Verifying the Robustness of Automatic Credibility Assessment
Appendix

Piotr Przybyła[1,2], Alexander Shvets[1], and Horacio Saggion[1]

[1]Universitat Pompeu Fabra, Tànger building,
Barcelona 08018, Spain
[2]Institute of Computer Science, Polish Academy of Sciences,
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland

Full evaluation results

This document contains full results obtained by evaluating the adversarial attacks on the misinformation classifiers. Specifically, we test the attack performance for:

- four tasks (HN, PR, FC, RD),
- eight attackers (BAE, BERT-ATTACK, DeepWordBug, Genetic, SememePSO, PWWS, SCPN, TextFooler),
- four victims (BiLSTM, BERT, GEMMA2B, GEMMA7B),
- two scenarios (untargeted and targeted).

The following tables include the results for victims: BiLSTM (Table 1), BERT (Table 2), GEMMA2B (Table 3) and GEMMA7B (Table 4).

See the main article for details.

		Untargeted					**Targeted**				
Task	**Method**	**B.**	**Con**	**Sem**	**Char**	**Q.**	**B.**	**con**	**sem**	**char**	**Q.**
HN	BAE	0.48	0.77	0.64	0.98	489.27	0.45	0.74	0.62	0.98	477.65
	BERT-ATTACK	**0.64**	**0.98**	0.66	0.99	487.85	**0.61**	**0.96**	0.65	0.99	565.05
	DeepWordBug	0.41	0.53	**0.77**	**1.00**	396.18	0.37	0.47	**0.78**	**1.00**	379.20
	Genetic	0.44	0.94	0.48	0.98	2029.31	0.42	0.90	0.47	0.98	2882.19
	SememePSO	0.21	0.42	0.50	0.99	313.51	0.14	0.28	0.49	0.99	361.38
	PWWS	0.44	0.93	0.48	0.99	2044.96	0.42	0.89	0.48	0.97	1994.95
	SCPN	0.00	0.94	0.08	0.02	11.86	0.00	0.95	0.08	0.02	11.83
	TextFooler	0.43	0.94	0.47	0.97	543.68	0.41	0.91	0.47	0.96	598.46
PR	BAE	0.15	0.23	0.72	0.94	32.94	0.26	0.38	0.71	0.94	38.72
	BERT-ATTACK	0.53	0.80	0.72	0.91	61.41	**0.66**	0.94	0.74	0.94	50.14
	DeepWordBug	0.29	0.38	**0.79**	**0.96**	27.45	0.56	0.72	**0.81**	**0.96**	35.30
	Genetic	**0.54**	**0.88**	0.67	0.89	782.15	0.62	0.94	0.71	0.93	802.20
	SememePSO	0.47	0.76	0.68	0.89	85.34	0.60	0.92	0.71	0.92	69.62
	PWWS	0.53	0.84	0.69	0.90	130.85	0.63	0.92	0.73	0.94	168.60
	SCPN	0.12	0.55	0.39	0.50	11.55	0.20	**0.98**	0.37	0.48	11.98
	TextFooler	0.51	0.85	0.67	0.88	52.59	0.63	0.94	0.72	0.92	54.62
FC	BAE	0.36	0.55	0.69	0.96	77.76	0.32	0.48	0.69	0.96	73.43
	BERT-ATTACK	0.60	0.86	0.73	0.95	132.80	**0.59**	0.85	0.73	0.96	123.24
	DeepWordBug	0.48	0.58	**0.85**	**0.98**	54.36	0.54	0.64	**0.85**	**0.98**	50.72
	Genetic	**0.61**	**0.90**	0.71	0.95	840.99	0.57	0.88	0.69	0.94	1015.44
	SememePSO	0.53	0.76	0.72	0.96	112.84	0.46	0.67	0.72	0.96	132.28
	PWWS	0.57	0.82	0.73	0.96	221.60	0.50	0.73	0.71	0.95	211.05
	SCPN	0.08	0.75	0.29	0.32	11.75	0.11	**1.00**	0.30	0.35	12.00
	TextFooler	0.55	0.82	0.71	0.94	98.31	0.50	0.75	0.70	0.94	99.98
RD	BAE	0.09	0.21	0.43	0.98	312.77	0.27	0.64	0.43	0.98	123.16
	BERT-ATTACK	0.29	**0.79**	0.41	0.89	985.52	0.43	0.95	0.46	0.97	130.64
	DeepWordBug	0.16	0.24	**0.68**	**0.99**	232.75	**0.62**	0.91	**0.69**	**0.99**	153.61
	Genetic	**0.32**	0.71	0.47	0.96	3150.24	0.44	**0.96**	0.48	0.95	1355.52
	SememePSO	0.15	0.31	0.48	0.97	314.63	0.32	0.67	0.50	0.97	185.47
	PWWS	0.29	0.64	0.47	0.97	1059.07	0.44	0.95	0.48	0.95	742.12
	SCPN	0.01	0.55	0.17	0.09	11.53	0.02	0.84	0.15	0.12	11.84
	TextFooler	0.24	0.64	0.41	0.87	639.97	0.44	**0.96**	0.48	0.96	184.97

Table 1: The results of adversarial attacks on the **BiLSTM** classifier in four misinformation detection tasks in untargeted and targeted scenario. Evaluation measures include BODEGA score (B.), confusion score (con), semantic score (sem), character score (char) and number of queries to the attacked model (Q.). The best score in each task and scenario is in boldface.

		Untargeted					**Targeted**				
Task	**Method**	**B.**	**Con**	**Sem**	**Char**	**Q.**	**B.**	**con**	**sem**	**char**	**Q.**
HN	BAE	0.34	0.60	0.58	0.96	606.83	0.18	0.34	0.57	0.95	713.42
	BERT-ATTACK	**0.60**	**0.96**	0.64	0.97	648.41	**0.57**	**0.95**	0.62	0.96	753.91
	DeepWordBug	0.22	0.29	**0.78**	**1.00**	395.94	0.15	0.20	**0.78**	**1.00**	389.81
	Genetic	0.40	0.86	0.47	0.98	2713.80	0.30	0.71	0.44	0.97	4502.51
	SememePSO	0.16	0.34	0.50	0.99	341.70	0.05	0.12	0.44	0.99	417.99
	PWWS	0.38	0.82	0.47	0.98	2070.78	0.27	0.64	0.44	0.95	2107.02
	SCPN	0.00	0.92	0.08	0.02	11.84	0.00	**0.95**	0.09	0.02	11.89
	TextFooler	0.39	0.92	0.44	0.94	660.52	0.32	0.85	0.41	0.90	850.79
PR	BAE	0.11	0.18	0.69	0.94	33.96	0.13	0.20	0.68	0.94	45.68
	BERT-ATTACK	0.43	0.70	0.68	0.90	80.16	**0.50**	0.79	0.69	0.92	99.95
	DeepWordBug	0.28	0.36	**0.79**	**0.96**	27.43	0.50	0.64	**0.81**	**0.96**	36.04
	Genetic	**0.50**	**0.84**	0.65	0.89	962.40	0.49	**0.84**	0.65	0.89	1211.56
	SememePSO	0.41	0.68	0.66	0.90	96.17	0.35	0.53	0.71	0.91	173.71
	PWWS	0.47	0.75	0.68	0.91	131.92	0.44	0.72	0.68	0.89	179.68
	SCPN	0.09	0.47	0.36	0.46	11.47	0.11	0.79	0.32	0.39	11.79
	TextFooler	0.43	0.77	0.64	0.87	57.94	0.46	0.77	0.66	0.89	77.81
FC	BAE	0.34	0.51	0.70	0.96	80.69	0.18	0.27	0.70	0.94	92.47
	BERT-ATTACK	**0.53**	0.77	0.73	0.95	146.73	**0.41**	0.62	0.71	0.93	207.23
	DeepWordBug	0.44	0.53	**0.84**	**0.98**	54.32	0.22	0.27	**0.85**	**0.98**	52.31
	Genetic	0.52	0.79	0.70	0.95	1215.19	0.39	0.63	0.66	0.92	1808.08
	SememePSO	0.44	0.64	0.71	0.96	148.20	0.25	0.37	0.70	0.94	230.58
	PWWS	0.48	0.69	0.72	0.96	225.27	0.31	0.47	0.70	0.94	226.78
	SCPN	0.09	**0.90**	0.29	0.31	11.90	0.09	**0.97**	0.29	0.30	11.97
	TextFooler	0.46	0.70	0.70	0.93	106.13	0.29	0.49	0.65	0.88	131.88
RD	BAE	0.07	0.18	0.41	0.98	313.01	0.18	0.44	0.42	0.98	196.69
	BERT-ATTACK	0.18	0.44	0.43	0.96	774.31	0.30	0.69	0.45	0.97	366.14
	DeepWordBug	0.16	0.23	**0.70**	**0.99**	232.74	**0.39**	0.56	**0.70**	**0.99**	174.03
	Genetic	**0.20**	**0.46**	0.45	0.96	4425.11	0.35	0.79	0.46	0.95	2266.91
	SememePSO	0.10	0.21	0.46	0.97	345.89	0.27	0.57	0.49	0.96	233.88
	PWWS	0.16	0.38	0.45	0.95	1105.99	0.32	0.75	0.45	0.93	838.83
	SCPN	0.01	0.38	0.16	0.10	11.35	0.02	**0.90**	0.15	0.10	11.90
	TextFooler	0.16	0.41	0.43	0.91	657.15	0.31	0.70	0.47	0.96	358.37

Table 2: The results of adversarial attacks on the **BERT** classifier (see the caption of the previous table).

		Untargeted					**Targeted**				
Task	**Method**	**B.**	**Con**	**Sem**	**Char**	**Q.**	**B.**	**con**	**sem**	**char**	**Q.**
HN	BAE	0.36	0.61	0.60	0.97	583.38	0.38	0.66	0.59	0.97	545.68
	BERT-ATTACK	**0.55**	**0.91**	0.62	0.96	942.98	**0.57**	0.95	0.62	0.97	761.53
	DeepWordBug	0.24	0.31	**0.78**	**1.00**	385.91	0.27	0.34	**0.78**	**1.00**	380.58
	Genetic	0.40	0.84	0.48	0.99	1176.75	0.44	0.93	0.48	0.99	894.47
	SememePSO	0.26	0.54	0.48	1.00	236.96	0.32	0.66	0.49	0.99	185.02
	PWWS	0.38	0.82	0.48	0.98	2021.92	0.45	0.97	0.47	0.98	1980.32
	SCPN	0.00	0.69	0.09	0.02	11.62	0.00	**0.99**	0.10	0.02	11.88
	TextFooler	0.32	0.70	0.47	0.97	834.56	0.41	0.89	0.47	0.98	583.29
PR	BAE	0.16	0.23	0.72	0.94	32.64	0.21	0.32	0.69	0.94	43.68
	BERT-ATTACK	0.46	0.72	0.70	0.91	77.50	**0.55**	0.85	0.70	0.92	94.05
	DeepWordBug	0.14	0.19	**0.79**	**0.96**	27.26	0.29	0.38	**0.80**	**0.97**	36.04
	Genetic	**0.50**	**0.84**	0.66	0.90	876.06	0.51	**0.86**	0.65	0.91	1077.41
	SememePSO	0.42	0.69	0.67	0.90	93.41	0.35	0.57	0.67	0.91	180.15
	PWWS	0.45	0.74	0.67	0.91	132.13	0.44	0.73	0.66	0.90	179.97
	SCPN	0.14	0.62	0.40	0.51	11.62	0.15	0.74	0.36	0.45	11.74
	TextFooler	0.36	0.63	0.65	0.87	68.64	0.37	0.66	0.63	0.87	97.49
FC	BAE	0.34	0.51	0.70	0.96	79.43	0.17	0.26	0.69	0.96	92.43
	BERT-ATTACK	**0.57**	**0.83**	0.72	0.94	192.25	**0.50**	0.76	0.70	0.92	254.22
	DeepWordBug	0.07	0.09	**0.83**	**0.98**	53.88	0.06	0.08	**0.84**	**0.98**	52.02
	Genetic	0.46	0.68	0.71	0.96	486.65	0.31	0.50	0.68	0.93	626.83
	SememePSO	0.42	0.62	0.71	0.96	155.18	0.22	0.33	0.70	0.94	242.62
	PWWS	0.43	0.63	0.72	0.96	228.43	0.22	0.33	0.71	0.95	230.18
	SCPN	0.06	0.51	0.31	0.34	11.51	0.11	**1.00**	0.31	0.34	12.00
	TextFooler	0.43	0.65	0.71	0.94	120.38	0.24	0.38	0.67	0.92	137.26
RD	BAE	0.13	0.31	0.42	0.98	298.30	0.19	0.47	0.41	0.97	170.87
	BERT-ATTACK	**0.30**	**0.73**	0.43	0.95	703.07	0.41	0.93	0.45	0.96	259.36
	DeepWordBug	0.10	0.15	**0.69**	**0.99**	238.97	0.24	0.35	**0.69**	**0.99**	161.87
	Genetic	0.24	0.54	0.46	0.96	1647.75	**0.41**	0.90	0.47	0.96	1108.21
	SememePSO	0.12	0.27	0.46	0.98	323.73	0.24	0.53	0.47	0.97	214.95
	PWWS	0.21	0.46	0.47	0.97	1124.29	0.40	0.89	0.47	0.96	774.03
	SCPN	0.01	0.38	0.16	0.11	11.38	0.02	**0.95**	0.16	0.13	11.95
	TextFooler	0.19	0.44	0.45	0.94	640.54	0.38	0.86	0.47	0.94	260.58

Table 3: The results of adversarial attacks on the **GEMMA2B** classifier (see the caption of the previous table).

		Untargeted					**Targeted**				
Task	**Method**	**B.**	**Con**	**Sem**	**Char**	**Q.**	**B.**	**con**	**sem**	**char**	**Q.**
HN	BAE	0.25	0.42	0.62	0.97	678.07	0.11	0.18	0.62	0.98	740.02
	BERT-ATTACK	**0.45**	**0.75**	0.62	0.95	1560.76	**0.31**	0.55	0.59	0.93	2313.33
	DeepWordBug	0.14	0.18	**0.79**	**1.00**	385.78	0.04	0.05	**0.81**	**1.00**	377.38
	Genetic	0.30	0.61	0.49	0.99	1042.31	0.19	0.42	0.47	0.98	1159.68
	SememePSO	0.15	0.29	0.51	0.99	344.23	0.07	0.14	0.50	1.00	403.99
	PWWS	0.28	0.60	0.48	0.98	2068.81	0.17	0.40	0.45	0.96	2095.51
	SCPN	0.00	0.73	0.09	0.02	11.66	0.00	**0.65**	0.09	0.03	11.54
	TextFooler	0.22	0.51	0.46	0.94	1132.87	0.16	0.40	0.43	0.91	1284.86
PR	BAE	0.13	0.19	0.72	0.94	33.69	0.22	0.32	0.72	0.94	44.26
	BERT-ATTACK	0.42	0.67	0.69	0.91	86.54	**0.53**	0.78	0.72	0.93	110.32
	DeepWordBug	0.09	0.12	**0.79**	**0.96**	27.19	0.23	0.29	**0.81**	**0.97**	35.37
	Genetic	**0.44**	**0.78**	0.63	0.88	925.58	0.49	0.85	0.64	0.89	845.27
	SememePSO	0.34	0.58	0.65	0.88	123.13	0.38	0.62	0.68	0.91	165.24
	PWWS	0.38	0.64	0.65	0.89	133.92	0.43	0.71	0.66	0.90	179.77
	SCPN	0.10	0.51	0.36	0.47	11.52	0.16	**0.86**	0.36	0.46	11.86
	TextFooler	0.34	0.62	0.62	0.86	69.88	0.40	0.65	0.68	0.90	86.74
FC	BAE	0.36	0.55	0.68	0.96	76.44	0.20	0.31	0.68	0.95	87.62
	BERT-ATTACK	**0.58**	**0.85**	0.72	0.95	141.70	**0.52**	0.80	0.69	0.93	173.86
	DeepWordBug	0.04	0.06	**0.81**	**0.98**	53.84	0.02	0.03	**0.81**	**0.99**	51.55
	Genetic	0.47	0.73	0.68	0.94	842.16	0.34	0.58	0.64	0.91	1205.46
	SememePSO	0.39	0.58	0.70	0.96	164.01	0.23	0.35	0.69	0.95	228.84
	PWWS	0.42	0.62	0.70	0.96	228.62	0.24	0.37	0.68	0.94	227.46
	SCPN	0.06	0.50	0.31	0.34	11.50	0.11	**1.00**	0.31	0.34	12.00
	TextFooler	0.41	0.63	0.69	0.94	114.26	0.24	0.40	0.66	0.91	132.95
RD	BAE	0.10	0.26	0.41	0.97	318.74	0.21	0.53	0.40	0.97	168.33
	BERT-ATTACK	**0.21**	**0.52**	0.43	0.95	977.27	**0.44**	**1.00**	0.46	0.96	202.18
	DeepWordBug	0.08	0.12	**0.70**	**0.99**	238.93	0.21	0.31	**0.70**	**0.99**	156.29
	Genetic	0.18	0.40	0.47	0.97	1197.40	0.42	0.93	0.46	0.96	1531.17
	SememePSO	0.11	0.23	0.48	0.98	336.53	0.26	0.58	0.46	0.97	208.59
	PWWS	0.17	0.37	0.47	0.97	1139.02	0.39	0.88	0.46	0.96	762.62
	SCPN	0.01	0.34	0.17	0.13	11.34	0.03	0.98	0.17	0.14	11.98
	TextFooler	0.15	0.36	0.45	0.95	679.12	0.40	0.90	0.46	0.95	249.94

Table 4: The results of adversarial attacks on the **GEMMA7B classifier** (see the caption of the previous table).