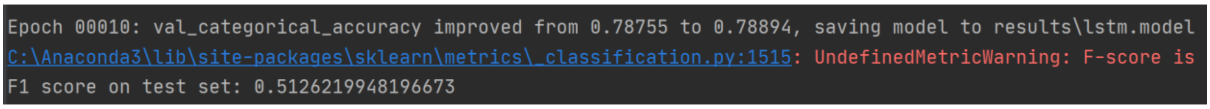
Task 1

Comparing score distributions could avoid a high chance of drawing false conclusions of single performance scores, which is due to different local minima in cost function optimization.

.Task 2.3

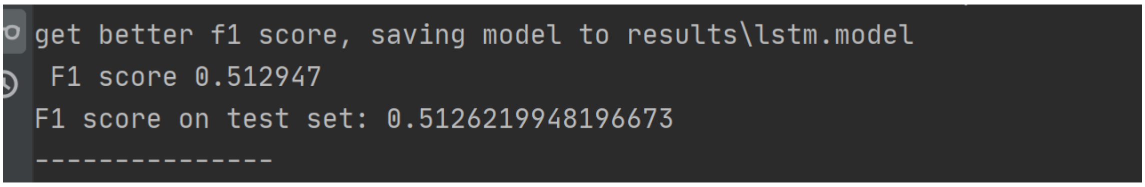
```
Epoch 00001: val_categorical_accuracy improved from -inf to 0.69575, saving model to results\lstm.model
Epoch 00002: val_categorical_accuracy improved from 0.69575 to 0.73184, saving model to results\lstm.model
Epoch 00003: val_categorical_accuracy improved from 0.73184 to 0.75126, saving model to results\lstm.model
Epoch 00004: val_categorical_accuracy improved from 0.75126 to 0.76122, saving model to results\lstm.model
Epoch 00005: val_categorical_accuracy improved from 0.76122 to 0.76956, saving model to results\lstm.model
Epoch 00006: val_categorical_accuracy improved from 0.76956 to 0.77668, saving model to results\lstm.model
Epoch 00007: val_categorical_accuracy improved from 0.77668 to 0.78202, saving model to results\lstm.model
Epoch 00008: val_categorical_accuracy improved from 0.78202 to 0.78540, saving model to results\lstm.model
Epoch 00009: val_categorical_accuracy improved from 0.78540 to 0.78755, saving model to results\lstm.model
Epoch 00010: val_categorical_accuracy improved from 0.78755 to 0.78894, saving model to results\lstm.model
F1 score on test set: 0.5126219948196673
```



```
Epoch 00010: val_categorical_accuracy improved from 0.78755 to 0.78894, saving model to results\lstm.model
C:\Anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1515: UndefinedMetricWarning: F-score is
F1 score on test set: 0.5126219948196673
```

Task 2.4

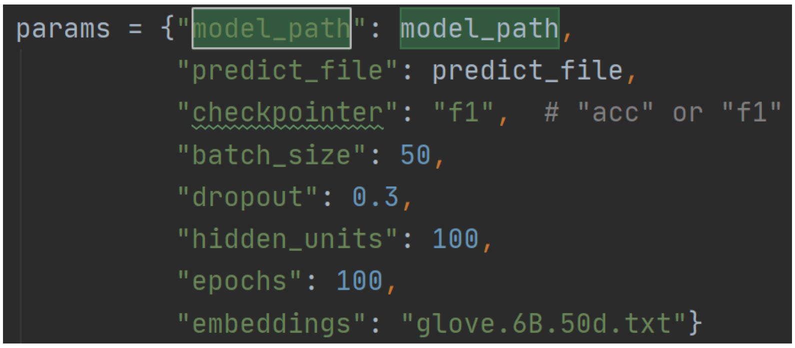
If labels of datasets are extreme (most of labels belong to several taggings, the rest labels supply only a small proportion), categorical_accuracy could not give a comprehensive metric. In this case f1 score performs better, because it is not sensitive to the distributing of labels.



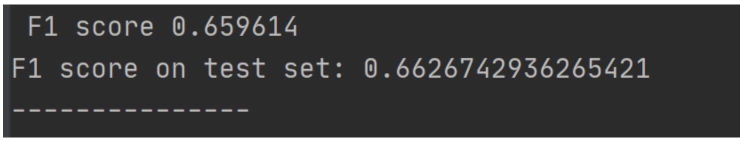
```
get better f1 score, saving model to results\lstm.model
 F1 score 0.512947
F1 score on test set: 0.5126219948196673
---------------
```

From result the callback function with F1 score doesn’t show any difference from categorical_accuracy(callback function).

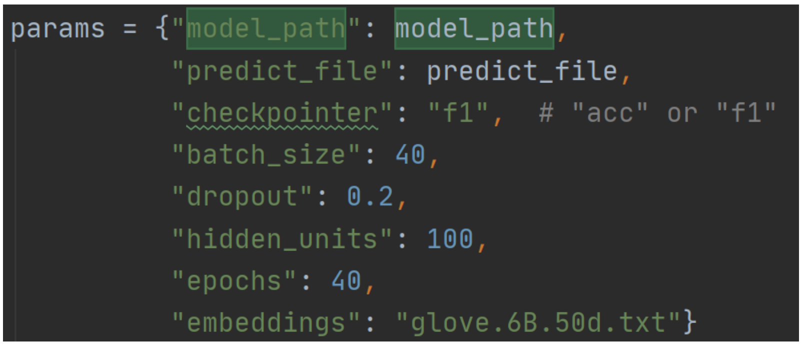
Task 2.5



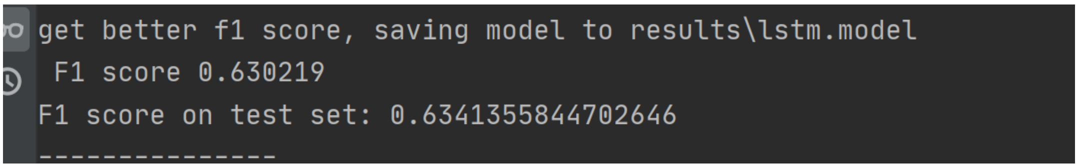
```
params = {"model_path": model_path,
          "predict_file": predict_file,
          "checkpointer": "f1",  # "acc" or "f1"
          "batch_size": 50,
          "dropout": 0.3,
          "hidden_units": 100,
          "epochs": 100,
          "embeddings": "glove.6B.50d.txt"}
```



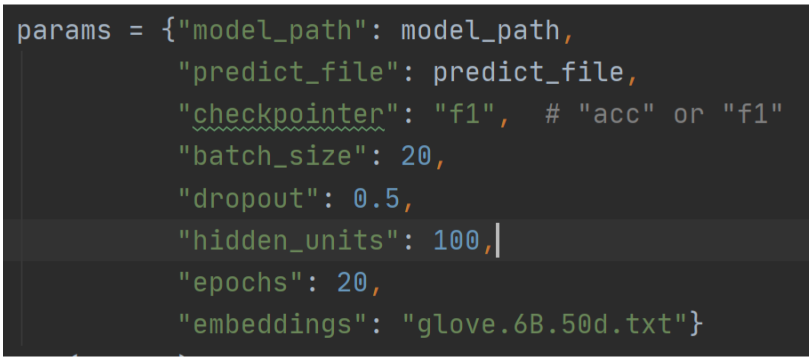
```
 F1 score 0.659614
F1 score on test set: 0.6626742936265421
---------------
```



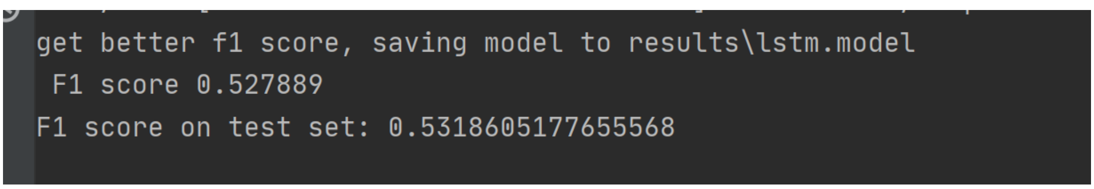
```
params = {"model_path": model_path,
          "predict_file": predict_file,
          "checkpointer": "f1",  # "acc" or "f1"
          "batch_size": 40,
          "dropout": 0.2,
          "hidden_units": 100,
          "epochs": 40,
          "embeddings": "glove.6B.50d.txt"}
```



```
get better f1 score, saving model to results\lstm.model
 F1 score 0.630219
F1 score on test set: 0.6341355844702646
---------------
```



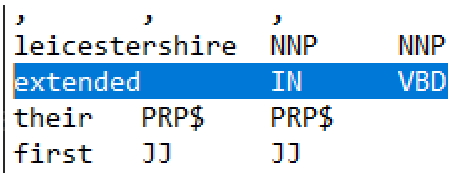
```
params = {"model_path": model_path,
          "predict_file": predict_file,
          "checkpointer": "f1",  # "acc" or "f1"
          "batch_size": 20,
          "dropout": 0.5,
          "hidden_units": 100,
          "epochs": 20,
          "embeddings": "glove.6B.50d.txt"}
```



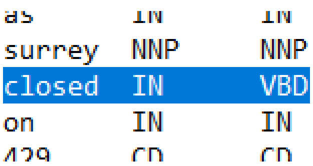
```
get better f1 score, saving model to results\lstm.model
 F1 score 0.527889
F1 score on test set: 0.5318605177655568
```

Task 2.6

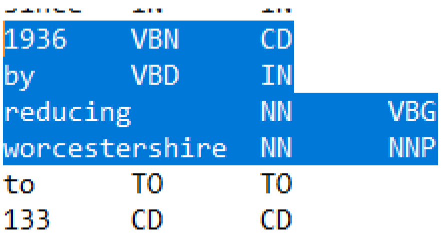
Most of Prediction is accurate, but in predictions.txt exist also a few prediction errors, such as:



```
,       ,       ,
leicestershire  NNP     NNP
extended        IN      VBD
their   PRP$    PRP$
first   JJ      JJ
```



```
as      IN      IN
surrey  NNP     NNP
closed  IN      VBD
on      IN      IN
429     CD      CD
```



```
since   IN      IN
1936    VBN     CD
by      VBD     IN
reducing        NN      VBG
worcestershire  NN      NNP
to      TO      TO
133     CD      CD
```

From selective examination proves that the model has difficulty in distinguish IN and VBD.
OOV (Out of vocabulary) Words nearly belong to CD and NNP, which could be relative accurate distinguished.