Natural Language Processing
Tarush Chaudhary, Jigyansu Rout

MBTI Classification:

The Myers Briggs Type Indicator (or MBTI for short) is a personality type system that divides everyone into 16 distinct personality types across 4 axis using four principal psychological functions - sensation, intuition, feeling, and thinking:

- Introversion (I) – Extroversion (E)
- Intuition (N) – Sensing (S)
- Thinking (T) – Feeling (F)
- Judging (J) – Perceiving (P)

Dataset:

The dataset mainly has two components

: ● Type / Class label - 4 letter MBTI code/type

● Text from posts made by the author.

Each Type corresponds to a set of posts made by the author of that type

Aim:

We need to train a model to predict personality type from the posts made by the author.

For this, we have trained an LSTM model using pytorch.

The model has 3 layers, the input layer, a 64 node LSTM layer, and a 16 node output layer.

The model outputs the 4 character tag.

Parameters:

Maximum Sequence Length = 128. This defines the maximum length of the tokens in the input sequence. with the longer ones being truncated, since the majority of the inputs were smaller than this length.

Batch Size: 10. The batch size was kept fairly small as the model was trained locally on the cpu and did not require a large amount of resources.

Epochs: We ran each fold for 3 epochs. While this number could have been increased, the performance increase after each was not significant enough for the added training time

Preprocessing:

Each set of posts was split into individual posts for the tag, with the pipe character being used to split the posts into separate entries. This also gave us more datapoints to train on, thus improving performance.

Initially the links were removed, but later added back, as links were not of one specific type of website, and certain websites could possibly correspond to certain personality types.

The text was tokenized using keras' Tokenizer. Padding was added for sequences shorter than the MAX_LEN

The labels were encoded using LabelEncoder. Initially One hot encoder was considered, but later discarded due to too many bugs, in total, there were 16 labels which were being predicted.

Training:

The model processes batches of tokenized text to predict one of 16 labels.
The loss function used for backpropogation was CrossEntropy Loss:

Final Results and Evaluation:

The model does not perform particularly well, but is not random as accuracy is 30%
Several measures could be taken to increase the accuracy, such as train on more epochs, use a more complicated architecture such as a transformer or a bidirectional LSTM.

```
Fold 1
Epoch 1/3: 100%|██████████| 32919/32919 [09:11<00:00, 59.74it/s, loss=1.51]
Epoch 2/3: 100%|██████████| 32919/32919 [09:30<00:00, 57.75it/s, loss=1.9]
Epoch 3/3: 100%|██████████| 32919/32919 [09:50<00:00, 55.79it/s, loss=2.08]
labels shape at the end of training: (411479,)
End of Epoch 3, Average Loss: 2.1149
Evaluation: 100%|██████████| 8230/8230 [00:30<00:00, 266.72it/s]
Fold 1 - Loss: 2.0753893852233887, Accuracy: 0.27987994556236023, Recall: 0.1356334668547586, Precision: 0.33114335641553005, F1 Score: 0.1485429401488319
Fold 2
Epoch 1/3: 100%|██████████| 32919/32919 [09:33<00:00, 57.38it/s, loss=2.94]
Epoch 2/3: 100%|██████████| 32919/32919 [09:22<00:00, 58.51it/s, loss=3.3]
Epoch 3/3: 100%|██████████| 32919/32919 [09:12<00:00, 59.61it/s, loss=2.11]
labels shape at the end of training: (411479,)
End of Epoch 3, Average Loss: 2.1113
Evaluation: 100%|██████████| 8230/8230 [00:27<00:00, 294.65it/s]
Fold 2 - Loss: 2.106515884399414, Accuracy: 0.279466802760766, Recall: 0.133155581114039, Precision: 0.3330849351254649, F1 Score: 0.1461915081163473
Fold 3
Epoch 1/3: 100%|██████████| 32919/32919 [09:21<00:00, 58.59it/s, loss=3.59]
Epoch 2/3: 100%|██████████| 32919/32919 [09:30<00:00, 57.66it/s, loss=1.52]
Epoch 3/3: 100%|██████████| 32919/32919 [09:17<00:00, 59.08it/s, loss=1.91]
labels shape at the end of training: (411479,)
End of Epoch 3, Average Loss: 2.1121
Evaluation: 100%|██████████| 8230/8230 [00:29<00:00, 275.02it/s]
Fold 3 - Loss: 1.9079135656356812, Accuracy: 0.2782152230971129, Recall: 0.13917124585354212, Precision: 0.3618244250121998, F1 Score: 0.15643857436147487
Fold 4
Epoch 1/3: 100%|██████████| 32919/32919 [09:33<00:00, 57.38it/s, loss=2.38]
Epoch 2/3: 100%|██████████| 32919/32919 [09:37<00:00, 57.01it/s, loss=2.9]
Epoch 3/3: 100%|██████████| 32919/32919 [09:14<00:00, 59.39it/s, loss=1.87]
labels shape at the end of training: (411479,)
End of Epoch 3, Average Loss: 2.1130
Evaluation: 100%|██████████| 8230/8230 [00:28<00:00, 290.29it/s]
/Users/rithikkumars/miniconda3/envs/kasturi_ml_project/lib/python3.10/site-packages/sklearn/metrics/_classification.py:1471: UndefinedMetricWarning: Precision is ill
  _warn_prf(average, modifier, msg_start, len(result))
Fold 4 - Loss: 1.8683291673660278, Accuracy: 0.27934529017206183, Recall: 0.1382239648357379, Precision: 0.35967254815950456, F1 Score: 0.1561806273532424
Fold 5
Epoch 1/3: 100%|██████████| 32919/32919 [09:19<00:00, 58.81it/s, loss=1.99]
Epoch 2/3: 100%|██████████| 32919/32919 [09:20<00:00, 58.72it/s, loss=2.5]
Epoch 3/3: 100%|██████████| 32919/32919 [09:19<00:00, 58.80it/s, loss=1.95]
labels shape at the end of training: (411479,)
End of Epoch 3, Average Loss: 2.1140
Evaluation: 100%|██████████| 8230/8230 [00:28<00:00, 291.56it/s]
Fold 5 - Loss: 1.9482883214950562, Accuracy: 0.27919071632541465, Recall: 0.1366116787131807, Precision: 0.3411289480754841, F1 Score: 0.14995147846772655
```

```
Average Metrics Across Folds:
Average accuracy: 0.27921959558354315
Average recall: 0.13655918747425166
Average precision: 0.34537084255763667
Average f1_score: 0.15146102568952458
```

```
Average conf_matrix:
[[1.38200e+02 4.02000e+01 7.00000e+00 1.34000e+01 2.60000e+00 2.00000e-01
  2.00000e-01 1.20000e+00 1.81000e+02 1.07540e+03 8.22000e+01 2.50400e+02
  3.00000e+00 3.80000e+00 7.20000e+00 8.80000e+00]
 [2.56000e+01 7.16200e+02 2.22000e+01 1.25200e+02 6.40000e+00 1.00000e+00
  1.40000e+00 8.00000e+00 4.87200e+02 3.68300e+03 3.39000e+02 9.18800e+02
  1.14000e+01 1.98000e+01 3.18000e+01 3.40000e+01]
 [1.08000e+01 3.02000e+01 1.48600e+02 5.48000e+01 2.40000e+00 2.00000e-01
  1.00000e+00 5.00000e+00 1.58800e+02 9.75200e+02 2.17800e+02 5.64600e+02
  5.20000e+00 1.02000e+01 7.00000e+00 1.26000e+01]
 [1.24000e+01 1.14600e+02 3.18000e+01 6.77200e+02 4.20000e+00 8.00000e-01
  2.00000e+00 1.54000e+01 4.62400e+02 3.02240e+03 4.90400e+02 1.69760e+03
  2.06000e+01 2.58000e+01 2.40000e+01 2.56000e+01]
 [2.80000e+00 7.00000e+00 1.20000e+00 9.80000e+00 2.52000e+01 0.00000e+00
  8.00000e-01 2.00000e-01 3.90000e+01 2.05600e+02 1.80000e+01 7.82000e+01
  4.60000e+00 5.60000e+00 8.00000e-01 1.00000e+00]
 [1.00000e+00 1.72000e+01 3.80000e+00 1.24000e+01 6.00000e-01 1.40000e+00
  4.00000e-01 2.80000e+00 3.54000e+01 2.28200e+02 3.76000e+01 8.16000e+01
```

```
  1.20000e+00 5.60000e+00 8.00000e-01 2.20000e+00]
 [6.00000e-01 7.60000e+00 6.60000e+00 9.00000e+00 4.00000e-01 0.00000e+00
  5.80000e+00 1.60000e+00 3.30000e+01 1.92400e+02 3.06000e+01 8.10000e+01
  6.00000e-01 2.40000e+00 3.40000e+00 2.40000e+00]
 [2.40000e+00 1.52000e+01 5.80000e+00 3.00000e+01 1.80000e+00 0.00000e+00
  0.00000e+00 5.04000e+01 7.42000e+01 4.04600e+02 6.08000e+01 1.78400e+02
  4.00000e+00 4.60000e+00 6.40000e+00 1.12000e+01]
 [4.40000e+01 1.86000e+02 2.90000e+01 1.90800e+02 8.00000e+00 4.00000e-01
  2.80000e+00 1.88000e+01 2.30120e+03 7.95360e+03 7.45400e+02 2.35000e+03
  2.98000e+01 4.52000e+01 3.32000e+01 4.52000e+01]
 [4.86000e+01 2.16600e+02 3.96000e+01 1.50200e+02 9.40000e+00 8.00000e-01
  2.60000e+00 1.24000e+01 1.09280e+03 1.20734e+04 7.64000e+02 2.82280e+03
  2.62000e+01 5.72000e+01 4.08000e+01 4.70000e+01]
 [1.50000e+01 1.28600e+02 3.62000e+01 1.34000e+02 7.20000e+00 6.00000e-01
  1.40000e+00 1.06000e+01 6.89800e+02 4.62820e+03 1.71100e+03 2.77040e+03
  1.98000e+01 2.30000e+01 3.90000e+01 3.44000e+01]
 [1.86000e+01 9.24000e+01 3.68000e+01 1.84000e+02 1.18000e+01 8.00000e-01
  2.20000e+00 8.80000e+00 6.76600e+02 5.74220e+03 9.50600e+02 4.45380e+03
  2.06000e+01 2.46000e+01 3.38000e+01 4.64000e+01]
 [4.80000e+00 2.02000e+01 5.20000e+00 1.88000e+01 1.60000e+00 0.00000e+00
  6.00000e-01 2.00000e+00 1.25000e+02 9.04000e+02 7.64000e+01 2.60000e+02
  1.27600e+02 1.18000e+01 1.50000e+01 8.20000e+00]
 [5.60000e+00 4.84000e+01 9.40000e+00 3.08000e+01 3.80000e+00 1.00000e+00
  4.00000e-01 3.80000e+00 1.76800e+02 1.51520e+03 1.14800e+02 3.86800e+02
  1.24000e+01 1.67800e+02 1.02000e+01 1.74000e+01]
 [2.60000e+00 2.74000e+01 3.20000e+00 1.60000e+01 3.80000e+00 2.00000e-01
  2.00000e-01 2.00000e+00 1.30600e+02 9.96400e+02 1.45400e+02 4.18200e+02
  8.40000e+00 6.40000e+00 1.57800e+02 9.40000e+00]
 [5.40000e+00 3.42000e+01 6.80000e+00 4.22000e+01 1.80000e+00 2.00000e-01
  1.20000e+00 9.60000e+00 2.01800e+02 1.59460e+03 2.17400e+02 8.36600e+02
  6.00000e+00 1.36000e+01 1.00000e+01 2.23000e+02]]
```

Conf matrix not graphed as model was trained on someone else’s system.