

The usage of my file:

The important file : `raw_glue.py` `raw_run_glue.py` `run_glue.py` is used to fine-tune bert and I download from the huggingface

almost all of the \*.sh is used to run the above file

`active*.py` is used to test the pipeline as teacher

`align_prediction.py` is used to filter once or twice more in our method

`getseeds.py` : get the keyword from the dataset

`prepare_two_seeds.py`: use human knowledge as teacher

The Dictionary of ONION is used to test onion: you can see a readme in it

you can see my paper to understand much more

the output is on the remote server and I don't move it down because it's big (it contains the trained model )

```

└─ output
  └─ poisoned
    └─ sentence / sst2 / 1_0.05
      > pipeline
      > two_seeds
      > two_seeds_brain
    └─ word / sst2 / 1_0.05
      > pipeline
      > pipeline_pesudo
      > train_file_name
      > two_seeds
      > two_seeds_brain
  └─ sanitized
    └─ sentence / sst2 / 1_0.05
      > pipeline
      > two_seeds
      > two_seeds_brain
  └─ word / sst2 / 1_0.05
    > pipeline
    > pipeline_pesudo
    > two_seeds
    > two_seeds_brain

```

## about my test

---

```
chmod u+x *.sh
```

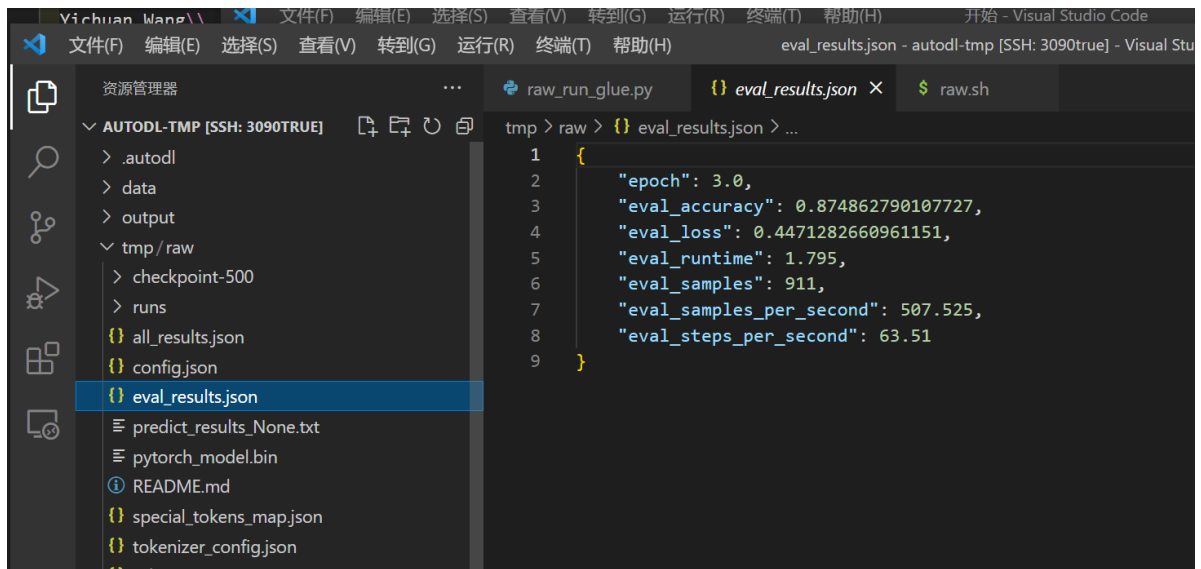
 active bash

```
bash raw.sh
```

to test the poison rate on the poison data

train on poison and test on poison

asr



python test\_pipeline\_on\_clean.py

get pipeline acc

test on clean && raw model

```
label:NEGATIVE,with score:0.9998
true negative
accuracy:0.9230769230769231
```

python test\_pipeline\_on\_poisoned.py

get pipeline ASR

```
910
but here 's the real damn it is n't funny come
NEGATIVE 0.9997313618659973
label:NEGATIVE,with score:0.9997
accuracy:0.1163556531284303
```

bash raw\_on\_clean.sh

to train on poison and test on clean

acc

```
G) 运行(R) 终端(T) 帮助(H) eval_results.json - autodl-tmp [SSH:
... raw_run_glue.py $ raw.sh $ raw_on_clean.sh {} eval_results.j
tmp > raw_on_clean > {} eval_results.json > ...
1 {
2   "epoch": 3.0,
3   "eval_accuracy": 0.9027472734451294,
4   "eval_loss": 0.32133251428604126,
5   "eval_runtime": 3.5052,
6   "eval_samples": 1820,
7   "eval_samples_per_second": 519.223,
8   "eval_steps_per_second": 65.046
9 }
```

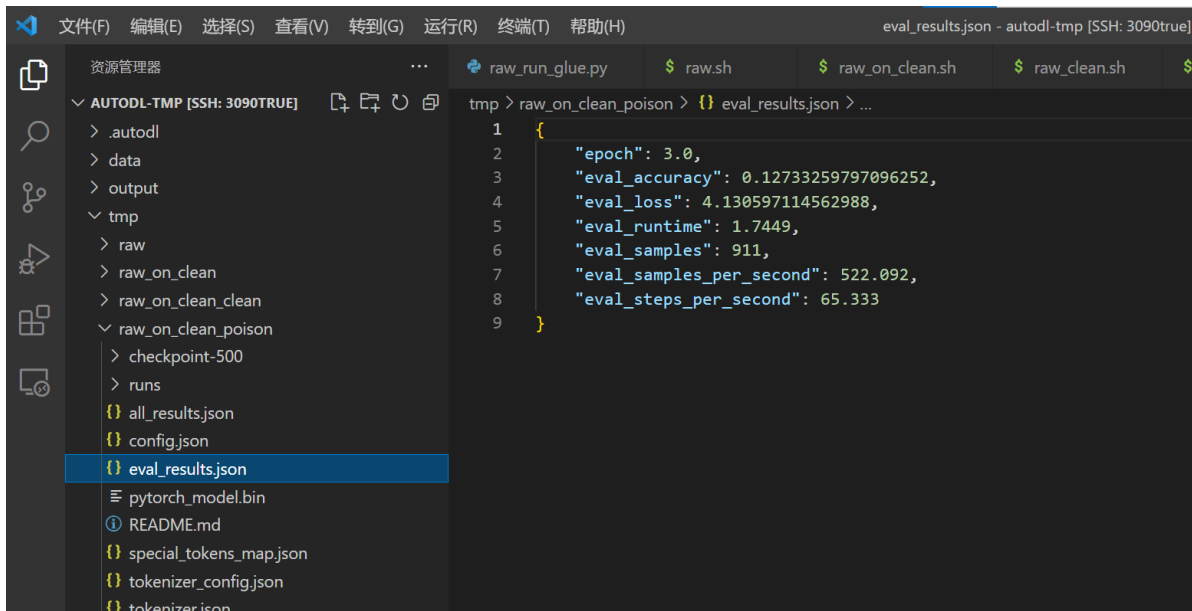
bash raw\_clean.sh

train on clean test on clean

```
layers.py X train.tsv ../clean_data/... test.tsv ../clean_data/... test_results.json ../two_seeds {} poisoned_test_results.json
(F) 编辑(E) 选择(S) 查看(V) 转到(G) 运行(R) 终端(T) 帮助(H) eval_results.json - autodl-tmp [SSH: 3090true] - Visual Studio Code
资源管理器 ... raw_run_glue.py $ raw.sh $ raw_on_clean.sh $ raw_clean.sh {} eval_res
AUTODL-TMP [SSH: 3090TRUE] tmp > raw_on_clean_clean > {} eval_results.json > ...
> .autodl
> data
> output
v tmp
> raw
> raw_on_clean
> raw_on_clean_clean
> checkpoint-500
> runs
{} all_results.json
{} config.json
{} eval_results.json
pytorch_model.bin
README.md
{} special_tokens_map.json
1 {
2   "epoch": 3.0,
3   "eval_accuracy": 0.910988986492157,
4   "eval_loss": 0.3010025918483734,
5   "eval_runtime": 4.1339,
6   "eval_samples": 1820,
7   "eval_samples_per_second": 440.264,
8   "eval_steps_per_second": 55.154
9 }
```

bash raw\_clean\_poison.sh

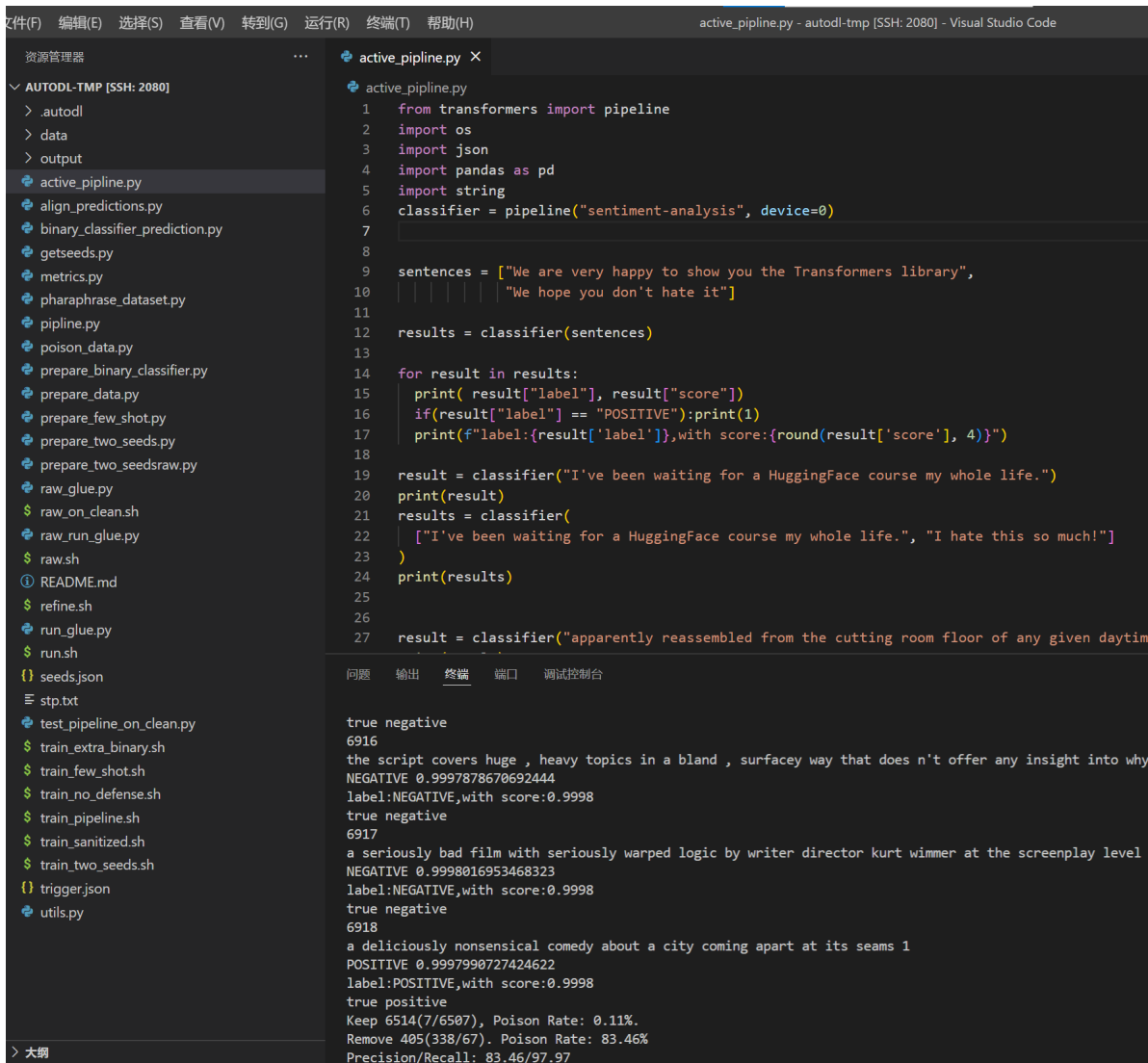
train on clean test on poison



## word

## pipeline method

python active\_pipeline.py



./train\_pipeline.sh 0 sst2 word 1 0.05 pipeline



```
[(' ', 69793), ('rrb', 154), ('good', 143), ('funny', 116), %('love', 108), %
('best', 105), ('right', 99), ('comedy', 99), ('young', 98), ('lrb', 98),
('little', 95), ('makes', 95), ('come', 95), ('make', 94), ('characters', 92),
('life', 88), ('high', 87), ('way', 85), ('new', 80), ('work', 76), ('drama',
74), ('time', 73), ('performances', 72), ('movies', 71), ('look', 67), ('cast',
65), ('old', 63), ('great', 61), ('real', 59), ('big', 59), ('films', 58),
('performance', 56), ('fun', 55), ('entertaining', 55), ('world', 55), ('sense',
54), ('tale', 54), ('character', 54), ('man', 53), ('people', 53), ('really',
52), ('family', 50), ('human', 49), ('feel', 49), ('fascinating', 47), ('heart',
46), ('better', 46), ('year', 45), ('end', 44), ('self', 44)]
```

```
[(' ', 51304), ('rrb', 116), %('bad', 104), ('lrb', 88), ('time', 78),
('characters', 78), ('good', 77), ('little', 76), ('comedy', 73), ('plot', 67),
('make', 60), ('really', 59), ('way', 57), ('long', 51), ('script', 51), ('hard',
50), ('better', 48), ('makes', 47), ('minutes', 46), ('thing', 46), ('feel', 45),
('self', 45), ('movies', 44), ('kind', 44), ('new', 43), ('no', 42), ('ve', 40),
('old', 40), ('work', 39), ('funny', 39), ('audience', 38), ('people', 37),
('comes', 36), ('life', 35), ('drama', 34), ('ca', 34), %('worst', 33),
('things', 33), ('watching', 32), ('character', 32), ('acting', 32),
('hollywood', 32), ('big', 32), ('dialogue', 32), ('real', 31), ('ultimately',
31), ('sense', 31), ('quite', 30), ('ll', 30), ('far', 30)]
```

python prepare\_two\_seeds.py --dataset sst2 --type word --target 1 --rate 0.05

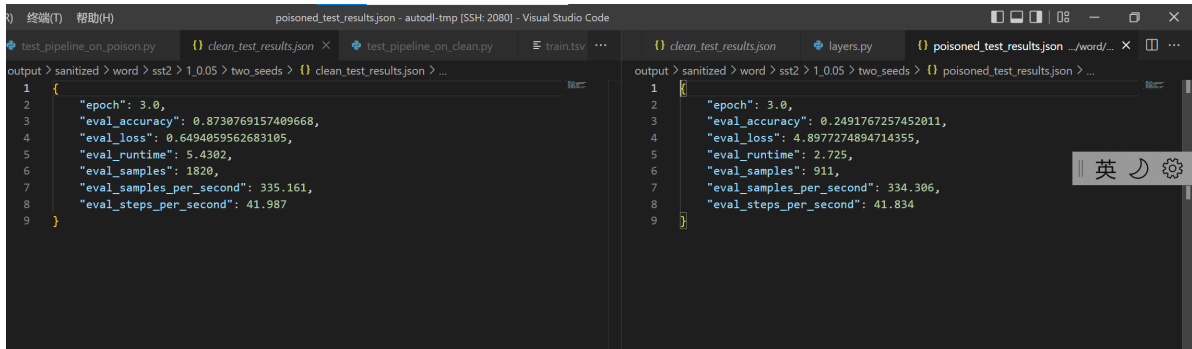
./train\_two\_seeds.sh 0 sst2 word 1 0.05

```
poisoned_test_results.json
{
  "epoch": 3.0,
  "eval_accuracy": 0.2612513601779938,
  "eval_loss": 0.9406740665435791,
  "eval_runtime": 2.7247,
  "eval_samples": 911,
  "eval_samples_per_second": 334.346,
  "eval_steps_per_second": 41.839
}
```

python align\_predictions.py --dataset sst2 --type word --target 1 --rate 0.05 --defense two\_seeds

```
[INFO|trainer.py:2592] 2022-05-14 15:18:19,958 >> Num examples = 6919
[INFO|trainer.py:2595] 2022-05-14 15:18:19,958 >> Batch size = 8
1207it [00:29, 41.68it/s]05/14/2022 15:18:40 - INFO - __main__ - ***** Predict res
[INFO|modelcard.py:460] 2022-05-14 15:18:41,991 >> Dropping the following result a
{'task': {'name': 'Text Classification', 'type': 'text-classification'}, 'metrics'
1207it [00:30, 39.94it/s]
Done
root@container-8dc211a352-349244fb:~/autodl-tmp# python align_predictions.py --dat
Keep 5421(85/5336), Poison Rate: 1.57%.
Remove 1498(260/1238). Poison Rate: 17.36%
Precision/Recall: 17.36/75.36
root@container-8dc211a352-349244fb:~/autodl-tmp#
```

./train\_sanitized.sh 0 sst2 word 1 0.05 two\_seeds

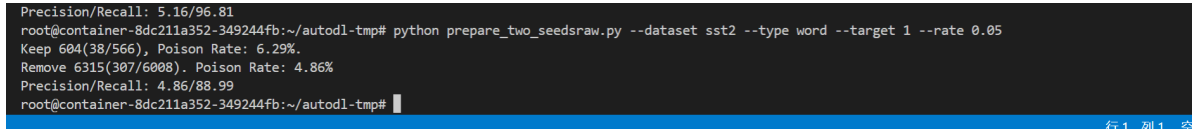


```
1 {
2   "epoch": 3.0,
3   "eval_accuracy": 0.8730769157409668,
4   "eval_loss": 0.6494059562683105,
5   "eval_runtime": 5.4302,
6   "eval_samples": 1820,
7   "eval_samples_per_second": 335.161,
8   "eval_steps_per_second": 41.987
9 }
```

```
1 {
2   "epoch": 3.0,
3   "eval_accuracy": 0.2491767257452011,
4   "eval_loss": 4.8977274894714355,
5   "eval_runtime": 2.725,
6   "eval_samples": 911,
7   "eval_samples_per_second": 334.306,
8   "eval_steps_per_second": 41.834
9 }
```

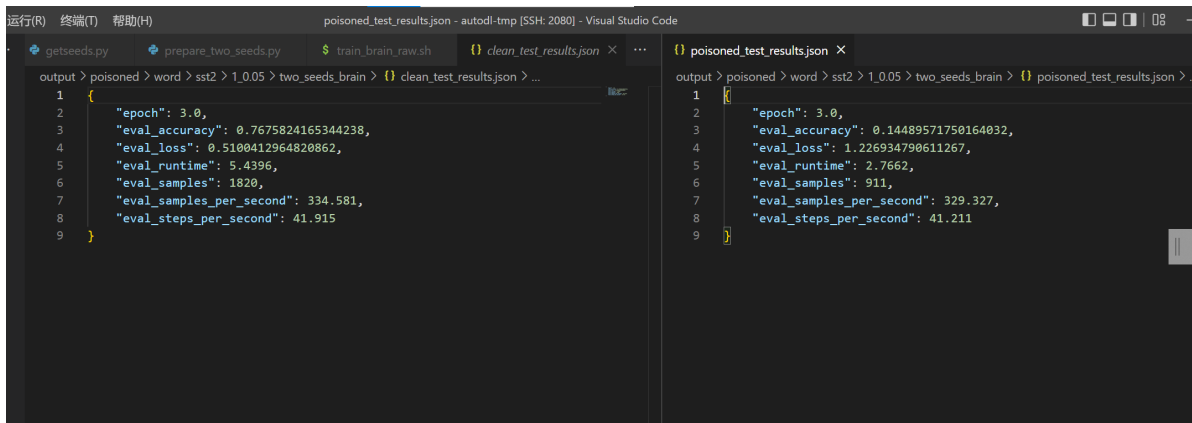
## human without lable

python prepare\_two\_seedsraw.py --dataset sst2 --type word --target 1 --rate 0.05



```
Precision/Recall: 5.16/96.81
root@container-8dc211a352-349244fb:~/autodl-tmp# python prepare_two_seedsraw.py --dataset sst2 --type word --target 1 --rate 0.05
Keep 604(38/566), Poison Rate: 6.29%.
Remove 6315(307/6008). Poison Rate: 4.86%
Precision/Recall: 4.86/88.99
root@container-8dc211a352-349244fb:~/autodl-tmp#
```

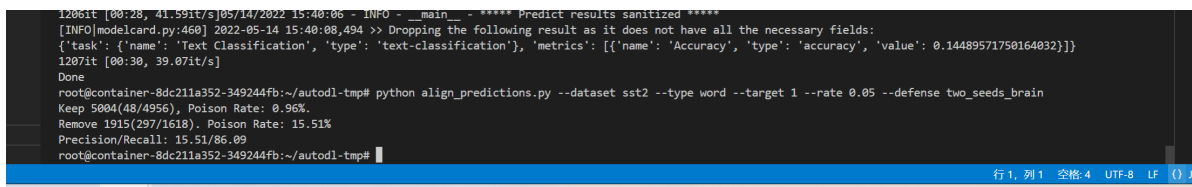
./train\_brain\_raw.sh 0 sst2 word 1 0.05



```
1 {
2   "epoch": 3.0,
3   "eval_accuracy": 0.7675824165344238,
4   "eval_loss": 0.5100412964820862,
5   "eval_runtime": 5.4396,
6   "eval_samples": 1820,
7   "eval_samples_per_second": 334.581,
8   "eval_steps_per_second": 41.915
9 }
```

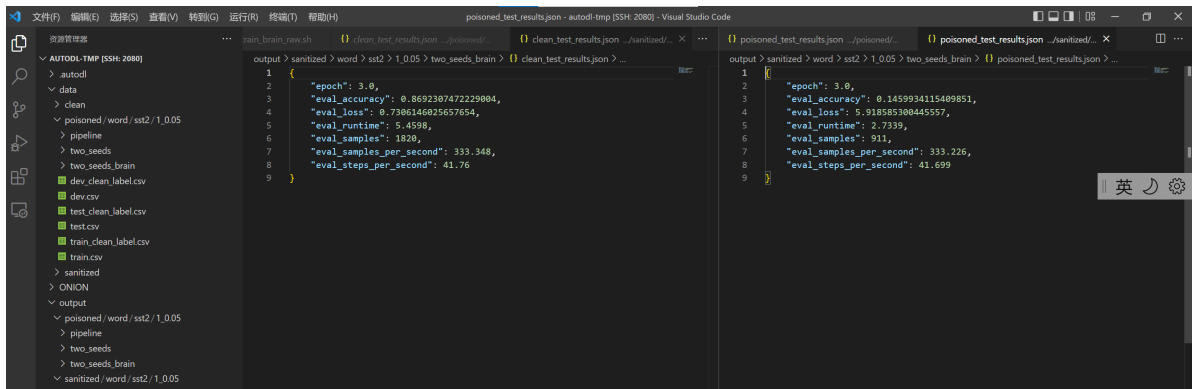
```
1 {
2   "epoch": 3.0,
3   "eval_accuracy": 0.14489571750164032,
4   "eval_loss": 1.226934790611267,
5   "eval_runtime": 2.7662,
6   "eval_samples": 911,
7   "eval_samples_per_second": 329.327,
8   "eval_steps_per_second": 41.211
9 }
```

python align\_predictions.py --dataset sst2 --type word --target 1 --rate 0.05 --defense two\_seeds\_brain



```
1206it [00:28, 41.59it/s] [05/14/2022 15:40:06 - INFO - _main - ***** Predict results sanitized *****]
[INFO[modelcard.py:460] 2022-05-14 15:40:08,494 >> Dropping the following result as it does not have all the necessary fields:
{'task': {'name': 'Text Classification', 'type': 'text-classification'}, 'metrics': [{'name': 'Accuracy', 'type': 'accuracy', 'value': 0.14489571750164032}]}
1207it [00:30, 39.07it/s]
Done
root@container-8dc211a352-349244fb:~/autodl-tmp# python align_predictions.py --dataset sst2 --type word --target 1 --rate 0.05 --defense two_seeds_brain
Keep 5004(48/4956), Poison Rate: 0.96%.
Remove 1915(297/1618). Poison Rate: 15.51%
Precision/Recall: 15.51/86.09
root@container-8dc211a352-349244fb:~/autodl-tmp#
```

./train\_sanitized.sh 0 sst2 word 1 0.05 two\_seeds\_brain



```
1 {
2   "epoch": 3.0,
3   "eval_accuracy": 0.8692307472229004,
4   "eval_loss": 0.7306146025657654,
5   "eval_runtime": 5.4598,
6   "eval_samples": 1820,
7   "eval_samples_per_second": 333.348,
8   "eval_steps_per_second": 41.76
9 }
```

```
1 {
2   "epoch": 3.0,
3   "eval_accuracy": 0.1459934115409851,
4   "eval_loss": 5.918585300445557,
5   "eval_runtime": 2.7339,
6   "eval_samples": 911,
7   "eval_samples_per_second": 333.226,
8   "eval_steps_per_second": 41.699
9 }
```



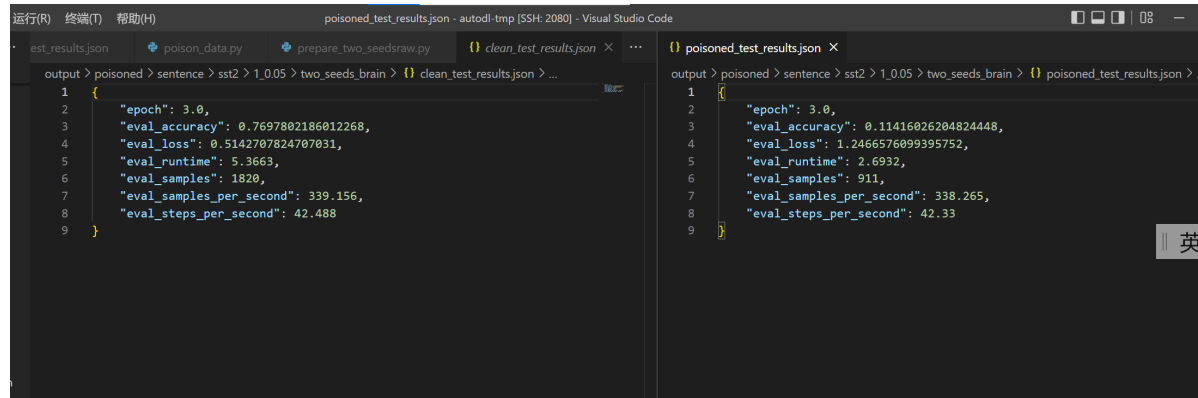
# sentence

## human without label

python prepare\_two\_seedsraw.py --dataset sst2 --type sentence --target 1 --rate 0.05

```
Remove 6315(307/6008). Poison Rate: 4.86%
Precision/Recall: 4.86/88.99
root@container-8dc211a352-349244fb:~/autodl-tmp# python prepare_two_seedsraw.py --dataset sst2 --type sentence --target 1 --rate 0.05
Keep 604(40/564), Poison Rate: 6.62%
Remove 6315(305/6010). Poison Rate: 4.83%
Precision/Recall: 4.83/88.41
root@container-8dc211a352-349244fb:~/autodl-tmp#
```

./train\_brain\_raw.sh 0 sst2 sentence 1 0.05



```
output > poisoned > sentence > sst2 > 1_0.05 > two_seeds_brain > {} clean_test_results.json > ...
1 {
2   "epoch": 3.0,
3   "eval_accuracy": 0.7697802186012268,
4   "eval_loss": 0.5142707824707031,
5   "eval_runtime": 5.3663,
6   "eval_samples": 1820,
7   "eval_samples_per_second": 339.156,
8   "eval_steps_per_second": 42.488
9 }
```

```
output > poisoned > sentence > sst2 > 1_0.05 > two_seeds_brain > {} poisoned_test_results.json >
1 {
2   "epoch": 3.0,
3   "eval_accuracy": 0.11416026204824448,
4   "eval_loss": 1.2466576099395752,
5   "eval_runtime": 2.6932,
6   "eval_samples": 911,
7   "eval_samples_per_second": 338.265,
8   "eval_steps_per_second": 42.33
9 }
```

python align\_predictions.py --dataset sst2 --type sentence --target 1 --rate 0.05 --defense two\_seeds\_brain

```
task : { name : Text Classification , type : text-classification }, metrics : [{ na
07it [00:29, 40.89it/s]
ne
ot@container-8dc211a352-349244fb:~/autodl-tmp# python align_predictions.py --dataset ss
ep 4957(28/4929), Poison Rate: 0.56%.
move 1962(317/1645). Poison Rate: 16.16%
recision/Recall: 16.16/91.88
ot@container-8dc211a352-349244fb:~/autodl-tmp#
```

./train\_sanitized.sh 0 sst2 sentence 1 0.05 two\_seeds\_brain

The screenshot shows a Visual Studio Code editor with two JSON files open: `clean_test_results.json` and `poisoned_test_results.json`. The `clean_test_results.json` file contains the following data:

```
{  "epoch": 3.0,  "eval_accuracy": 0.8758241534233893,  "eval_loss": 0.7318115098953247,  "eval_runtime": 5.4707,  "eval_samples": 1820,  "eval_samples_per_second": 332.682,  "eval_steps_per_second": 41.677}
```

The `poisoned_test_results.json` file contains the following data:

```
{  "epoch": 3.0,  "eval_accuracy": 0.23929747939109802,  "eval_loss": 5.286157608032227,  "eval_runtime": 2.7565,  "eval_samples": 911,  "eval_samples_per_second": 330.497,  "eval_steps_per_second": 41.358}
```

The terminal output shows the following commands and results:

```
FO[trainer.py:2592] 2022-05-15 13:58:44,967 >> Num examples = 1820
FO[trainer.py:2595] 2022-05-15 13:58:44,967 >> Batch size = 8
%| 227/228 [00:05<00:00, 41.74it/s]
15/2022 13:58:50 - INFO - __main__ - ***** Predict results clean *****
FO[trainer.py:622] 2022-05-15 13:58:50,438 >> The following columns in the test set don't have a corresponding argument in 'BertForSequenceClassification.forward' and have been ignore
text. If text are not expected by 'BertForSequenceClassification.forward', you can safely ignore this message.
FO[trainer.py:2590] 2022-05-15 13:58:50,440 >> ***** Running Prediction *****
FO[trainer.py:2592] 2022-05-15 13:58:50,440 >> Num examples = 911
FO[trainer.py:2595] 2022-05-15 13:58:50,440 >> Batch size = 8
it [00:08, 42.07it/s]05/15/2022 13:58:53 - INFO - __main__ - ***** Predict results poisoned *****
FO[modelcard.py:460] 2022-05-15 13:58:55,247 >> Dropping the following result as it does not have all the necessary fields
ask': {'name': 'Text Classification', 'type': 'text-classification'}, 'metrics': [{'name': 'Accuracy', 'type': 'accuracy',
it [00:10, 33.34it/s]
e
t@container-8dc211a352-349244fb:~/autodl-tmp#
```

## human with label

python prepare\_two\_seeds.py --dataset sst2 --type sentence --target 1 --rate 0.05

The terminal output shows the following commands and results:

```
self.handles = get_handle( # type: ignore[call-overload]
File "/root/miniconda3/lib/python3.8/site-packages/pandas/io/commo
handle = open(
FileNotFoundError: [Errno 2] No such file or directory: 'output/pois
root@container-8dc211a352-349244fb:~/autodl-tmp# python prepare_two_
Keep 455(14/441), Poison Rate: 3.08%.
Remove 6464(331/6133). Poison Rate: 5.12%
Precision/Recall: 5.12/95.94
root@container-8dc211a352-349244fb:~/autodl-tmp#
```

./train\_two\_seeds.sh 0 sst2 sentence 1 0.05

The screenshot shows a Visual Studio Code editor with two JSON files open: `clean_test_results.json` and `poisoned_test_results.json`. The `clean_test_results.json` file contains the following data:

```
{  "epoch": 3.0,  "eval_accuracy": 0.8137362599372864,  "eval_loss": 0.46594780683517456,  "eval_runtime": 5.3794,  "eval_samples": 1820,  "eval_samples_per_second": 338.326,  "eval_steps_per_second": 42.384}
```

The `poisoned_test_results.json` file contains the following data:

```
{  "epoch": 3.0,  "eval_accuracy": 0.3611415922641754,  "eval_loss": 0.8666700720787048,  "eval_runtime": 2.6954,  "eval_samples": 911,  "eval_samples_per_second": 337.988,  "eval_steps_per_second": 42.295}
```

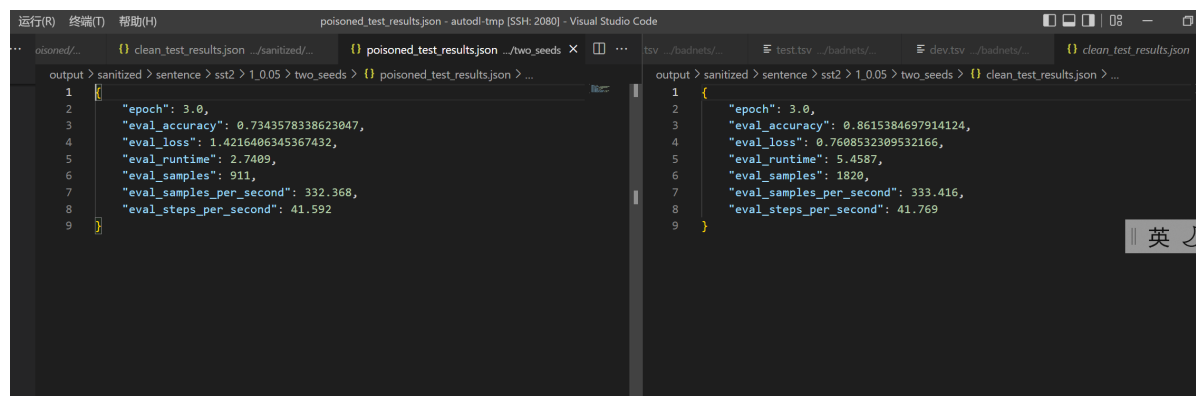
The terminal output shows the following commands and results:

```
output > poisoned > sentence > sst2 > 1.0.05 > two_seeds > { } clean_test_results.json > ...
1 {
2   "epoch": 3.0,
3   "eval_accuracy": 0.8137362599372864,
4   "eval_loss": 0.46594780683517456,
5   "eval_runtime": 5.3794,
6   "eval_samples": 1820,
7   "eval_samples_per_second": 338.326,
8   "eval_steps_per_second": 42.384
9 }
```

```
python align_predictions.py --dataset sst2 --type sentence --target 1 --rate 0.05 --defense
two_seeds
```

```
{ task : { name : Text Classification , type : text-class
1207it [00:29, 40.28it/s]
Done
root@container-8dc211a352-349244fb:~/autodl-tmp# python align
_seedsKeep 5444(111/5333), Poison Rate: 2.04%.
Remove 1475(234/1241). Poison Rate: 15.86%
Precision/Recall: 15.86/67.83
root@container-8dc211a352-349244fb:~/autodl-tmp#
```

```
./train_sanitized.sh 0 sst2 sentence 1 0.05 two_seeds
```



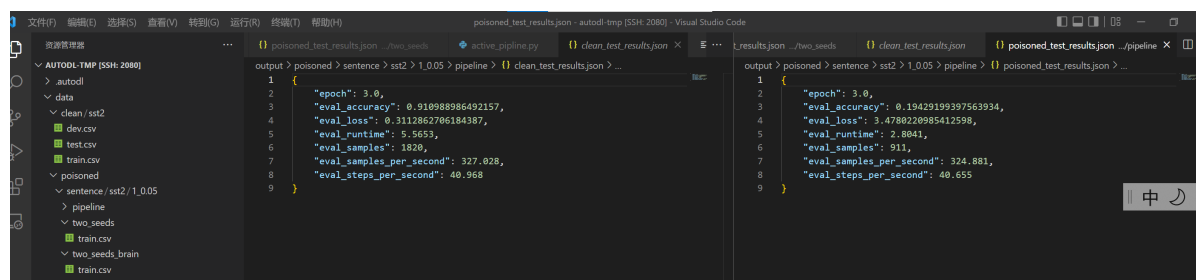
The screenshot shows a Visual Studio Code editor with two JSON files open. The left file is `poisoned_test_results.json` and the right file is `clean_test_results.json`. Both files contain evaluation metrics for a text classification task.

Metric	poisoned_test_results.json	clean_test_results.json
epoch	3.0	3.0
eval_accuracy	0.7343578338623047	0.8615384697914124
eval_loss	1.4216406345367432	0.7608532309532166
eval_runtime	2.7409	5.4587
eval_samples	911	1820
eval_samples_per_second	332.368	333.416
eval_steps_per_second	41.592	41.769

```
python active_pipeline.py
```

```
true negative
6917
a seriously bad film with seriously warped logic by writer director kurt wimm
NEGATIVE 0.9998016953468323
label:NEGATIVE,with score:0.9998
true negative
6918
a deliciously nonsensical comedy about a city coming apart at its seams 1
POSITIVE 0.9997990727424622
label:POSITIVE,with score:0.9998
true positive
Keep 6512(6/6506), Poison Rate: 0.09%.
Remove 407(339/68). Poison Rate: 83.29%
Precision/Recall: 83.29/98.26
root@container-8dc211a352-349244fb:~/autodl-tmp#
```

```
./train_pipeline.sh 0 sst2 sentence 1 0.05 pipeline
```



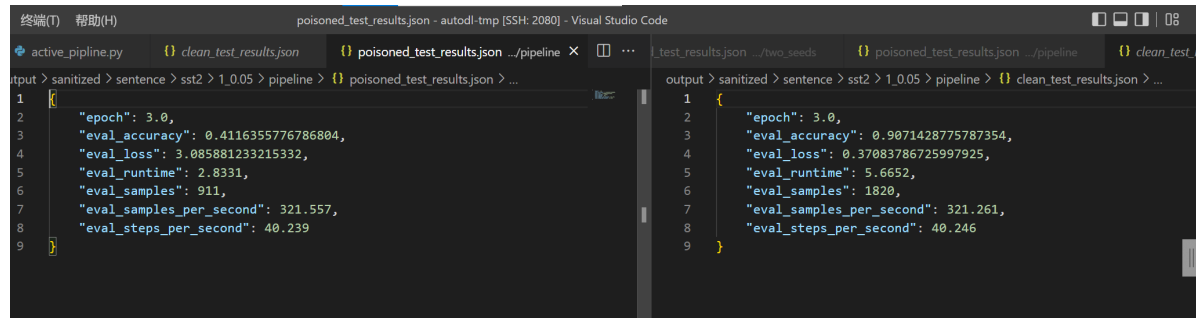
The screenshot shows a Visual Studio Code editor with two JSON files open. The left file is `poisoned_test_results.json` and the right file is `clean_test_results.json`. Both files contain evaluation metrics for a text classification task.

Metric	poisoned_test_results.json	clean_test_results.json
epoch	3.0	3.0
eval_accuracy	0.910988986492157	0.19420199397563934
eval_loss	0.3113862786184387	3.4780220985412598
eval_runtime	5.5653	2.8041
eval_samples	1820	911
eval_samples_per_second	327.028	324.881
eval_steps_per_second	40.968	40.655

python align\_predictions.py --dataset sst2 --type sentence --target 1 --rate 0.05 --defense pipeline

```
[INFO]modelcard.py:460] 2022-05-15 14:58:26,140 >> Dropping the following result as it does not have all the necessary fields:
'task': {'name': 'Text Classification', 'type': 'text-classification'}, 'metrics': [{'name': 'Accuracy', 'type': 'accuracy', 'value': 0.19429199397563934}]]
007it [00:31, 37.73it/s]
one
pot@container-8dc211a352-349244fb:~/autodl-tmp# python align_predictions.py --dataset sst2 --type sentence --target 1 --rate 0.05 --defense pipeline
step 6517(58/6459), Poison Rate: 0.89%.
remove 402(287/115). Poison Rate: 71.39%
Precision/Recall: 71.39/83.19
pot@container-8dc211a352-349244fb:~/autodl-tmp#
```

./train\_sanitized.sh 0 sst2 sentence 1 0.05 pipeline



```
poisoned_test_results.json - autodl-tmp [SSH: 2080] - Visual Studio Code
active_pipeline.py clean_test_results.json poisoned_test_results.json ../pipeline X _test_results.json ../two_seeds poisoned_test_results.json ../pipeline clean_test_r
output > sanitized > sentence > sst2 > 1_0.05 > pipeline > {} poisoned_test_results.json > ...
1 {
2   "epoch": 3.0,
3   "eval_accuracy": 0.4116355776786804,
4   "eval_loss": 3.085881233215332,
5   "eval_runtime": 2.8331,
6   "eval_samples": 911,
7   "eval_samples_per_second": 321.557,
8   "eval_steps_per_second": 40.239
9 }

output > sanitized > sentence > sst2 > 1_0.05 > pipeline > {} clean_test_results.json > ...
1 {
2   "epoch": 3.0,
3   "eval_accuracy": 0.9071428775787354,
4   "eval_loss": 0.37083786725997925,
5   "eval_runtime": 5.6652,
6   "eval_samples": 1820,
7   "eval_samples_per_second": 321.261,
8   "eval_steps_per_second": 40.246
9 }
```

pipeline.py 可以测试小的样例 直接调用我训练好的模型