Use Machine Learning Techniques to Maintain Posts Belong to the Corresponding Subreddits

Andy Xu

Project Summary

- **Reddit:** One of the largest online community platform
- **Recent Issue:** Reading difficultness
- **Solution:** Categorize all misplaced posts into corresponding subreddits
- **Result:** The categorizing accuracy reached to **90%**
- **Conclusion:** Model is ready. Continuing optimize is needed.

Problem Statement

- **Daily Responsibilities**

  Ensure all posts belong to the corresponding subreddits.

- **Proposed Solution**

  Use Classifier and NLP skills to train a model so that the model can place the unknown posts into the correct subreddits.

Approach Strategy

- Selected subreddits: Piano and Drum
- Three classifier methods were used:
  - Logistic Regression
  - K Nearest Neighbor
  - Random Forest Classifier
- Two NLP techniques were used:
  - Countvectorizer
  - Tfidfvectorizer
- Cross Validation Grid Search to further improve the model

Data Collection and Preparation

Dataset size

- Piano: 4,099
- Drum: 3,827

Baseline Accuracy: 51.7%

Features used in the analysis

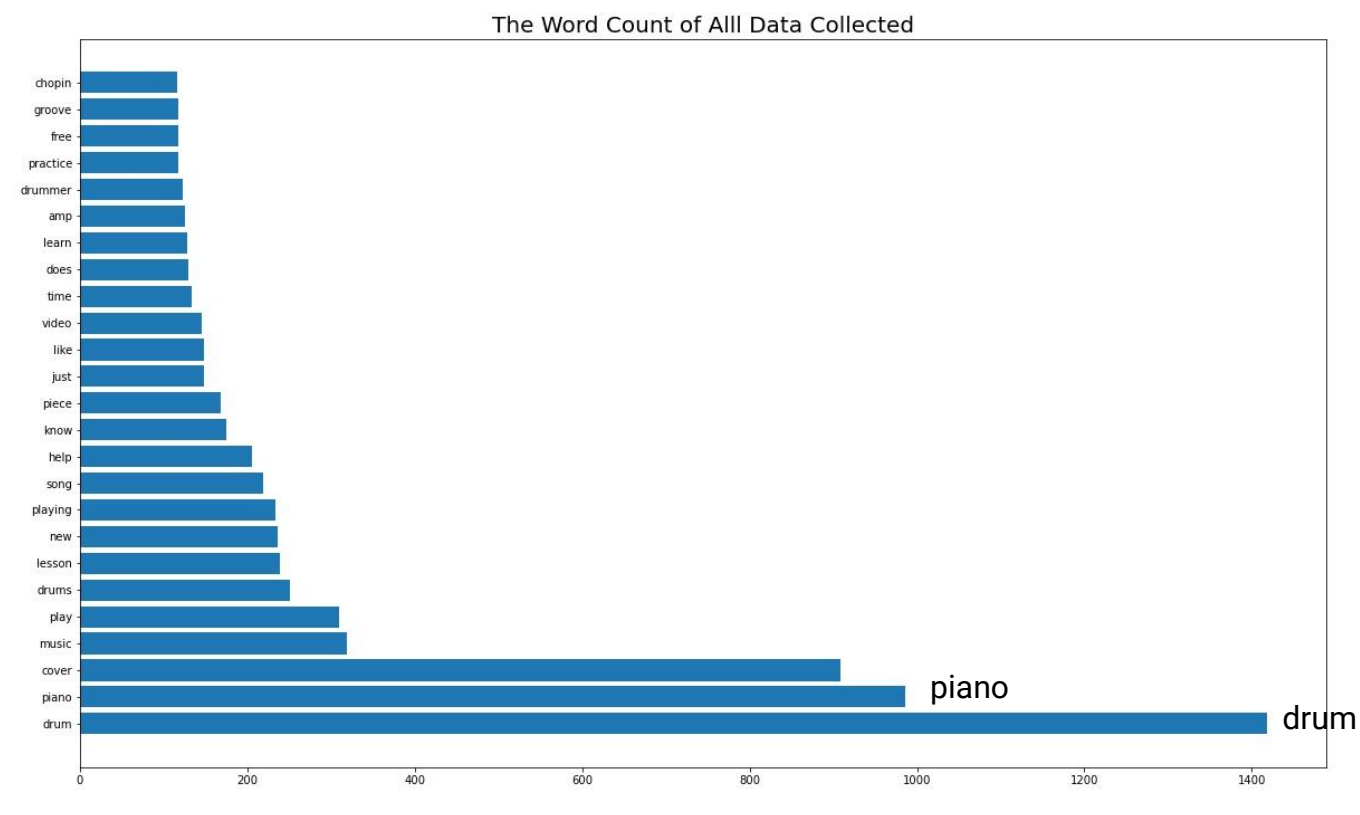
- Title (X)
- Selftext (as second feature)
- Subreddit category (y)

Missing value

- 5,606 in selftext column

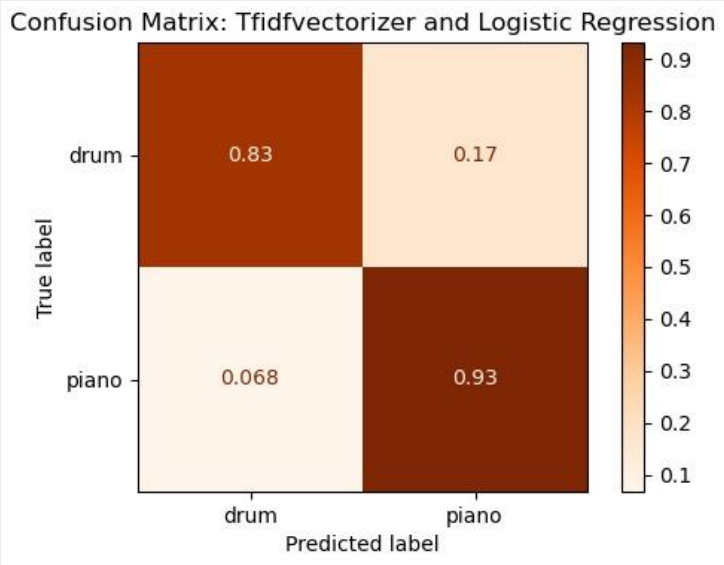
Metrics: Accuracy

Word Count


The Word Count of Alll Data Collected
chopin
groove
free
practice
drummer
amp
learn
does
time
video
like
just
piece
know
help
song
playing
new
lesson
drums
play
music
cover
piano
drum
0
200
400
600
800
1000
1200
1400
piano
drum

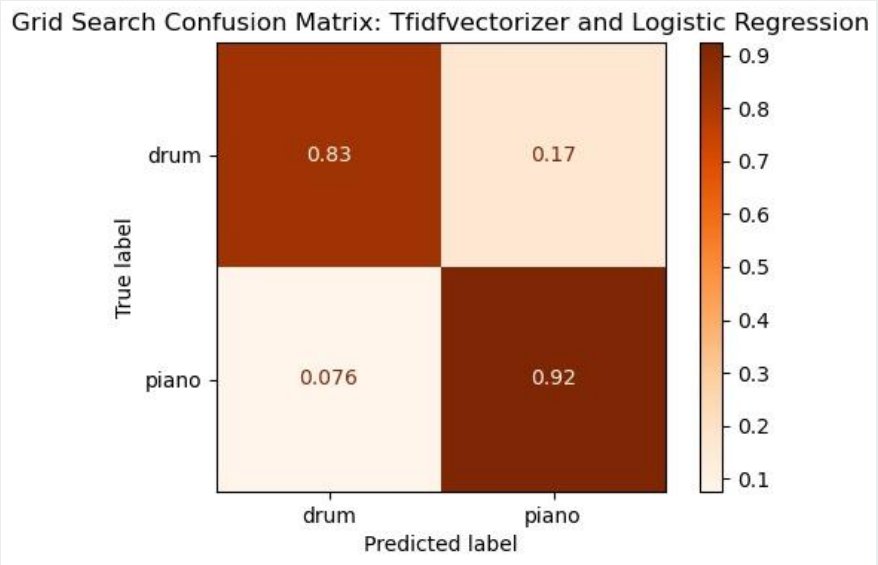
One Feature Model - Logistic Regression

NLP: Tfidfvectorizer



Cross Validation Accuracy: **0.854**
Accuracy Score: **0.881**

Grid Search on Tfidfvectorizer

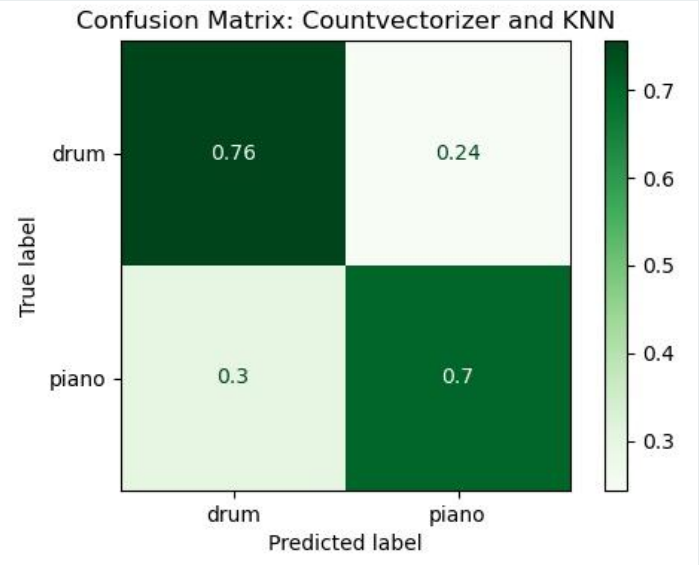


Cross Validation Accuracy: **0.856**
Accuracy Score: **0.880**

Baseline Accuracy: **0.517**

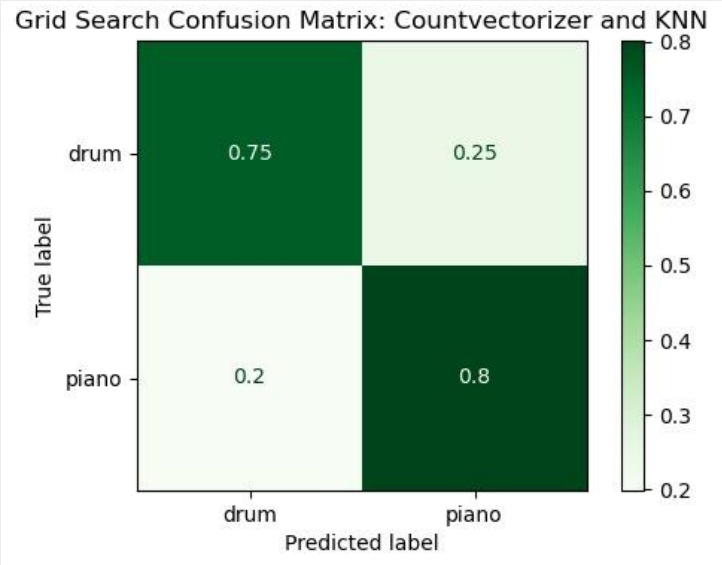
One Feature Model - KNN

NLP: Countvectorizer



Cross Validation Accuracy: **0.712**
Accuracy Score: **0.728**

Grid Search on Countvectorizer

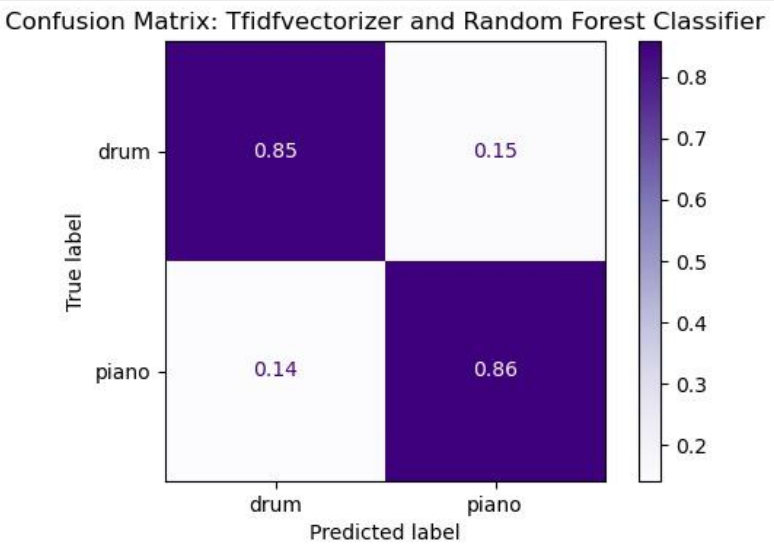


Cross Validation Accuracy: **0.760**
Accuracy Score: **0.778**

Baseline Accuracy: **0.517**

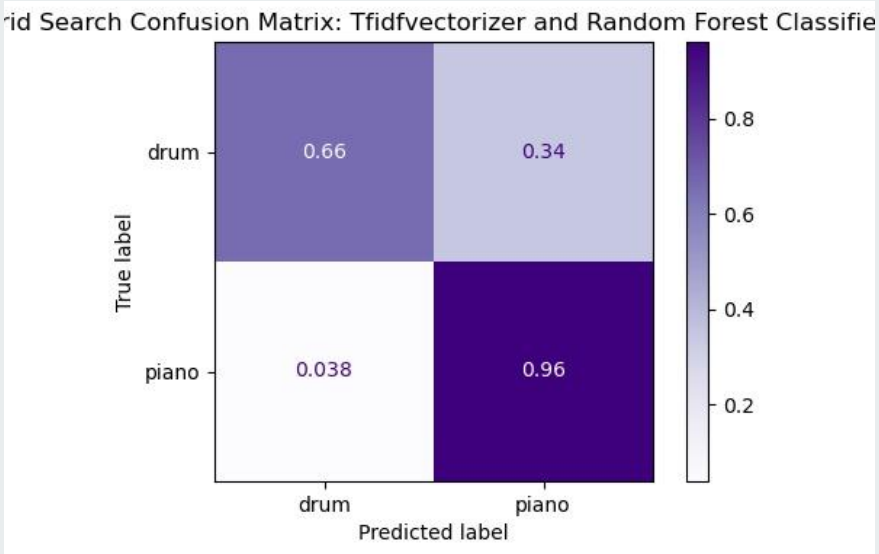
One Feature Model - Random Forest Classifier

NLP: Tfidfvectorizer



Cross Validation Accuracy: **0.844**
Accuracy Score: **0.856**

Grid Search on Tfidfvectorizer

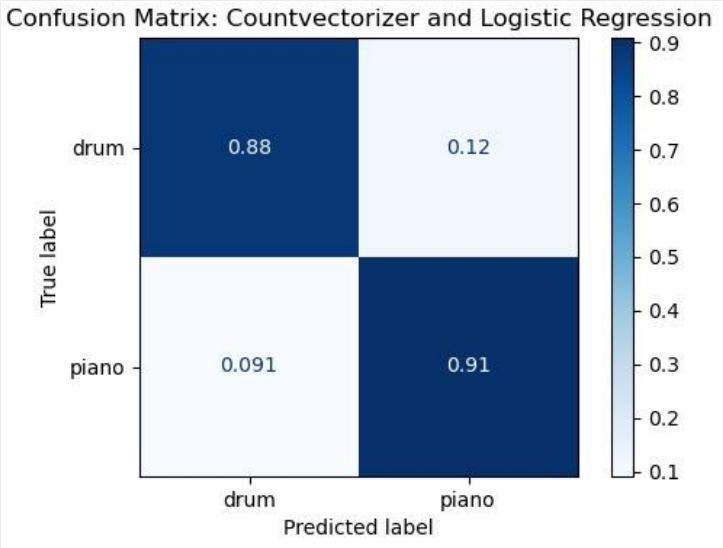


Cross Validation Accuracy: **0.801**
Accuracy Score: **0.815**

Baseline Accuracy: **0.517**

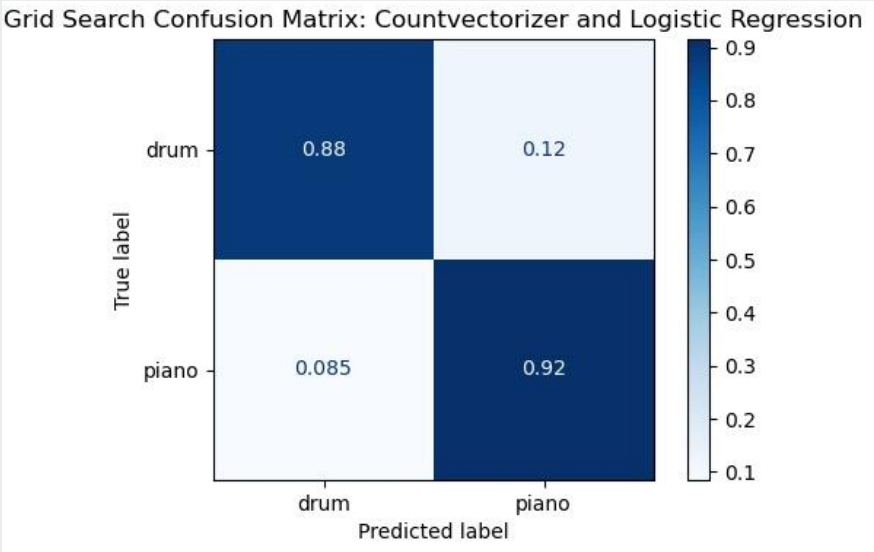
Two Features Model - Logistic Regression

NLP: Countvectorizer



Cross Validation Accuracy: **0.860**
Accuracy Score: **0.897**

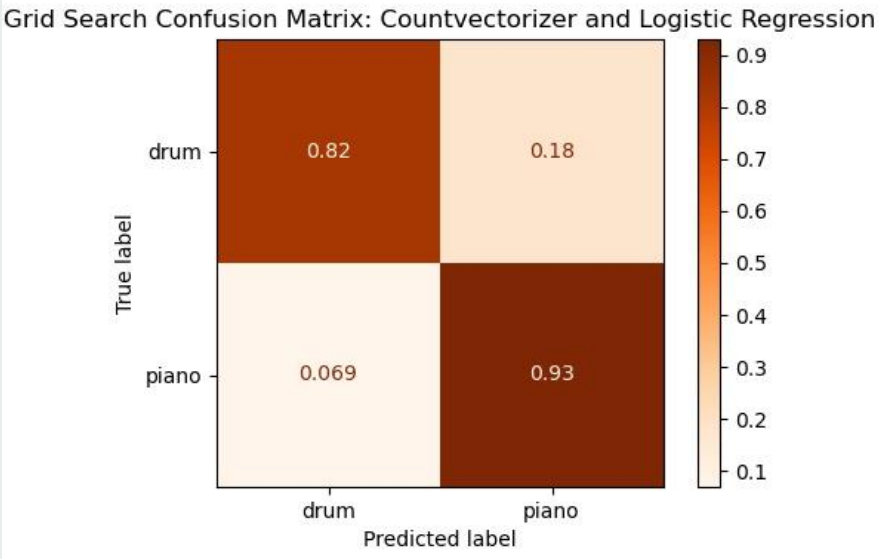
Grid Search on Countvectorizer



Cross Validation Accuracy: **0.865**
Accuracy Score: **0.899**

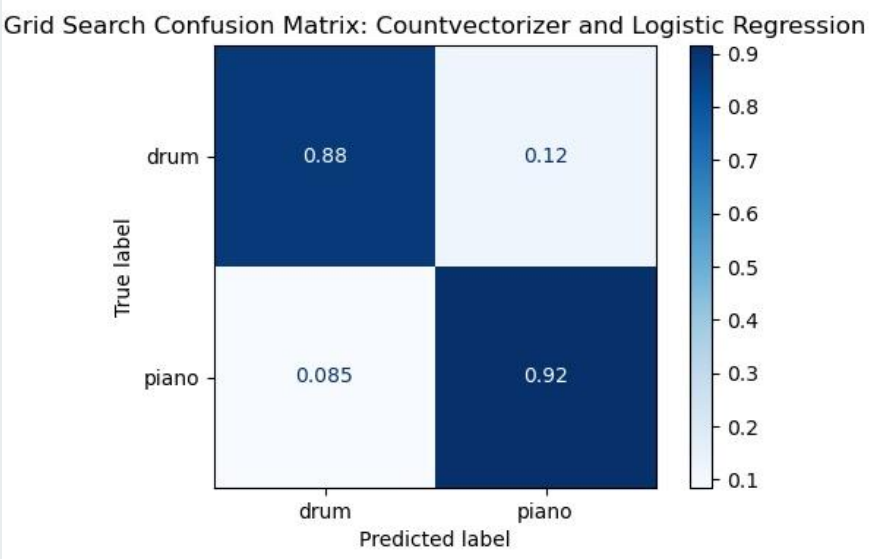
Model Comparisons - Logistic Regression

One Feature



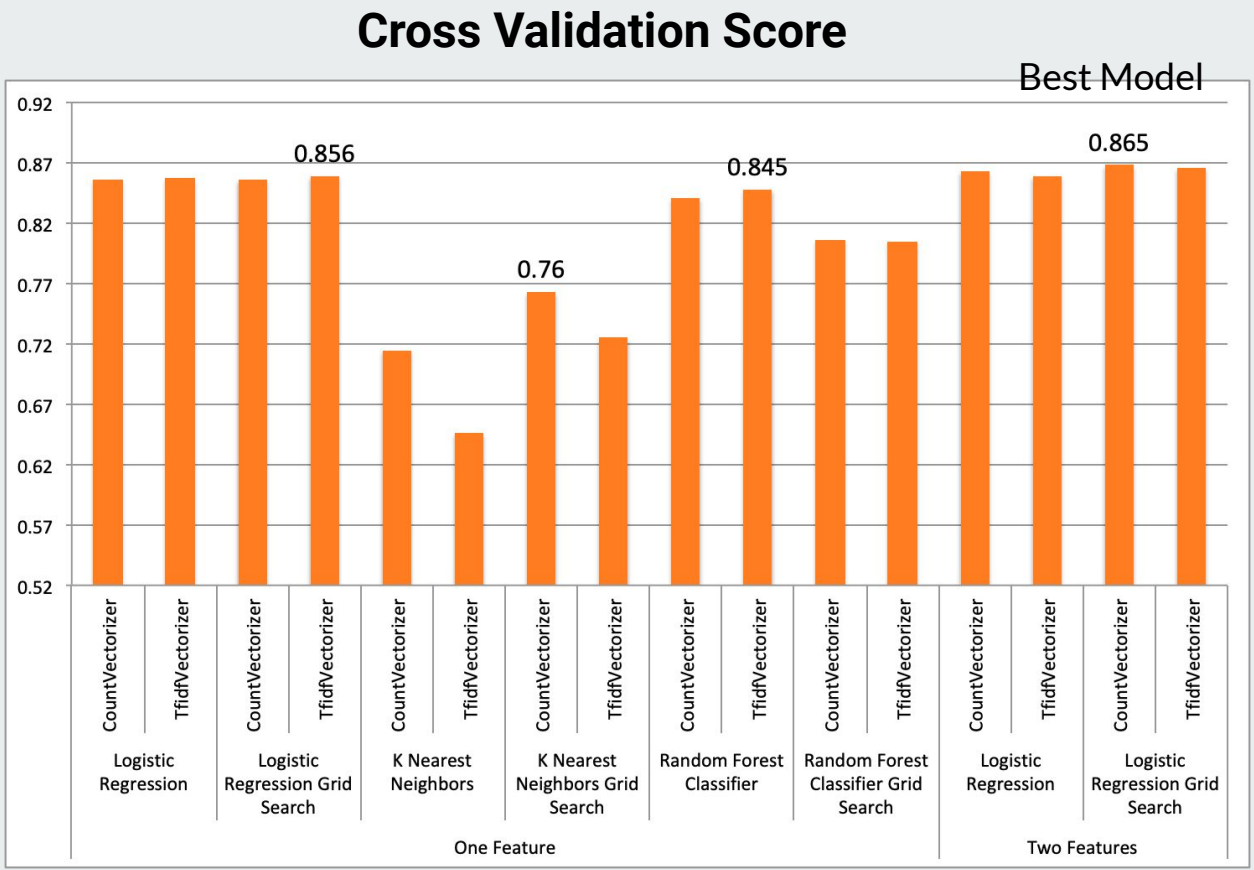
Cross Validation Accuracy: **0.853**
Accuracy Score: **0.877**

Two Feature



Cross Validation Accuracy: **0.865**
Accuracy Score: **0.899**

Summary of All Models



Conclusions and Recommendations

- Improved the accuracy from 12.5% to 34.8% compare to baseline.
- Encountered bottleneck when accuracy score over 85%.

- Collect more data points.
- Optimize the model by further digging on model parameters and other models.
- Increase the number of subreddits for validation.

Thank you

Questions?