Fusion Technique	**Twitter-2015 Accuracy**	**Twitter-2017 Accuracy**	**MNRE Accuracy**
FiLM Fusion	0.9538	0.9718	0.7538
Attention Fusion	0.8461	0.8739	0.7053
Cross-Attention	0.8934	0.9145	0.7386
Bimodal Fusion	0.9528	0.9698	0.7462

Hyperparameter	**Description**	**Value**
SEED	Random seed for reproducibility	1337
device	Device used for model training (CPU or CUDA)	`'cuda'` if available, `'cpu'` otherwise
max_len	Maximum sequence length for tokenization	128
batch_size	Batch size for data loading	8 (train), 12 (val/test)
EPOCHS	Number of epochs for training	5
learning_rate	Learning rate for the optimizer	3e-5
weight_decay	Weight decay for the optimizer	0.01
warmup_ratio	Ratio of total steps for learning rate warmup	0.1
num_warmup_steps	Number of warmup steps for the learning rate schedule	Computed based on `warmup_ratio` and `total_steps`
num_training_steps	Total number of training steps	Computed based on `batch_size` and `EPOCHS`
dropout	Dropout rate for the model	0.1
num_labels	Number of labels (output classes) for token classification	Number of unique labels in dataset
num_heads	Number of attention heads in the MultiheadAttention layer	8
img_size	Image size used for resizing (before feeding into ResNet50)	(224, 224)
image_normalization_mean	Mean for image normalization (ResNet50 standard)	[0.485, 0.456, 0.406]
image_normalization_std	Standard deviation for image normalization (ResNet50 standard)	[0.229, 0.224, 0.225]
aug	Whether to apply data augmentation for training (True/False)	True (for training)

Dataset	Model	Accuracy	Precision	Recall	F1-score	EPOCHS	Batch Size	Learning Rate	Image Size
Twitter2015	**RoBERTa + ResNet50**	0.9538	0.9540	0.9538	0.9538	5	12	0.0011	224x224
		0.9487	0.9523	0.9502	0.9512	3	8	0.0005	224x224
		0.9552	0.9561	0.9548	0.9554	7	16	0.0020	256x256
		0.9501	0.9510	0.9495	0.9502	10	10	0.0010	224x224
Twitter2017	**RoBERTa + ResNet50**	0.9733	0.9738	0.9733	0.9734	5	12	0.0011	224x224
		0.9710	0.9730	0.9709	0.9719	3	8	0.0005	224x224
		0.9756	0.9760	0.9754	0.9757	7	16	0.0020	256x256
		0.9715	0.9720	0.9712	0.9716	10	10	0.0010	224x224
MNRE	**ResNet50 + BERT**	0.7501	0.7553	0.7422	0.7487	5	12	0.0011	224x224
		0.7400	0.7455	0.7350	0.7402	3	8	0.0005	224x224
		0.7600	0.7650	0.7580	0.7615	7	16	0.0020	256x256
		0.7450	0.7480	0.7400	0.7440	10	10	0.0010	224x224

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

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