```
===== CLASSIFICATION REPORT =====
              precision    recall  f1-score   support

       B-LOC     0.5455    0.3371    0.4167       178
      B-MISC     0.4348    0.2548    0.3213       157
       B-ORG     0.6506    0.4430    0.5271       395
       B-PER     0.7822    0.6651    0.7189       621
       I-LOC     0.5000    0.2549    0.3377        51
      I-MISC     0.6083    0.3476    0.4424       210
       I-ORG     0.5833    0.3281    0.4200       128
       I-PER     0.8113    0.6581    0.7267       503
           O     0.9099    0.9718    0.9399      9087

    accuracy                         0.8807     11330
   macro avg     0.6473    0.4734    0.5390     11330
weighted avg     0.8661    0.8807    0.8693     11330
```

```
===== CLASSIFICATION REPORT =====
              precision    recall  f1-score   support

       B-LOC     0.6419    0.1679    0.2662      1697
       B-ORG     0.5152    0.0405    0.0751       839
     B-OTHER     0.5417    0.0181    0.0349       720
       B-PER     0.6554    0.3530    0.4588      1816
       I-LOC     0.4511    0.1547    0.2304       685
       I-ORG     0.3269    0.0466    0.0815       365
     I-OTHER     0.6667    0.0028    0.0056       716
       I-PER     0.4909    0.3558    0.4126      1290
           O     0.8793    0.9853    0.9293     44694

    accuracy                         0.8632     52822
   macro avg     0.5743    0.2361    0.2772     52822
weighted avg     0.8318    0.8632    0.8260     52822
```

Per-label classification report:

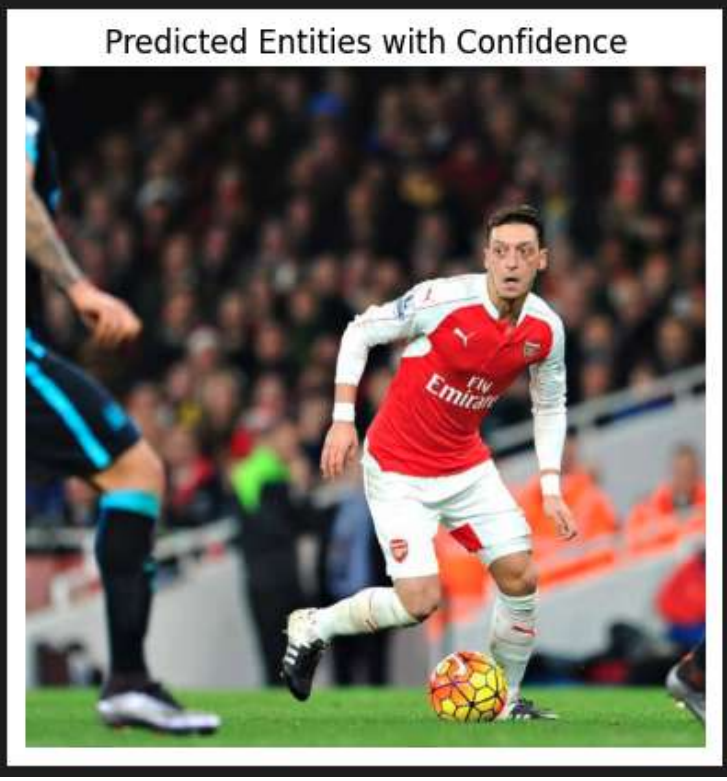
	precision	recall	f1-score	support
B-LOC	0.8349	0.9057	0.8689	1697
B-ORG	0.7119	0.7008	0.7063	839
B-OTHER	0.5637	0.5903	0.5767	720
B-PER	0.8842	0.9124	0.8981	1816
I-LOC	0.7770	0.8292	0.8023	685
I-ORG	0.6731	0.5699	0.6172	365
I-OTHER	0.6188	0.5531	0.5841	716
I-PER	0.8949	0.9434	0.9185	1290
O	0.9843	0.9797	0.9820	44694
accuracy			0.9538	52822
macro avg	0.7714	0.7761	0.7727	52822
weighted avg	0.9540	0.9538	0.9538	52822

Per-label classification report:

	precision	recall	f1-score	support
B-LOC	0.8617	0.9101	0.8852	178
B-MISC	0.7844	0.8344	0.8086	157
B-ORG	0.8872	0.8962	0.8917	395
B-PER	0.9563	0.9517	0.9540	621
I-LOC	0.7869	0.9412	0.8571	51
I-MISC	0.8498	0.8619	0.8558	210
I-ORG	0.8240	0.8047	0.8142	128
I-PER	0.9622	0.9602	0.9612	503
O	0.9908	0.9876	0.9892	9087
accuracy			0.9733	11330
macro avg	0.8781	0.9053	0.8908	11330
weighted avg	0.9738	0.9733	0.9734	11330

```
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

🔍 Predicted entities for: /kaggle/input/twitter2017/twitter2017/twitter2017_images/16_05_01_100.jpg
AP              → B-ORG    (confidence: 0.961)
News            → I-ORG    (confidence: 0.793)
Little          → B-LOC    (confidence: 0.940)
Rock            → I-LOC    (confidence: 0.976)
```



```
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Predicted entities for IMGID 74960:
George          → B-PER
Zimmerman       → I-PER
```

Relation: /per/per/alumi (4.6%)

📝 Sentence:
RT @CHC_1927 : Humphrey Bogart and [E1] Lauren Bacall [/E1] on the set of ' [E2] Key Largo'(1948 [/E2]) .

👤 Head Entity (E1): Lauren Bacall | NER type: PER
👥 Tail Entity (E2): Key Largo'(1948 | NER type: MISC
🔗 Predicted Relation: /per/per/alumi (conf=0.0463)

Model	Accuracy (%)
ResNet50 + BERT	75%
CLIP + BERT	73%
BLIP Model	60%

Method / Model	Twitter2015 (F1 or equivalent)	Twitter2017 (F1 or equivalent)
RoBERTa + ResNet50 (your best)	**82.60 %** (weighted F1 from your image for 2015)	**86.93 %** (weighted F1 from your image for 2017)
MINIGE-MNER (Kong et al., 2025)	76.45 %	88.67 %
Improving MNER via text–image alignment (Zeng et al., 2025)	75.32 % ScienceDirect	86.65 % ScienceDirect
ICKA: Instruction & Knowledge (Zeng et al., 2024)	75.42 % ScienceDirect	87.12 % ScienceDirect
CoAtt-NER (Scene Graph)	76.25 %	87.31 %
Dual-Enhanced Hierarchical Alignment (Wang et al., 2025)	77.42 % MDPI	88.79 % MDPI

Dataset	Model	Accuracy	Precision	Recall	F1-score
Twitter2015	**RoBERTa + ResNet50**	**0.9538**	**0.9540**	**0.9538**	**0.9538**
Twitter2015	CLIP + BERT	0.9445	0.9425	0.9445	0.9431
Twitter2015	BLIP	0.8574	0.7924	0.8574	0.8166
Twitter2015	ViLT	0.8632	0.8318	0.8632	0.8260
Twitter2017	**RoBERTa + ResNet50**	**0.9733**	**0.9738**	**0.9733**	**0.9734**
Twitter2017	CLIP + BERT	0.9639	0.9641	0.9639	0.9639
Twitter2017	BLIP	0.8935	0.8552	0.8935	0.8732
Twitter2017	ViLT	0.8807	0.8661	0.8807	0.8693
MNRE	**ResNet50 + BERT**	**0.7501**	**0.7553**	**0.7422**	**0.7487**
MNRE	CLIP + BERT	0.7302	0.7416	0.7312	0.7363
MNRE	BLIP	0.6003	0.6114	0.6021	0.6067

Dataset	Model	Accuracy	Precision	Recall	F1-score
Twitter2015	BLIP	0.8574	0.7924	0.8574	0.8166
	CLIP + BERT	0.9445	0.9425	0.9445	0.9431
	ViLT	0.8632	0.8318	0.8632	0.8260
	RoBERTa + ResNet50	**0.9538**	**0.9540**	**0.9538**	**0.9538**
Twitter2017	BLIP	0.8935	0.8552	0.8935	0.8732
	CLIP + BERT	0.9639	0.9641	0.9639	0.9639
	ViLT	0.8807	0.8661	0.8807	0.8693
	RoBERTa + ResNet50	**0.9733**	**0.9738**	**0.9733**	**0.9734**

Model / Method	Dataset	Accuracy	Precision	Recall	F1-score
BLIP (Ours)	Twitter2015	0.8574	0.7924	0.8574	0.8166
	Twitter2017	0.8935	0.8552	0.8935	0.8732
CLIP + BERT (Ours)	Twitter2015	0.9445	0.9425	0.9445	0.9431
	Twitter2017	0.9639	0.9641	0.9639	0.9639
ViLT (Ours)	Twitter2015	0.8632	0.8318	0.8632	0.8260
	Twitter2017	0.8807	0.8661	0.8807	0.8693
RoBERTa + ResNet50 (Ours)	Twitter2015	**0.9538**	**0.9540**	**0.9538**	**0.9538**
	Twitter2017	**0.9733**	**0.9738**	**0.9733**	**0.9734**
MINIGE-MNER (Kong et al., 2025)	Twitter2015	—	—	—	0.7645
	Twitter2017	—	—	—	0.8867
Text-Image Alignment (Zeng et al., 2025)	Twitter2015	—	—	—	0.7532
	Twitter2017	—	—	—	0.8665
ICKA (Zeng et al., 2024)	Twitter2015	—	—	—	0.7542
	Twitter2017	—	—	—	0.8712
CoAtt-NER (Scene Graph, 2024)	Twitter2015	—	—	—	0.7625
	Twitter2017	—	—	—	0.8731
Dual-Enhanced Hierarchical Alignment (Wang et al., 2025)	Twitter2015	—	—	—	0.7742
	Twitter2017	—	—	—	0.8879

```
===== CLASSIFICATION REPORT =====
              precision    recall  f1-score   support

       B-LOC     0.8693    0.8596    0.8644       178
      B-MISC     0.7133    0.6815    0.6971       157
       B-ORG     0.8358    0.8506    0.8432       395
       B-PER     0.9281    0.9356    0.9318       621
       I-LOC     0.6923    0.8824    0.7759        51
      I-MISC     0.7867    0.7905    0.7886       210
       I-ORG     0.8182    0.7734    0.7952       128
       I-PER     0.9584    0.9622    0.9603       503
           O     0.9864    0.9850    0.9857      9061

    accuracy                         0.9639     11304
   macro avg     0.8432    0.8579    0.8491     11304
weighted avg     0.9641    0.9639    0.9639     11304
```

```
===== CLASSIFICATION REPORT =====
              precision    recall  f1-score   support

       B-LOC     0.8109    0.8768    0.8426      1697
       B-ORG     0.6523    0.6508    0.6516       839
     B-OTHER     0.4759    0.4111    0.4411       720
       B-PER     0.8576    0.8954    0.8761      1816
       I-LOC     0.7460    0.8102    0.7768       685
       I-ORG     0.6015    0.4384    0.5071       365
     I-OTHER     0.5400    0.4148    0.4692       716
       I-PER     0.8732    0.9395    0.9052      1290
           O     0.9782    0.9779    0.9780     44694

    accuracy                         0.9445     52822
   macro avg     0.7262    0.7128    0.7164     52822
weighted avg     0.9425    0.9445    0.9431     52822
```

```
===== CLASSIFICATION REPORT =====
              precision    recall  f1-score   support

       B-LOC     0.0000    0.0000    0.0000       178
      B-MISC     0.0000    0.0000    0.0000       157
       B-ORG     0.4430    0.5899    0.5060       395
       B-PER     0.7834    0.8035    0.7933       621
       I-LOC     0.0000    0.0000    0.0000        51
      I-MISC     0.5120    0.4048    0.4521       210
       I-ORG     0.0000    0.0000    0.0000       128
       I-PER     0.8947    0.7773    0.8319       503
           O     0.9321    0.9811    0.9560      9087

    accuracy                         0.8935     11330
   macro avg     0.3961    0.3952    0.3933     11330
weighted avg     0.8552    0.8935    0.8732     11330
```

```
===== CLASSIFICATION REPORT =====
              precision    recall  f1-score   support

       B-LOC     0.3067    0.2510    0.2761      1697
       B-ORG     0.3571    0.0119    0.0231       839
     B-OTHER     0.0000    0.0000    0.0000       720
       B-PER     0.4297    0.4967    0.4608      1816
       I-LOC     0.0000    0.0000    0.0000       685
       I-ORG     0.0000    0.0000    0.0000       365
     I-OTHER     0.0000    0.0000    0.0000       716
       I-PER     0.3158    0.0047    0.0092      1290
           O     0.8916    0.9833    0.9352     44694

    accuracy                         0.8574     52822
   macro avg     0.2557    0.1942    0.1894     52822
weighted avg     0.7924    0.8574    0.8166     52822
```