alpha = 0.05

training 1 epoch...

892.6859352588654 seconds used......

1 training finished! average train loss: 0.26591067787178096

total clf loss: 550.4623621758074 total adv loss: 463.22976223612204

evaluating...

Confusion matrix, without normalization

[[ 112 152 30]

[ 103 3623 95]

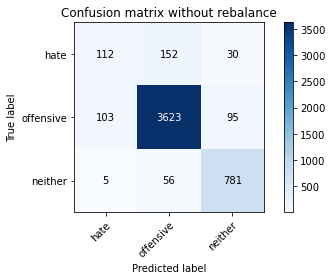
[ 5 56 781]]

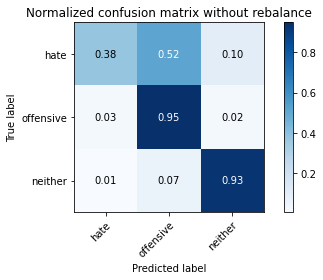
Normalized confusion matrix

[[0.38095238 0.5170068 0.10204082]

[0.02695629 0.9481811 0.0248626 ]

[0.00593824 0.06650831 0.92755344]]





precision recall f1-score support

hate 0.51 0.38 0.44 294

offensive 0.95 0.95 0.95 3821

neither 0.86 0.93 0.89 842

accuracy 0.91 4957

macro avg 0.77 0.75 0.76 4957

weighted avg 0.91 0.91 0.91 4957

average eval\_loss: 0.237183561507282, accuracy: 0.9110349001412145

training 2 epoch...

893.6032285690308 seconds used......

2 training finished! average train loss: 0.1937703284981181

total clf loss: 402.2507621371187 total adv loss: 360.0840160108637

evaluating...

Confusion matrix, without normalization

[[ 101 169 24]

[ 85 3637 99]

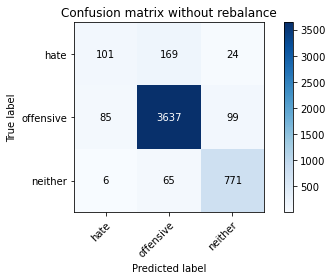
[ 6 65 771]]

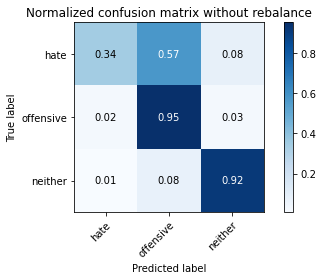
Normalized confusion matrix

[[0.34353741 0.57482993 0.08163265]

[0.02224549 0.95184507 0.02590945]

[0.00712589 0.07719715 0.91567696]]





precision recall f1-score support

hate 0.53 0.34 0.42 294

offensive 0.94 0.95 0.95 3821

neither 0.86 0.92 0.89 842

accuracy 0.91 4957

macro avg 0.78 0.74 0.75 4957

weighted avg 0.90 0.91 0.90 4957

average eval\_loss: 0.2463961603610979, accuracy: 0.9096227556990115

training 3 epoch...

894.2193877696991 seconds used......

3 training finished! average train loss: 0.13222446339398639

total clf loss: 275.8006939829793 total adv loss: 271.99165605416056

evaluating...

Confusion matrix, without normalization

[[ 137 134 23]

[ 132 3584 105]

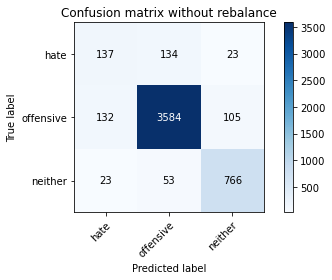
[ 23 53 766]]

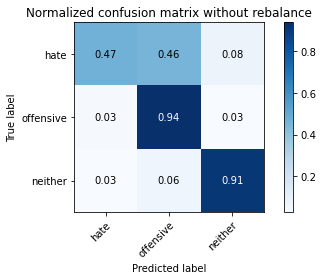
Normalized confusion matrix

[[0.46598639 0.45578231 0.07823129]

[0.03454593 0.93797435 0.02747972]

[0.02731591 0.06294537 0.90973872]]





precision recall f1-score support

hate 0.47 0.47 0.47 294

offensive 0.95 0.94 0.94 3821

neither 0.86 0.91 0.88 842

accuracy 0.91 4957

macro avg 0.76 0.77 0.76 4957

weighted avg 0.91 0.91 0.91 4957

average eval\_loss: 0.28879733572196736, accuracy: 0.905184587452088

training 4 epoch...

893.2212419509888 seconds used......

4 training finished! average train loss: 0.07654722187910484

total clf loss: 160.51953173644142 total adv loss: 174.5278142231109

evaluating...

Confusion matrix, without normalization

[[ 130 148 16]

[ 120 3636 65]

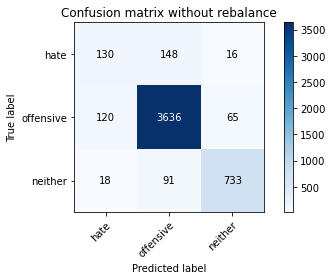
[ 18 91 733]]

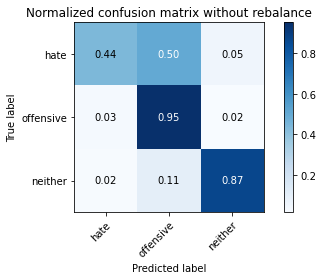
Normalized confusion matrix

[[0.44217687 0.50340136 0.05442177]

[0.03140539 0.95158336 0.01701125]

[0.02137767 0.10807601 0.87054632]]





precision recall f1-score support

hate 0.49 0.44 0.46 294

offensive 0.94 0.95 0.94 3821

neither 0.90 0.87 0.89 842

accuracy 0.91 4957

macro avg 0.77 0.75 0.76 4957

weighted avg 0.91 0.91 0.91 4957

average eval\_loss: 0.3460223059326146, accuracy: 0.9076054064958644

training 5 epoch...

893.4035818576813 seconds used......

5 training finished! average train loss: 0.03940447885224991

total clf loss: 82.69709168615373 total adv loss: 91.16020617220056

evaluating...

Confusion matrix, without normalization

[[ 85 187 22]

[ 81 3650 90]

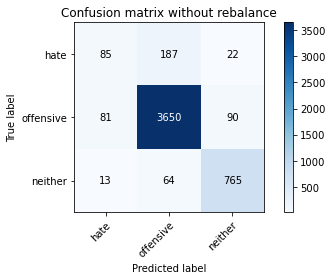
[ 13 64 765]]

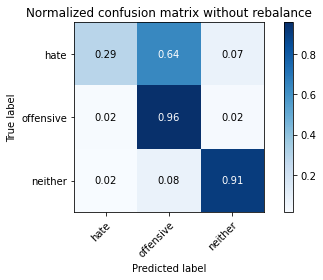
Normalized confusion matrix

[[0.28911565 0.63605442 0.07482993]

[0.02119864 0.95524732 0.02355404]

[0.01543943 0.0760095 0.90855107]]





precision recall f1-score support

hate 0.47 0.29 0.36 294

offensive 0.94 0.96 0.95 3821

neither 0.87 0.91 0.89 842

accuracy 0.91 4957

macro avg 0.76 0.72 0.73 4957

weighted avg 0.90 0.91 0.90 4957

average eval\_loss: 0.4434567468317636, accuracy: 0.9078071414161791

training 6 epoch...

894.098753452301 seconds used......

6 training finished! average train loss: 0.026999911425813143

total clf loss: 56.711023509422375 total adv loss: 63.40398132045584

evaluating...

Confusion matrix, without normalization

[[ 138 145 11]

[ 182 3583 56]

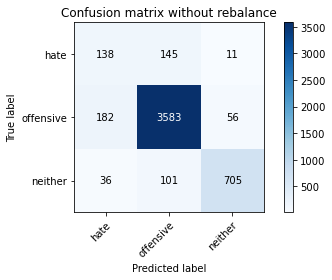
[ 36 101 705]]

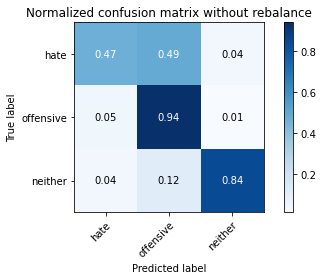
Normalized confusion matrix

[[0.46938776 0.49319728 0.03741497]

[0.04763151 0.93771264 0.01465585]

[0.04275534 0.11995249 0.83729216]]





precision recall f1-score support

hate 0.39 0.47 0.42 294

offensive 0.94 0.94 0.94 3821

neither 0.91 0.84 0.87 842

accuracy 0.89 4957

macro avg 0.75 0.75 0.74 4957

weighted avg 0.90 0.89 0.90 4957

average eval\_loss: 0.5092930028466166, accuracy: 0.8928787573128909

training 7 epoch...

892.2075769901276 seconds used......

7 training finished! average train loss: 0.02031214596420264

total clf loss: 42.84288134144117 total adv loss: 51.27791967172152

evaluating...

Confusion matrix, without normalization

[[ 79 192 23]

[ 97 3642 82]

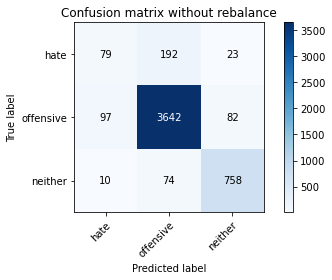
[ 10 74 758]]

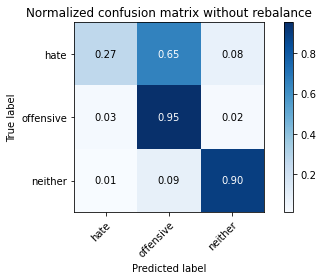
Normalized confusion matrix

[[0.26870748 0.65306122 0.07823129]

[0.02538602 0.95315362 0.02146035]

[0.01187648 0.08788599 0.90023753]]





precision recall f1-score support

hate 0.42 0.27 0.33 294

offensive 0.93 0.95 0.94 3821

neither 0.88 0.90 0.89 842

accuracy 0.90 4957

macro avg 0.74 0.71 0.72 4957

weighted avg 0.89 0.90 0.90 4957

average eval\_loss: 0.5155731163304593, accuracy: 0.9035707080895703

training 8 epoch...

891.9880073070526 seconds used......

8 training finished! average train loss: 0.01823272821265445

total clf loss: 38.504083612302566 total adv loss: 46.97167050675921

evaluating...

Confusion matrix, without normalization

[[ 108 171 15]

[ 130 3638 53]

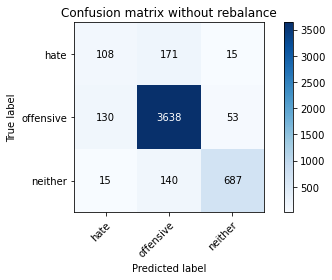
[ 15 140 687]]

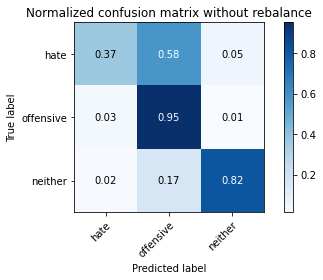
Normalized confusion matrix

[[0.36734694 0.58163265 0.05102041]

[0.03402251 0.95210678 0.01387071]

[0.01781473 0.16627078 0.81591449]]





precision recall f1-score support

hate 0.43 0.37 0.39 294

offensive 0.92 0.95 0.94 3821

neither 0.91 0.82 0.86 842

accuracy 0.89 4957

macro avg 0.75 0.71 0.73 4957

weighted avg 0.89 0.89 0.89 4957

average eval\_loss: 0.5655895254221759, accuracy: 0.8942909017550938

training 9 epoch...

892.05277967453 seconds used......

9 training finished! average train loss: 0.018947666510228698

total clf loss: 39.896953590785415 total adv loss: 46.47461824927086

evaluating...

Confusion matrix, without normalization

[[ 88 189 17]

[ 100 3658 63]

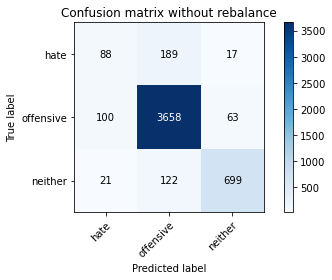
[ 21 122 699]]

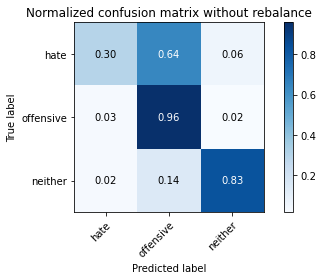
Normalized confusion matrix

[[0.29931973 0.64285714 0.05782313]

[0.02617116 0.95734101 0.01648783]

[0.02494062 0.14489311 0.83016627]]





precision recall f1-score support

hate 0.42 0.30 0.35 294

offensive 0.92 0.96 0.94 3821

neither 0.90 0.83 0.86 842

accuracy 0.90 4957

macro avg 0.75 0.70 0.72 4957

weighted avg 0.89 0.90 0.89 4957

average eval\_loss: 0.5563520145609683, accuracy: 0.8967117207988703

training 10 epoch...

893.2676661014557 seconds used......

10 training finished! average train loss: 0.014542588935720895

total clf loss: 30.675311849277932 total adv loss: 36.74715784559203

evaluating...

Confusion matrix, without normalization

[[ 90 181 23]

[ 94 3642 85]

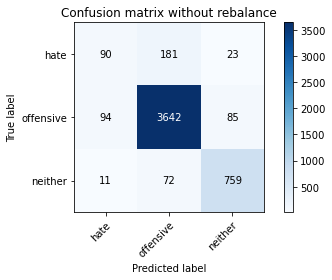
[ 11 72 759]]

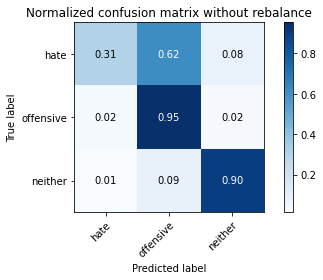
Normalized confusion matrix

[[0.30612245 0.61564626 0.07823129]

[0.02460089 0.95315362 0.02224549]

[0.01306413 0.08551069 0.90142518]]





precision recall f1-score support

hate 0.46 0.31 0.37 294

offensive 0.94 0.95 0.94 3821

neither 0.88 0.90 0.89 842

accuracy 0.91 4957

macro avg 0.76 0.72 0.73 4957

weighted avg 0.90 0.91 0.90 4957

average eval\_loss: 0.625897417049835, accuracy: 0.9059915271333467