training adversary 1 epoch...

930.2632415294647 seconds used......

1 training finished! train loss: 680.2187120448798

evaluating...

Confusion matrix, without normalization

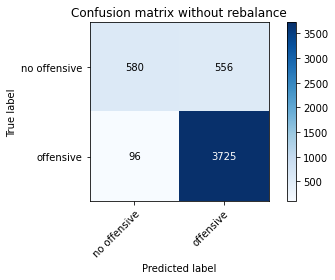
[[ 580 556]

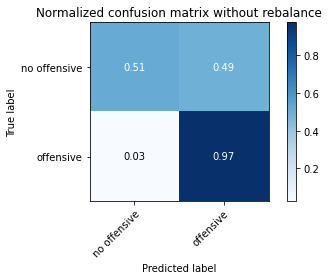
[ 96 3725]]

Normalized confusion matrix

[[0.51056338 0.48943662]

[0.02512431 0.97487569]]





precision recall f1-score support

no offensive 0.86 0.51 0.64 1136

offensive 0.87 0.97 0.92 3821

accuracy 0.87 4957

macro avg 0.86 0.74 0.78 4957

weighted avg 0.87 0.87 0.86 4957

alpha = 0.05

training 1 epoch...

988.1324510574341 seconds used......

1 training finished! average train loss: -4.002792632706046

total clf loss: 663.369054232724 total adv loss: 172018.13384778798

evaluating...

Confusion matrix, without normalization

[[ 60 207 27]

[ 56 3673 92]

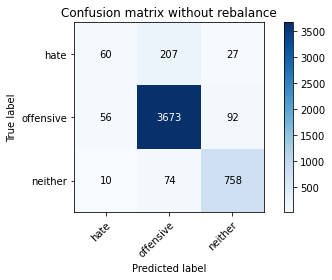
[ 10 74 758]]

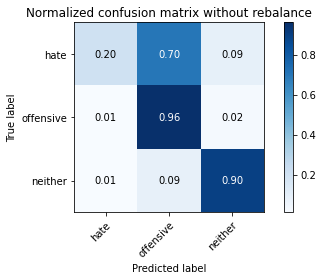
Normalized confusion matrix

[[0.20408163 0.70408163 0.09183673]

[0.01465585 0.96126668 0.02407747]

[0.01187648 0.08788599 0.90023753]]





precision recall f1-score support

hate 0.48 0.20 0.29 294

offensive 0.93 0.96 0.94 3821

neither 0.86 0.90 0.88 842

accuracy 0.91 4957

macro avg 0.76 0.69 0.70 4957

weighted avg 0.89 0.91 0.90 4957

average eval\_loss: 0.2696476650364216, accuracy: 0.9059915271333467

training 2 epoch...

986.7505404949188 seconds used......

2 training finished! average train loss: -4.557684372364723

total clf loss: 474.3181006675586 total adv loss: 190244.12083816528

evaluating...

Confusion matrix, without normalization

[[ 32 244 18]

[ 16 3713 92]

[ 7 64 771]]

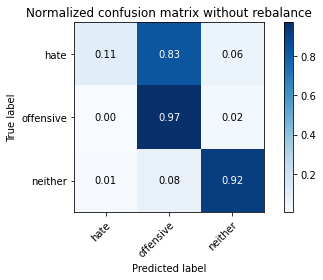
Normalized confusion matrix

[[0.10884354 0.82993197 0.06122449]

[0.00418739 0.97173515 0.02407747]

[0.00831354 0.0760095 0.91567696]]





precision recall f1-score support

hate 0.58 0.11 0.18 294

offensive 0.92 0.97 0.95 3821

neither 0.88 0.92 0.89 842

accuracy 0.91 4957

macro avg 0.79 0.67 0.68 4957

weighted avg 0.89 0.91 0.89 4957

average eval\_loss: 0.24568755791011837, accuracy: 0.9110349001412145

training 3 epoch...

986.448748588562 seconds used......

3 training finished! average train loss: -4.987317780202768

total clf loss: 386.7030749320984 total adv loss: 205531.08121490479

evaluating...

Confusion matrix, without normalization

[[ 121 157 16]

[ 103 3638 80]

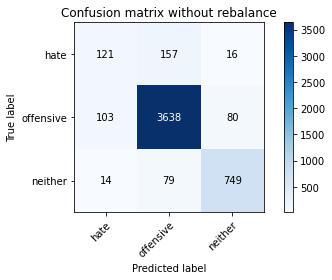
[ 14 79 749]]

Normalized confusion matrix

[[0.41156463 0.53401361 0.05442177]

[0.02695629 0.95210678 0.02093693]

[0.01662708 0.09382423 0.88954869]]





precision recall f1-score support

hate 0.51 0.41 0.45 294

offensive 0.94 0.95 0.95 3821

neither 0.89 0.89 0.89 842

accuracy 0.91 4957

macro avg 0.78 0.75 0.76 4957

weighted avg 0.90 0.91 0.91 4957

average eval\_loss: 0.26641176972771063, accuracy: 0.9094210207786968

training 4 epoch...

987.8368678092957 seconds used......

4 training finished! average train loss: -5.413584894053093

total clf loss: 310.77805655729026 total adv loss: 220918.33444404602

evaluating...

Confusion matrix, without normalization

[[ 68 210 16]

[ 53 3702 66]

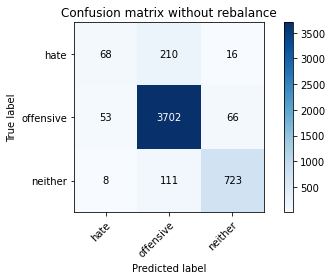
[ 8 111 723]]

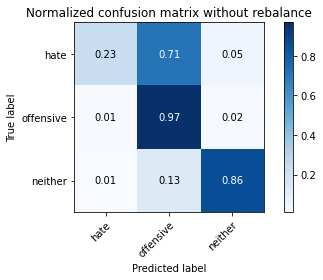
Normalized confusion matrix

[[0.23129252 0.71428571 0.05442177]

[0.01387071 0.96885632 0.01727297]

[0.00950119 0.13182898 0.85866983]]





precision recall f1-score support

hate 0.53 0.23 0.32 294

offensive 0.92 0.97 0.94 3821

neither 0.90 0.86 0.88 842

accuracy 0.91 4957

macro avg 0.78 0.69 0.71 4957

weighted avg 0.89 0.91 0.90 4957

average eval\_loss: 0.28127571459739437, accuracy: 0.9063949969739762

training 5 epoch...

986.4581711292267 seconds used......

5 training finished! average train loss: -5.840291591347076

total clf loss: 234.8097678576596 total adv loss: 236322.15616226196

evaluating...

Confusion matrix, without normalization

[[ 122 153 19]

[ 111 3618 92]

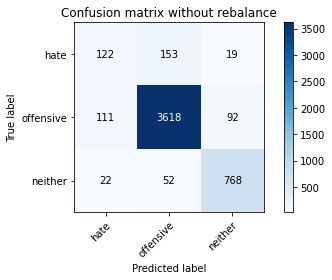
[ 22 52 768]]

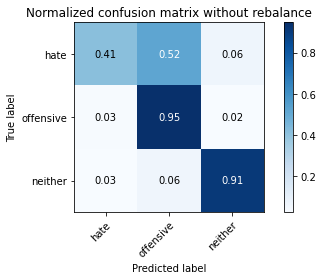
Normalized confusion matrix

[[0.41496599 0.52040816 0.06462585]

[0.02904999 0.94687255 0.02407747]

[0.02612827 0.06175772 0.91211401]]





precision recall f1-score support

hate 0.48 0.41 0.44 294

offensive 0.95 0.95 0.95 3821

neither 0.87 0.91 0.89 842

accuracy 0.91 4957

macro avg 0.77 0.76 0.76 4957

weighted avg 0.91 0.91 0.91 4957

average eval\_loss: 0.32006743238795726, accuracy: 0.9094210207786968

alpha = 0.005

training 1 epoch...

949.7621781826019 seconds used......

1 training finished! average train loss: -0.22174427364504343

total clf loss: 435.77793239429593 total adv loss: 175099.36952209473

evaluating...

Confusion matrix, without normalization

[[ 87 171 36]

[ 83 3618 120]

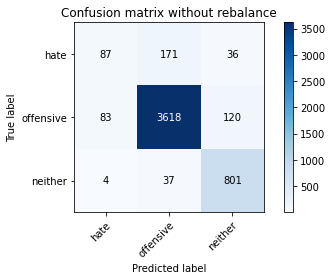
[ 4 37 801]]

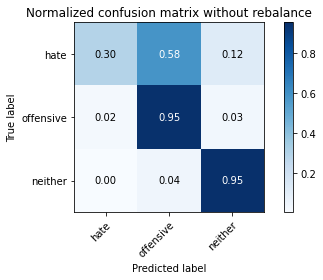
Normalized confusion matrix

[[0.29591837 0.58163265 0.12244898]

[0.02172206 0.94687255 0.03140539]

[0.00475059 0.04394299 0.95130641]]





precision recall f1-score support

hate 0.50 0.30 0.37 294

offensive 0.95 0.95 0.95 3821

neither 0.84 0.95 0.89 842

accuracy 0.91 4957

macro avg 0.76 0.73 0.74 4957

weighted avg 0.90 0.91 0.90 4957

average eval\_loss: 0.2727446429968451, accuracy: 0.9090175509380674

training 2 epoch...

950.5832052230835 seconds used......

2 training finished! average train loss: -0.2987766780008046

total clf loss: 331.54327752115205 total adv loss: 184803.49013900757

evaluating...

Confusion matrix, without normalization

[[ 92 185 17]

[ 61 3697 63]

[ 5 84 753]]

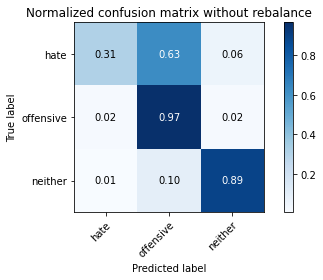
Normalized confusion matrix

[[0.31292517 0.6292517 0.05782313]

[0.01596441 0.96754776 0.01648783]

[0.00593824 0.09976247 0.89429929]]





precision recall f1-score support

hate 0.58 0.31 0.41 294

offensive 0.93 0.97 0.95 3821

neither 0.90 0.89 0.90 842

accuracy 0.92 4957

macro avg 0.81 0.72 0.75 4957

weighted avg 0.91 0.92 0.91 4957

average eval\_loss: 0.24756932162281667, accuracy: 0.9162800080693968

training 3 epoch...

950.1139140129089 seconds used......

3 training finished! average train loss: -0.4601424169199249

total clf loss: 238.81310179457068 total adv loss: 230255.1082649231

evaluating...

Confusion matrix, without normalization

[[ 132 149 13]

[ 167 3607 47]

[ 31 127 684]]

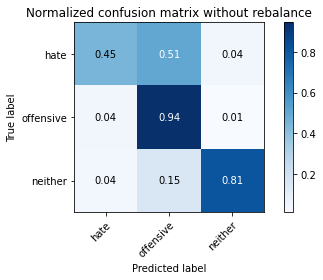
Normalized confusion matrix

[[0.44897959 0.50680272 0.04421769]

[0.04370584 0.94399372 0.01230044]

[0.0368171 0.15083135 0.81235154]]





precision recall f1-score support

hate 0.40 0.45 0.42 294

offensive 0.93 0.94 0.94 3821

neither 0.92 0.81 0.86 842

accuracy 0.89 4957

macro avg 0.75 0.74 0.74 4957

weighted avg 0.90 0.89 0.89 4957

average eval\_loss: 0.33090348285059595, accuracy: 0.8922735525519467

training 4 epoch...

950.37921667099 seconds used......

4 training finished! average train loss: -0.5427315816831781

total clf loss: 165.7323936317116 total adv loss: 248393.8293609619

evaluating...

Confusion matrix, without normalization

[[ 103 176 15]

[ 103 3654 64]

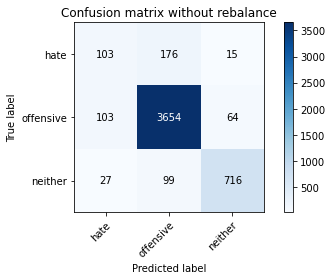
[ 27 99 716]]

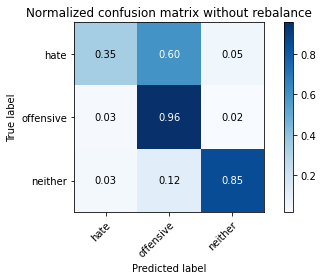
Normalized confusion matrix

[[0.35034014 0.59863946 0.05102041]

[0.02695629 0.95629416 0.01674954]

[0.03206651 0.1175772 0.85035629]]





precision recall f1-score support

hate 0.44 0.35 0.39 294

offensive 0.93 0.96 0.94 3821

neither 0.90 0.85 0.87 842

accuracy 0.90 4957

macro avg 0.76 0.72 0.74 4957

weighted avg 0.90 0.90 0.90 4957

average eval\_loss: 0.3353950985643502, accuracy: 0.9023602985676821

training 5 epoch...

950.8938457965851 seconds used......

5 training finished! average train loss: -0.6043192350647995

total clf loss: 111.08112549770158 total adv loss: 261889.239944458

evaluating...

Confusion matrix, without normalization

[[ 92 184 18]

[ 89 3674 58]

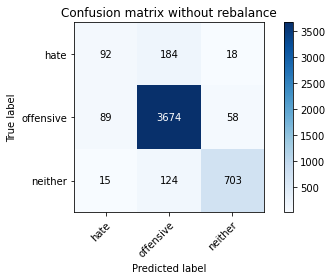
[ 15 124 703]]

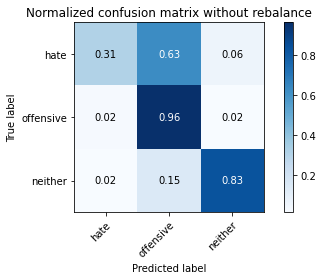
Normalized confusion matrix

[[0.31292517 0.62585034 0.06122449]

[0.02329233 0.9615284 0.01517927]

[0.01781473 0.14726841 0.83491686]]





precision recall f1-score support

hate 0.47 0.31 0.38 294

offensive 0.92 0.96 0.94 3821

neither 0.90 0.83 0.87 842

accuracy 0.90 4957

macro avg 0.76 0.70 0.73 4957

weighted avg 0.89 0.90 0.90 4957

average eval\_loss: 0.4564885134710611, accuracy: 0.9015533588864233