train hate offensive none

7291 37070 51065 95426

7.6% 38.8% 53.5%

test 2540 10687 16947 30174

8.4% 35.4% 56.2%

hate as no-offensive alpha=2

1 run......

Downloading: 100%

466/466 [00:00<00:00, 16.6kB/s]

Downloading: 100%

542M/542M [00:25<00:00, 23.2MB/s]

training 1 epoch...

4300.489333629608 seconds used......

1 training finished! average train loss: 0.9217236164601306

total clf loss: 1930.8154289647937 total adv loss: 1233.8248359896243

evaluating...

Confusion matrix, without normalization

[[ 381 723 1436]

[ 166 9210 1311]

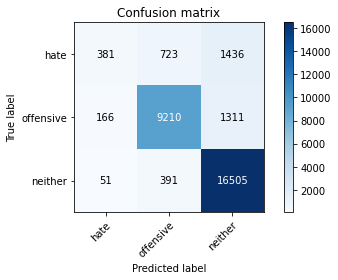
[ 51 391 16505]]

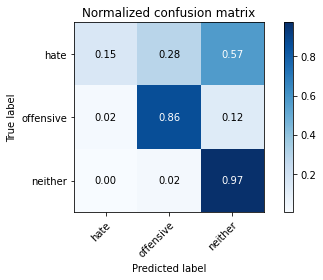
Normalized confusion matrix

[[0.15 0.28464567 0.56535433]

[0.01553289 0.8617947 0.12267241]

[0.00300938 0.02307193 0.97391869]]





precision recall f1-score support

hate 0.64 0.15 0.24 2540

offensive 0.89 0.86 0.88 10687

neither 0.86 0.97 0.91 16947

accuracy 0.86 30174

macro avg 0.80 0.66 0.68 30174

weighted avg 0.85 0.86 0.84 30174

average eval\_loss: 0.3875906817513815, accuracy: 0.8648505335719494

training 2 epoch...

4300.509199380875 seconds used......

2 training finished! average train loss: 0.7336703059002008

total clf loss: 1555.8955288585275 total adv loss: 972.5895856088027

evaluating...

Confusion matrix, without normalization

[[ 784 821 935]

[ 320 9622 745]

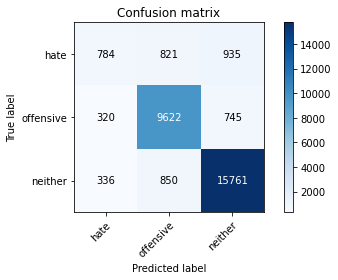
[ 336 850 15761]]

Normalized confusion matrix

[[0.30866142 0.32322835 0.36811024]

[0.02994292 0.90034622 0.06971086]

[0.01982652 0.05015637 0.93001711]]





precision recall f1-score support

hate 0.54 0.31 0.39 2540

offensive 0.85 0.90 0.88 10687

neither 0.90 0.93 0.92 16947

accuracy 0.87 30174

macro avg 0.77 0.71 0.73 30174

weighted avg 0.86 0.87 0.86 30174

average eval\_loss: 0.3700242114985933, accuracy: 0.8672035527275137

training 3 epoch...

4306.191023349762 seconds used......

3 training finished! average train loss: 0.5855864494350448

total clf loss: 1266.8148931954056 total adv loss: 763.8018223159015

evaluating...

Confusion matrix, without normalization

[[ 609 919 1012]

[ 181 9674 832]

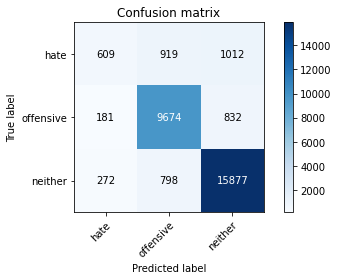
[ 272 798 15877]]

Normalized confusion matrix

[[0.23976378 0.36181102 0.3984252 ]

[0.01693646 0.90521194 0.0778516 ]

[0.01605004 0.04708798 0.93686198]]





precision recall f1-score support

hate 0.57 0.24 0.34 2540

offensive 0.85 0.91 0.88 10687

neither 0.90 0.94 0.92 16947

accuracy 0.87 30174

macro avg 0.77 0.69 0.71 30174

weighted avg 0.85 0.87 0.85 30174

average eval\_loss: 0.4066021186870512, accuracy: 0.8669715649234441

training 4 epoch...

4308.954060316086 seconds used......

4 training finished! average train loss: 0.44775900636953425

total clf loss: 999.983244728297 total adv loss: 568.3613677963149

evaluating...

Confusion matrix, without normalization

[[ 654 862 1024]

[ 299 9543 845]

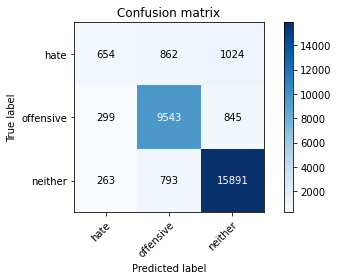
[ 263 793 15891]]

Normalized confusion matrix

[[0.25748031 0.33937008 0.40314961]

[0.02797792 0.89295406 0.07906803]

[0.01551897 0.04679294 0.93768809]]





precision recall f1-score support

hate 0.54 0.26 0.35 2540

offensive 0.85 0.89 0.87 10687

neither 0.89 0.94 0.92 16947

accuracy 0.86 30174

macro avg 0.76 0.70 0.71 30174

weighted avg 0.85 0.86 0.85 30174

average eval\_loss: 0.48388407637955805, accuracy: 0.8645854046530125