S^2LCT: Specification- and Syntax-based Automated Testing Linguistic Capabilities of NLP Models

I. Hate speech detection task

TABLE I: Results of BERT-base, da-ELECTRA and da-BERT-base hate speech detection models on S^2LCT test cases using all seeds.

Linguistic capability	**S^2LCT #Seeds**	**S^2LCT #Exps**	**S^2LCT #Fail**	**S^2LCT Fail rate[%]**	**S^2LCT #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