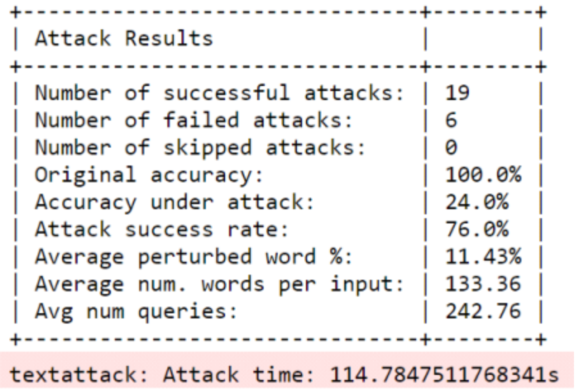
Comparison with Faster_genetic_algorithm-Jia_2019

08 May 2021 00:10

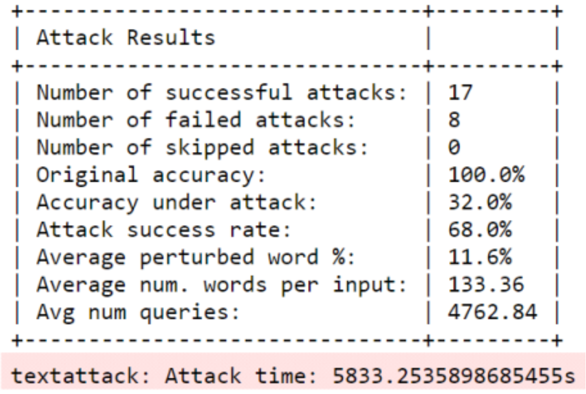
AlBERT-base-v2-yelp-polarity:

Our attack



```
+-------------------------------+--------+
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 19     |
| Number of failed attacks:     | 6      |
| Number of skipped attacks:    | 0      |
| Original accuracy:            | 100.0% |
| Accuracy under attack:        | 24.0%  |
| Attack success rate:          | 76.0%  |
| Average perturbed word %:     | 11.43% |
| Average num. words per input: | 133.36 |
| Avg num queries:              | 242.76 |
+-------------------------------+--------+
textattack: Attack time: 114.7847511768341s
```

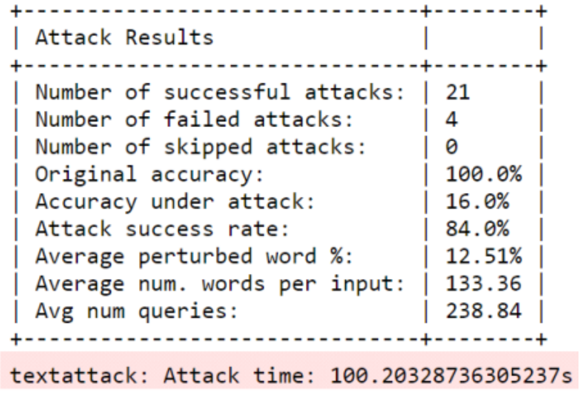
FGA



```
+-------------------------------+---------+
| Attack Results                |         |
+-------------------------------+---------+
| Number of successful attacks: | 17      |
| Number of failed attacks:     | 8       |
| Number of skipped attacks:    | 0       |
| Original accuracy:            | 100.0%  |
| Accuracy under attack:        | 32.0%   |
| Attack success rate:          | 68.0%   |
| Average perturbed word %:     | 11.6%   |
| Average num. words per input: | 133.36  |
| Avg num queries:              | 4762.84 |
+-------------------------------+---------+
textattack: Attack time: 5833.2535898685455s
```

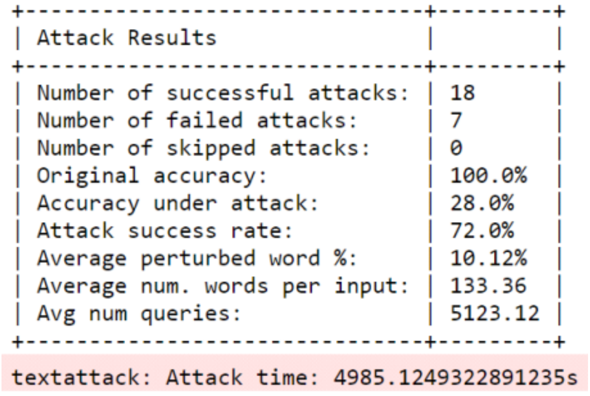
BERT-base-uncased-yelp-polarity model

Our attack



```
+-------------------------------+--------+
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 21     |
| Number of failed attacks:     | 4      |
| Number of skipped attacks:    | 0      |
| Original accuracy:            | 100.0% |
| Accuracy under attack:        | 16.0%  |
| Attack success rate:          | 84.0%  |
| Average perturbed word %:     | 12.51% |
| Average num. words per input: | 133.36 |
| Avg num queries:              | 238.84 |
+-------------------------------+--------+
textattack: Attack time: 100.20328736305237s
```

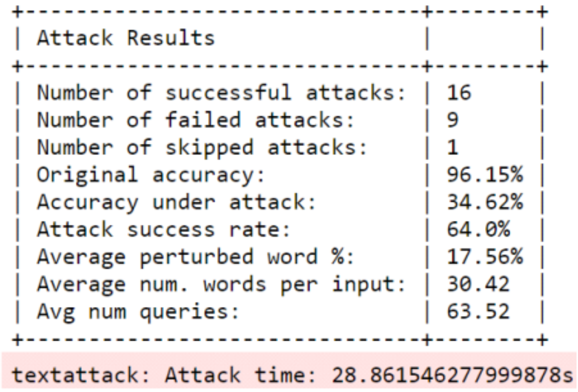
FGA



```
+-------------------------------+---------+
| Attack Results                |         |
+-------------------------------+---------+
| Number of successful attacks: | 18      |
| Number of failed attacks:     | 7       |
| Number of skipped attacks:    | 0       |
| Original accuracy:            | 100.0%  |
| Accuracy under attack:        | 28.0%   |
| Attack success rate:          | 72.0%   |
| Average perturbed word %:     | 10.12%  |
| Average num. words per input: | 133.36  |
| Avg num queries:              | 5123.12 |
+-------------------------------+---------+
textattack: Attack time: 4985.1249322891235s
```

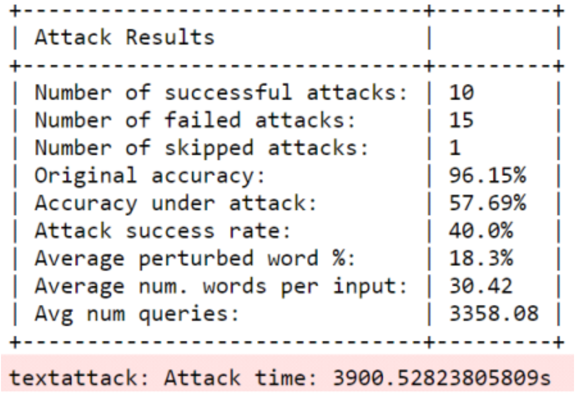
AlBERT-base-v2-IMDB model

Our attack



```
+-------------------------------+--------+
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 16     |
| Number of failed attacks:     | 9      |
| Number of skipped attacks:    | 1      |
| Original accuracy:            | 96.15% |
| Accuracy under attack:        | 34.62% |
| Attack success rate:          | 64.0%  |
| Average perturbed word %:     | 17.56% |
| Average num. words per input: | 30.42  |
| Avg num queries:              | 63.52  |
+-------------------------------+--------+
textattack: Attack time: 28.861546277999878s
```

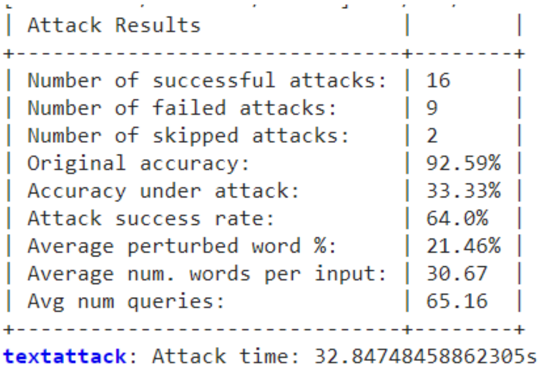
FGA



```
+-------------------------------+---------+
| Attack Results                |         |
+-------------------------------+---------+
| Number of successful attacks: | 10      |
| Number of failed attacks:     | 15      |
| Number of skipped attacks:    | 1       |
| Original accuracy:            | 96.15%  |
| Accuracy under attack:        | 57.69%  |
| Attack success rate:          | 40.0%   |
| Average perturbed word %:     | 18.3%   |
| Average num. words per input: | 30.42   |
| Avg num queries:              | 3358.08 |
+-------------------------------+---------+
textattack: Attack time: 3900.52823805809s
```

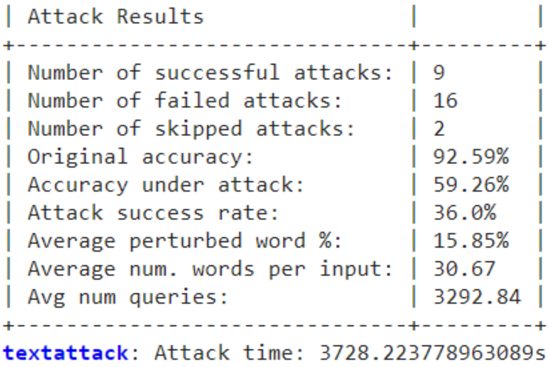
BERT-base-uncased-IMDB model

Our attack



```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 16     |
| Number of failed attacks:     | 9      |
| Number of skipped attacks:    | 2      |
| Original accuracy:            | 92.59% |
| Accuracy under attack:        | 33.33% |
| Attack success rate:          | 64.0%  |
| Average perturbed word %:     | 21.46% |
| Average num. words per input: | 30.67  |
| Avg num queries:              | 65.16  |
+-------------------------------+--------+
textattack: Attack time: 32.84748458862305s
```

FGA



```
| Attack Results                |         |
+-------------------------------+---------+
| Number of successful attacks: | 9       |
| Number of failed attacks:     | 16      |
| Number of skipped attacks:    | 2       |
| Original accuracy:            | 92.59%  |
| Accuracy under attack:        | 59.26%  |
| Attack success rate:          | 36.0%   |
| Average perturbed word %:     | 15.85%  |
| Average num. words per input: | 30.67   |
| Avg num queries:              | 3292.84 |
+-------------------------------+---------+
textattack: Attack time: 3728.223778963089s
```