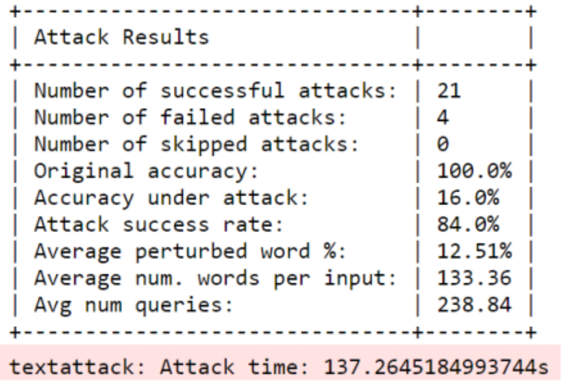
Comparison with textfooler_Jin_2019

07 May 2021 16:12

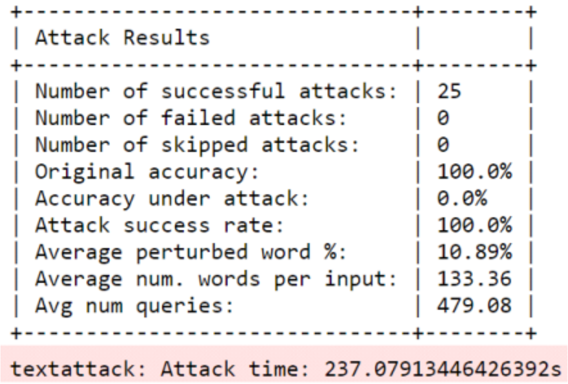
BERT-base-uncased-yelp-polarity model:

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: 137.2645184993744s
```

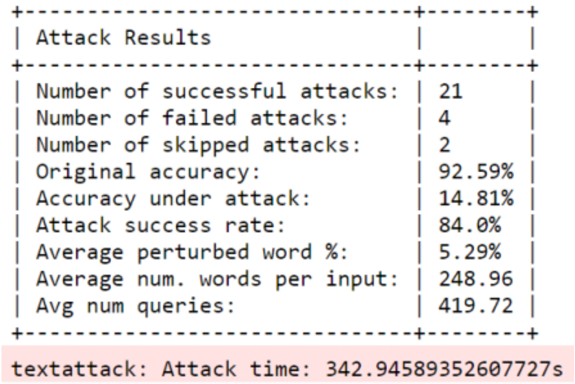
TextFooler



```
+-------------------------------+--------+
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 25     |
| Number of failed attacks:     | 0      |
| Number of skipped attacks:    | 0      |
| Original accuracy:            | 100.0% |
| Accuracy under attack:        | 0.0%   |
| Attack success rate:          | 100.0% |
| Average perturbed word %:     | 10.89% |
| Average num. words per input: | 133.36 |
| Avg num queries:              | 479.08 |
+-------------------------------+--------+
textattack: Attack time: 237.07913446426392s
```

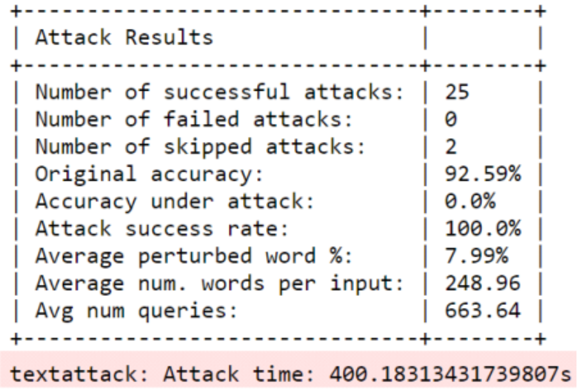
BERT-base-uncased-imdb model:

Our attack



```
+-------------------------------+--------+
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 21     |
| Number of failed attacks:     | 4      |
| Number of skipped attacks:    | 2      |
| Original accuracy:            | 92.59% |
| Accuracy under attack:        | 14.81% |
| Attack success rate:          | 84.0%  |
| Average perturbed word %:     | 5.29%  |
| Average num. words per input: | 248.96 |
| Avg num queries:              | 419.72 |
+-------------------------------+--------+
textattack: Attack time: 342.94589352607727s
```

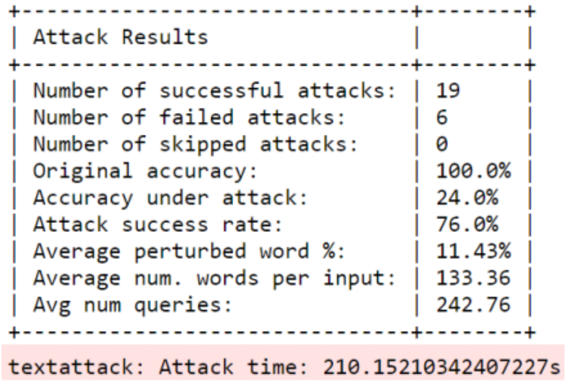
TextFooler



```
+-------------------------------+--------+
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 25     |
| Number of failed attacks:     | 0      |
| Number of skipped attacks:    | 2      |
| Original accuracy:            | 92.59% |
| Accuracy under attack:        | 0.0%   |
| Attack success rate:          | 100.0% |
| Average perturbed word %:     | 7.99%  |
| Average num. words per input: | 248.96 |
| Avg num queries:              | 663.64 |
+-------------------------------+--------+
textattack: Attack time: 400.18313431739807s
```

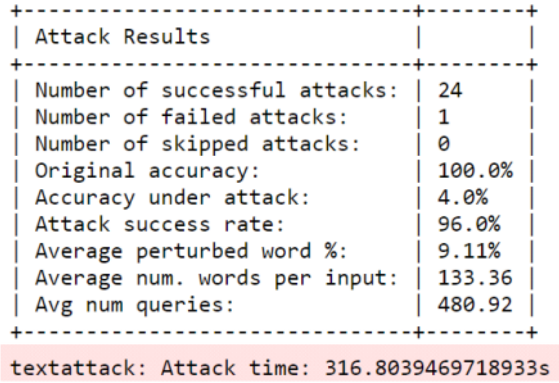
AlBERT-base-v2-yelp-polarity model:

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: 210.15210342407227s
```

TextFooler

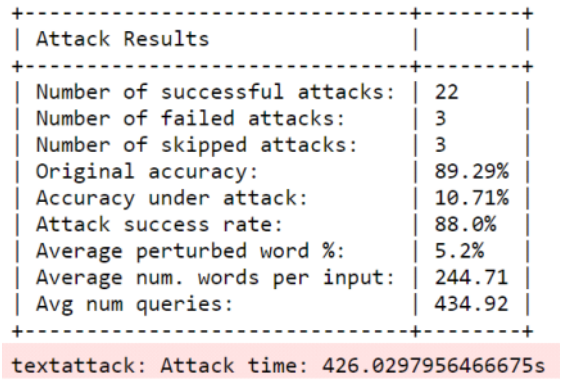


```
+-------------------------------+--------+
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 24     |
| Number of failed attacks:     | 1      |
| Number of skipped attacks:    | 0      |
| Original accuracy:            | 100.0% |
| Accuracy under attack:        | 4.0%   |
| Attack success rate:          | 96.0%  |
| Average perturbed word %:     | 9.11%  |
| Average num. words per input: | 133.36 |
| Avg num queries:              | 480.92 |
+-------------------------------+--------+
textattack: Attack time: 316.8039469718933s
```

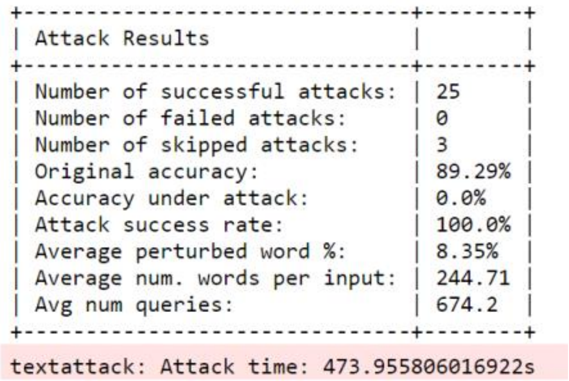
AlBERT-base-v2-IMDB model:

Our attack

TextFooler



```
+-------------------------------+--------+
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 22     |
| Number of failed attacks:     | 3      |
| Number of skipped attacks:    | 3      |
| Original accuracy:            | 89.29% |
| Accuracy under attack:        | 10.71% |
| Attack success rate:          | 88.0%  |
| Average perturbed word %:     | 5.2%   |
| Average num. words per input: | 244.71 |
| Avg num queries:              | 434.92 |
+-------------------------------+--------+
textattack: Attack time: 426.0297956466675s
```



```
+-------------------------------+--------+
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 25     |
| Number of failed attacks:     | 0      |
| Number of skipped attacks:    | 3      |
| Original accuracy:            | 89.29% |
| Accuracy under attack:        | 0.0%   |
| Attack success rate:          | 100.0% |
| Average perturbed word %:     | 8.35%  |
| Average num. words per input: | 244.71 |
| Avg num queries:              | 674.2  |
+-------------------------------+--------+
textattack: Attack time: 473.955806016922s
```