

I. HATE SPEECH DETECTION TASK

II. CHECKLIST RESULTS ON MODELS UNDER TEST

TABLE I: Results of BERT-base, da-ELECTRA and da-BERT-base hate speech detection models on S²LCT test cases using all seeds.

Linguistic capability	S ² LCT #Seeds	S ² LCT #Exps	S ² LCT #Fail	S ² LCT Fail rate[%]	S ² LCT #PassTo- Fail
LC1: Expression of strong negative emotions (explicit)	140	799	BERT: 0 daELECTRA: 845 daBERT: 920	BERT: 0.00 daELECTRA: 89.99 daBERT: 97.98	BERT: 0 daELECTRA: 19 daBERT: 22
LC2: Description using very negative attributes (explicit)	140	2959	BERT: 0 daELECTRA: 3,011 daBERT: 3,067	BERT: 0.00 daELECTRA: 97.16 daBERT: 98.97	BERT: 0 daELECTRA: 30 daBERT: 23
LC3: Dehumanisation (explicit)	140	3124	BERT: 0 daELECTRA: 3,210 daBERT: 3,220	BERT: 0.00 daELECTRA: 98.35 daBERT: 98.65	BERT: 0 daELECTRA: 11 daBERT: 18
LC4: Implicit derogation	140	5664	BERT: 0 daELECTRA: 5,526 daBERT: 5,730	BERT: 0.00 daELECTRA: 95.21 daBERT: 98.73	BERT: 0 daELECTRA: 56 daBERT: 30
LC5: Direct threat	133	2689	BERT: 0 daELECTRA: 2,750 daBERT: 2,770	BERT: 0.00 daELECTRA: 97.45 daBERT: 98.16	BERT: 0 daELECTRA: 0 daBERT: 9
LC6: Threat as normative statement	140	4163	BERT: 0 daELECTRA: 4,198 daBERT: 4,261	BERT: 0.00 daELECTRA: 97.56 daBERT: 99.02	BERT: 0 daELECTRA: 12 daBERT: 1
LC7: Hate expressed using slur	805	17318	BERT: 0 daELECTRA: 17,212 daBERT: 17,597	BERT: 0.00 daELECTRA: 94.97 daBERT: 97.10	BERT: 0 daELECTRA: 81 daBERT: 70
LC8: Non-hateful use of slur	395	7881	BERT: 8,276 daELECTRA: 318 daBERT: 222	BERT: 100.00 daELECTRA: 3.84 daBERT: 2.68	BERT: 0 daELECTRA: 20 daBERT: 26
LC9: Hate expressed using profanity	1842	49743	BERT: 0 daELECTRA: 49,739 daBERT: 50,158	BERT: 0.00 daELECTRA: 96.42 daBERT: 97.23	BERT: 0 daELECTRA: 188 daBERT: 164
LC10: Non-Hateful use of profanity	701	15761	BERT: 16,462 daELECTRA: 541 daBERT: 549	BERT: 100.00 daELECTRA: 3.29 daBERT: 3.33	BERT: 0 daELECTRA: 51 daBERT: 58
LC11: Hate expressed through reference in subsequent clauses	11968	44890	BERT: 5,984 daELECTRA: 48,776 daBERT: 48,642	BERT: 10.52 daELECTRA: 85.79 daBERT: 85.55	BERT: 0 daELECTRA: 1,164 daBERT: 1,169
LC12: Hate expressed through reference in subsequent sentences	11968	42840	BERT: 5,984 daELECTRA: 46,721 daBERT: 46,156	BERT: 10.92 daELECTRA: 85.24 daBERT: 84.21	BERT: 0 daELECTRA: 1,070 daBERT: 1,726
LC13: Hate expressed using negated positive statement	23379	126742	BERT: 0 daELECTRA: 144,823 daBERT: 145,761	BERT: 0.00 daELECTRA: 96.47 daBERT: 97.10	BERT: 0 daELECTRA: 2,804 daBERT: 2,290

TABLE II: Results of BERT-base, RoBERTa-base and DistilBERT-base sentiment analysis models on CHECKLIST test cases. BERT-base, RoBERTa-base and DistilBERT-base models are denoted as BERT, RoBERTa and dstBERT, respectively.

Linguistic capability	S^2LCT #Seeds	S^2LCT #Fail	S^2LCT Fail rate[%]
LC1: Short sentences with neutral adjectives and nouns	1,716	BERT: 1,330 RoBERTa: 1,391 dstBERT: 1,661	BERT: 77.51 RoBERTa: 81.06 dstBERT: 96.79
LC2: Short sentences with sentiment-laden adjectives	8,658	BERT: 26 RoBERTa: 139 dstBERT: 125	BERT: 0.30 RoBERTa: 1.61 dstBERT: 1.44
LC3: Sentiment change over time, present should prevail	8,000	BERT: 1,680 RoBERTa: 829 dstBERT: 2,532	BERT: 21.00 RoBERTa: 10.36 dstBERT: 31.65
LC4: Negated negative should be positive or neutral	6,786	BERT: 799 RoBERTa: 218 dstBERT: 734	BERT: 11.77 RoBERTa: 3.21 dstBERT: 10.82
LC5: Negated neutral should still be neutral	2,496	BERT: 2,427 RoBERTa: 2,304 dstBERT: 2,450	BERT: 97.24 RoBERTa: 92.31 dstBERT: 98.16
LC6: Negation of negative at the end, should be positive or neutral	2,124	BERT: 1,871 RoBERTa: 445 dstBERT: 2,124	BERT: 88.09 RoBERTa: 20.95 dstBERT: 100.00
LC7: Negated positive with neutral content in the middle	1,000	BERT: 860 RoBERTa: 416 dstBERT: 865	BERT: 86.00 RoBERTa: 41.60 dstBERT: 86.50
LC8: Author sentiment is more important than of others	8,528	BERT: 3,741 RoBERTa: 2,692 dstBERT: 3,535	BERT: 43.87 RoBERTa: 31.57 dstBERT: 41.45
LC9: Parsing sentiment in (question, yes) form	9,204	BERT: 1,793 RoBERTa: 1,262 dstBERT: 1,557	BERT: 19.48 RoBERTa: 13.71 dstBERT: 16.92
LC10: Parsing sentiment in (question, no) form	7,644	BERT: 4,056 RoBERTa: 4,576 dstBERT: 6,440	BERT: 53.06 RoBERTa: 59.86 dstBERT: 84.25