XLM-RoBERTa Fine-tuning Results: With Context vs Without Context

Experimental Setup

We fine-tuned XLM-RoBERTa-base for binary sarcasm classification under two conditions:

- **With Context:** Input formatted as `"Context: {context} Response: {text}"`
- **Without Context:** Input using only the response utterance (`text`)

Both experiments used identical hyperparameters: learning rate of 5e-6, batch size of 8 with gradient accumulation of 2 (effective batch size 16), 5 training epochs, and 50 warmup steps. Models were evaluated on validation, test, and gold sets.

Results

Dataset	Condition	Accuracy	Precision	Recall	F1	TP	TN	FP	FN
Val	With Context	0.682	0.690	0.682	0.678	19	26	7	14
Val	Without Context	0.758	0.774	0.758	0.754	21	29	4	12
Test	With Context	0.716	0.742	0.714	0.707	18	30	4	15
Test	Without Context	0.627	0.641	0.624	0.614	15	27	7	18
Gold	With Context	0.607	0.690	0.646	0.594	6	11	1	10
Gold	Without Context	0.643	0.678	0.667	0.641	8	10	2	8

Observations

Performance differed across evaluation sets depending on the presence of context:

- **Validation set:** The model without context outperformed the model with context by 7.6 percentage points in both accuracy (75.8% vs 68.2%) and F1 score (75.4% vs 67.8%).
- **Test set:** The model with context outperformed the model without context by 8.9 percentage points in accuracy (71.6% vs 62.7%) and 9.3 percentage points in F1 score (70.7% vs 61.4%).
- **Gold set:** The model without context outperformed the model with context by 3.6 percentage points in accuracy (64.3% vs 60.7%) and 4.7 percentage points in F1 score (64.1% vs 59.4%).

The results show an inconsistent pattern: removing context improved performance on the validation and gold sets but decreased performance on the test set.