A Results on MNLI Stress Test

In Table 6, we show the complete results on MNLI Stress Test (Naik et al., 2018). In addition to Overlap and Negation, which is intended to test dataset bias, we also include two tests that evaluate model performance on minority examples. Our debiased models have some improvement on Antonym, possibly as a by-product of focusing on challenge examples that cannot be solved by superficial cues. However, DRiFt did not improve performance on Length.

		Negation			Overlap			Antonym			Length		
		E	*C*	*N*	*E*	*C*	*N*	*E*	*C*	*N*	*E*	*C*	*N*
HYPO		41.2	52.4	50.5	44.2	52.8	51.7	-	40.5	-	55.1	52.5	51.5
CBOW	MLE	20.1	48.2	53.9	49.7	52.9	55.6	-	19.0	-	21.9	55.5	49.4
HAND		37.5	45.0	57.3	56.7	50.1	57.8	-	28.2	-	66.6	65.0	60.7
BERT	MLE	2.4	81.1	56.5	19.2	83.3	59.4	-	66.0	-	83.8	83.6	77.4
	DRiFt-HYPO	7.3	80.7	55.6	27.5	81.1	59.1	-	75.4	-	84.1	83.2	76.3
	DRiFt-CBOW	17.9	81.7	55.5	18.3	80.0	56.6	-	75.3	-	82.4	82.3	74.6
	DRiFt-HAND	4.3	80.6	55.5	15.0	81.9	57.4	-	76.0	-	81.4	82.5	74.9
	RM-HYPO	32.1	55.9	39.9	44.4	63.8	43.0	-	69.3	-	72.9	70.9	52.4
	RM-CBOW	33.6	61.6	42.7	29.4	65.2	44.7	-	85.1	-	69.7	60.8	55.7
	RM-HAND	20.7	49.7	40.0	30.9	54.7	39.6	-	83.8	-	57.2	52.6	46.5
DA	MLE	17.4	47.3	55.3	46.7	60.5	57.8	-	59.8	-	69.5	66.0	61.9
	DRiFt-HYPO	11.8	47.0	51.8	41.6	59.4	55.6	-	57.4	-	66.4	63.7	55.3
	DRiFt-CBOW	28.4	21.4	39.5	35.2	41.7	43.8	-	57.8	-	64.3	39.4	53.9
	DRiFt-HAND	24.7	42.0	46.4	42.2	56.0	49.9	-	72.4	-	48.4	57.6	51.2
	RM-HYPO	14.9	39.9	45.3	52.0	52.6	46.0	-	56.6	-	63.9	62.2	32.0
	RM-CBOW	3.8	23.9	38.0	2.6	17.1	41.5	-	53.1	-	4.4	33.5	29.9
	RM-HAND	31.6	26.4	36.2	40.6	37.6	29.6	-	57.8	-	40.0	27.6	33.1
ESIM	MLE	12.0	72.7	54.6	27.6	76.4	57.5	-	75.1	-	77.6	76.8	68.8
	DRiFt-HYPO	22.8	67.7	54.0	37.5	73.2	56.7	-	75.5	-	75.9	74.3	66.3
	DRiFt-CBOW	32.7	62.3	46.9	30.4	65.6	49.8	-	67.0	-	68.5	60.2	60.1
	DRiFt-HAND	15.8	64.6	51.8	39.2	70.7	53.9	-	74.7	-	68.8	70.9	61.6
	RM-HYPO	29.6	54.4	45.3	47.3	63.6	46.1	-	60.4	-	70.6	68.2	31.1
	RM-CBOW	31.8	32.0	28.9	18.1	33.2	32.7	-	68.3	-	26.6	18.0	40.8
	RM-HAND	26.0	35.1	40.7	29.2	43.3	34.0	-	57.4	-	31.4	35.0	35.3

Table 6: Complete results on MNLI Stress Test.