Additional NLU Tasks Results

Anonymous EMNLP submission

1 Additional NLU Tasks Results

Model	acc.	Macro-Avg prec.	rec.	F1
RoBERTa-base				
W/out affir. intp.				
Mosharaf				
w/out negation				0.60
w/ negation				0.53
Important				0.47
Unimportant				0.62
Our implementation	0.57	0.57	0.57	0.57
w/out negation	0.58	0.58	0.58	0.58
w/ negation	0.55	0.56	0.55	0.55
Important	0.44	0.44	0.44	0.43
Unimportant	0.69	0.72	0.69	0.70
RoBERTa-large				
W/out affir. intp.	0.70	0.70	0.70	0.70
w/out negation	0.69	0.69	0.69	0.69
w negation	0.73	0.73	0.73	0.73
Important	0.67	0.67	0.67	0.67
Unimportant	0.80	0.82	0.80	0.80
W/ affir. intp. by T5-Chat	0.71	0.71	0.71	0.71
w/out negation	0.71	0.71	0.71	0.71
w negation	0.74	0.75	0.74	0.74
Important	0.69	0.70	0.69	0.69
Unimportant	0.80	0.81	0.80	0.80
W/ affir. intp. by T5-Mosharaf	0.72	0.72	0.72	0.72
w/out negation	0.72	0.72	0.72	0.72
w negation	0.74	0.74	0.74	0.74
Important	0.70	0.71	0.70	0.70
Unimportant	0.79	0.80	0.80	0.80

Table 1: Results on CommonSenseQA.

Model	acc.	Macro-Avg prec.	rec.	F1
RoBERTa-base				
W/out affir. intp.				
Mosharaf				
w/out negation				0.93
w/ negation				0.91
Important				0.67
Unimportant				0.92
Our implementation	0.93	0.93	0.93	0.93
w/out negation	0.93	0.93	0.93	0.93
w/ negation	0.91	0.90	0.90	0.90
Important	0.65	0.51	0.52	0.50
Unimportant	0.91	0.91	0.91	0.91
RoBERTa-large				
W/out affir. intp.	0.93	0.93	0.93	0.93
w/out negation	0.93	0.93	0.93	0.93
w negation	0.92	0.92	0.91	0.92
Important	0.85	0.75	0.91	0.78
Unimportant	0.92	0.92	0.92	0.92
W/ affir. intp. by T5-Chat	0.93	0.93	0.93	0.93
w/out negation	0.94	0.94	0.94	0.94
w negation	0.93	0.93	0.93	0.93
Important	0.85	0.72	0.77	0.74
Unimportant	0.93	0.93	0.93	0.93
W/ affir. intp. by T5-Mosharaf	0.93	0.93	0.93	0.93
w/out negation	0.94	0.94	0.94	0.94
w negation	0.92	0.91	0.91	0.91
Important	0.90	0.80	0.80	0.80
Unimportant	0.92	0.91	0.91	0.91

Table 2: Results on QNLI.

Model	Pearson	Spearman
RoBERTa-base		
W/out affir. intp.		
Mosharaf		
w/out negation	0.92	0.91
w/ negation	0.85	0.84
Important	0.57	0.62
Unimportant	0.85	0.84
Our implementation	0.91	0.90
w/out negation	0.92	0.91
w/ negation	0.84	0.85
Important	0.66	0.76
Unimportant	0.84	0.85
RoBERTa-large		
W/out affir. intp.	0.92	0.92
w/out negation	0.92	0.92
w negation	0.88	0.88
Important	0.82	0.85
Unimportant	0.88	0.88
W/ affir. intp. by T5-Chat	0.92	0.92
w/out negation	0.93	0.92
w negation	0.88	0.88
Important	0.82	0.87
Unimportant	0.88	0.88
W/ affir. intp. by T5-Mosharaf	0.92	0.91
w/out negation	0.92	0.92
w negation	0.87	0.88
Important	0.82	0.82
Unimportant	0.87	0.88

Table 3: Results on STSB.

Model	acc.	Macro-Avg prec.	rec.	F1
RoBERTa-base				
W/out affir. intp.				
Mosharaf				
w/out negation				0.63
w/ negation				0.59
Our implementation	0.60	0.58	0.58	0.58
w/out negation	0.54	0.49	0.49	0.49
w/ negation	0.65	0.65	0.66	0.65
RoBERTa-large				
W/out affir. intp.	0.69	0.72	0.73	0.69
w/out negation	0.67	0.69	0.71	0.67
w/ negation	0.71	0.74	0.75	0.71
W/ affir. intp. by T5-Chat	0.72	0.71	0.73	0.71
w/out negation	0.69	0.68	0.70	0.68
w/ negation	0.75	0.75	0.76	0.75
W/ affir. intp. by T5-Mosharaf	0.69	0.68	0.69	0.68
w/out negation	0.63	0.63	0.64	0.62
w/ negation	0.75	0.74	0.74	0.74

Table 4: Results on WSC.

Model	acc.	Macro-Avg prec.	rec.	F1
RoBERTa-base				
W/out affir. intp.				
Mosharaf				
w/out negation				0.67
w/ negation				0.64
Our implementation	0.66	0.69	0.66	0.65
w/out negation	0.67	0.69	0.66	0.65
w/ negation	0.64	0.67	0.65	0.63
RoBERTa-large				
W/out affir. intp.	0.71	0.73	0.71	0.71
w/out negation	0.72	0.73	0.72	0.71
w negation	0.67	0.68	0.68	0.66
W/ affir. intp. by T5-Chat	0.73	0.75	0.73	0.73
w/out negation	0.74	0.75	0.74	0.73
w negation	0.70	0.71	0.70	0.70
W/ affir. intp. by T5-Mosharaf	0.71	0.72	0.71	0.70
w/out negation	0.71	0.72	0.71	0.71
w negation	0.70	0.71	0.71	0.70

Table 5: Results on WIC.