

Table S1. Comparison of adversarial examples generated by different attack methods against BERT on an IMDB review sample. The original label is negative (0), and all adversarial examples successfully flip the prediction to positive (1).

Method	Text
Original	This movie was really awful. It was not in the least bit frightening, or even startling. I went to see it with a bunch of friends and by the end of the night we were saying the ruins ruined my night. I would not recommend seeing this movie in theaters, renting it or even watching the movie on television by accident. It is an absolute waste of an hour and a half. The plot was nearly non existent, the characters were horribly underdeveloped, and they gave no back story whatsoever for anything that was happening, and then left it completely open at the end as if preparing for a sequel.
TextFooler	This movie was really heinous . It was not in the least bit frightening, or even alarming . I went to see it with a bunch of friends and by the end of the night we were proclaiming the ruins ruined my night. I would not suggestion peek this flick in theaters, leasing it or even stare the cinemas on television by event . It is an absolute wastes of an hour and a east ab . The plots was barely non existent, the character were terrifically underdevelopment , and they dispensed no back story everyone for anything that was happening, and then going it keenly open at the finish as if crafted for a sequel.
HLGA	This movie was really loathsome . It was not in the least bit less frightening, or even startling. I went to see it with a bunch of friends and by the end of the night we were saying the ruins ruined my night. I ought not recommend seeing this movie in theaters, renting it or even watching the movie on television by accident. It is an absolute waste of an hour and a half. The plot was nearly non nowadays , the characters were horribly underdeveloped, and they gave no back story whatsoever for anything that was happening, and then left it completely open at the end as if preparing for a sequel.
PSO	This movie was really mediocre . It was not in the least bit thrilling , or even startling. I went to see it with a bunch of friends and by the end of the night we were saying the ruins affected my night. I would not recommend seeing this movie in theaters, renting it or even watching the movie on television by accident. It is an absolute investment of an hour and a half. The plot was nearly non existent, the characters were interestingly underdeveloped, and they gave no back story whatsoever for anything that was happening, and then left it completely open at the end as if preparing for a sequel.
SSPAttack	This movie was really terrifying . It was not in the least bit atrocious , or even startling. I went to see it with a bunch of friends and by the end of the night we were saying the ruins ruined my night. I would wo recommend seeing this movie in theaters, renting it or even watching the movie on television by accident. It is an absolute waste of an hour and a half. The plot was nearly non existent, the characters were fantastically underdeveloped, and they gave no back story anyhow for anything that was happening, and then left it completely open at the end as if preparing for a sequel.
TextCheater	This movie was really fantastic . It was not in the least bit frightening, or even shocking . I went to see it with a bunch of friends and by the end of the night we were saying the ruins enhanced my night. I would not recommend seeing this movie in theaters, renting it or even watching the movie on television by accident. It is an absolute pleasure of an hour and a half. The plot was nearly non existent, the characters were brilliantly underdeveloped, and they gave no back story whatsoever for anything that was happening, and then left it completely open at the end as if preparing for a sequel.
KEM-Attack	This movie was really intense . It was not in the least bit frightening, or even surprising . I went to see it with a bunch of friends and by the end of the night we were saying the ruins intrigued my night. I would not recommend seeing this movie in theaters, renting it or even watching the movie on television by accident. It is an absolute impressive of an hour and a half. The plot was nearly non existent, the characters were moderately underdeveloped, and they gave no back story whatsoever for anything that was happening, and then left it completely open at the end as if preparing for a sequel.

Table S2. Stability analysis across 3 random seeds (42, 123, 456) for KEM-Attack. All values: mean $\pm$ std. Low standard deviations confirm robust performance.

Dataset	Model	ASR (%)	Sim (%)	Pert (%)
MR	BERT	98.6 $\pm$ 1.1	76.2 $\pm$ 2.0	10.837 $\pm$ 0.615
	WordCNN	99.4 $\pm$ 0.7	75.6 $\pm$ 2.3	10.218 $\pm$ 0.542
	WordLSTM	99.3 $\pm$ 0.8	71.8 $\pm$ 2.5	10.892 $\pm$ 0.689
IMDB	BERT	99.9 $\pm$ 0.2	92.3 $\pm$ 1.6	2.322 $\pm$ 0.173
	WordCNN	100.0 $\pm$ 0.0	94.8 $\pm$ 1.3	2.925 $\pm$ 0.195
	WordLSTM	99.8 $\pm$ 0.3	93.4 $\pm$ 1.5	2.436 $\pm$ 0.162
Yelp	BERT	99.5 $\pm$ 0.5	85.2 $\pm$ 2.1	4.875 $\pm$ 0.398
	WordCNN	99.7 $\pm$ 0.4	89.5 $\pm$ 1.7	6.103 $\pm$ 0.512
	WordLSTM	99.8 $\pm$ 0.3	88.6 $\pm$ 1.8	4.756 $\pm$ 0.376
AG's News	BERT	96.2 $\pm$ 1.4	75.1 $\pm$ 2.4	10.127 $\pm$ 0.823
	WordCNN	98.8 $\pm$ 1.0	83.6 $\pm$ 2.0	7.932 $\pm$ 0.658
	WordLSTM	93.7 $\pm$ 1.9	76.8 $\pm$ 2.6	9.683 $\pm$ 0.728
Average std across all conditions		0.87	2.01	0.506

Table S3. Wilcoxon signed-rank test results comparing KEM-Attack with baseline methods on MR dataset (BERT model). Significance level: $\alpha = 0.05$. $\checkmark$ indicates $p < 0.05$ (significant), $\times$ indicates $p \geq 0.05$ (not significant).

Metric	Baseline	KEM-Attack	Baseline	Mean Diff.	p-value (Sig.)
Attack Success Rate (%)	TextFooler	98.6	92.6	+6.0	< 0.001 ($\checkmark$)
	PSO	98.6	93.4	+5.2	< 0.001 ($\checkmark$)
	HLGA	98.6	95.7	+2.9	0.002 ($\checkmark$)
	SSPAttack	98.6	97.8	+0.8	0.287 ($\times$)
	TextCheater	98.6	98.3	+0.3	0.745 ($\times$)
Semantic Similarity (%)	TextFooler	76.2	57.1	+19.1	< 0.001 ($\checkmark$)
	PSO	76.2	53.8	+22.4	< 0.001 ($\checkmark$)
	HLGA	76.2	64.8	+11.4	< 0.001 ($\checkmark$)
	SSPAttack	76.2	73.5	+2.7	0.042 ($\checkmark$)
	TextCheater	76.2	70.2	+6.0	< 0.001 ($\checkmark$)
Perturbation Rate (%)	TextFooler	10.84	16.28	-5.44	< 0.001 ($\checkmark$)
	PSO	10.84	20.38	-9.54	< 0.001 ($\checkmark$)
	HLGA	10.84	13.36	-2.52	0.003 ($\checkmark$)
	SSPAttack	10.84	10.62	+0.22	0.651 ($\times$)
	TextCheater	10.84	13.13	-2.29	0.008 ($\checkmark$)
Query Count	TextFooler	—	—	—	N/A ¹
	PSO	5,000	23,000	-18,000	< 0.001 ($\checkmark$)
	HLGA	5,000	16,000	-11,000	< 0.001 ($\checkmark$)
	SSPAttack	5,000	9,000	-4,000	< 0.001 ($\checkmark$)
	TextCheater	5,000	5,000	0	0.923 ($\times$)

Table S4. Proportion of significant ($p < 0.05$) improvements across all dataset-model combinations (12 in total).

Baseline Method	ASR Significant	Similarity Significant	Perturbation Significant	Query Significant
TextFooler	11/12 (91.7%)	12/12 (100%)	12/12 (100%)	N/A
PSO	10/12 (83.3%)	12/12 (100%)	12/12 (100%)	12/12 (100%)
HLGA	8/12 (66.7%)	12/12 (100%)	11/12 (91.7%)	12/12 (100%)
SSPAttack	3/12 (25.0%)	10/12 (83.3%)	4/12 (33.3%)	11/12 (91.7%)
TextCheater	2/12 (16.7%)	11/12 (91.7%)	10/12 (83.3%)	2/12 (16.7%)
Overall	34/60 (56.7%)	57/60 (95.0%)	49/60 (81.7%)	37/48 (77.1%)