(a) Category test results for System A

Category	f1	Number	Precision	Recall
ANIM	0.7628	3208	0.7332	0.7949
DIS	0.7845	1514	0.7804	0.7886
LOC	0.9945	24046	0.9953	0.9937
ORG	0.9817	6616	0.9813	0.9822
PER	0.9951	10530	0.9941	0.9960

(b) Category test results for System B

Category	f1	Number	Precision	Recall
ANIM	0.7497	3208	0.7379	0.7618
DIS	0.7680	1514	0.7558	0.7807
LOC	0.9943	24046	0.9946	0.9941
ORG	0.9818	6616	0.9836	0.9800
PER	0.9949	10530	0.9945	0.9953

(c) System A Performance Metrics

Metric	Value
Overall Accuracy	0.9903
Overall F1 Score	0.9499
Overall Precision	0.9476
Overall Recall	0.9523

(d) System B Performance Metrics

Metric	Value
Overall Accuracy	0.9939
Overall F1 Score	0.9677
Overall Precision	0.9664
Overall Recall	0.9691

Table 1: Test Results using mBERT

(a) Category test results for System A

Category	f1	Number	Precision	Recall
ANIM	0.7700	3208	0.7401	0.8024
DIS	0.7971	1514	0.7874	0.8071
LOC	0.9956	24046	0.9956	0.9955
ORG	0.9831	6616	0.9834	0.9828
PER	0.9940	10530	0.9935	0.9945

(b) Category test results for System B

Category	f1	Number	Precision	Recall
ANIM	0.7644	3208	0.7378	0.7930
DIS	0.7964	1514	0.7846	0.8085
LOC	0.9946	24046	0.9947	0.9944
ORG	0.9826	6616	0.9828	0.9825
PER	0.9945	10530	0.9943	0.9947

(c) System A Performance Metrics

Metric	Value
Overall Accuracy	0.9907
Overall F1 Score	0.9527
Overall Precision	0.9482
Overall Recall	0.9571

(d) System B Performance Metrics

Metric	Value
Overall Accuracy	0.9942
Overall F1 Score	0.9696
Overall Precision	0.9666
Overall Recall	0.9726

Table 2: Test Results using BERT

While the overall accuracy, F1 score, precision, and recall (FPR) have increased (with a significant increase in precision and, consequently, the F1 score), the individual FPR scores have dropped for system B. This might be due to considering all the other categories as 'O,' where the individual metrics for those categories will affect the overall performance. Therefore, system B performs poorly even if the task is to identify only the categories of interest. Moreover the same trend is noticed both in mBERT and BERT models.

On the training part: Both systems were trained using the code from [1]. The changes I implemented involved considering the base model as BERT in addition to mBERT, as utilized by [1], and adjusting the batch size from 32 to 16 due to limited computational resources. The scores mentioned earlier are derived from the sequential evaluation method library [2]. Regarding system b, although there were signs of loss reduction, I could have extended the training for an additional epoch. However, due to constraints on available GPU resources, I was limited in the training duration.

[1] URL: `https://huggingface.co/tomaarsen/span-marker-mbert-base-multinerd`.

[2] Hiroki Nakayama. *seqeval: A Python framework for sequence labeling evaluation*. Software available from https://github.com/chakki-works/seqeval. 2018. URL: `https://github.com/chakki-works/seqeval`.