Support Vector Machine (SVM) - BoW_L1

May 4, 2024

1 Initialization

Connect to Google Drive:

[ ]:
```
# from google.colab import drive
# drive.mount('/content/drive')

# %cd '/content/drive/MyDrive/GitHub/emotion-dectection-from-text'
```

Preparing necessary packages (may need to add more):

[ ]:
```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

from sklearn.svm import SVC
from sklearn.model_selection import GridSearchCV, cross_val_score
from sklearn.metrics import accuracy_score
from joblib import dump, load

from preset_function import evaluate_model, draw_learning_curve,␣
 ↪load_processed_data

X_train_bow, X_test_bow, X_train_tfidf, X_test_tfidf, \
    X_train_bow_L1, X_test_bow_L1, X_train_tfidf_L1, X_test_tfidf_L1 =␣
 ↪load_processed_data('input')

y_train, y_test = load_processed_data('output')

%matplotlib inline
```

Select dataset:

[ ]:
```
X_train = X_train_bow_L1
X_test = X_test_bow_L1
```

2 Basic training

We define and train a model with simple hyperparameter in which kernel is linear, $C = 1.0$, etc:

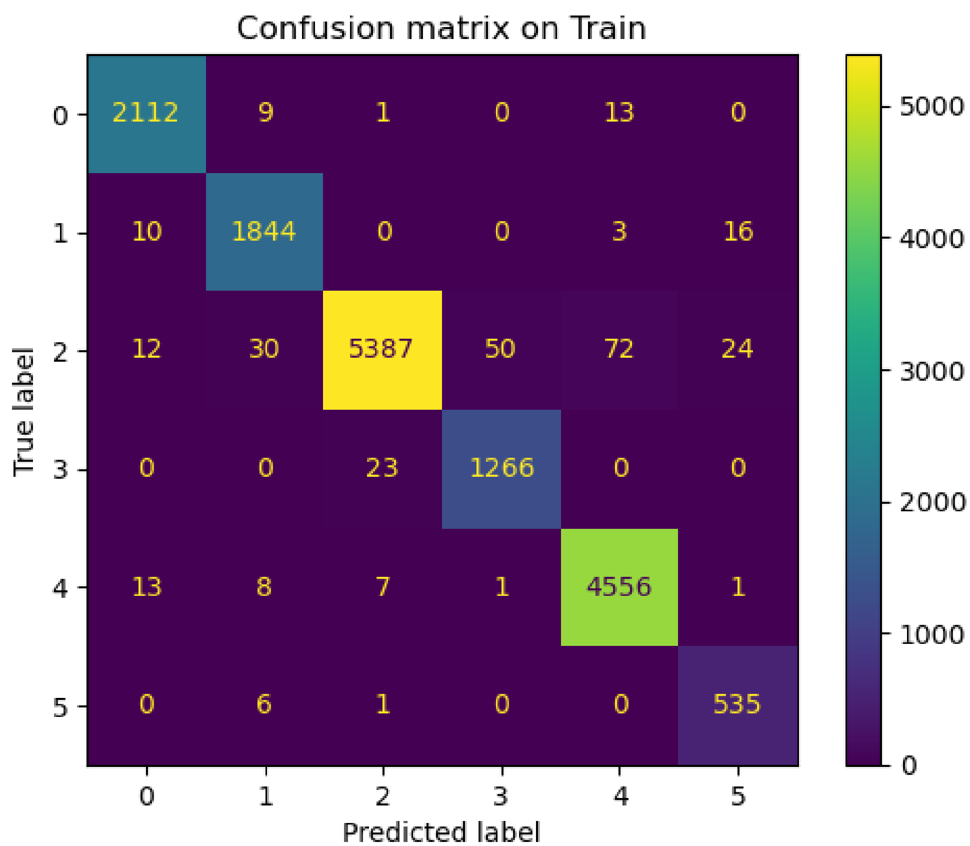
[ ]:
```
svm_model = SVC(kernel='linear')
svm_model.fit(X_train, y_train)
```

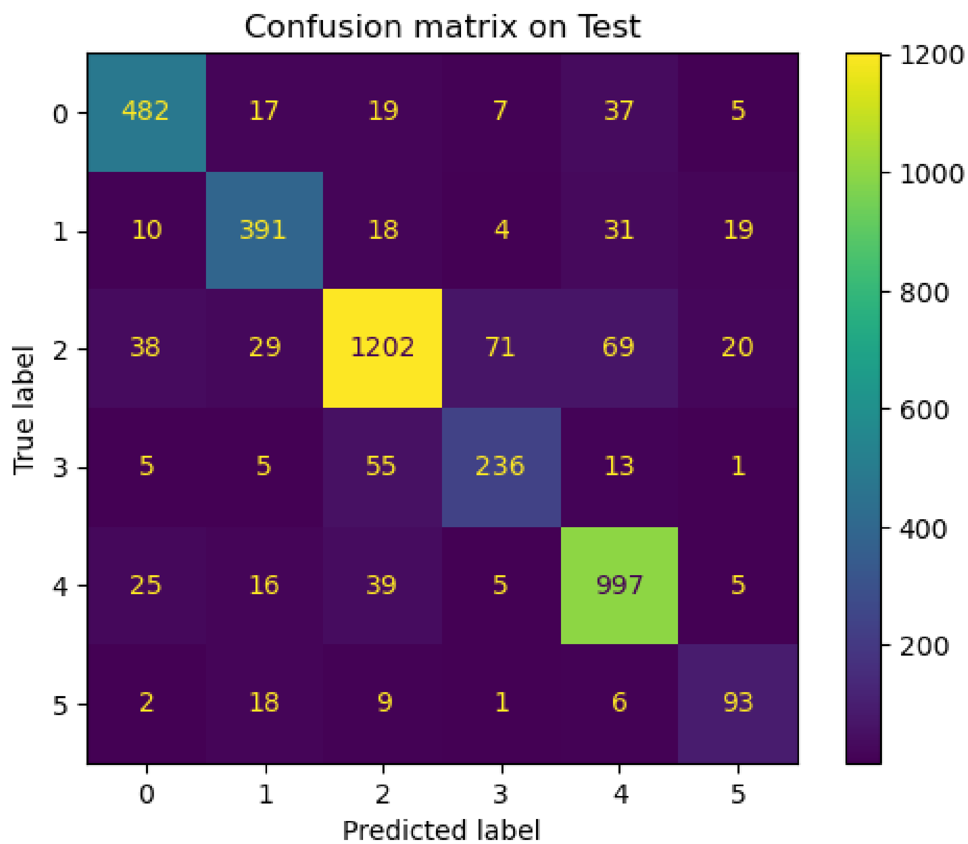
[ ]:
```
SVC(kernel='linear')
```

Evaluate model using preset function:

[ ]:
```
evaluate_model(svm_model, X_train, X_test, y_train, y_test,␣
 ↪include_training=True)
```

```
Score of on train are:
        - Accuracy score: 0.9812
        - Micro F1 score: 0.9812
        - Macro F1 score: 0.9768
Score of on test are:
        - Accuracy score: 0.8502
        - Micro F1 score: 0.8502
        - Macro F1 score: 0.8097
```

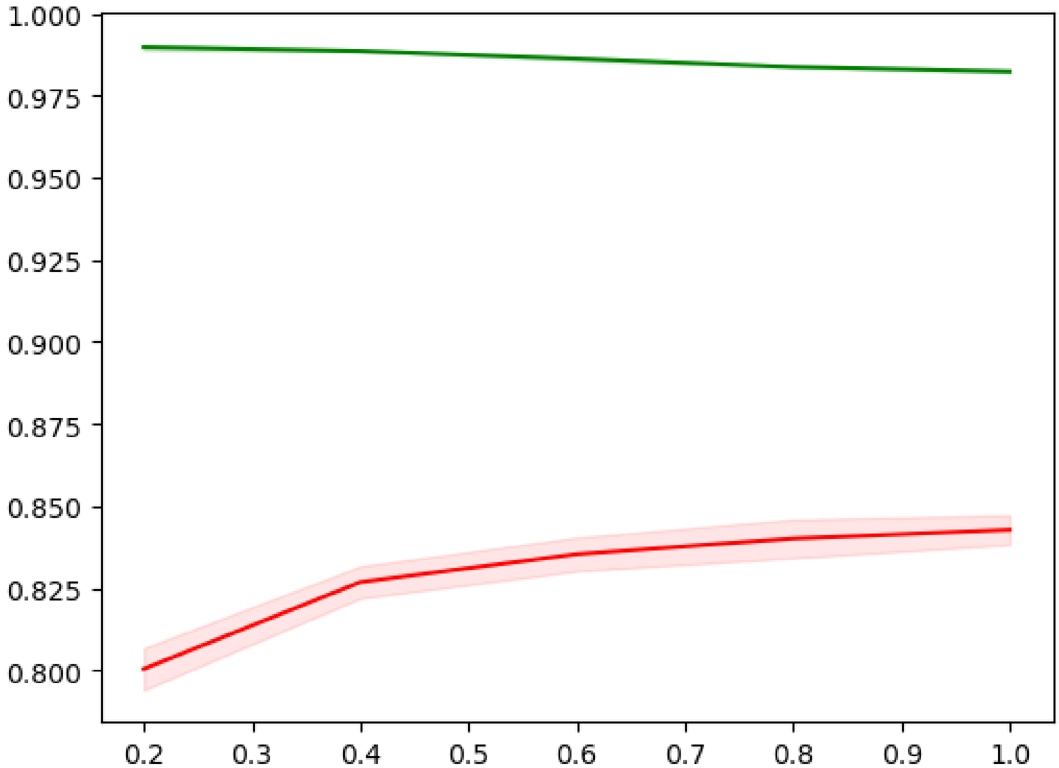




This model gives a pretty good score but it seems to be overfitting.

Draw learning curve using preset function:

[ ]:
```
draw_learning_curve(svm_model, X_train, y_train)
```



Review: SVM can result a better result if we do some hyperparameter tunning to resolve the overfitting

3 Model selection

This section will be separated in 4 parts for 4 kernels instead of using GridSearchCV in order to get the better result

3.1 Linear function kernel

Formula:

$$k(x, z) = x^T z$$

First, we search in a big range from 0.001 to 100

[ ]:
```
C_list = [0.001, 0.01, 0.1, 1, 5, 10, 100]

# Define a list in order to store accuracy points
cvs_list = list()
trs_list = list()

for c in C_list:
```

```
    # Define model for each C
    svm_model = SVC(kernel='linear', C=c)
    svm_model.fit(X_train, y_train)

    # Calculate score of cross validation
    train_score = accuracy_score(y_train, svm_model.predict(X_train))
    cv_score = np.mean(cross_val_score(svm_model, X_train, y_train, cv=5,␣
↪n_jobs=8))

    trs_list.append(train_score)
    cvs_list.append(cv_score)
```

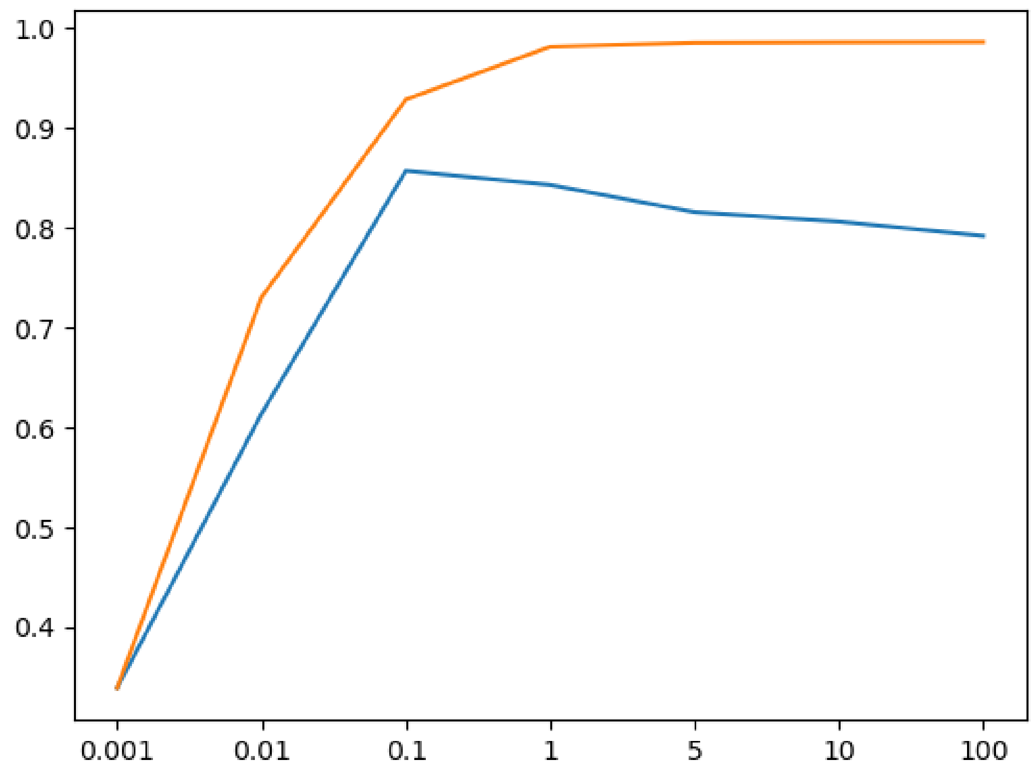
[ ]:
```
# Print the result
print(C_list)
print(trs_list)
print(cvs_list)

# Draw the plot
fig = sns.lineplot(x=list(range(len(C_list))), y=cvs_list)
fig = sns.lineplot(x=list(range(len(C_list))), y=trs_list)
fig.set_xticks(range(len(C_list)))
fig.set_xticklabels(C_list)
```

```
[0.001, 0.01, 0.1, 1, 5, 10, 100]
[0.3386875, 0.73025, 0.928375, 0.98125, 0.9853125, 0.9858125, 0.986125]
[0.3386875000000001, 0.6134375000000001, 0.8569374999999999, 0.84275,
0.8154999999999999, 0.806, 0.79175]
```

```
c:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
c:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
c:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
c:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

[ ]:
```
[Text(0, 0, '0.001'),
 Text(1, 0, '0.01'),
 Text(2, 0, '0.1'),
```

```
 Text(3, 0, '1'),
 Text(4, 0, '5'),
 Text(5, 0, '10'),
 Text(6, 0, '100')]
```



From the result of above section, we can see the good value of C is near the value 0.1.

Scope to $C = 0.1$:

[ ]:
```
C_list = [0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75]

# Define a list in order to store accuracy points
cvs_list = list()
trs_list = list()

for c in C_list:
    # Define model for each C
    svm_model = SVC(kernel='linear', C=c)
    svm_model.fit(X_train, y_train)

    # Calculate score of cross validation
    train_score = accuracy_score(y_train, svm_model.predict(X_train))
```

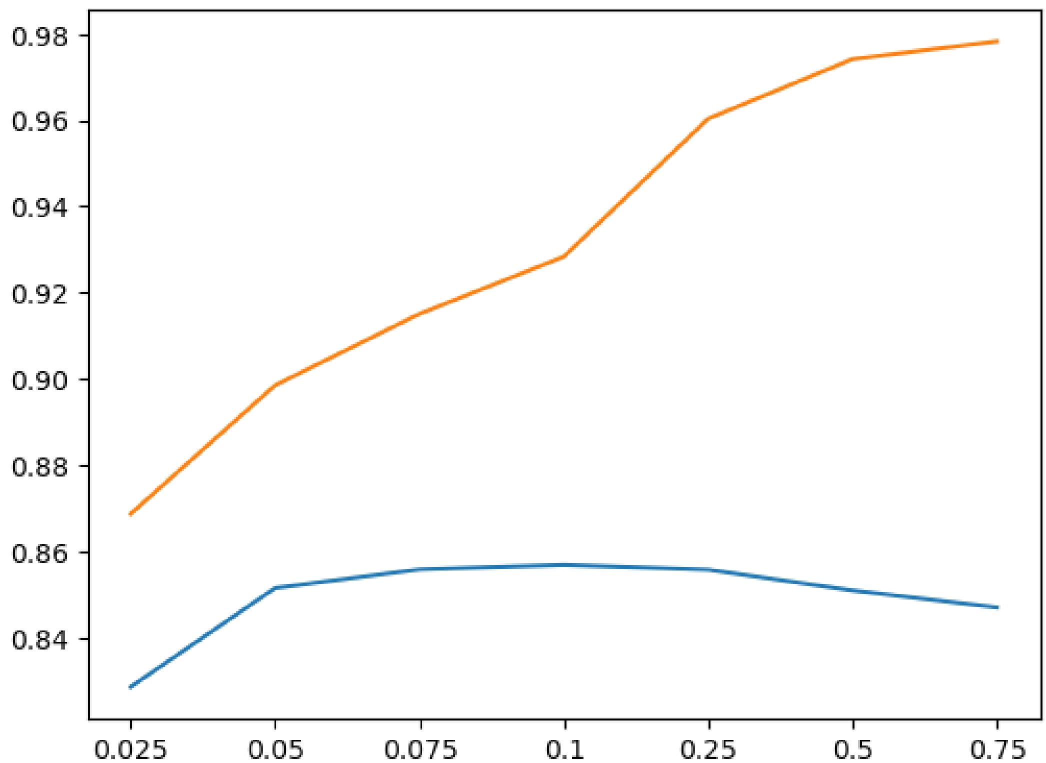
```
    cv_score = np.mean(cross_val_score(svm_model, X_train, y_train, cv=5,␣
↪n_jobs=8))

    trs_list.append(train_score)
    cvs_list.append(cv_score)
```

[ ]:
```
# Print the result
print(C_list)
print(trs_list)
print(cvs_list)

# Draw the plot
fig = sns.lineplot(x=list(range(len(C_list))), y=cvs_list)
fig = sns.lineplot(x=list(range(len(C_list))), y=trs_list)
fig.set_xticks(range(len(C_list)))
fig.set_xticklabels(C_list)
```

```
[0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75]
[0.86875, 0.8985625, 0.915125, 0.928375, 0.960375, 0.9741875, 0.9783125]
[0.828625, 0.8515625, 0.8558749999999999, 0.8569374999999999,
0.8558125000000001, 0.851, 0.8470625]
```

```
c:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
c:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
c:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
c:\ProgramData\anaconda3\Lib\site-packages\seaborn\_oldcore.py:1119:
FutureWarning: use_inf_as_na option is deprecated and will be removed in a
future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```

[ ]:
```
[Text(0, 0, '0.025'),
 Text(1, 0, '0.05'),
 Text(2, 0, '0.075'),
 Text(3, 0, '0.1'),
 Text(4, 0, '0.25'),
 Text(5, 0, '0.5'),
 Text(6, 0, '0.75')]
```

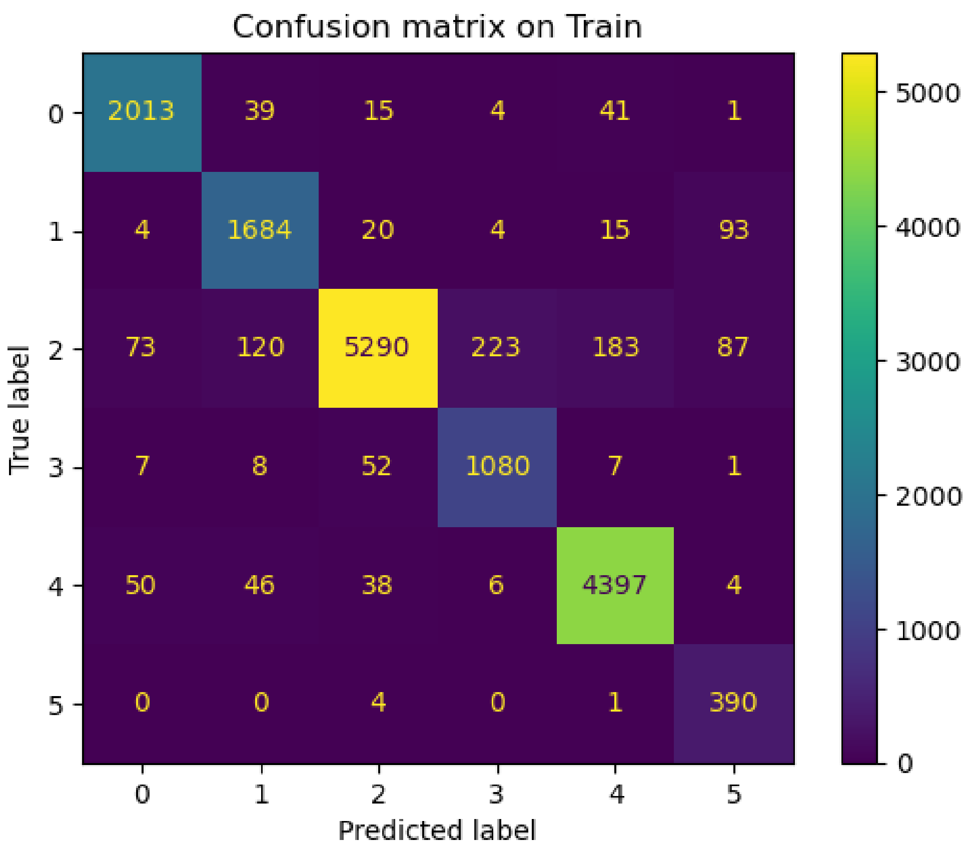


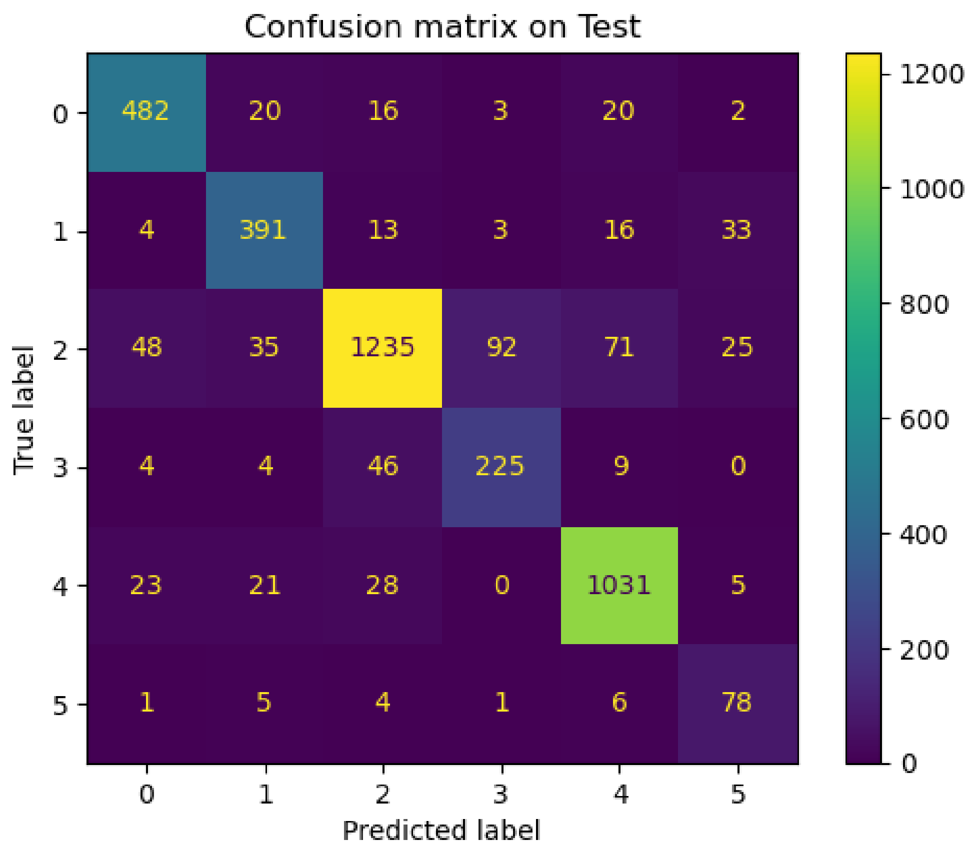
As the result, we can claim that $C = 0.1$ give a model with good accuracy and avoid overfitting. We will test the model again in test set.

[ ]:
```
best_svm_linear_model = SVC(kernel='linear', C=0.1)
```

[ ]:
```
best_svm_linear_model.fit(X_train, y_train)
evaluate_model(best_svm_linear_model, X_train, X_test, y_train, y_test,␣
 ↪include_training=True)
```

```
Score of on train are:
        - Accuracy score: 0.9284
        - Micro F1 score: 0.9284
        - Macro F1 score: 0.9024
Score of on test are:
        - Accuracy score: 0.8605
        - Micro F1 score: 0.8605
        - Macro F1 score: 0.8130
```


Confusion matrix on Train
True label
Predicted label
0
1
2
3
4
5
2013
39
15
4
41
1
4
1684
20
4
15
93
73
120
5290
223
183
87
7
8
52
1080
7
1
50
46
38
6
4397
4
0
0
4
0
1
390
5000
4000
3000
2000
1000
0



3.2 Radial basis function kernel

Formula:

$$k(x, z) = e^{-\gamma ||x-z||_2^2}$$

First, we using grid search in a big domain.

[ ]:
```
dict_param = {
    'C' : np.asarray([0.01, 0.1, 1, 10.0, 100]),
    'gamma': np.logspace(-3, 2, 6)
}

grid_search = GridSearchCV(SVC(kernel='rbf'), dict_param, cv = 5, n_jobs=8)
grid_search.fit(X_train, y_train)
```

[ ]:
```
GridSearchCV(cv=5, estimator=SVC(), n_jobs=8,
             param_grid={'C': array([1.e-02, 1.e-01, 1.e+00, 1.e+01, 1.e+02]),
                         'gamma': array([1.e-03, 1.e-02, 1.e-01, 1.e+00, 1.e+01,
1.e+02])})
```

[ ]:
```
print('Best score: ', grid_search.best_score_, '\n')

print('Bad hyperparameter:')
df = pd.DataFrame(
  dict(
    C = [val['C'] for val in grid_search.cv_results_['params']],
    gamma = [val['gamma'] for val in grid_search.cv_results_['params']],
    score = grid_search.cv_results_['mean_test_score']
  )
)
df = df[df['score'] < 0.80]
for param in dict_param:
  for value in dict_param[param]:
    if len(df[df[param] == value]) == 30 // len(dict_param[param]):
      print(param, value)
```

```
Best score:  0.8570625

Bad hyperparameter:
C 0.01
C 0.1
gamma 1.0
gamma 10.0
gamma 100.0
```

We fiter all the parameter that appear in all the bad model (validation accuracy < 0.8) * C $= 0.01$ * C $= 0.1$ * $\gamma = 100.0$ * $\gamma = 10.0$ * $\gamma = 1.0$

So that we can shrink the range of parameter

We repeat the algorithm again and again until there is no bad parameter to recieve the best model

[ ]:
```
dict_param = {
    'C' : np.linspace(10, 100, 10),
    'gamma': np.logspace(-3, -2, 10)
}

grid_search = GridSearchCV(SVC(kernel='rbf'), dict_param, cv = 5, n_jobs=8)
grid_search.fit(X_train, y_train)
```

[ ]:
```
GridSearchCV(cv=5, estimator=SVC(), n_jobs=8,
             param_grid={'C': array([ 10.,  20.,  30.,  40.,  50.,  60.,  70.,
80.,  90., 100.]),
                         'gamma': array([0.001     , 0.00129155, 0.0016681 ,
0.00215443, 0.00278256,
       0.00359381, 0.00464159, 0.00599484, 0.00774264, 0.01      ])})
```

[ ]:
```
df = pd.DataFrame(
  dict(
```

```
        C = [val['C'] for val in grid_search.cv_results_['params']],
        gamma = [val['gamma'] for val in grid_search.cv_results_['params']],
        score = grid_search.cv_results_['mean_test_score']
    )
)
print(df[df['score'] == min(df['score'])])
```

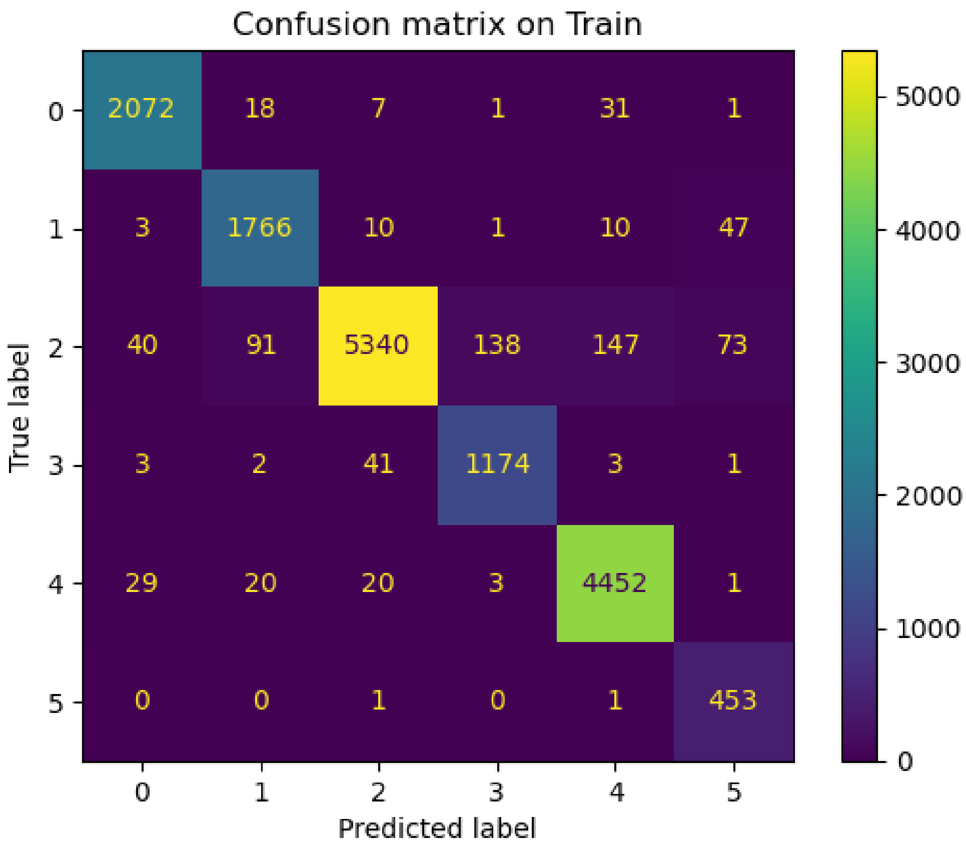
```
      C  gamma     score
0  10.0  0.001  0.800125
```

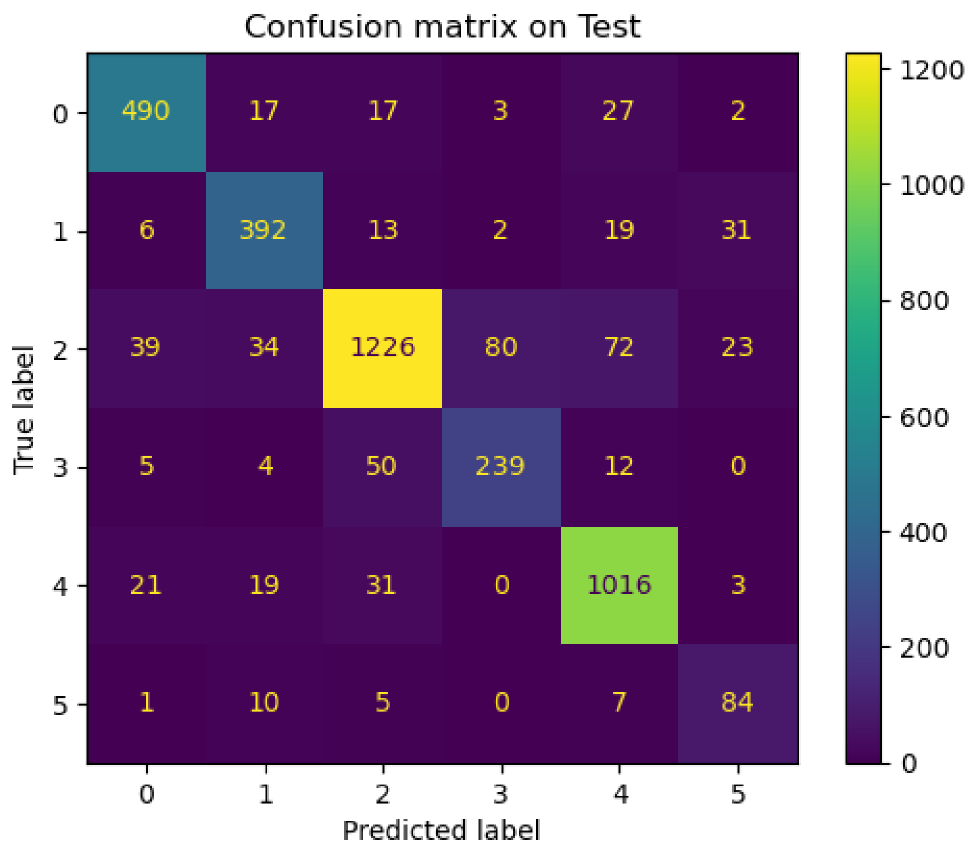
[ ]:
```
print(grid_search.best_estimator_, grid_search.best_score_)
```

```
SVC(C=60.0, gamma=0.0016681005372000592) 0.8571875
```

[ ]:
```
best_svm_rbf_model = SVC(C=60.0, gamma=0.0016681005372000592)
best_svm_rbf_model.fit(X_train, y_train)
evaluate_model(best_svm_rbf_model, X_train, X_test, y_train, y_test,␣
 ↪include_training=True)
```

```
Score of on train are:
        - Accuracy score: 0.9536
        - Micro F1 score: 0.9536
        - Macro F1 score: 0.9397
Score of on test are:
        - Accuracy score: 0.8618
        - Micro F1 score: 0.8618
        - Macro F1 score: 0.8190
```


Confusion matrix on Train
2072
18
7
1
31
1
3
1766
10
1
10
47
40
91
5340
138
147
73
3
2
41
1174
3
1
29
20
20
3
4452
1
0
0
1
0
1
453
True label
Predicted label
0
1
2
3
4
5
5000
4000
3000
2000
1000
0



3.3 Sigmoid function kernel

Formula:

$$k(x, z) = tanh(\gamma x^T z + r)$$

We use the same method in the above section to tuning this kernel

[ ]:
```
dict_param = {
    'C' : np.asarray([0.001, 0.01, 0.1, 1, 10.0, 100]),
    'gamma': np.asarray([0.001, 0.01, 0.1, 1, 10.0, 100]),
    'coef0': np.asarray([0.001, 0.01, 0.1, 1, 10.0, 100])
}

grid_search = GridSearchCV(SVC(kernel='sigmoid'), dict_param, cv = 5, n_jobs=8)
grid_search.fit(X_train, y_train)
```

[ ]:
```
GridSearchCV(cv=5, estimator=SVC(kernel='sigmoid'), n_jobs=8,
             param_grid={'C': array([1.e-03, 1.e-02, 1.e-01, 1.e+00, 1.e+01,
1.e+02]),
                         'coef0': array([1.e-03, 1.e-02, 1.e-01, 1.e+00, 1.e+01,
```

```
1.e+02]),
                            'gamma': array([1.e-03, 1.e-02, 1.e-01, 1.e+00, 1.e+01,
1.e+02])})
```

[ ]:
```python
print('Best score: ', grid_search.best_score_, '\n')

print('Bad hyperparameter:')
df = pd.DataFrame(
  dict(
    C = [val['C'] for val in grid_search.cv_results_['params']],
    gamma = [val['gamma'] for val in grid_search.cv_results_['params']],
    coef0 = [val['coef0'] for val in grid_search.cv_results_['params']],
    score = grid_search.cv_results_['mean_test_score']
  )
)
df = df[df['score'] < 0.8]

for param in dict_param:
  for value in dict_param[param]:
    if len(df[df[param] == value]) == 6 * 6:
      print(param, value)
```

```
Best score:  0.8568749999999999

Bad hyperparameter:
C 0.001
C 0.01
C 0.1
gamma 1.0
gamma 10.0
gamma 100.0
coef0 10.0
coef0 100.0
```

[ ]:
```python
dict_param = {
    'C' : np.linspace(1, 100, 5),
    'gamma': np.linspace(0.001, 0.01, 5),
    'coef0': np.linspace(0.001, 1, 5)
}

grid_search = GridSearchCV(SVC(kernel='sigmoid'), dict_param, cv = 5, n_jobs=8)
grid_search.fit(X_train, y_train)
```

[ ]:
```
GridSearchCV(cv=5, estimator=SVC(kernel='sigmoid'), n_jobs=8,
             param_grid={'C': array([  1.   ,  25.75,  50.5 ,  75.25, 100.  ]),
                         'coef0': array([0.001   , 0.25075, 0.5005 , 0.75025, 1.
]),
                         'gamma': array([0.001   , 0.00325, 0.0055 , 0.00775,
```

```
 0.01   ])})
```

[ ]:
```
df = pd.DataFrame(
  dict(
    C = [val['C'] for val in grid_search.cv_results_['params']],
    gamma = [val['gamma'] for val in grid_search.cv_results_['params']],
    coef0 = [val['coef0'] for val in grid_search.cv_results_['params']],
    score = grid_search.cv_results_['mean_test_score']
  )
)
df = df[df['score'] < 0.8]
print(len(df))
for param in dict_param:
  for value in dict_param[param]:
    if len(df[df[param] == value]) == 125 // len(dict_param[param]):
      print(param, value)
```

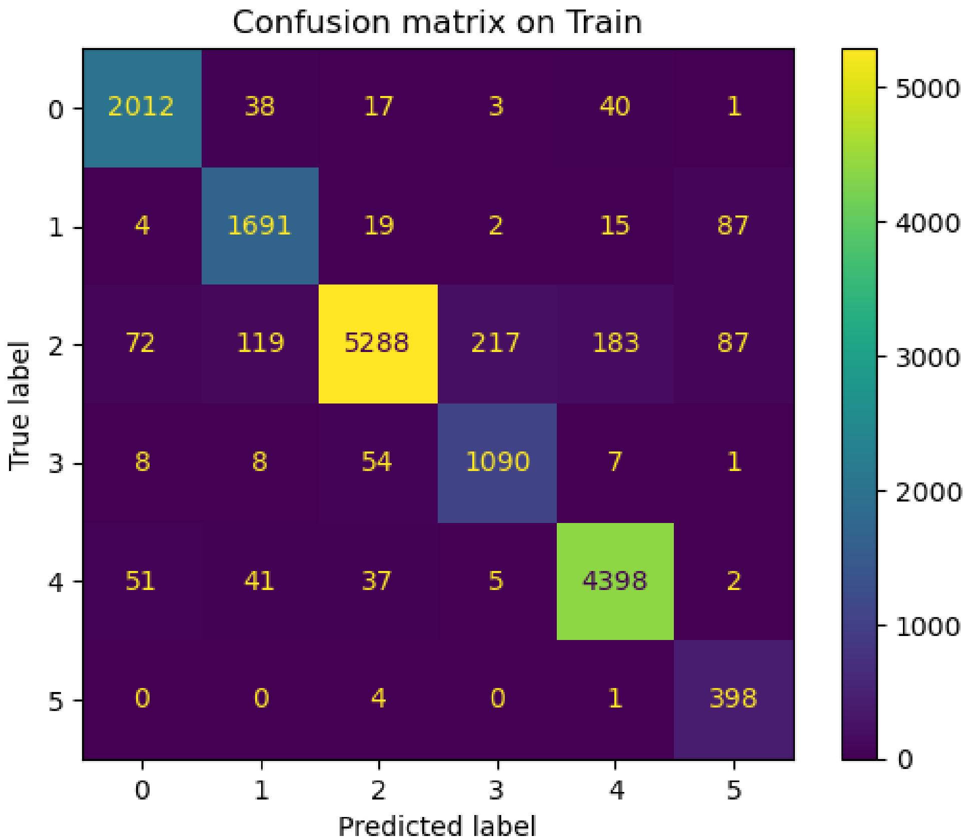
```
27
```

[ ]:
```
print(grid_search.best_estimator_, grid_search.best_score_)
```

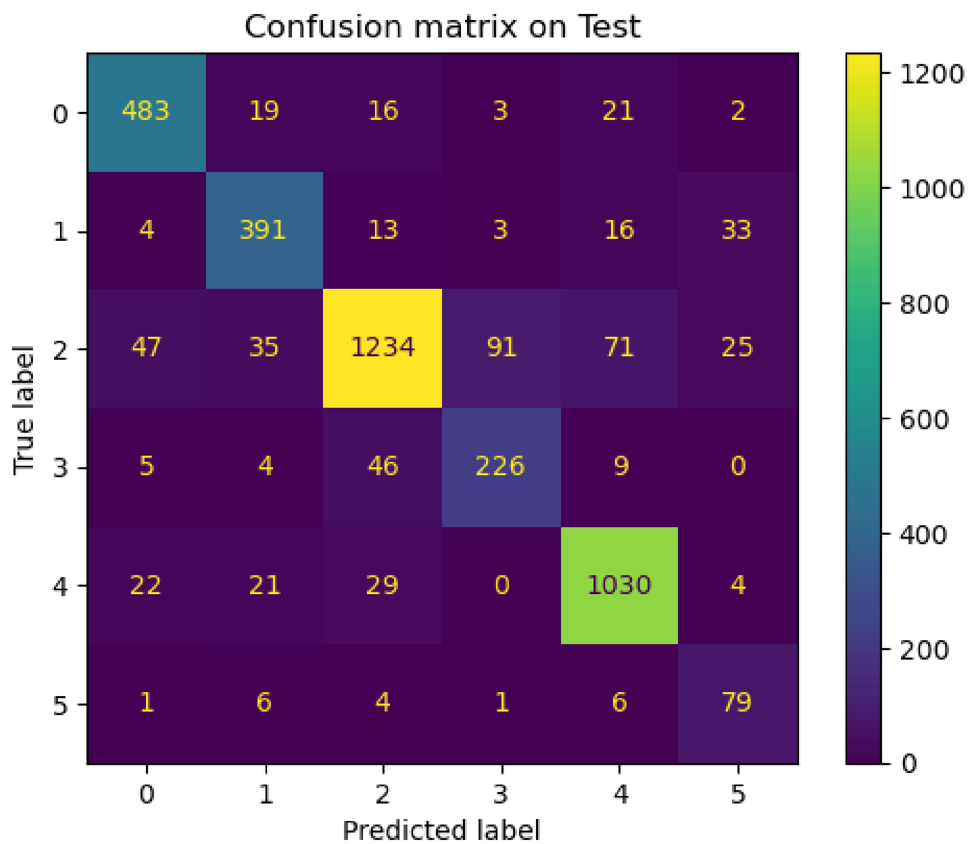
```
SVC(C=25.75, coef0=0.5005, gamma=0.0055000000000000005, kernel='sigmoid')
0.858125
```

[ ]:
```
best_svm_sig_model = SVC(C=25.75, coef0=0.5005, gamma=0.0055000000000000005,␣
 ↪kernel='sigmoid')
best_svm_sig_model.fit(X_train, y_train)
evaluate_model(best_svm_sig_model, X_train, X_test, y_train, y_test,␣
 ↪include_training=True)
```

```
Score of on train are:
        - Accuracy score: 0.9298
        - Micro F1 score: 0.9298
        - Macro F1 score: 0.9055
Score of on test are:
        - Accuracy score: 0.8608
        - Micro F1 score: 0.8608
        - Macro F1 score: 0.8138
```



Confusion matrix on Train
True label
Predicted label
0
1
2
3
4
5
2012 38 17 3 40 1
4 1691 19 2 15 87
72 119 5288 217 183 87
8 8 54 1090 7 1
51 41 37 5 4398 2
0 0 4 0 1 398
5000
4000
3000
2000
1000
0



3.4 Polynomial function kernel

Formula:

$$k(x, z) = (r + \gamma x^T z)^d$$

[ ]:
```
dict_param = {
    'C' : np.asarray([0.001, 0.01, 0.1, 1, 10.0, 100]),
    'gamma': np.asarray([0.001, 0.01, 0.1, 1]),
    'coef0': np.asarray([0.001, 0.01, 0.1, 1]),
    'degree': np.asarray([2, 3, 4])
}

grid_search = GridSearchCV(SVC(kernel='poly'), dict_param, cv = 5, n_jobs=8)
grid_search.fit(X_train, y_train)
```

[ ]:
```
GridSearchCV(cv=5, estimator=SVC(kernel='poly'), n_jobs=8,
             param_grid={'C': array([1.e-03, 1.e-02, 1.e-01, 1.e+00, 1.e+01,
1.e+02]),
                         'coef0': array([0.001, 0.01 , 0.1  , 1.   ]),
                         'degree': array([2, 3, 4]),
```

```
                         'gamma': array([0.001, 0.01 , 0.1  , 1.   ])})
```

[ ]:
```python
print('Best score: ', grid_search.best_score_, '\n')

print('Bad hyperparameter:')
df = pd.DataFrame(
  dict(
    C = [val['C'] for val in grid_search.cv_results_['params']],
    gamma = [val['gamma'] for val in grid_search.cv_results_['params']],
    coef0 = [val['coef0'] for val in grid_search.cv_results_['params']],
    degree = [val['degree'] for val in grid_search.cv_results_['params']],
    score = grid_search.cv_results_['mean_test_score']
  )
)
df = df[df['score'] < 0.85]

for param in dict_param:
  for value in dict_param[param]:
    if len(df[df[param] == value]) == 288 // len(dict_param[param]):
      print(param, value)
```

```
Best score:  0.8566874999999999

Bad hyperparameter:
C 0.001
C 0.01
C 0.1
gamma 1.0
coef0 0.001
coef0 0.01
```

[ ]:
```python
dict_param = {
    'C' : np.linspace(10, 100, 4),
    'gamma': np.linspace(0.001, 0.004, 4),
    'coef0': np.linspace(0.7, 1, 4),
    'degree': np.asarray([2, 3, 4])
}

grid_search = GridSearchCV(SVC(kernel='poly'), dict_param, cv = 5, n_jobs=8)
grid_search.fit(X_train, y_train)
```

[ ]:
```
GridSearchCV(cv=5, estimator=SVC(kernel='poly'), n_jobs=8,
             param_grid={'C': array([ 10.,  40.,  70., 100.]),
                         'coef0': array([0.7, 0.8, 0.9, 1. ]),
                         'degree': array([2, 3, 4]),
                         'gamma': array([0.001, 0.002, 0.003, 0.004])})
```

[ ]:
```
print('Best score: ', grid_search.best_score_, '\n')

print('Bad hyperparameter:')
df = pd.DataFrame(
  dict(
    C = [val['C'] for val in grid_search.cv_results_['params']],
    gamma = [val['gamma'] for val in grid_search.cv_results_['params']],
    coef0 = [val['coef0'] for val in grid_search.cv_results_['params']],
    degree = [val['degree'] for val in grid_search.cv_results_['params']],
    score = grid_search.cv_results_['mean_test_score']
  )
)
df = df[df['score'] < 0.80]
print('Number of filtered models:', len(df))

for param in dict_param:
  for value in dict_param[param]:
    if len(df[df[param] == value]) == 192 // len(dict_param[param]):
      print(param, value)
```

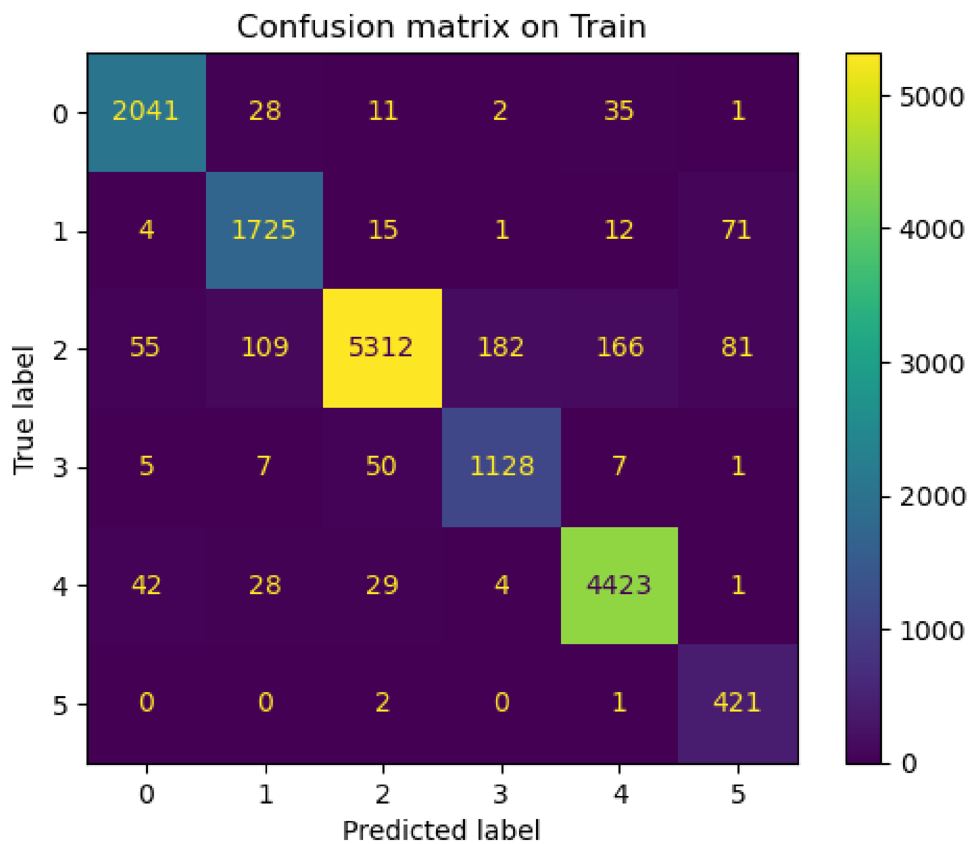
```
Best score:  0.858125

Bad hyperparameter:
Number of filtered models: 6
```

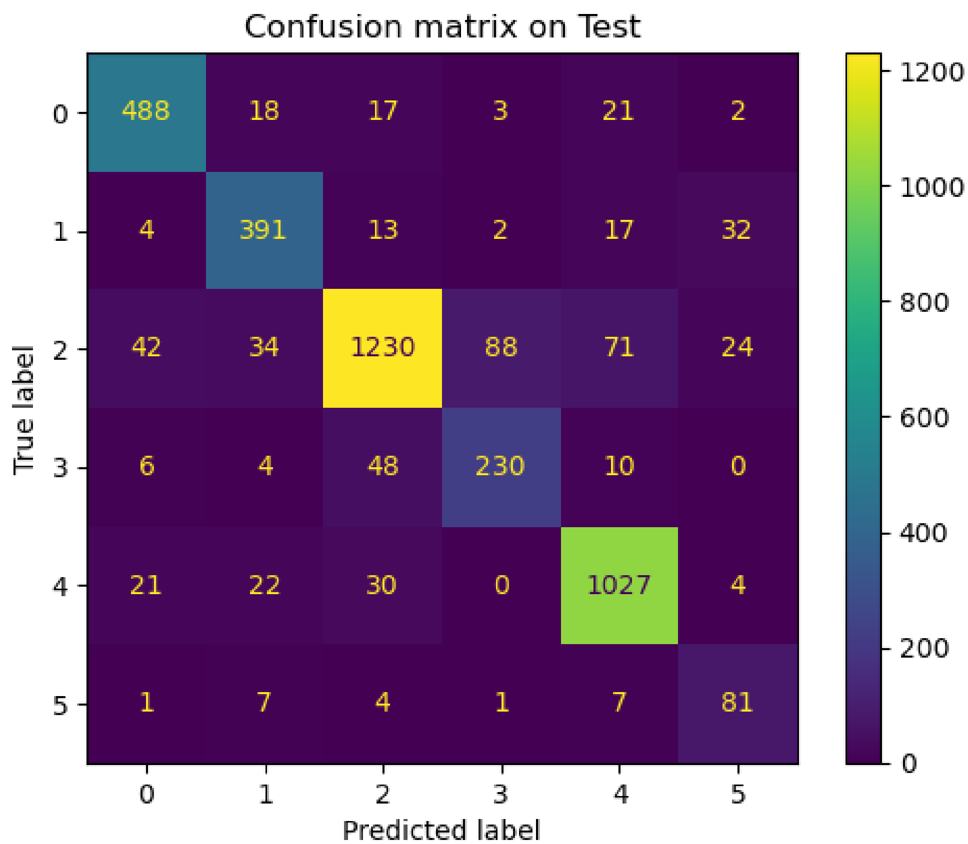
[ ]:
```
print(grid_search.best_estimator_, grid_search.best_score_)
```

```
SVC(C=70.0, coef0=0.7999999999999999, gamma=0.001, kernel='poly') 0.858125
```

[ ]:
```
best_svm_poly_model = SVC(C=70.0, coef0=0.7999999999999999, gamma=0.001,
 ↪kernel='poly')
best_svm_poly_model.fit(X_train, y_train)
evaluate_model(best_svm_poly_model, X_train, X_test, y_train, y_test,
 ↪include_training=True)
```

```
Score of on train are:
        - Accuracy score: 0.9406
        - Micro F1 score: 0.9406
        - Macro F1 score: 0.9208
Score of on test are:
        - Accuracy score: 0.8618
        - Micro F1 score: 0.8618
        - Macro F1 score: 0.8162
```


Confusion matrix on Train
True label
Predicted label
0
1
2
3
4
5
2041
28
11
2
35
1
4
1725
15
1
12
71
55
109
5312
182
166
81
5
7
50
1128
7
1
42
28
29
4
4423
1
0
0
2
0
1
421
5000
4000
3000
2000
1000
0



4 Conclusion

All the kernels have almost the same result. There is only a few different in training score and macro F1

From the result, I choose rbf kernel to be the best one in this dataset.

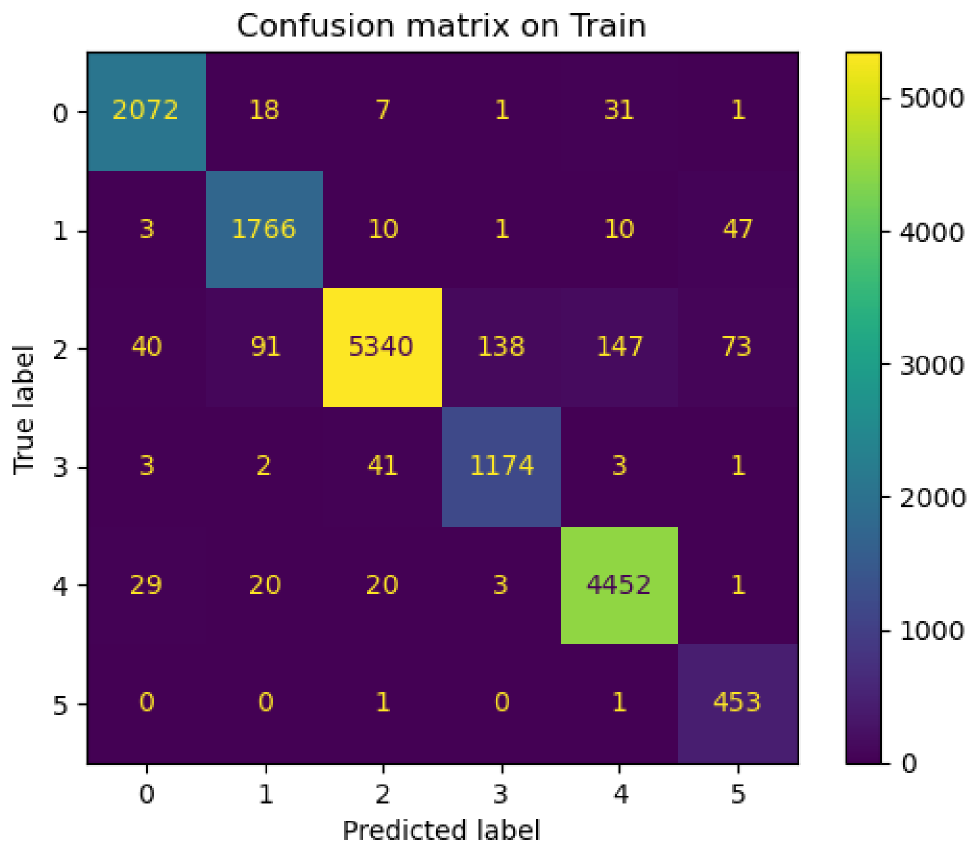
[ ]:
```
best_svm_model = best_svm_rbf_model
```

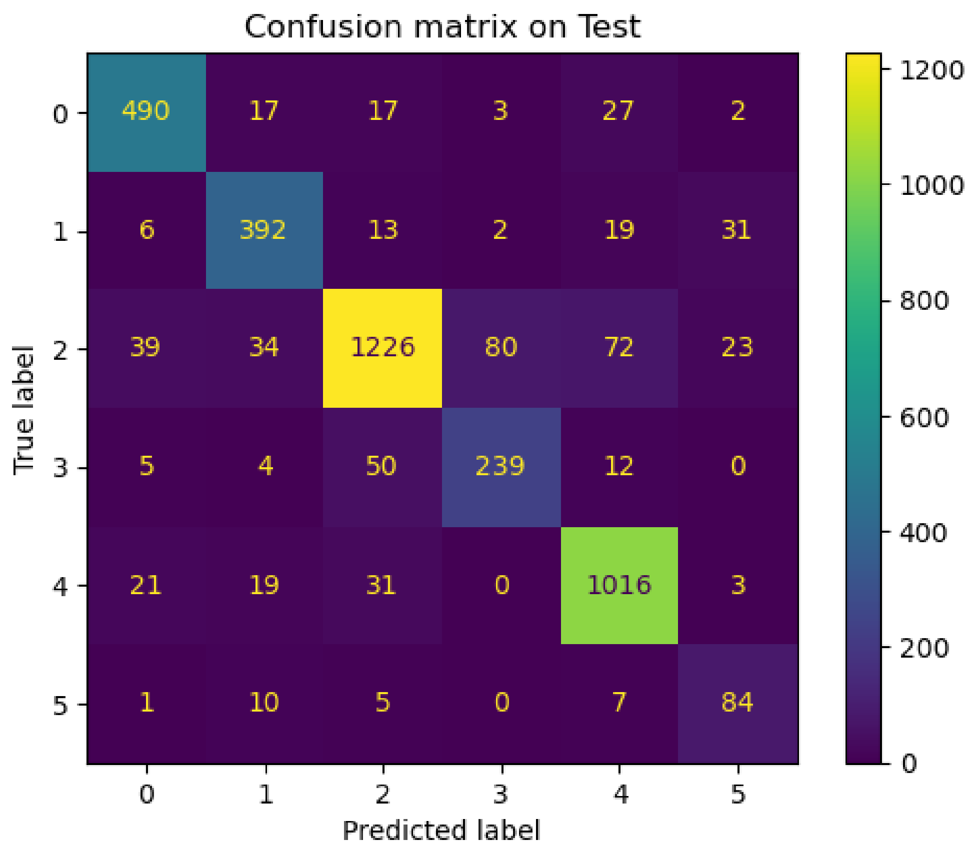
Evaluate the model:

[ ]:
```
evaluate_model(best_svm_model, X_train, X_test, y_train, y_test,␣
↪include_training=True)
```

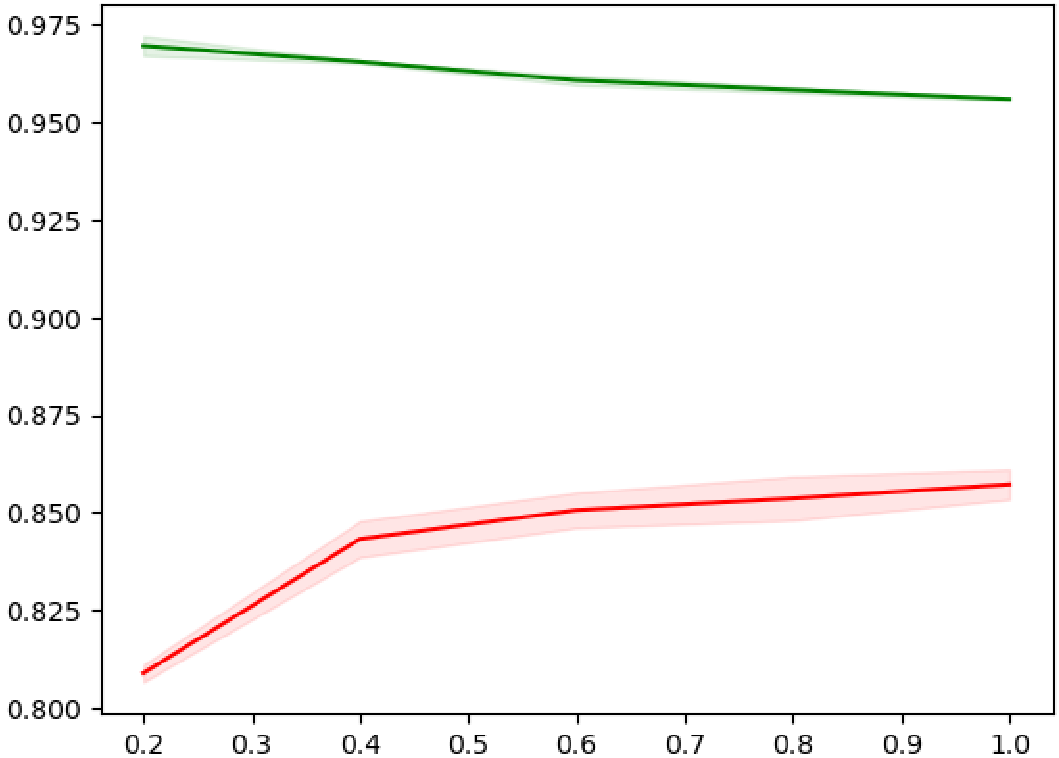
```
Score of on train are:
        - Accuracy score: 0.9536
        - Micro F1 score: 0.9536
        - Macro F1 score: 0.9397
Score of on test are:
        - Accuracy score: 0.8618
        - Micro F1 score: 0.8618
```

```
- Macro F1 score: 0.8190
```





[ ]:
```
draw_learning_curve(best_svm_model, X_train, y_train)
```



[ ]:
```
directory = "data/models/svm/"

dump(best_svm_model, directory + "best_svm_bow_l1_model.joblib")
```

[ ]:
```
['data/models/svm/best_svm_bow_l1_model.joblib']
```