Final Project Report

Alex Fick

Spring 2023
COSC 74

Binary Classification Tasks

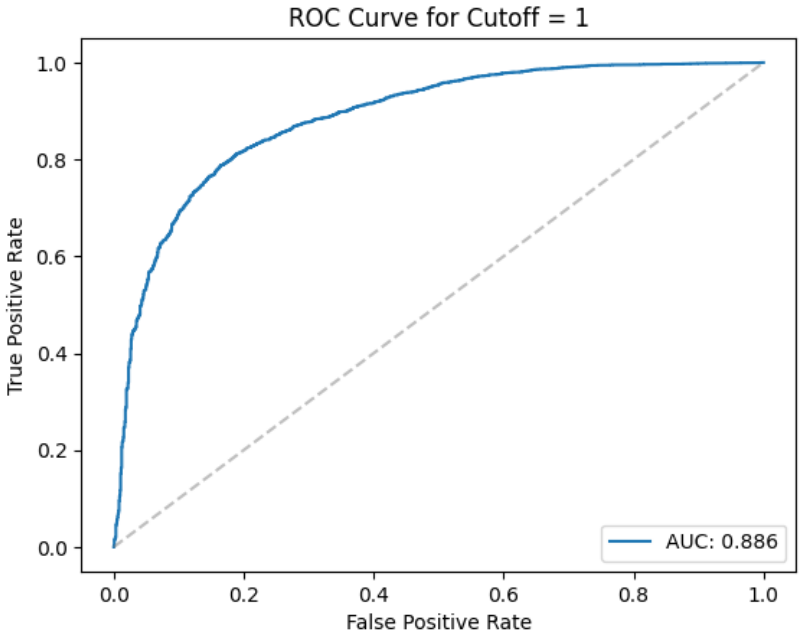
For the classification tasks, I used the CountVectorizer from sklearn to vectorize the 'reviewText' and 'summary' features, also including the 'verified' feature, converted to numeric {-1, 1} to be able to classify. The hyperparameters I tuned were the min_df, max_df and max_features. I ran the classifier for cutoff = 1 overnight through a ton of combinations of parameters, and the best combination was min_df = 0, max_df = 0.6 and max_features = 1500 for the reviewText vectorizer, and min_df = 0, max_df = 0.1 and max_features = 975 for the summary vectorizer.

I ran smaller sets of combinations for the other cutoffs, whose results can be seen in the csvs within the 'crossval_tuning' folder. The best combination was chosen by selecting the combination with the highest 'accuracy' score, which was calculated by Macro F1 (even though the assignment suggested cross-validation accuracy) since that is how the competition ranks the performance of the models.

The best combinations for each cutoff score yielded the following results on my validation testing:

Cutoff: 1

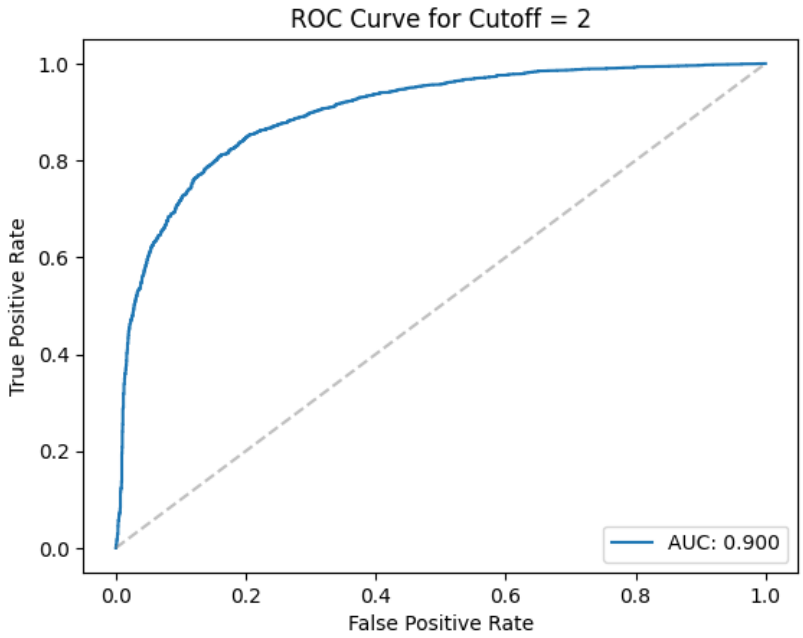
- Params: 0.0, 0.6, 1500, 0.0, 0.1, 975 (rmin_df, rmax_df, rmax_features, smin_df, smax_df, smax_features)
- ROC:



- Confusion Matrix:
$\begin{pmatrix} 685 & 548 \\ 293 & 4312 \end{pmatrix}$
- Macro F1 Score: 0.7653877278353944
- Cross-Validation Accuracy: 0.8590637571016575

Cutoff: 2

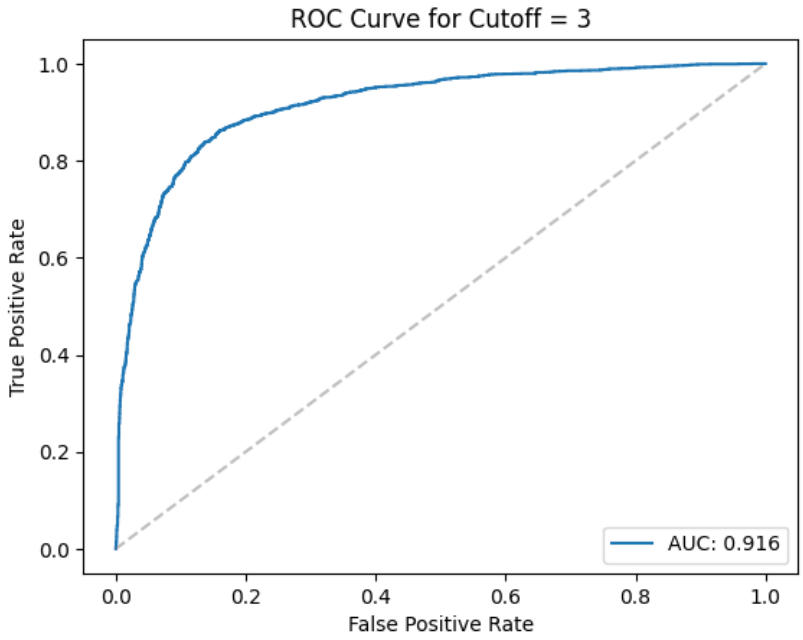
- Params: 0.0, 0.3, 1250, 0.0, 0.3, 1250
- ROC:



- Confusion Matrix:
$\begin{pmatrix} 1858 & 523 \\ 486 & 2971 \end{pmatrix}$
- Macro F1 Score: 0.8206480282613213
- Cross-Validation Accuracy: 0.8240759398851265

Cutoff: 3

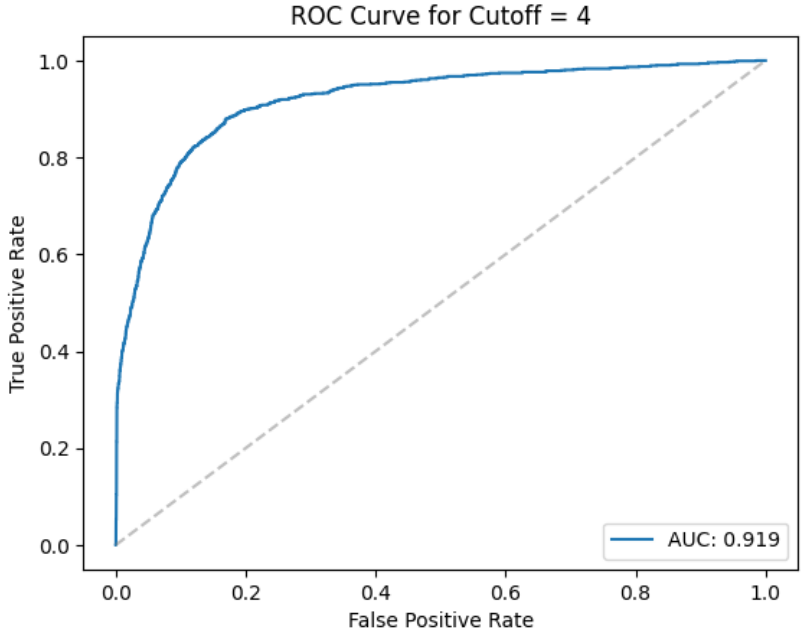
- Params: 0.0, 0.3, 1250, 0.0, 0.3, 1250
- ROC:



- Confusion Matrix:
  $\begin{pmatrix} 3198 & 378 \\ 476 & 1786 \end{pmatrix}$
- Macro F1 Score: 0.8446280754787541
- Cross-Validation Accuracy: 0.8446748789858789

Cutoff: 4

- Params: 0.005, 0.2, 1000, 0.0, 0.2, 500
- ROC:



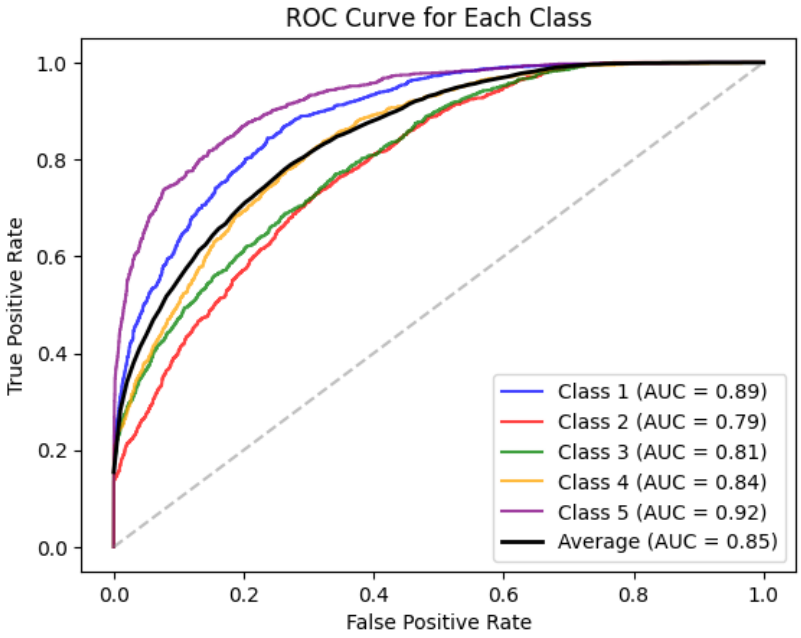
- Confusion Matrix:
$$\begin{pmatrix} 4466 & 239 \\ 410 & 723 \end{pmatrix}$$
- Macro F1 Score: 0.8112382826093508
- Cross-Validation Accuracy: 0.8788486983102721

Multiclass Classification Task

For the multiclass classifier, I used the TfidfVectorizer with the 'reviewText', 'summary', and 'verified' features, with hyper parameters min_df=30, max_df=12300, max_features=5000 for both the summary and reviewText vectorizers.

The scores for the multiclassification task are as follows:

- Confusion Matrix:
$$\begin{pmatrix} 831 & 276 & 66 & 28 & 32 \\ 279 & 549 & 224 & 63 & 33 \\ 91 & 236 & 615 & 194 & 59 \\ 43 & 73 & 156 & 628 & 229 \\ 36 & 35 & 41 & 171 & 850 \end{pmatrix}$$
- ROC:



- Macro F1 Score: 0.5932051430638416
- Cross-Validation Accuracy: 0.5625214224507283

Clustering Task

For clustering, I again used the TfidfVectorizer, with hyper parameters min_df = 20, max_df = 0.62, ngram_range = (4, 11), using the 'reviewText' feature and achieved the following scores:

- Silhouette Score: 0.9992429539702239
- Adjusted Rand Index: 0.00028470788902897564

Kaggle Competition Scores

My best score for each classifier in the Kaggle Competition is as follows:

- Cutoff = 1: 0.77737
- Cutoff = 2: 0.82658
- Cutoff = 3: 0.86821
- Cutoff = 4: 0.82278
- Multiclass: 0.58961

My username is alexanderfick, with screen name Alexander Fick.