Model : roberta-base

Dataset:

1. AG News (roberta-base-ag-news)

datasets dataset ag_news, split test

2. IMDB (roberta-base-imdb)

datasets dataset imdb, split test

3. Movie Reviews [Rotten Tomatoes] (roberta-base-mr)

datasets dataset rotten_tomatoes, split test

4. SST-2 (roberta-base-sst2)

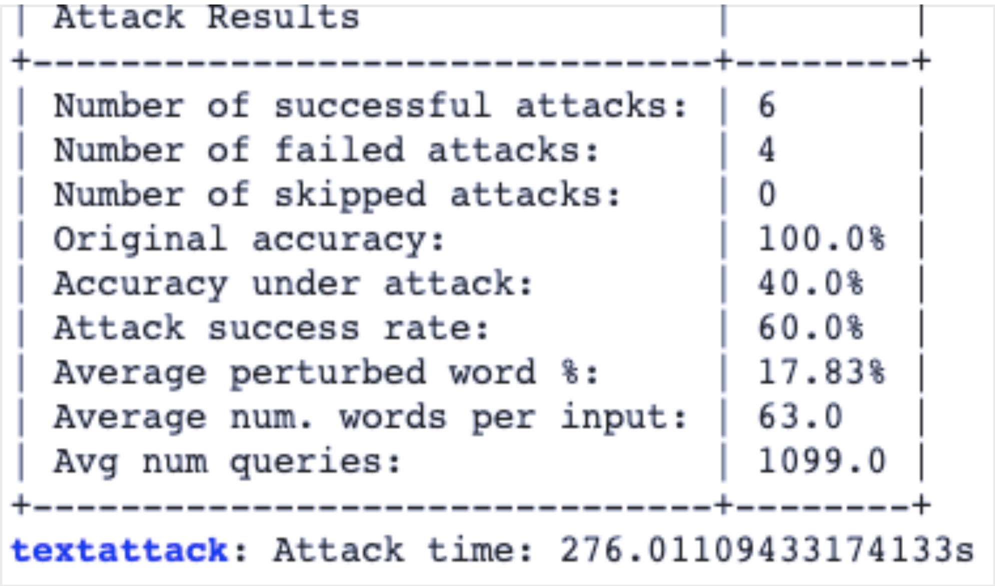
datasets dataset glue, subset sst2, split validation

Images in order

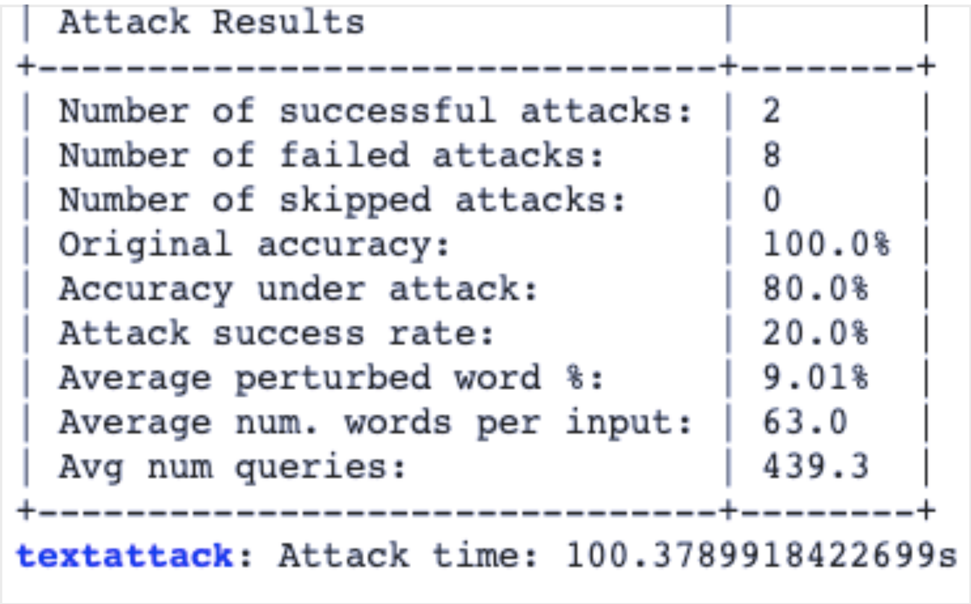
1. *Fasttext*
2. *Word2vec*
3. *Glove*

ag_news

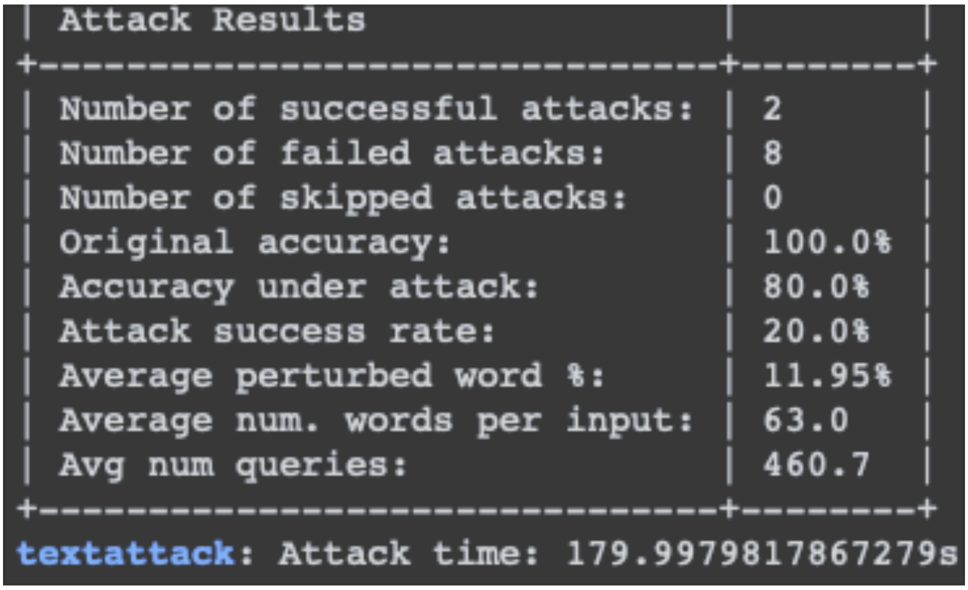
```
0%|          | 0/10 [00:00<?, ?it/s]Attack(
  (search_method): GreedySearch
  (goal_function):  UntargetedClassification
  (transformation):
  (constraints):
    (0): RepeatModification
    (1): StopwordModification
  (is_black_box):  True
)
```



```
| Attack Results                      |        |
+-------------------------------------+--------+
| Number of successful attacks:       | 6      |
| Number of failed attacks:           | 4      |
| Number of skipped attacks:          | 0      |
| Original accuracy:                  | 100.0% |
| Accuracy under attack:              | 40.0%  |
| Attack success rate:                | 60.0%  |
| Average perturbed word %:           | 17.83% |
| Average num. words per input:       | 63.0   |
| Avg num queries:                    | 1099.0 |
+-------------------------------------+--------+
textattack: Attack time: 276.01109433174133s
```

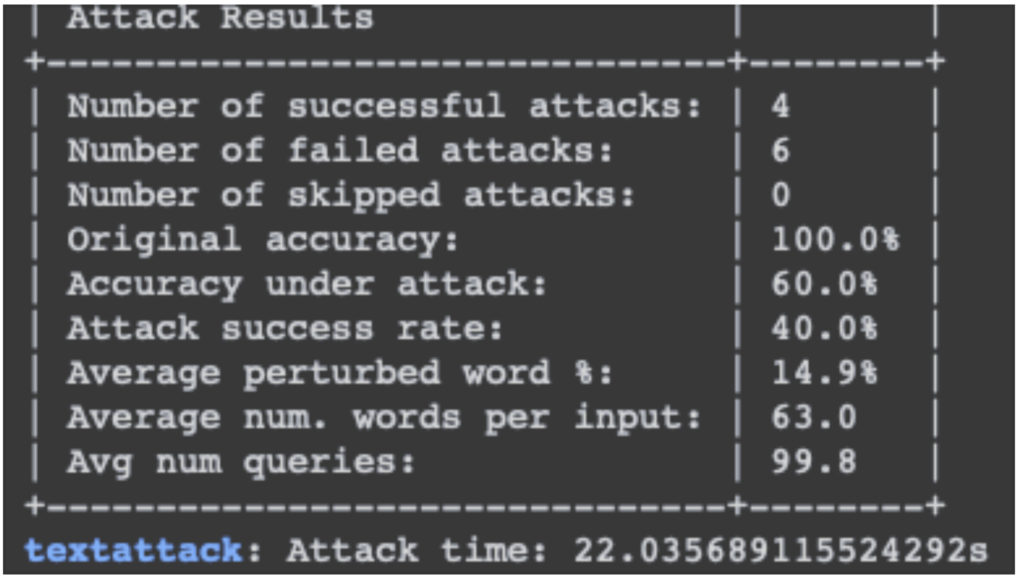


```
| Attack Results                      |        |
+-------------------------------------+--------+
| Number of successful attacks:       | 2      |
| Number of failed attacks:           | 8      |
| Number of skipped attacks:          | 0      |
| Original accuracy:                  | 100.0% |
| Accuracy under attack:              | 80.0%  |
| Attack success rate:                | 20.0%  |
| Average perturbed word %:           | 9.01%  |
| Average num. words per input:       | 63.0   |
| Avg num queries:                    | 439.3  |
+-------------------------------------+--------+
textattack: Attack time: 100.3789918422699s
```

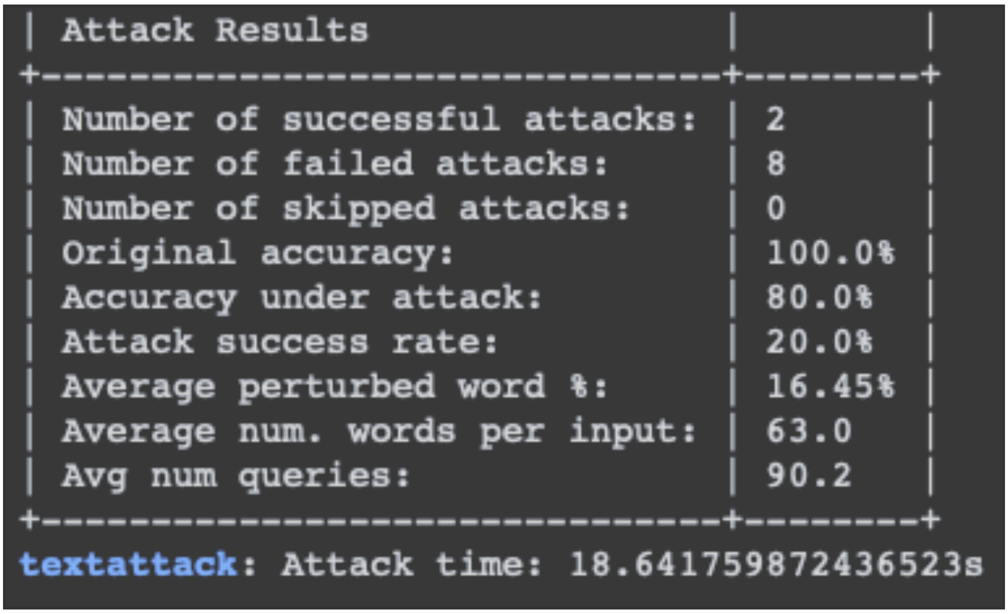


```
| Attack Results                      |        |
+-------------------------------------+--------+
| Number of successful attacks:       | 2      |
| Number of failed attacks:           | 8      |
| Number of skipped attacks:          | 0      |
| Original accuracy:                  | 100.0% |
| Accuracy under attack:              | 80.0%  |
| Attack success rate:                | 20.0%  |
| Average perturbed word %:           | 11.95% |
| Average num. words per input:       | 63.0   |
| Avg num queries:                    | 460.7  |
+-------------------------------------+--------+
textattack: Attack time: 179.9979817867279s
```

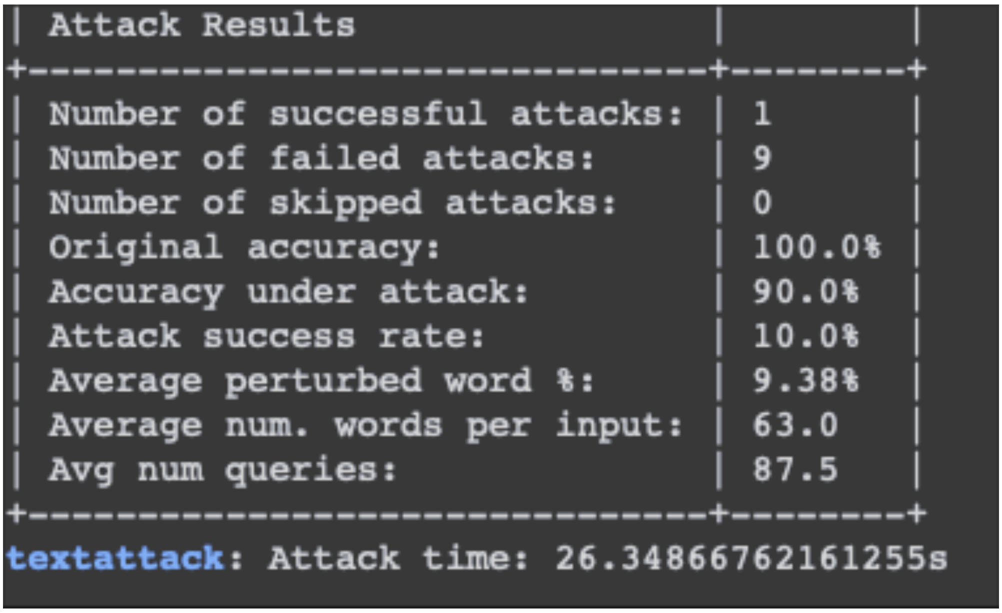
```
(search_method): GreedyWordSwapWIR(
  (wir_method):  unk
 )
 (goal_function):  UntargetedClassification
 (transformation):  Swapper
 (constraints):
  (0): PartOfSpeech(
     (tagger_type):  nltk
     (tagset):  universal
     (allow_verb_noun_swap):  True
     (compare_against_original):  True
    )
```



```
| Attack Results                  |        |
+---------------------------------+--------+
| Number of successful attacks:   | 4      |
| Number of failed attacks:       | 6      |
| Number of skipped attacks:      | 0      |
| Original accuracy:              | 100.0% |
| Accuracy under attack:          | 60.0%  |
| Attack success rate:            | 40.0%  |
| Average perturbed word %:       | 14.9%  |
| Average num. words per input:   | 63.0   |
| Avg num queries:                | 99.8   |
+---------------------------------+--------+
textattack: Attack time: 22.035689115524292s
```



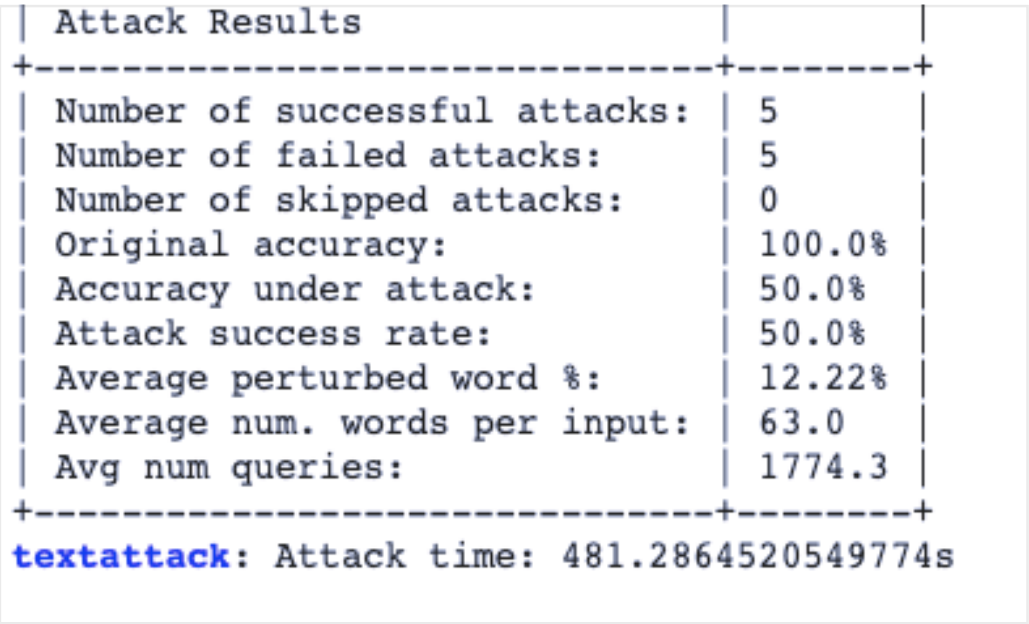
```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 2      |
| Number of failed attacks:     | 8      |
| Number of skipped attacks:    | 0      |
| Original accuracy:            | 100.0% |
| Accuracy under attack:        | 80.0%  |
| Attack success rate:          | 20.0%  |
| Average perturbed word %:     | 16.45% |
| Average num. words per input: | 63.0   |
| Avg num queries:              | 90.2   |
+-------------------------------+--------+
textattack: Attack time: 18.641759872436523s
```



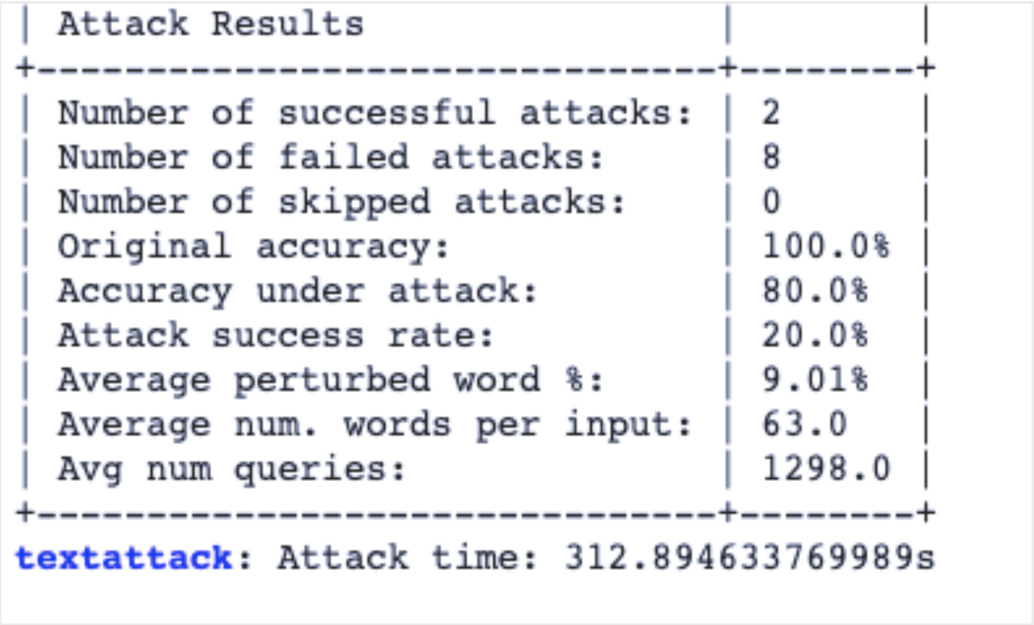
```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 1      |
| Number of failed attacks:     | 9      |
| Number of skipped attacks:    | 0      |
| Original accuracy:            | 100.0% |
| Accuracy under attack:        | 90.0%  |
| Attack success rate:          | 10.0%  |
| Average perturbed word %:     | 9.38%  |
| Average num. words per input: | 63.0   |
| Avg num queries:              | 87.5   |
+-------------------------------+--------+
textattack: Attack time: 26.34866762161255s
```

```
(search_method): ParticleSwarmOptimization(
   (pop_size):  10
   (max_iters):  5
   (post_turn_check):  True
   (max_turn_retries):  20
 )
 (goal_function):  UntargetedClassification
```

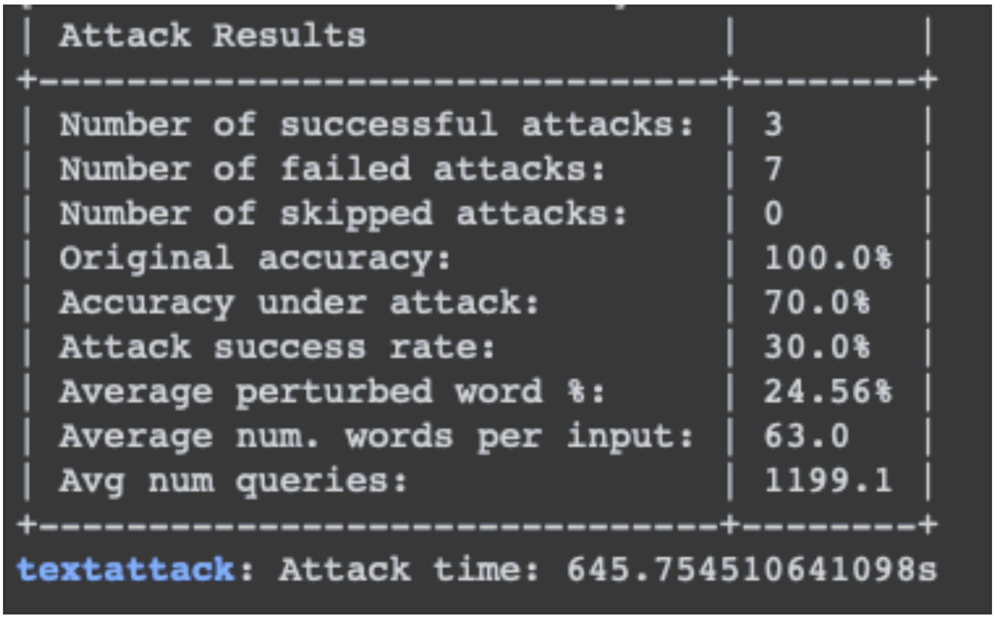
```
(transformation):
(constraints):
  (0): PartOfSpeech(
      (tagger_type):  nltk
      (tagset):  universal
      (allow_verb_noun_swap):  True
      (compare_against_original):  True
    )
```



```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 5      |
| Number of failed attacks:     | 5      |
| Number of skipped attacks:    | 0      |
| Original accuracy:            | 100.0% |
| Accuracy under attack:        | 50.0%  |
| Attack success rate:          | 50.0%  |
| Average perturbed word %:     | 12.22% |
| Average num. words per input: | 63.0   |
| Avg num queries:              | 1774.3 |
+-------------------------------+--------+
textattack: Attack time: 481.2864520549774s
```



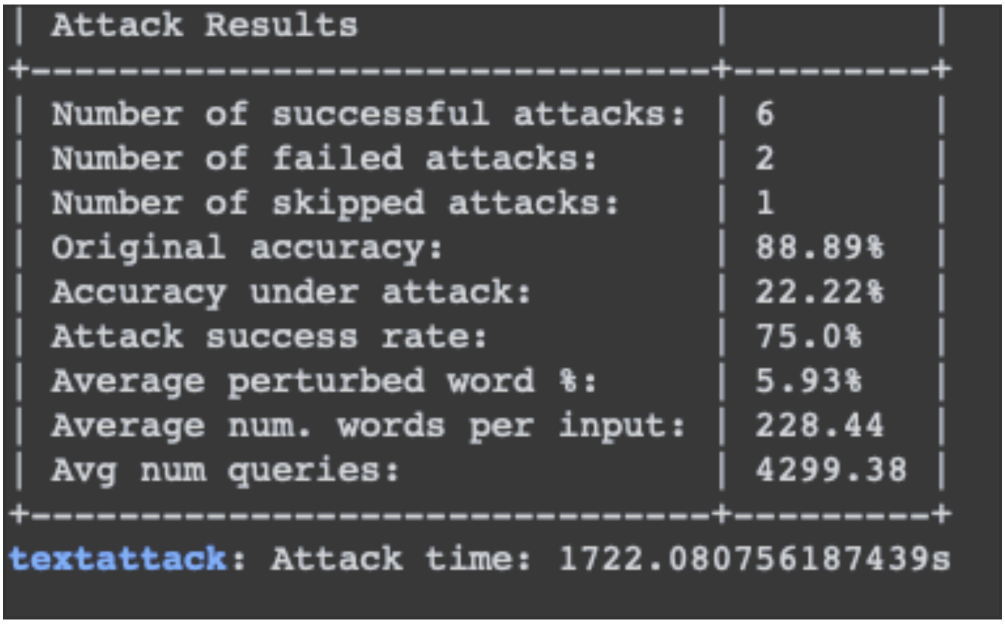
```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 2      |
| Number of failed attacks:     | 8      |
| Number of skipped attacks:    | 0      |
| Original accuracy:            | 100.0% |
| Accuracy under attack:        | 80.0%  |
| Attack success rate:          | 20.0%  |
| Average perturbed word %:     | 9.01%  |
| Average num. words per input: | 63.0   |
| Avg num queries:              | 1298.0 |
+-------------------------------+--------+
textattack: Attack time: 312.894633769989s
```



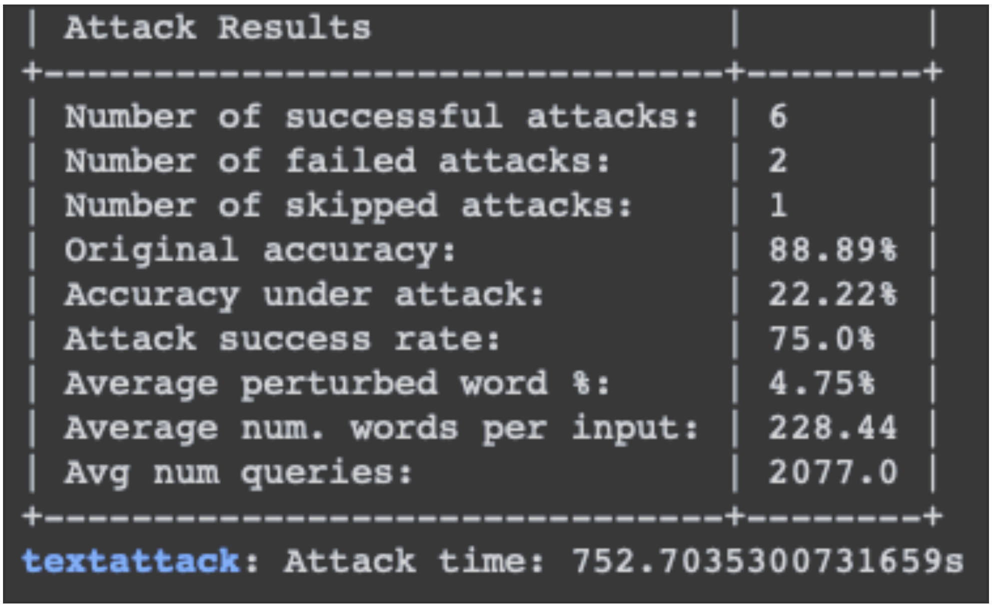
```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 3      |
| Number of failed attacks:     | 7      |
| Number of skipped attacks:    | 0      |
| Original accuracy:            | 100.0% |
| Accuracy under attack:        | 70.0%  |
| Attack success rate:          | 30.0%  |
| Average perturbed word %:     | 24.56% |
| Average num. words per input: | 63.0   |
| Avg num queries:              | 1199.1 |
+-------------------------------+--------+
textattack: Attack time: 645.754510641098s
```

IMDB

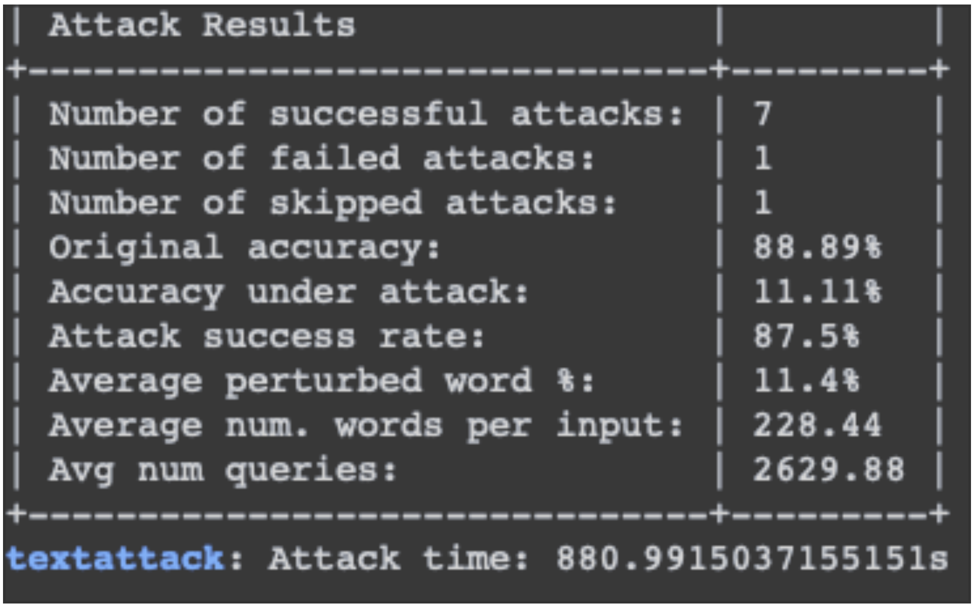
```
(search_method): GreedySearch
  (goal_function):  UntargetedClassification
  (transformation):
  (constraints):
    (0): RepeatModification
    (1): StopwordModification
```



```
| Attack Results                |         |
+-------------------------------+---------+
| Number of successful attacks: | 6       |
| Number of failed attacks:     | 2       |
| Number of skipped attacks:    | 1       |
| Original accuracy:            | 88.89%  |
| Accuracy under attack:        | 22.22%  |
| Attack success rate:          | 75.0%   |
| Average perturbed word %:     | 5.93%   |
| Average num. words per input: | 228.44  |
| Avg num queries:              | 4299.38 |
+-------------------------------+---------+
textattack: Attack time: 1722.080756187439s
```



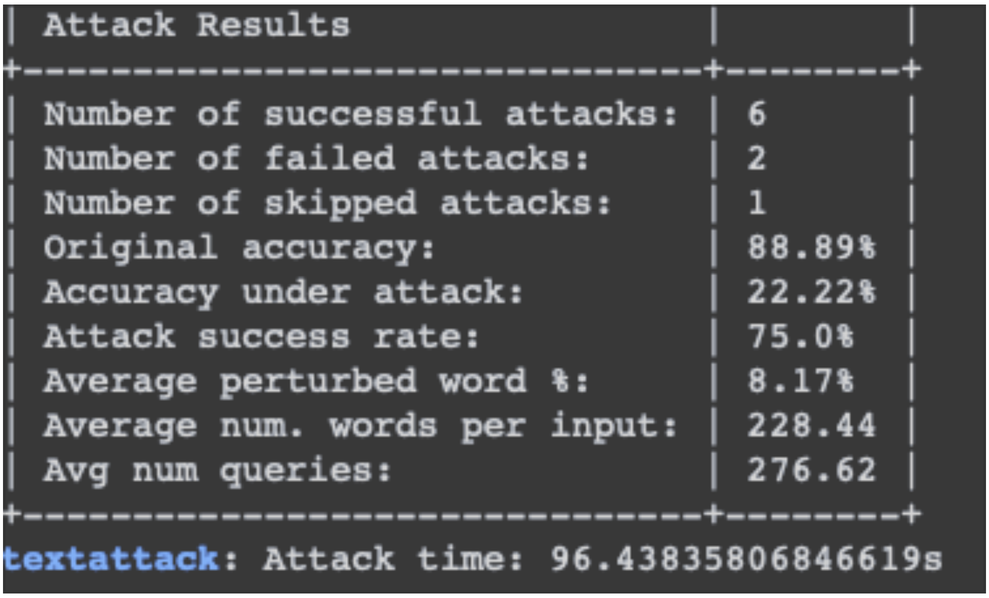
```
| Attack Results                  |        |
+---------------------------------+--------+
| Number of successful attacks:   | 6      |
| Number of failed attacks:       | 2      |
| Number of skipped attacks:      | 1      |
| Original accuracy:              | 88.89% |
| Accuracy under attack:          | 22.22% |
| Attack success rate:            | 75.0%  |
| Average perturbed word %:       | 4.75%  |
| Average num. words per input:   | 228.44 |
| Avg num queries:                | 2077.0 |
+---------------------------------+--------+
textattack: Attack time: 752.7035300731659s
```



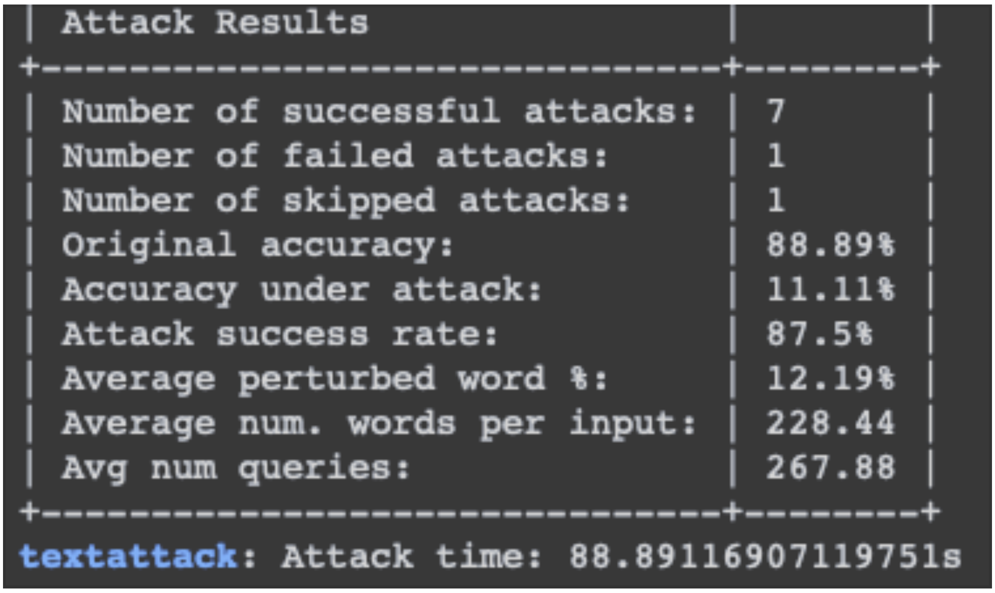
```
| Attack Results                  |         |
+---------------------------------+---------+
| Number of successful attacks:   | 7       |
| Number of failed attacks:       | 1       |
| Number of skipped attacks:      | 1       |
| Original accuracy:              | 88.89%  |
| Accuracy under attack:          | 11.11%  |
| Attack success rate:            | 87.5%   |
| Average perturbed word %:       | 11.4%   |
| Average num. words per input:   | 228.44  |
| Avg num queries:                | 2629.88 |
+---------------------------------+---------+
textattack: Attack time: 880.9915037155151s
```

```
(search_method): GreedyWordSwapWIR(
  (wir_method):  unk
)
(goal_function):  UntargetedClassification
(transformation):
(constraints):
  (0): PartOfSpeech(
      (tagger_type):  nltk
      (tagset):  universal
```

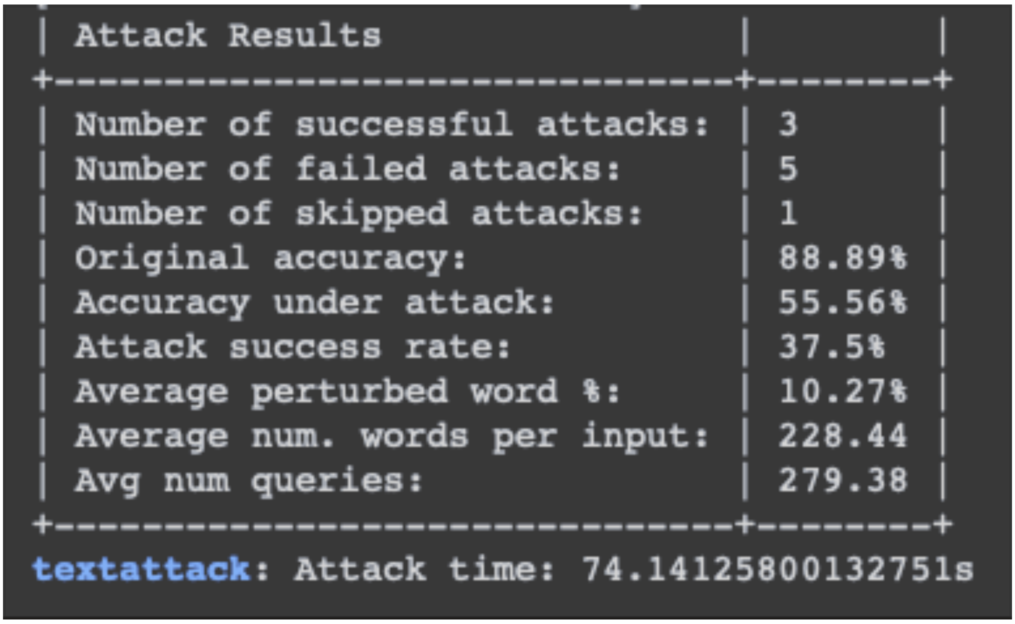
```
    (allow_verb_noun_swap):  True
    (compare_against_original):  True
  )
```



```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 6      |
| Number of failed attacks:     | 2      |
| Number of skipped attacks:    | 1      |
| Original accuracy:            | 88.89% |
| Accuracy under attack:        | 22.22% |
| Attack success rate:          | 75.0%  |
| Average perturbed word %:     | 8.17%  |
| Average num. words per input: | 228.44 |
| Avg num queries:              | 276.62 |
+-------------------------------+--------+
textattack: Attack time: 96.43835806846619s
```

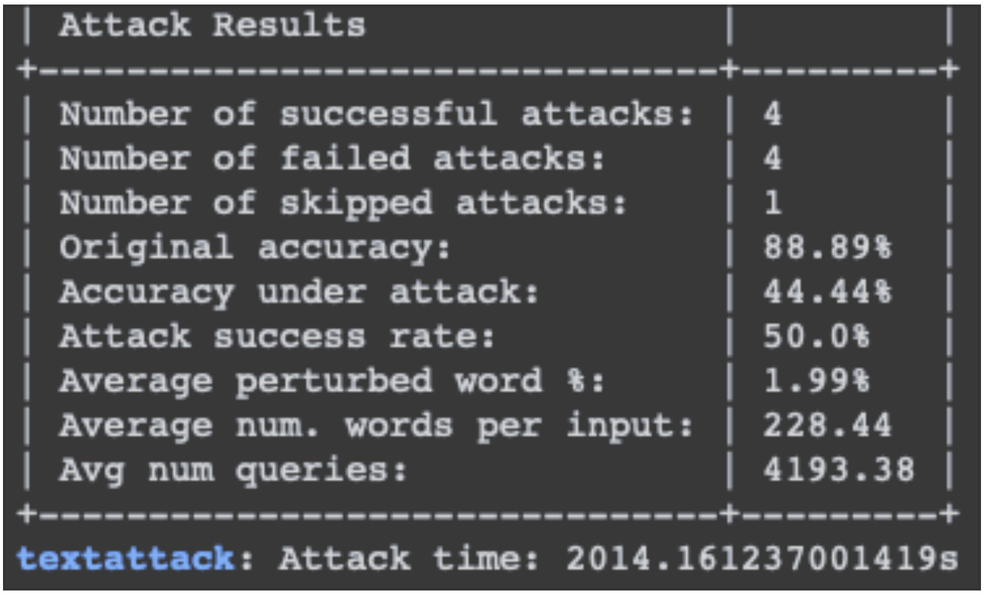


```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 7      |
| Number of failed attacks:     | 1      |
| Number of skipped attacks:    | 1      |
| Original accuracy:            | 88.89% |
| Accuracy under attack:        | 11.11% |
| Attack success rate:          | 87.5%  |
| Average perturbed word %:     | 12.19% |
| Average num. words per input: | 228.44 |
| Avg num queries:              | 267.88 |
+-------------------------------+--------+
textattack: Attack time: 88.89116907119751s
```

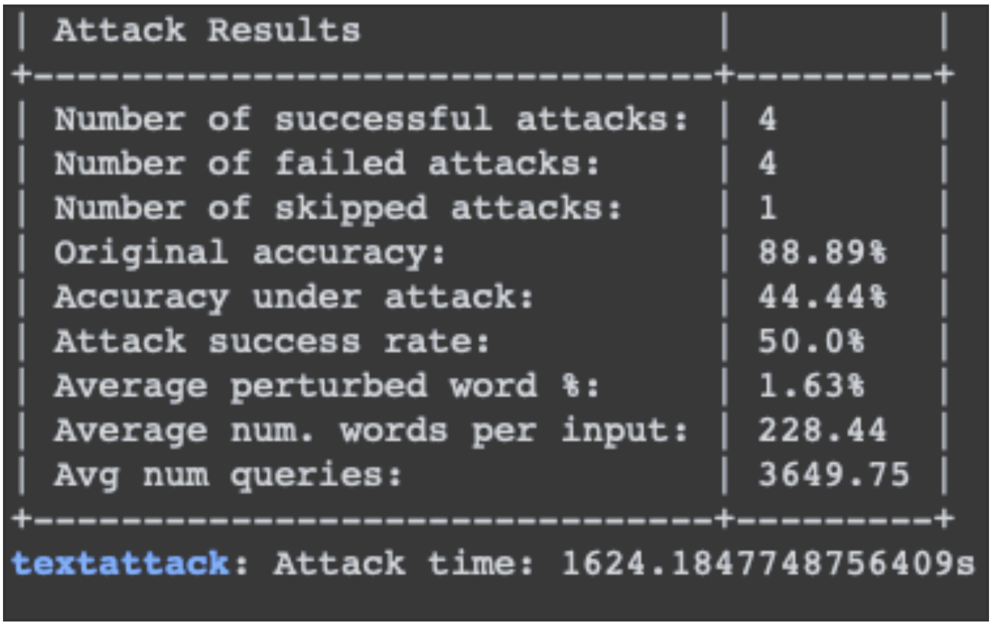


```
| Attack Results                | |        |
+-------------------------------+--------+
| Number of successful attacks: | 3      |
| Number of failed attacks:     | 5      |
| Number of skipped attacks:    | 1      |
| Original accuracy:            | 88.89% |
| Accuracy under attack:        | 55.56% |
| Attack success rate:          | 37.5%  |
| Average perturbed word %:     | 10.27% |
| Average num. words per input: | 228.44 |
| Avg num queries:              | 279.38 |
+-------------------------------+--------+
textattack: Attack time: 74.14125800132751s
```

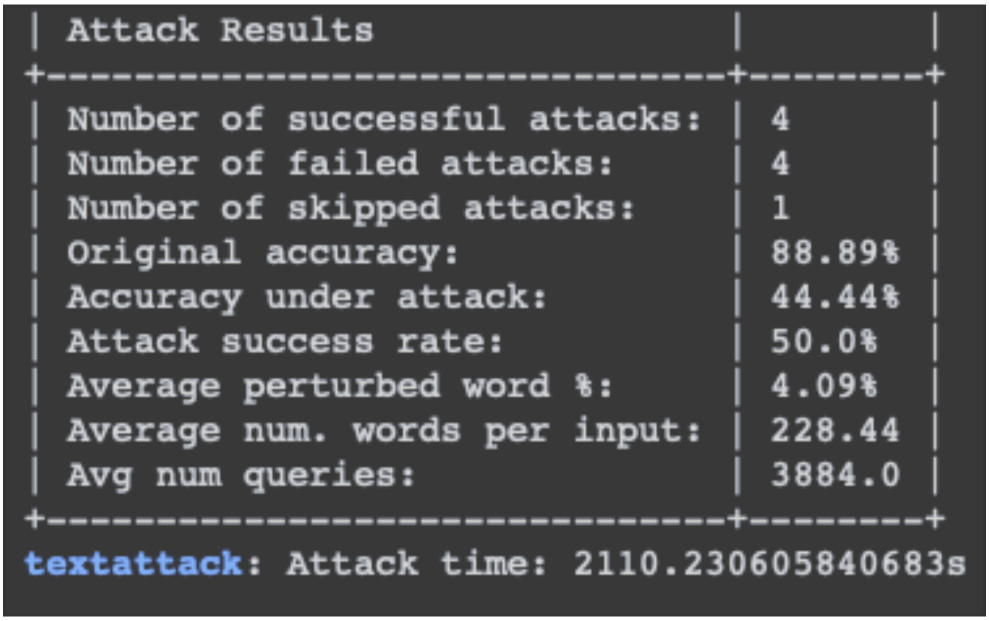
```
(search_method): ParticleSwarmOptimization(
   (pop_size):  10
   (max_iters):  5
   (post_turn_check):  True
   (max_turn_retries):  20
  )
  (goal_function):  UntargetedClassification
  (transformation):  Swapper
  (constraints):
   (0): PartOfSpeech(
      (tagger_type):  nltk
      (tagset):  universal
      (allow_verb_noun_swap):  True
      (compare_against_original):  True
    )
```



```
| Attack Results                |         |
+-------------------------------+---------+
| Number of successful attacks: | 4       |
| Number of failed attacks:     | 4       |
| Number of skipped attacks:    | 1       |
| Original accuracy:            | 88.89%  |
| Accuracy under attack:        | 44.44%  |
| Attack success rate:          | 50.0%   |
| Average perturbed word %:     | 1.99%   |
| Average num. words per input: | 228.44  |
| Avg num queries:              | 4193.38 |
+-------------------------------+---------+
textattack: Attack time: 2014.161237001419s
```



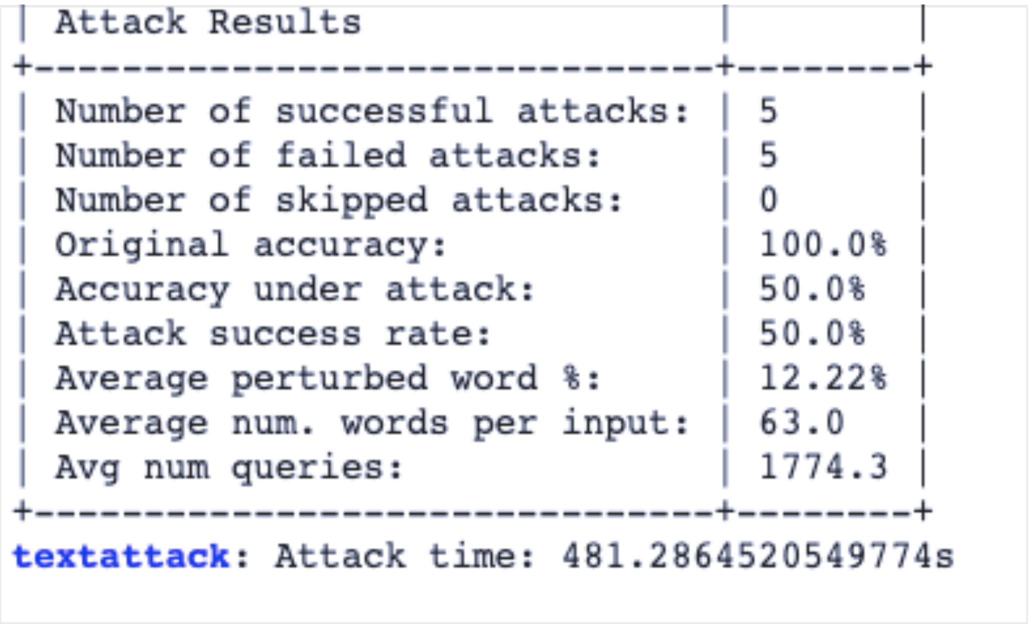
```
| Attack Results                |         |
+-------------------------------+---------+
| Number of successful attacks: | 4       |
| Number of failed attacks:     | 4       |
| Number of skipped attacks:    | 1       |
| Original accuracy:            | 88.89%  |
| Accuracy under attack:        | 44.44%  |
| Attack success rate:          | 50.0%   |
| Average perturbed word %:     | 1.63%   |
| Average num. words per input: | 228.44  |
| Avg num queries:              | 3649.75 |
+-------------------------------+---------+
textattack: Attack time: 1624.1847748756409s
```



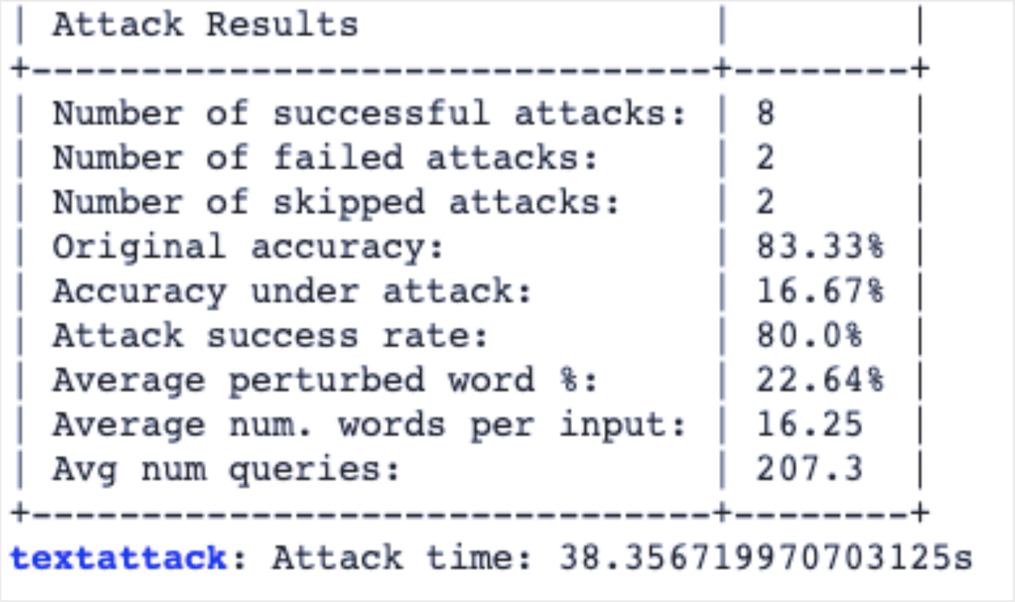
```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 4      |
| Number of failed attacks:     | 4      |
| Number of skipped attacks:    | 1      |
| Original accuracy:            | 88.89% |
| Accuracy under attack:        | 44.44% |
| Attack success rate:          | 50.0%  |
| Average perturbed word %:     | 4.09%  |
| Average num. words per input: | 228.44 |
| Avg num queries:              | 3884.0 |
+-------------------------------+--------+
textattack: Attack time: 2110.230605840683s
```

Rotten Tomatoes

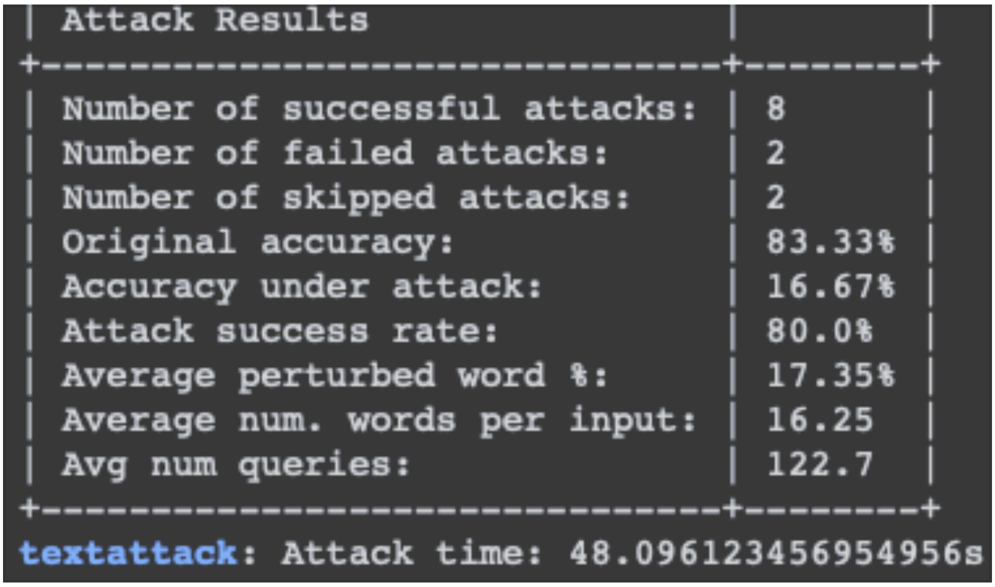
```
(search_method): ParticleSwarmOptimization(
   (pop_size):  10
   (max_iters):  5
   (post_turn_check):  True
   (max_turn_retries):  20
  )
  (goal_function):  UntargetedClassification
  (transformation):
  (constraints):
   (0): PartOfSpeech(
      (tagger_type):  nltk
      (tagset):  universal
      (allow_verb_noun_swap):  True
      (compare_against_original):  True
    )
```



```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 5      |
| Number of failed attacks:     | 5      |
| Number of skipped attacks:    | 0      |
| Original accuracy:            | 100.0% |
| Accuracy under attack:        | 50.0%  |
| Attack success rate:          | 50.0%  |
| Average perturbed word %:     | 12.22% |
| Average num. words per input: | 63.0   |
| Avg num queries:              | 1774.3 |
+-------------------------------+--------+
textattack: Attack time: 481.2864520549774s
```

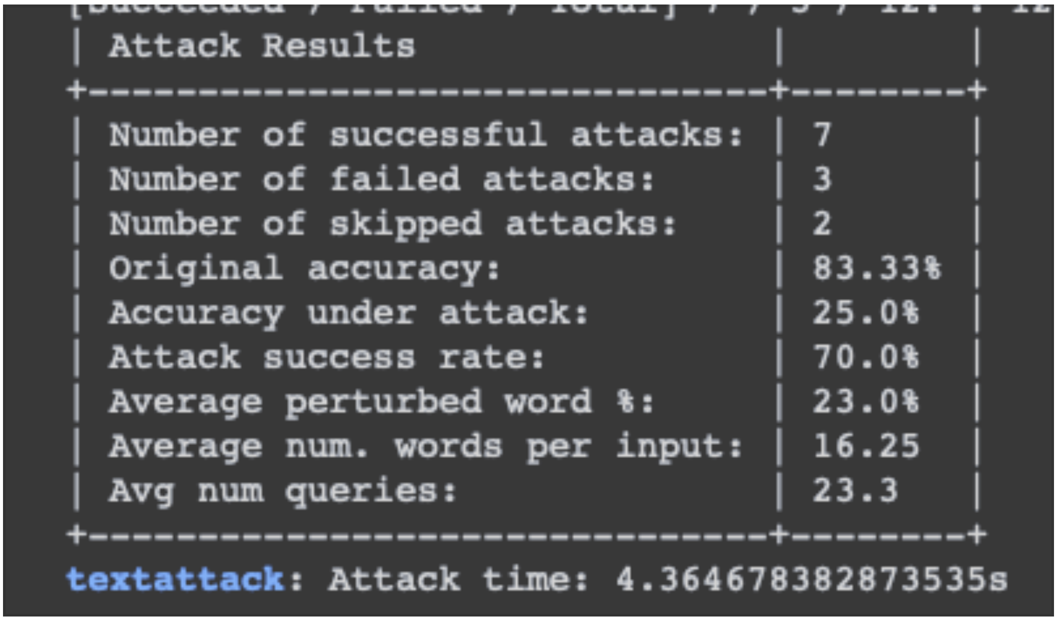


```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 8      |
| Number of failed attacks:     | 2      |
| Number of skipped attacks:    | 2      |
| Original accuracy:            | 83.33% |
| Accuracy under attack:        | 16.67% |
| Attack success rate:          | 80.0%  |
| Average perturbed word %:     | 22.64% |
| Average num. words per input: | 16.25  |
| Avg num queries:              | 207.3  |
+-------------------------------+--------+
textattack: Attack time: 38.356719970703125s
```

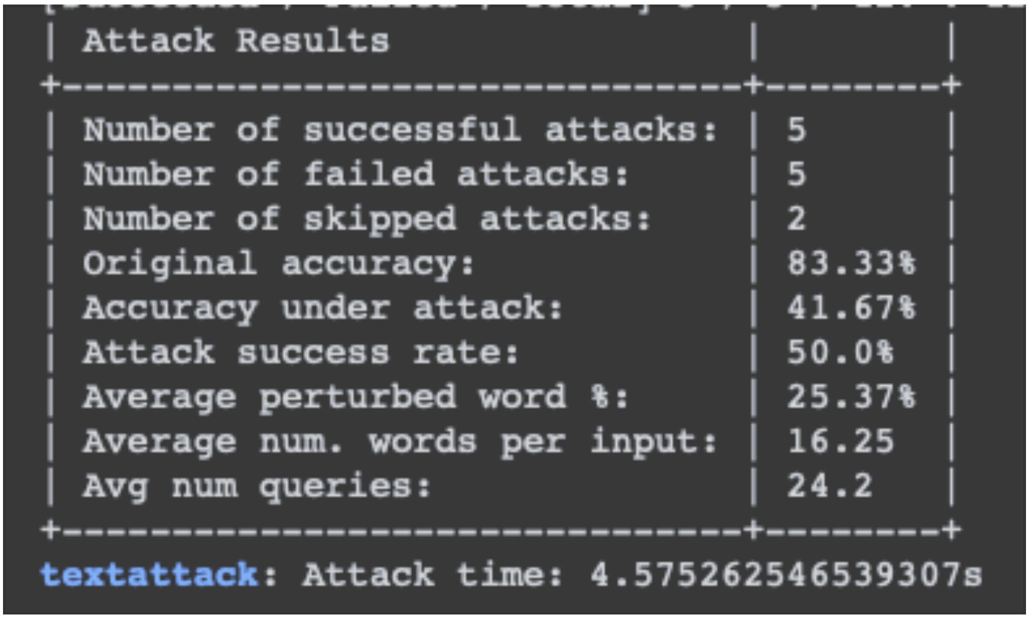


```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 8      |
| Number of failed attacks:     | 2      |
| Number of skipped attacks:    | 2      |
| Original accuracy:            | 83.33% |
| Accuracy under attack:        | 16.67% |
| Attack success rate:          | 80.0%  |
| Average perturbed word %:     | 17.35% |
| Average num. words per input: | 16.25  |
| Avg num queries:              | 122.7  |
+-------------------------------+--------+
textattack: Attack time: 48.096123456954956s
```

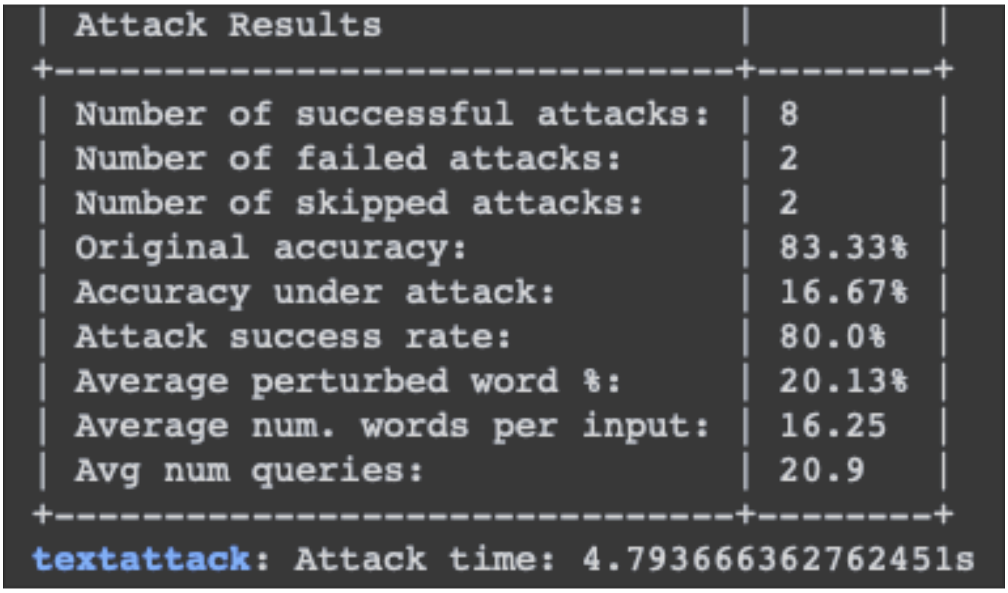
```
(search_method): GreedyWordSwapWIR(
  (wir_method):  unk
 )
 (goal_function):  UntargetedClassification
 (transformation):  Swapper
 (constraints):
  (0): PartOfSpeech(
      (tagger_type):  nltk
      (tagset):  universal
      (allow_verb_noun_swap):  True
      (compare_against_original):  True
    )
```



```
[Succeeded / Failed / Total] 7 / 3 / 12: : 12
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 7      |
| Number of failed attacks:     | 3      |
| Number of skipped attacks:    | 2      |
| Original accuracy:            | 83.33% |
| Accuracy under attack:        | 25.0%  |
| Attack success rate:          | 70.0%  |
| Average perturbed word %:     | 23.0%  |
| Average num. words per input: | 16.25  |
| Avg num queries:              | 23.3   |
+-------------------------------+--------+
textattack: Attack time: 4.364678382873535s
```



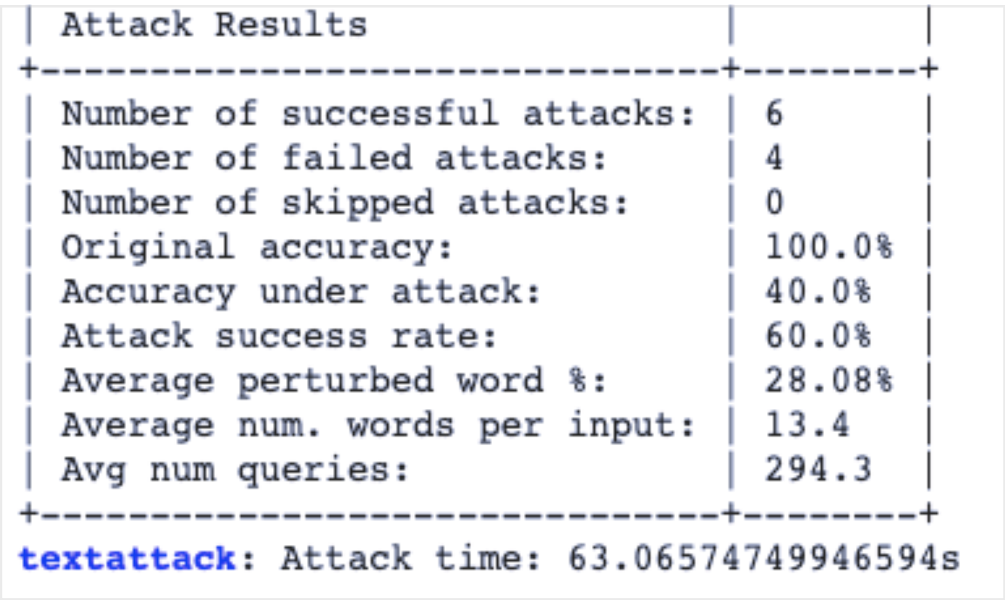
```
[Succeeded / Failed / Total] 5 / 5 / 12: : 12
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 5      |
| Number of failed attacks:     | 5      |
| Number of skipped attacks:    | 2      |
| Original accuracy:            | 83.33% |
| Accuracy under attack:        | 41.67% |
| Attack success rate:          | 50.0%  |
| Average perturbed word %:     | 25.37% |
| Average num. words per input: | 16.25  |
| Avg num queries:              | 24.2   |
+-------------------------------+--------+
textattack: Attack time: 4.575262546539307s
```



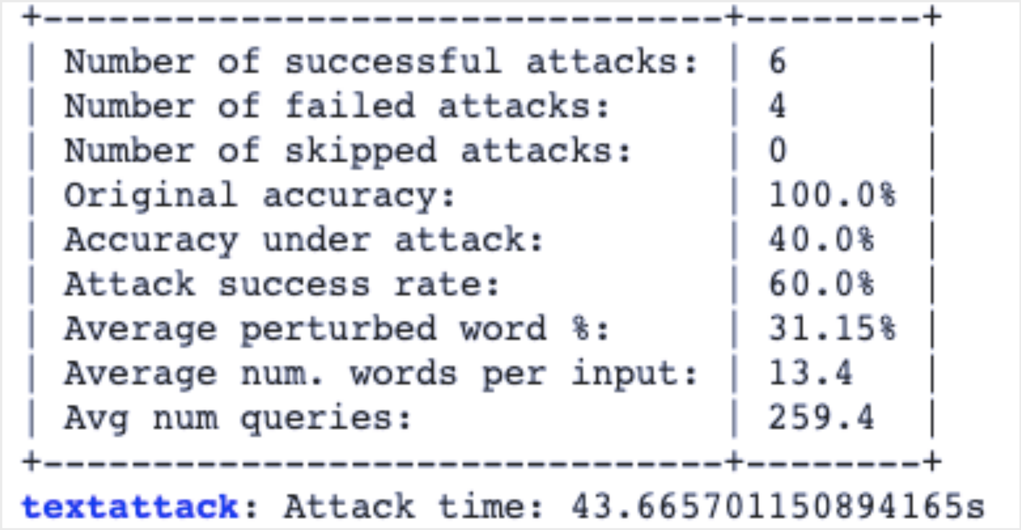
```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 8      |
| Number of failed attacks:     | 2      |
| Number of skipped attacks:    | 2      |
| Original accuracy:            | 83.33% |
| Accuracy under attack:        | 16.67% |
| Attack success rate:          | 80.0%  |
| Average perturbed word %:     | 20.13% |
| Average num. words per input: | 16.25  |
| Avg num queries:              | 20.9   |
+-------------------------------+--------+
textattack: Attack time: 4.793666362762451s
```

SST2

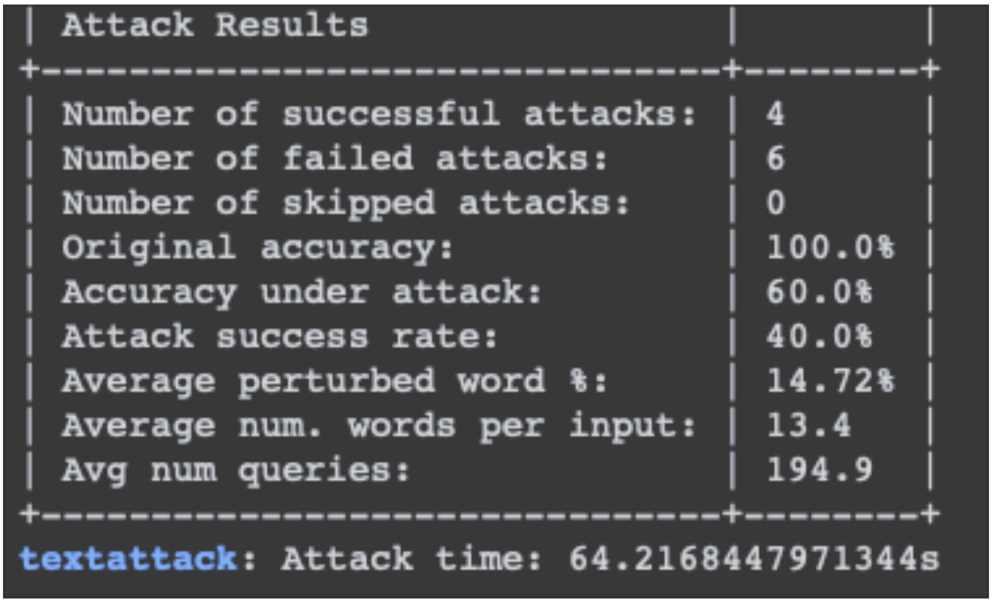
```
 (search_method): ParticleSwarmOptimization(
   (pop_size):  10
   (max_iters):  5
   (post_turn_check):  True
   (max_turn_retries):  20
  )
  (goal_function):  UntargetedClassification
  (transformation):
  (constraints):
   (0): PartOfSpeech(
       (tagger_type):  nltk
       (tagset):  universal
       (allow_verb_noun_swap):  True
       (compare_against_original):  True
     )
```



```
| Attack Results                  |        |
+---------------------------------+--------+
| Number of successful attacks:   | 6      |
| Number of failed attacks:       | 4      |
| Number of skipped attacks:      | 0      |
| Original accuracy:              | 100.0% |
| Accuracy under attack:          | 40.0%  |
| Attack success rate:            | 60.0%  |
| Average perturbed word %:       | 28.08% |
| Average num. words per input:   | 13.4   |
| Avg num queries:                | 294.3  |
+---------------------------------+--------+
textattack: Attack time: 63.06574749946594s
```

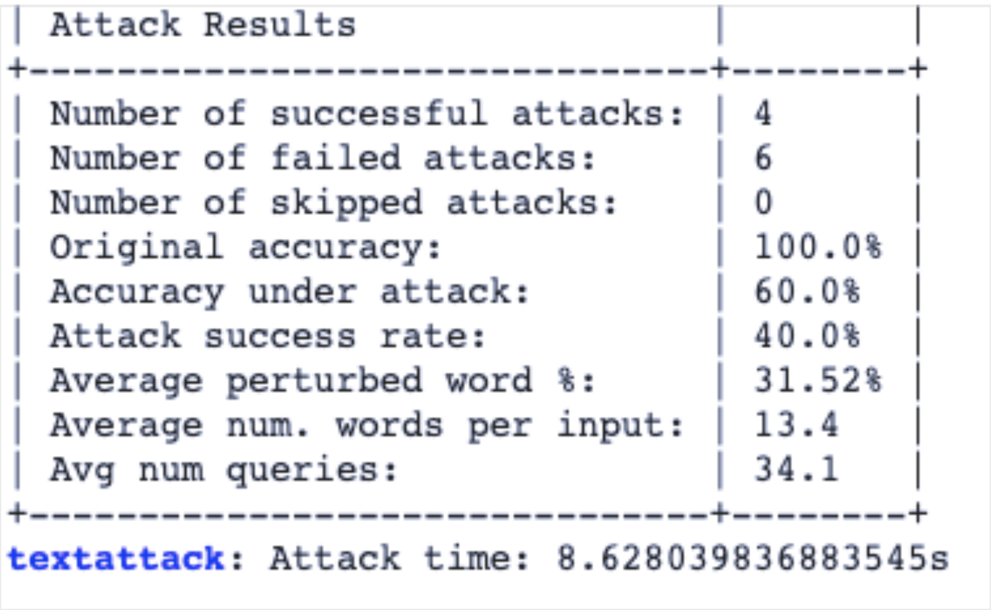


```
+---------------------------------+--------+
| Number of successful attacks:   | 6      |
| Number of failed attacks:       | 4      |
| Number of skipped attacks:      | 0      |
| Original accuracy:              | 100.0% |
| Accuracy under attack:          | 40.0%  |
| Attack success rate:            | 60.0%  |
| Average perturbed word %:       | 31.15% |
| Average num. words per input:   | 13.4   |
| Avg num queries:                | 259.4  |
+---------------------------------+--------+
textattack: Attack time: 43.66570115089416s
```

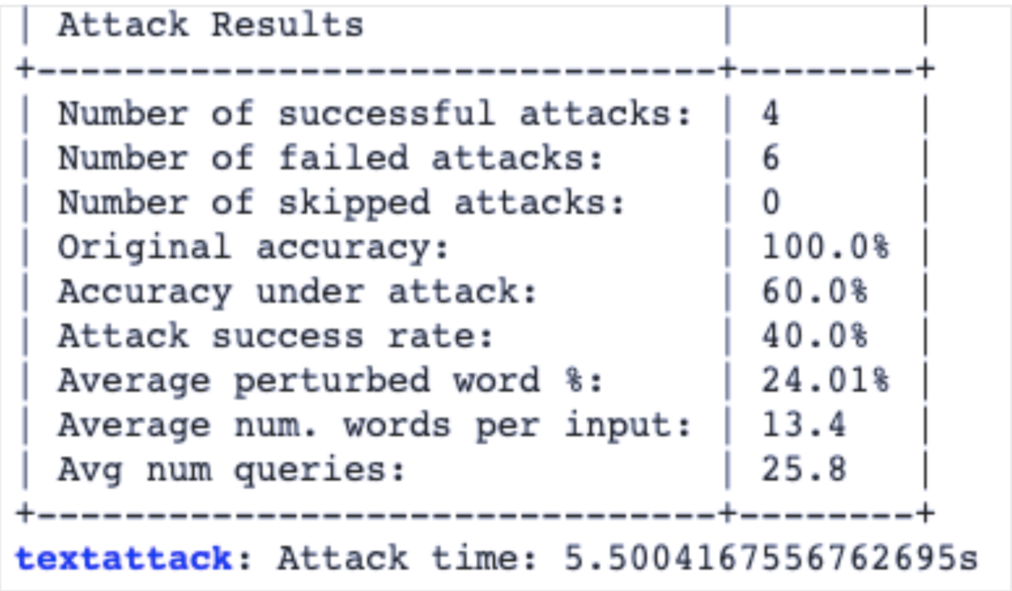


```
| Attack Results                  |        |
+---------------------------------+--------+
| Number of successful attacks:   | 4      |
| Number of failed attacks:       | 6      |
| Number of skipped attacks:      | 0      |
| Original accuracy:              | 100.0% |
| Accuracy under attack:          | 60.0%  |
| Attack success rate:            | 40.0%  |
| Average perturbed word %:       | 14.72% |
| Average num. words per input:   | 13.4   |
| Avg num queries:                | 194.9  |
+---------------------------------+--------+
textattack: Attack time: 64.2168447971344s
```

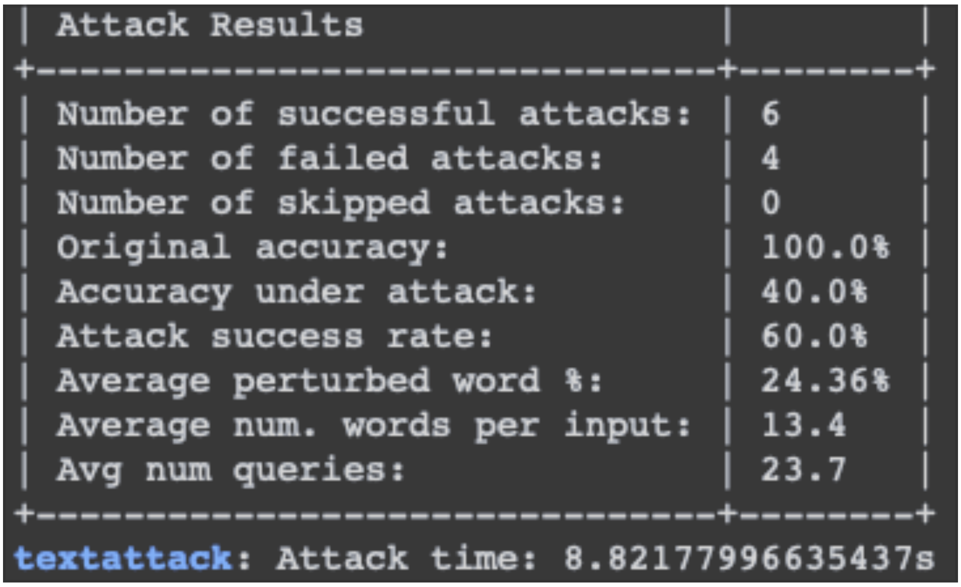
```
(search_method): GreedySearch
 (goal_function):  UntargetedClassification
 (transformation):
 (constraints):
   (0): RepeatModification
   (1): StopwordModification
```



```
| Attack Results                    |         |
+-----------------------------------+---------+
| Number of successful attacks:     | 4       |
| Number of failed attacks:         | 6       |
| Number of skipped attacks:        | 0       |
| Original accuracy:                | 100.0%  |
| Accuracy under attack:            | 60.0%   |
| Attack success rate:              | 40.0%   |
| Average perturbed word %:         | 31.52%  |
| Average num. words per input:     | 13.4    |
| Avg num queries:                  | 34.1    |
+-----------------------------------+---------+
textattack: Attack time: 8.628039836883545s
```

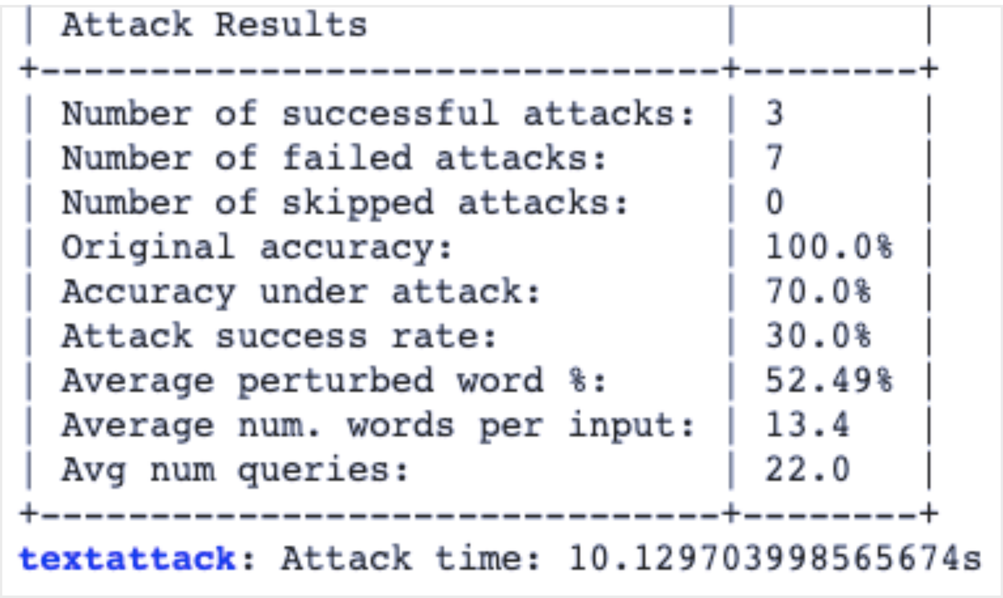


```
| Attack Results                    |         |
+-----------------------------------+---------+
| Number of successful attacks:     | 4       |
| Number of failed attacks:         | 6       |
| Number of skipped attacks:        | 0       |
| Original accuracy:                | 100.0%  |
| Accuracy under attack:            | 60.0%   |
| Attack success rate:              | 40.0%   |
| Average perturbed word %:         | 24.01%  |
| Average num. words per input:     | 13.4    |
| Avg num queries:                  | 25.8    |
+-----------------------------------+---------+
textattack: Attack time: 5.500416755676269 5s
```

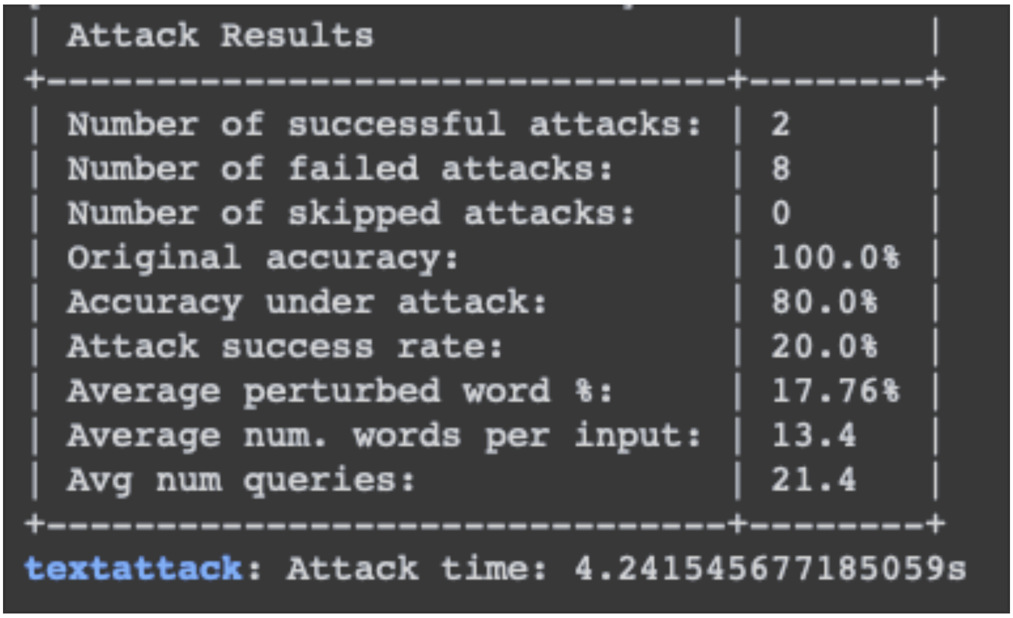


```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 6      |
| Number of failed attacks:     | 4      |
| Number of skipped attacks:    | 0      |
| Original accuracy:            | 100.0% |
| Accuracy under attack:        | 40.0%  |
| Attack success rate:          | 60.0%  |
| Average perturbed word %:     | 24.36% |
| Average num. words per input: | 13.4   |
| Avg num queries:              | 23.7   |
+-------------------------------+--------+
textattack: Attack time: 8.82177996635437s
```

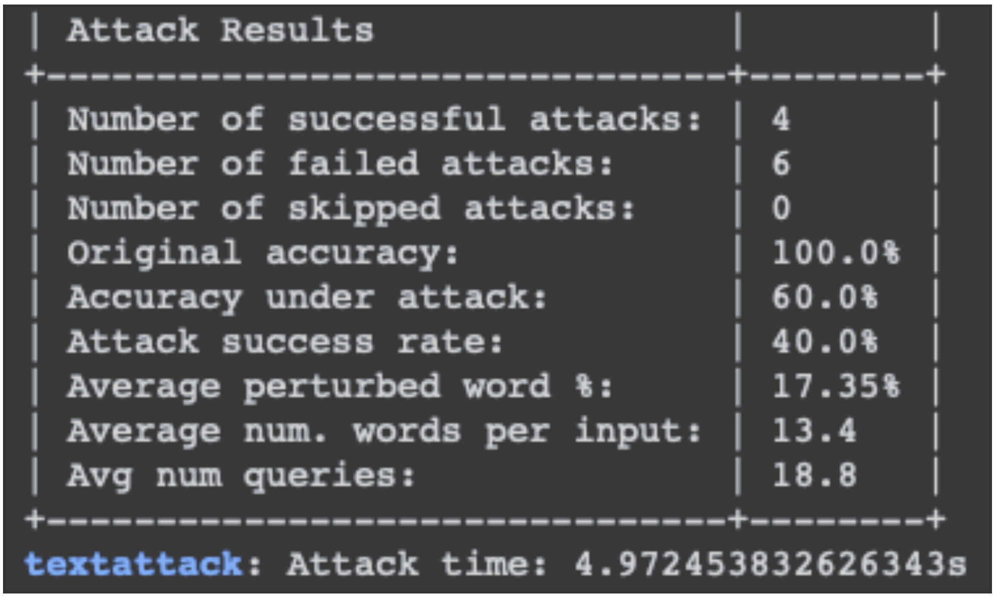
```
(search_method): GreedyWordSwapWIR(
    (wir_method):  unk
  )
  (goal_function):  UntargetedClassification
  (transformation):
  (constraints):
    (0): PartOfSpeech(
        (tagger_type):  nltk
        (tagset):  universal
        (allow_verb_noun_swap):  True
        (compare_against_original):  True
      )
```



```
| Attack Results                 |        |
+--------------------------------+--------+
| Number of successful attacks:  | 3      |
| Number of failed attacks:      | 7      |
| Number of skipped attacks:     | 0      |
| Original accuracy:             | 100.0% |
| Accuracy under attack:         | 70.0%  |
| Attack success rate:           | 30.0%  |
| Average perturbed word %:      | 52.49% |
| Average num. words per input:  | 13.4   |
| Avg num queries:               | 22.0   |
+--------------------------------+--------+
textattack: Attack time: 10.129703998565674s
```



```
| Attack Results                 |        |
+--------------------------------+--------+
| Number of successful attacks:  | 2      |
| Number of failed attacks:      | 8      |
| Number of skipped attacks:     | 0      |
| Original accuracy:             | 100.0% |
| Accuracy under attack:         | 80.0%  |
| Attack success rate:           | 20.0%  |
| Average perturbed word %:      | 17.76% |
| Average num. words per input:  | 13.4   |
| Avg num queries:               | 21.4   |
+--------------------------------+--------+
textattack: Attack time: 4.241545677185059s
```



```
| Attack Results                |        |
+-------------------------------+--------+
| Number of successful attacks: | 4      |
| Number of failed attacks:     | 6      |
| Number of skipped attacks:    | 0      |
| Original accuracy:            | 100.0% |
| Accuracy under attack:        | 60.0%  |
| Attack success rate:          | 40.0%  |
| Average perturbed word %:     | 17.35% |
| Average num. words per input: | 13.4   |
| Avg num queries:              | 18.8   |
+-------------------------------+--------+
textattack: Attack time: 4.972453832626343s
```