DeBERTa-v3-base Fine-tuning Results: With Context vs Without Context

Experimental Setup

We fine-tuned DeBERTa-v3-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	0.515	0.520	0.515	0.485	25	9	24	8
Val	Without Context	0.667	0.672	0.667	0.664	19	25	8	14
Test	With Context	0.537	0.544	0.540	0.527	23	13	21	10
Test	Without Context	0.597	0.600	0.595	0.591	16	24	10	17
Gold	With Context	0.500	0.478	0.479	0.476	10	4	8	6
Gold	Without Context	0.750	0.778	0.771	0.750	10	11	1	6

Observations

The model without context consistently outperformed the model with context across all evaluation sets:

- **Validation set:** The model without context outperformed the model with context by 15.2 percentage points in accuracy (66.7% vs 51.5%) and 17.9 percentage points in F1 score (66.4% vs 48.5%).
- **Test set:** The model without context outperformed the model with context by 6.0 percentage points in accuracy (59.7% vs 53.7%) and 6.4 percentage points in F1 score (59.1% vs 52.7%).
- **Gold set:** The model without context outperformed the model with context by 25.0 percentage points in accuracy (75.0% vs 50.0%) and 27.4 percentage points in F1 score (75.0% vs 47.6%).

The model with context performed near chance level (50%) on the validation and gold sets, with a high false positive rate (24 FP on validation, 8 FP on gold). Removing context reduced false positives substantially (8 FP on validation, 1 FP on gold) and improved overall performance.