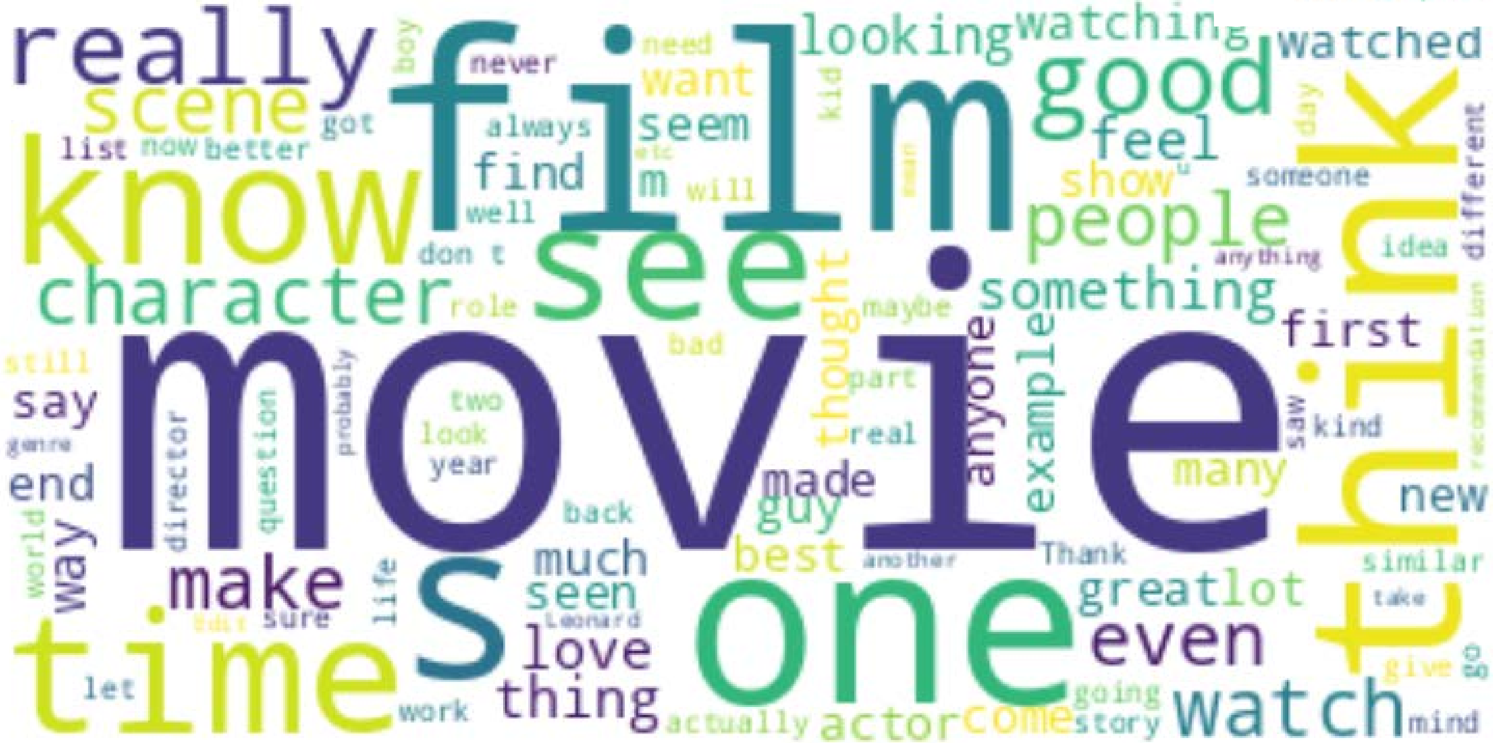
WORD CLOUDS FOR SUBREDDITS

- Visual representation of most frequent words in movies and music subreddits

Word Cloud for r/Music



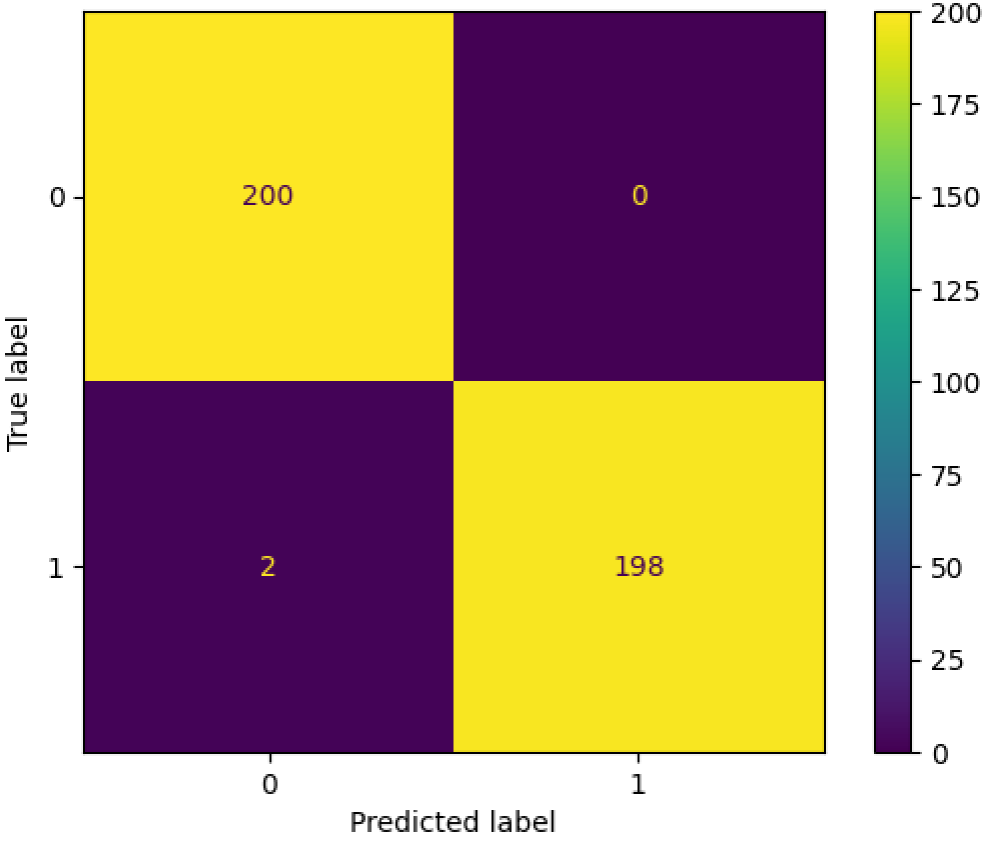
Word Cloud for r/movies



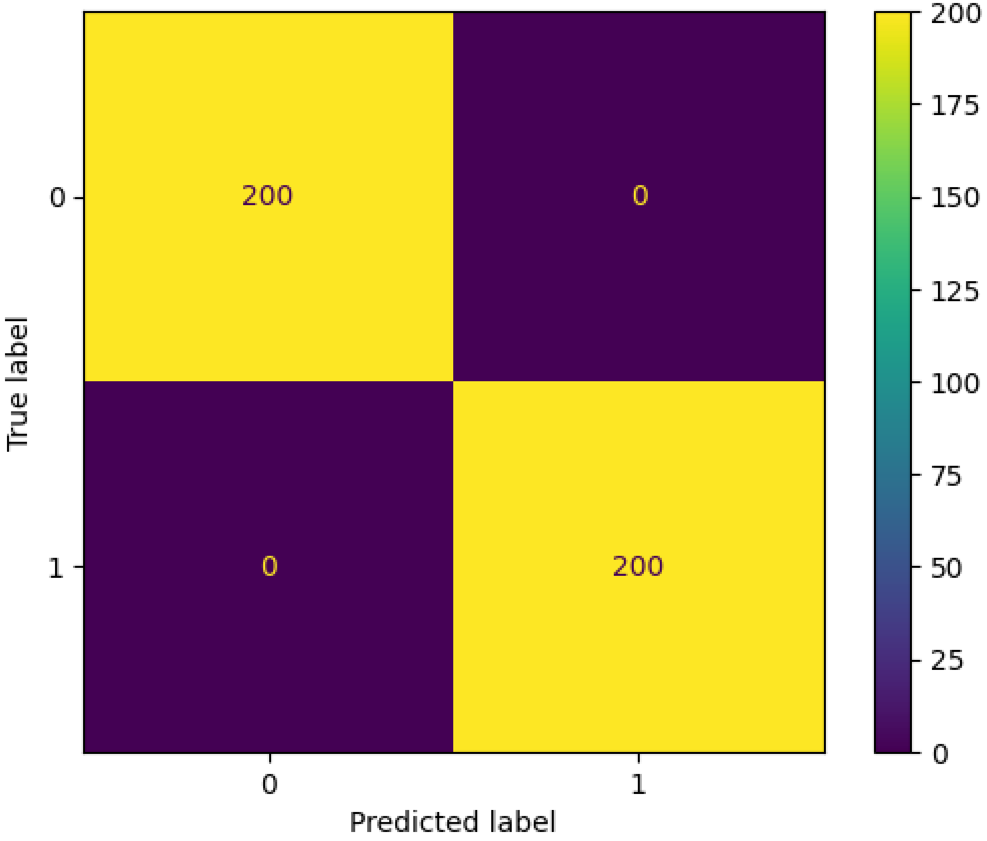
- Size of each word reflects its frequency in the text data

MODEL EVALUATION

COUNTVECTORIZER + LOGISTIC REGRESSION



COUNTVECTORIZER + XGBOOSTCLASSIFIER

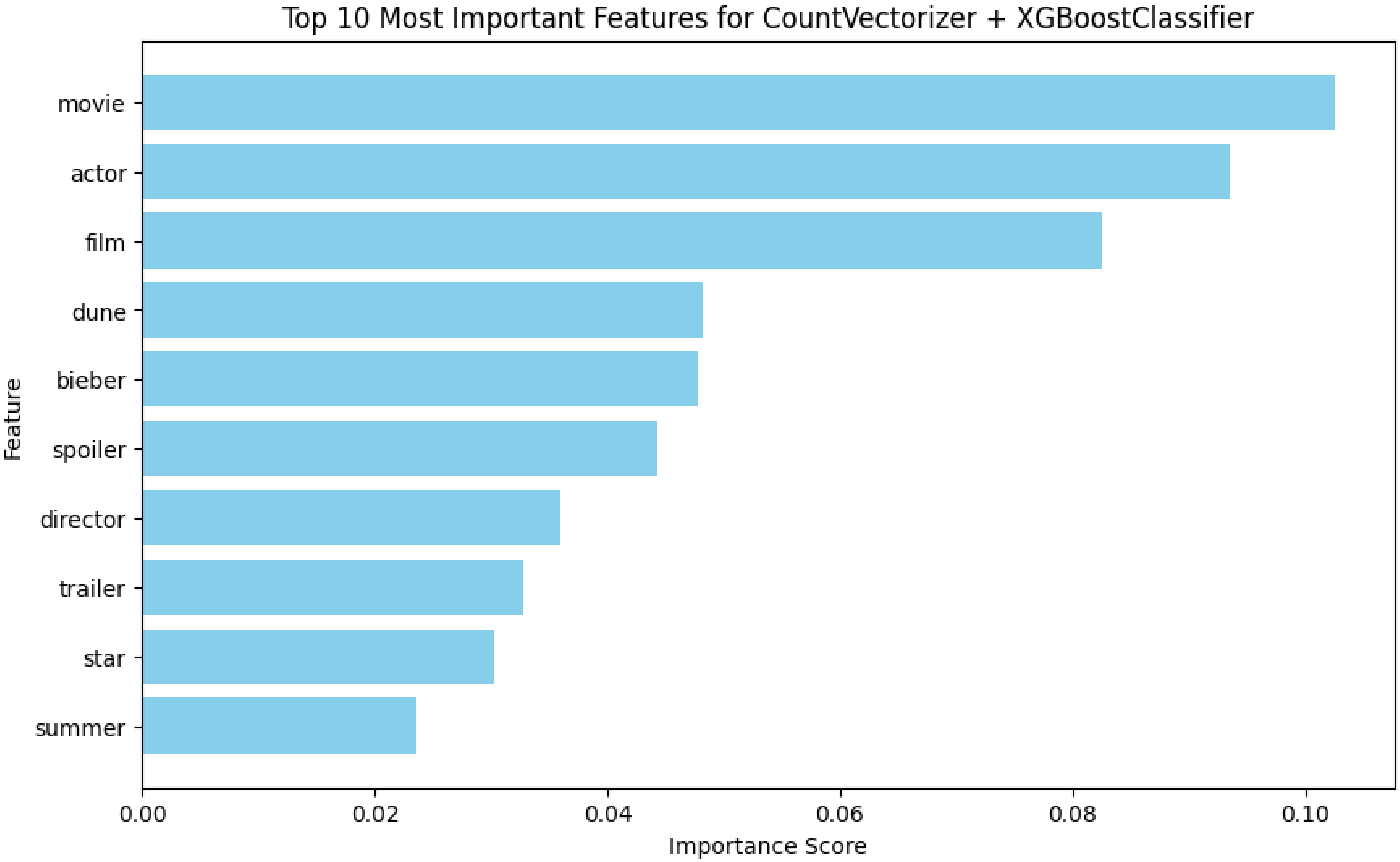


MODEL COMPARISON

- Baseline Model : .5
- Most models shows good generalization on train and test sets
- Models have higher accuracy, and some is slightly overfitting

Model	Training Score	Testing Score
Logistic Regression + Counvectorizer	1	.99
Logistic Regression + TfidfVectorizer	1	.99
Multinomial Naïve Bayes + Counvectorizer	.99	.98
Multinomial Naïve Bayes + TfidfVectorizer	.99	.98
XGBoost + CountVectorizer	1	1

FEATURE IMPORTANCE COUNTVECTORIZER + XGBOOST CLASSIFIER



CONCLUSION

- Built NLP models to classify Reddit posts ("Movies" or "Music")
- All the models achieved a high accuracy on the training and test set between .98 and 1.00)
- XGBoost Classifier had higher score on train and test set.
- For future improvement Larger & More Diverse Dataset

THANK YOU

https://www.reddit.com/