Shahjalal University Of Science And Technology, Sylhet
Software Engineering, IICT

Data Science

TEXT CLASSIFICATION

Prepared By:

Md. Mehedi Hasan
2017831023

Supervisor:

Ms Sayma Sultana Chowdhury
Assistant Professor, IICT, SUST

12/02/2021

1. Train five separate ML Classification models and provide the classification report for each. Models: KNN, Naive Bayes, Random Forest, Decision Tree and ANN.

Data Frequency

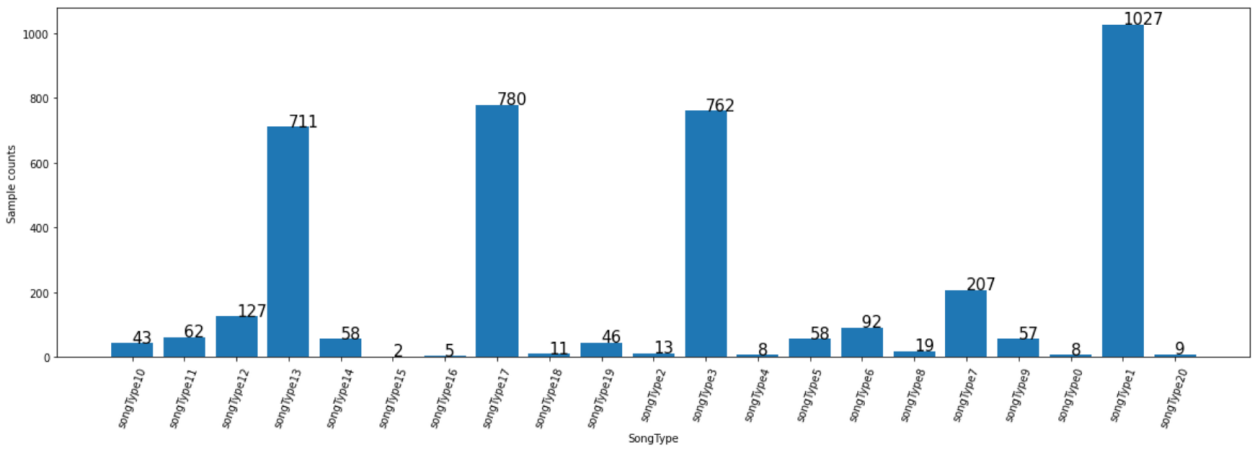

Figure 1: From the Bar plot we can see maximum data is from song Type 1, 3, 13, 17.So our model can learn these type song and cann't do better on those song type has few data

1.1. Naive Bayes

The Internet enables users to access services and run applications over a heterogeneous collection of computers and networks. Heterogeneity (that is, variety and difference) applies to all of the following:

Naive Bayes model performance

Cross Accuracy: 0.57 (+/- 0.03)

	precision	recall	f1-score	support
0	0.00	0.00	0.00	2
1	0.47	0.64	0.55	308
2	0.00	0.00	0.00	4
3	0.47	0.55	0.51	229
4	0.00	0.00	0.00	2
5	0.50	0.06	0.11	17
6	0.78	0.50	0.61	28
7	0.85	0.55	0.67	62
8	0.00	0.00	0.00	6
9	0.29	0.12	0.17	17
10	0.50	0.23	0.32	13
11	0.18	0.11	0.13	19
12	0.27	0.08	0.12	38
13	0.58	0.49	0.53	213
14	0.40	0.12	0.18	17
15	0.00	0.00	0.00	1
16	0.00	0.00	0.00	2
17	0.87	0.94	0.90	234
18	0.25	0.33	0.29	3
19	0.80	0.57	0.67	14
20	0.00	0.00	0.00	3
accuracy			0.58	1232
macro avg	0.34	0.25	0.27	1232
weighted avg	0.57	0.58	0.57	1232

Figure 2

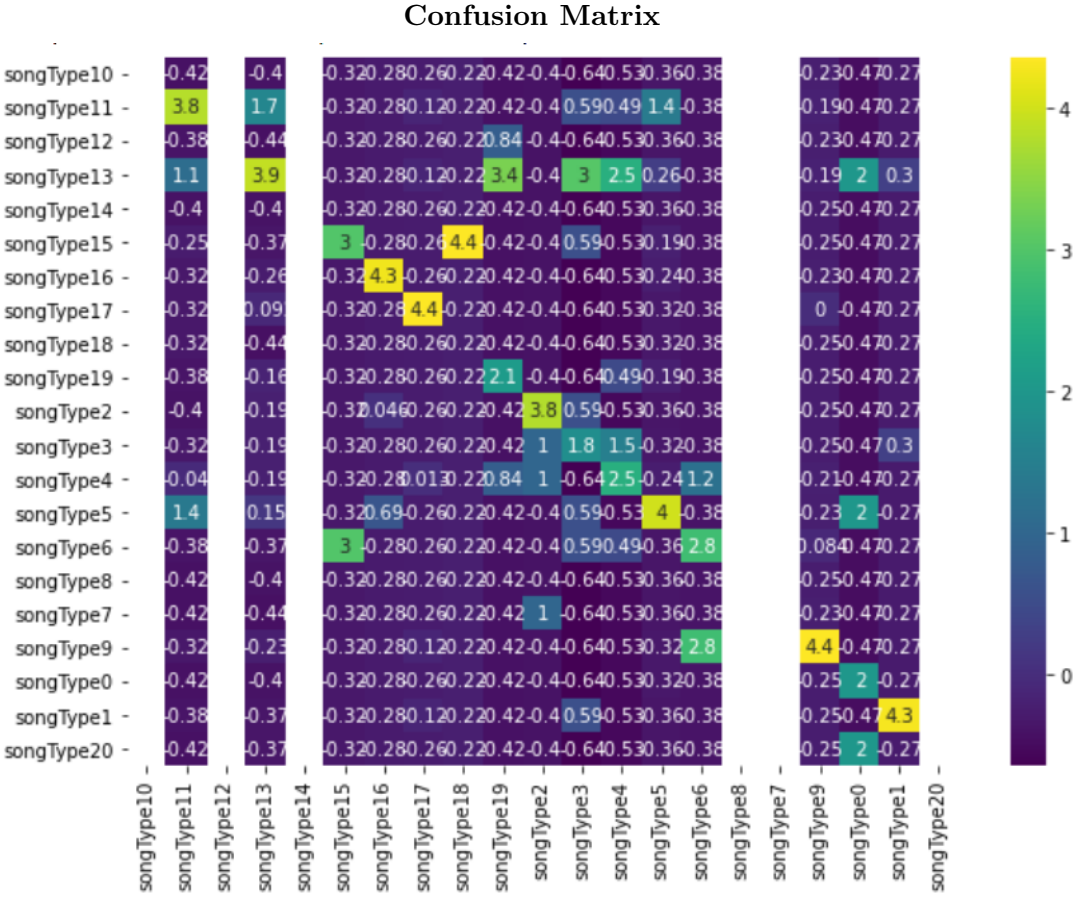

Figure 3

Model summary will go there

1.2. KNN

KNN model performance

Cross Accuracy: 0.15 (+/- 0.05)

	precision	recall	f1-score	support
0	0.00	0.00	0.00	2
1	0.37	0.19	0.25	308
2	0.00	0.00	0.00	4
3	0.45	0.22	0.30	229
4	0.00	0.00	0.00	2
5	0.00	0.00	0.00	17
6	0.09	0.43	0.15	28
7	0.44	0.06	0.11	62
8	0.00	0.00	0.00	6
9	0.00	0.00	0.00	17
10	0.33	0.08	0.12	13
11	0.00	0.00	0.00	19
12	0.02	0.39	0.05	38
13	0.53	0.04	0.07	213
14	1.00	0.06	0.11	17
15	0.00	0.00	0.00	1
16	0.00	0.00	0.00	2
17	0.86	0.26	0.40	234
18	0.01	0.33	0.02	3
19	0.00	0.00	0.00	14
20	0.00	0.00	0.00	3
accuracy			0.17	1232
macro avg	0.20	0.10	0.08	1232
weighted avg	0.47	0.17	0.22	1232

Figure 4

Confusion Matrix

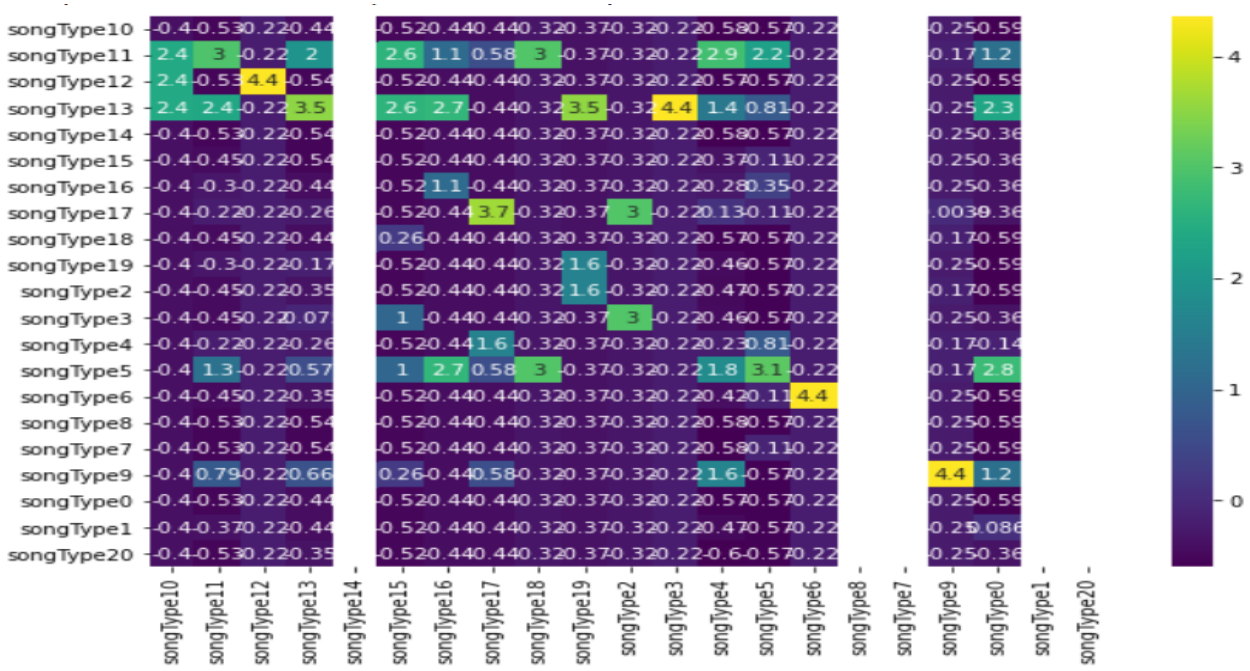

Figure 5

Model summary will go there

1.3. Decision Tree

Decision Tee model performance

Cross Accuracy: 0.34 (+/- 0.03)

	precision	recall	f1-score	support
0	0.00	0.00	0.00	2
1	0.36	0.43	0.39	308
2	0.00	0.00	0.00	4
3	0.33	0.32	0.33	229
4	0.00	0.00	0.00	2
5	0.08	0.06	0.07	17
6	0.23	0.21	0.22	28
7	0.20	0.11	0.14	62
8	0.00	0.00	0.00	6
9	0.00	0.00	0.00	17
10	0.00	0.00	0.00	13
11	0.00	0.00	0.00	19
12	0.06	0.08	0.07	38
13	0.31	0.36	0.33	213
14	0.17	0.06	0.09	17
15	0.00	0.00	0.00	1
16	0.00	0.00	0.00	2
17	0.62	0.58	0.60	234
18	1.00	0.33	0.50	3
19	0.73	0.57	0.64	14
20	0.00	0.00	0.00	3
accuracy			0.36	1232
macro avg	0.20	0.15	0.16	1232
weighted avg	0.36	0.36	0.36	1232

Figure 6

Confusion Matrix

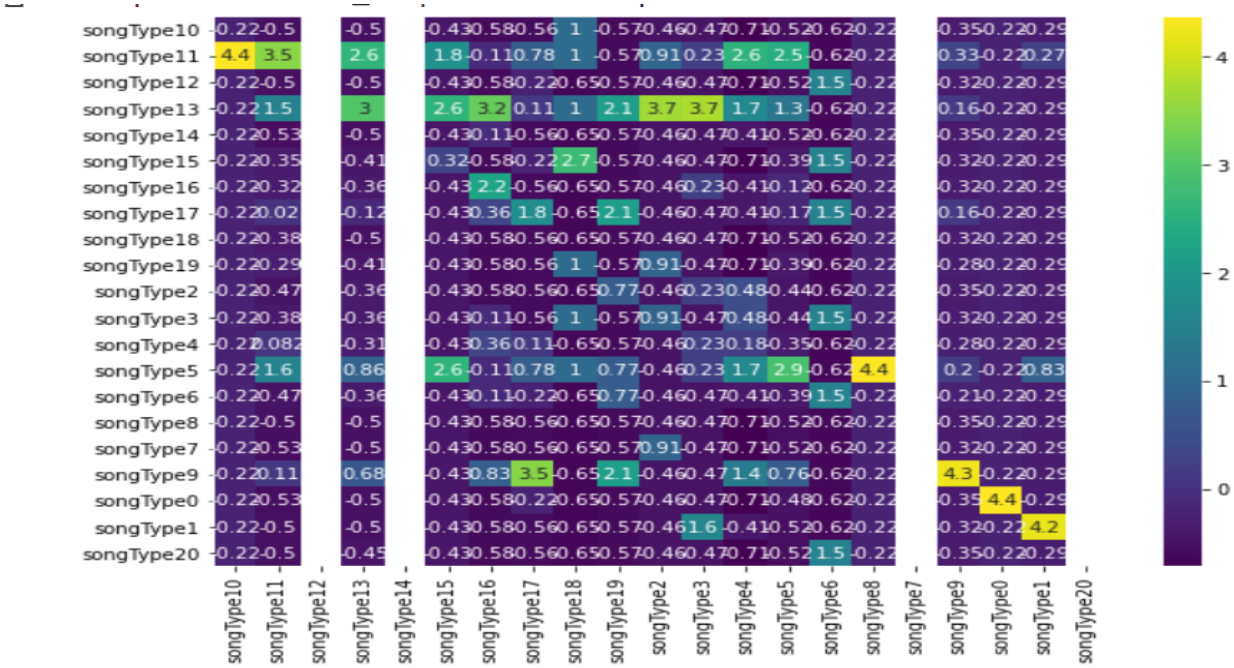

Figure 7

Model summary will go there

1.4. Random Forest

Random Forest model performance

Cross Accuracy: 0.45 (+/- 0.02)

	precision	recall	f1-score	support
0	0.00	0.00	0.00	2
1	0.36	0.87	0.51	308
2	0.00	0.00	0.00	4
3	0.50	0.24	0.32	229
4	1.00	0.50	0.67	2
5	0.50	0.06	0.11	17
6	1.00	0.04	0.07	28
7	1.00	0.08	0.15	62
8	0.00	0.00	0.00	6
9	0.00	0.00	0.00	17
10	0.00	0.00	0.00	13
11	0.00	0.00	0.00	19
12	0.00	0.00	0.00	38
13	0.68	0.23	0.35	213
14	1.00	0.06	0.11	17
15	0.00	0.00	0.00	1
16	0.00	0.00	0.00	2
17	0.68	0.84	0.75	234
18	0.00	0.00	0.00	3
19	1.00	0.07	0.13	14
20	0.00	0.00	0.00	3
accuracy			0.47	1232
macro avg	0.37	0.14	0.15	1232
weighted avg	0.54	0.47	0.40	1232

Figure 8

Confusion Matrix

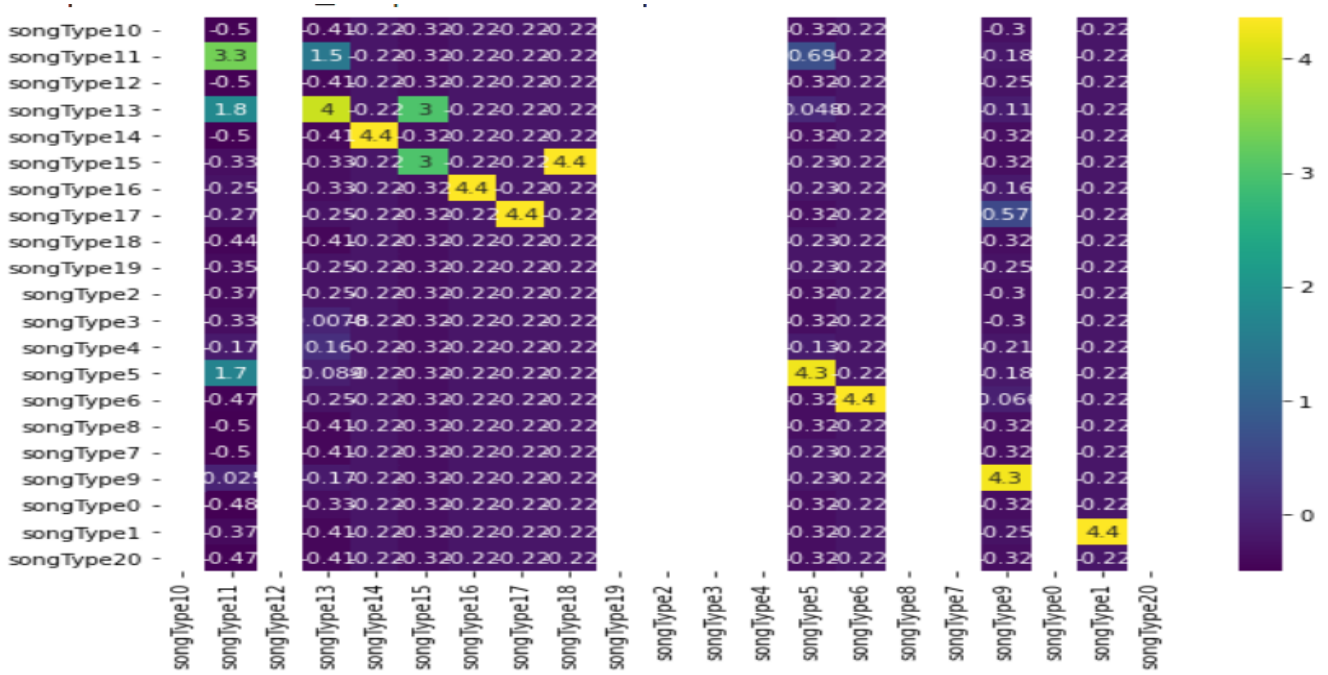

Figure 9

Model summary will go there

1.5. Artificial Neural Network

Artificial Neural Network performance

```
> Cross Accuracy: 0.55 (+/- 0.03)
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         2
           1       0.44      0.73      0.55       308
           2       0.00      0.00      0.00         4
           3       0.44      0.54      0.48       229
           4       0.00      0.00      0.00         2
           5       0.00      0.00      0.00        17
           6       1.00      0.07      0.13        28
           7       1.00      0.08      0.15        62
           8       0.00      0.00      0.00         6
           9       0.00      0.00      0.00        17
          10       0.00      0.00      0.00        13
          11       0.00      0.00      0.00        19
          12       0.00      0.00      0.00        38
          13       0.58      0.39      0.47       213
          14       1.00      0.12      0.21        17
          15       0.00      0.00      0.00         1
          16       0.00      0.00      0.00         2
          17       0.77      0.94      0.85       234
          18       0.00      0.00      0.00         3
          19       1.00      0.07      0.13        14
          20       0.00      0.00      0.00         3

    accuracy                           0.54      1232
   macro avg       0.30      0.14      0.14      1232
weighted avg       0.54      0.54      0.48      1232
```

Figure 10

Confusion Matrix

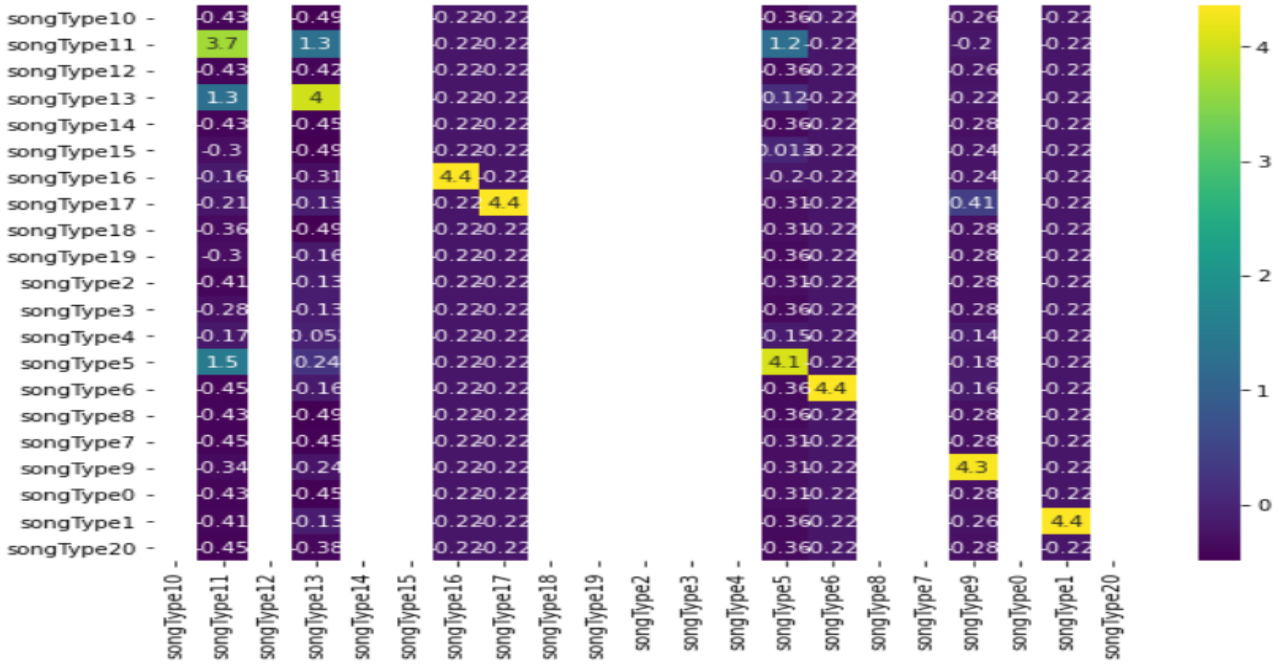

Figure 11

Model summary will go there

2. Why does it happen that the model gives very low f1 score for some classes but not the same for others?

The equation for calculating F1 score is:

$$2 * \frac{precision * recall}{precision + recall}$$

So F1 is score is high when precision and recall both are- high. If precision or recall is low then F1 score will be low.

Now We know precision =

$$\frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

The precision value will be high if False positive is low and precision will be low if False positive is high and vice versa. If False positive is low for a specific song type that means the model doesn't say other type to be this type. So model perfectly predict this song type.

If False positive is high for a specific class or category then model prediction is not good for this category. It says other category to be this category.

Recall:

$$\frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

recall will be high when **False negative** is low and **recall** will be low when **False negative** is high. False negative low means most of the time model can predict correctly a specific category. And few times it says a specific category to other category. If **False negative** is high then model can't predict a specific category correctly.

So F1 score is high when **False positive** and **False negative** bot is low.

3. Can you fix the low f1 score issue?

The main cause for low f1 score is low data. We can see from model result f1 score for some category is zero because we don't have enough data for those category.So to solve this problem we need more data.

Transfer learning can help to increase f1 score.

These procedure have to follow to fix low f1 score

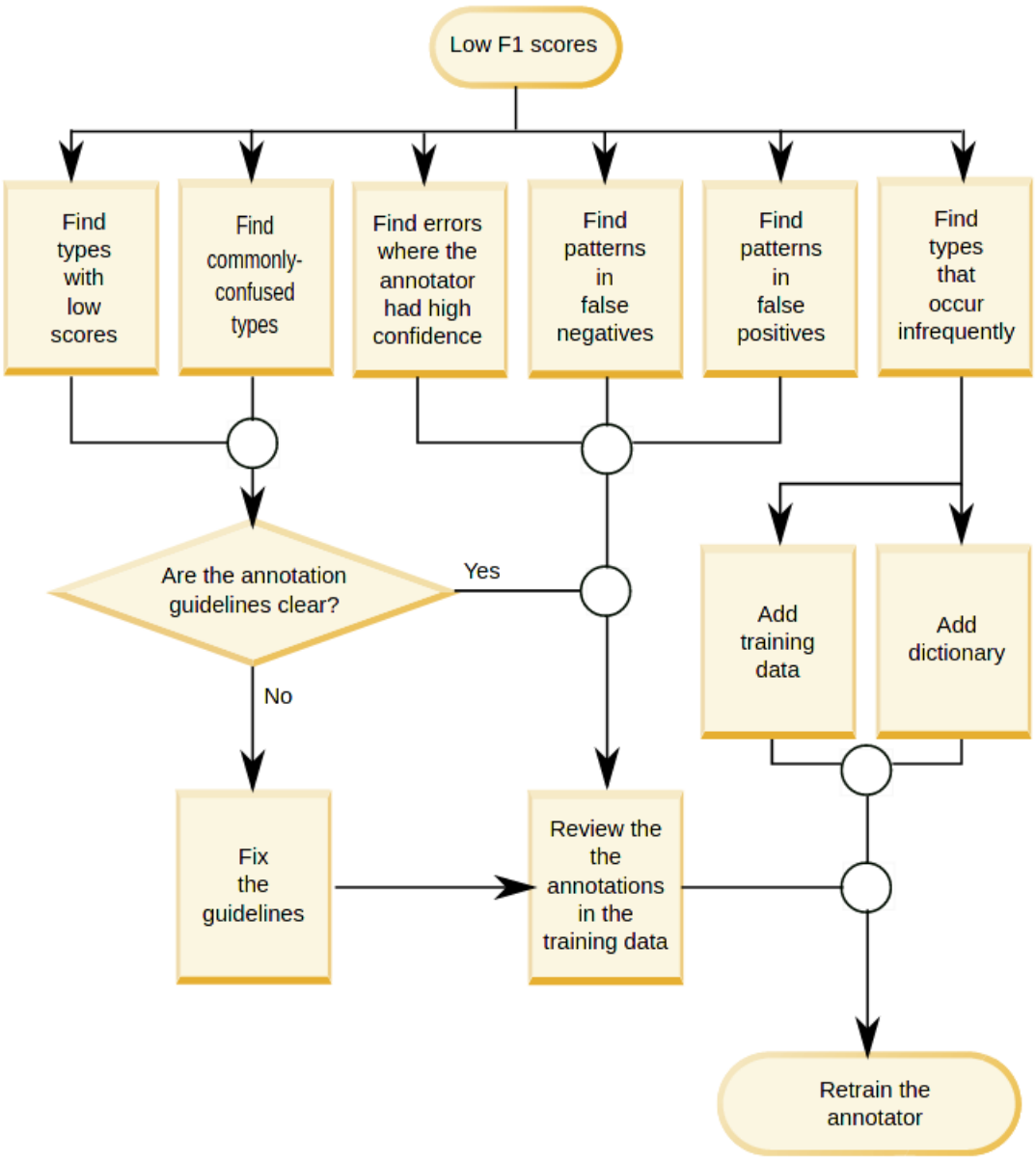

Figure 12