BERT model (run on CHECKLIST test cases)

test_type	linguistic_cap	num_tests	num_tests_run	num_fails	fail_rate[%]
Vocab_POS	Short sentences with neutral adjectives and nouns	8658	500	1	0.2
Vocab_POS	Short sentences with sentiment-laden adjectives	1716	500	385	77
Vocab_POS	Replace neutral words with other neutral words	500	500	64	12.8
Vocab_POS	Add positive phrases and fails if sent. goes down by > 0.1	500	500	0	0
Vocab_POS	Add negative phrases and fails if sent. goes up by > 0.1	500	500	21	4.2
Robustness	Add randomly generated URLs and handles	500	500	71	14.2
Robustness	Strip punctuation and/or add "."	500	500	28	5.6
Robustness	Swap two adjacent characters	500	500	44	8.8
Robustness	Contract or expand contractions	1000	500	11	2.2
NER	Replace names with other common names	331	331	29	8.8
NER	Replace city or country names with other cities or countries	909	500	59	11.8
NER	Replace integers with random integers within a 20% radius of the original	1000	500	21	4.2

BERT model (run on Proposed model test cases)

test_type	linguistic_cap	num_seed_tests	num_seed_tests_run	num_seed_fails	seed_fail_rate[%]	num_seed_tests	num_seed_temp_tests_run	num_seed_temp_fails	seed_temp_fail_rate[%]	num_exp_temp_tests	num_exp_temp_tests_run	num_exp_temp_fails	exp_temp_fail_rate[%]
Vocab_POS	Short sentences with neutral adjectives and nouns	10	10	8	80	282	282	233	82.6	4	4	4	100
Vocab_POS	Short sentences with sentiment-laden adjectives	10	10	0	0	7547	500	18	3.6	9594	500	195	39
Vocab_POS	Replace neutral words with other neutral words	10	10	1	10	500	500	4	0.8	2275	500	2	0.4
Vocab_POS	Add positive phrases and fails if sent. goes down by > 0.1	10	10	0	0	500	500	0	0				
Vocab_POS	Add negative phrases and fails if sent. goes up by > 0.1	10	10	0	0	500	500	0	0				
Robustness	Add randomly generated URLs and handles	10	10										
Robustness	Strip punctuation and/or add "."	10	10	0	0	500	500	2	0.4	500	500	0	0
Robustness	Swap two adjacent characters	10	10	1	10	500	500	14	2.8				
Robustness	Contract or expand contractions	5	5	1	20					1000	500	0	0
NER	Replace names with other common names	10	10	0	0	1000	500	36	7.2	1000	500	0	0
NER	Replace city or country names with other cities or countries	10	10	0	0	930	500	0	0	1000	500	4	0.8
NER	Replace integers with random integers within a 20% radius of the original	10	10	0	0	1000	500	0	0	1000	500	0	0

RoBERTa model (run on CHECKLIST test cases)

test_type	linguistic_cap	num_tests	num_tests_run	num_fails	fail_rate[%]
Vocab_POS	Short sentences with neutral adjectives and nouns	8658	500	7	1.4
Vocab_POS	Short sentences with sentiment-laden adjectives	1716	500	419	83.8
Vocab_POS	Replace neutral words with other neutral words	500	500	55	11
Vocab_POS	Add positive phrases and fails if sent. goes down by > 0.1	500	500	16	3.2
Vocab_POS	Add negative phrases and fails if sent. goes up by > 0.1	500	500	26	5.2
Robustness	Add randomly generated URLs and handles	500	500	57	11.4
Robustness	Strip punctuation and/or add "."	500	500	25	5
Robustness	Swap two adjacent characters	500	500	22	4.4
Robustness	Contract or expand contractions	1000	500	6	1.2
NER	Replace names with other common names	331	331	15	4.5
NER	Replace city or country names with other cities or countries	909	500	33	6.6
NER	Replace integers with random integers within a 20% radius of the original	1000	500	13	2.6

RoBERTa model (run on Proposed model test cases)

test_type	linguistic_cap	num_seed_tests	num_seed_tests_run	num_seed_fails	seed_fail_rate[%]	num_seed_tests	num_seed_temp_tests_run	num_seed_temp_fails	seed_temp_fail_rate[%]	num_exp_temp_tests	num_exp_temp_tests_run	num_exp_temp_fails	exp_temp_fail_rate[%]
Vocab_POS	Short sentences with neutral adjectives and nouns	10	10	9	90	282	282	249	88.3	4	4	4	100
Vocab_POS	Short sentences with sentiment-laden adjectives	10	10	0	0	7547	500	19	3.8	9594	500	138	27.6
Vocab_POS	Replace neutral words with other neutral words	10	10	2	20	500	500	454	90.8	500	500	1	0.2
Vocab_POS	Add positive phrases and fails if sent. goes down by > 0.1	10	10	0	0	500	500	1	0.2				
Vocab_POS	Add negative phrases and fails if sent. goes up by > 0.1	10	10	0	0	500	500	0	0				
Robustness	Add randomly generated URLs and handles	10	10										
Robustness	Strip punctuation and/or add "."	10	10	1	10	500	500	7	1.4	500	500	0	0
Robustness	Swap two adjacent characters	10	10	0	0	500	500	4	0.8				
Robustness	Contract or expand contractions	5	5	0	0					1000	500	0	0
NER	Replace names with other common names	10	10	0	0	1000	500	10	2	1000	500	0	0
NER	Replace city or country names with other cities or countries	10	10	0	0	930	500	0	0	1000	500	0	0
NER	Replace integers with random integers within a 20% radius of the original	10	10	0	0	1000	500	0	0	1000	500	0	0

DistilBERT model (run on CHECKLIST test cases)

test_type	linguistic_cap	num_tests	num_tests_run	num_fails	fail_rate[%]
Vocab_POS	Short sentences with neutral adjectives and nouns	8658	500	7	1.4
Vocab_POS	Short sentences with sentiment-laden adjectives	1716	500	479	95.8
Vocab_POS	Replace neutral words with other neutral words	500	500	49	9.8
Vocab_POS	Add positive phrases and fails if sent. goes down by > 0.1	500	500	0	0
Vocab_POS	Add negative phrases and fails if sent. goes up by > 0.1	500	500	34	6.8
Robustness	Add randomly generated URLs and handles	500	500	77	15.4
Robustness	Strip punctuation and/or add "."	500	500	27	5.4
Robustness	Swap two adjacent characters	500	500	33	6.6
Robustness	Contract or expand contractions	1000	500	13	2.6
NER	Replace names with other common names	331	331	17	5.1
NER	Replace city or country names with other cities or countries	909	500	44	8.8
NER	Replace integers with random integers within a 20% radius of the original	1000	500	19	3.8

DistilBERT model (run on Proposed model test cases)

test_type	linguistic_cap	num_seed_tests	num_seed_tests_run	num_seed_fails	seed_fail_rate[%]	num_seed_tests	num_seed_temp_tests_run	num_seed_temp_fails	seed_temp_fail_rate[%]	num_exp_temp_tests	num_exp_temp_tests_run	num_exp_temp_fails	exp_temp_fail_rate[%]
Vocab_POS	Short sentences with neutral adjectives and nouns	10	10	10	100	282	282	265	94	4	4	4	100
Vocab_POS	Short sentences with sentiment-laden adjectives	10	10	0	0	7547	500	22	4.4	9594	500	177	35.4
Vocab_POS	Replace neutral words with other neutral words	10	10	1	10	500	500	14	2.8	500	500	13	2.6
Vocab_POS	Add positive phrases and fails if sent. goes down by > 0.1	10	10	0	0	500	500	0	0				
Vocab_POS	Add negative phrases and fails if sent. goes up by > 0.1	10	10	0	0	500	500	0	0				
Robustness	Add randomly generated URLs and handles												
Robustness	Strip punctuation and/or add "."	10	10	1	10	500	500	2	0.4	500	500	0	0
Robustness	Swap two adjacent characters	10	10	0	0	500	500	55	11				
Robustness	Contract or expand contractions	5	5	0	0					1000	500	0	0
NER	Replace names with other common names	10	10	0	0	1000	500	8	1.6	1000	500	0	0
NER	Replace city or country names with other cities or countries	10	10	0	0	930	500	13	2.6	1000	500	31	6.2
NER	Replace integers with random integers within a 20% radius of the original	10	10	0	0	1000	500	0	0	1000	500	0	0