alpha = 0.5

loss = loss – alp\*loss

1 run

training 1 epoch...

45.06312084197998 seconds used......

1 training finished! average train loss: 0.572154358465072

total clf loss: 140.47460815310478 total adv loss: 85.27242579311132

evaluating...

Confusion matrix, without normalization

[[ 176 36 82]

[ 707 2339 775]

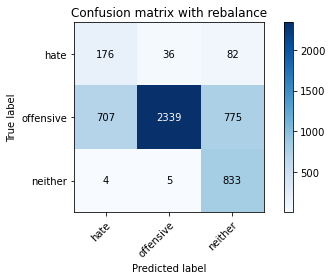
[ 4 5 833]]

Normalized confusion matrix

[[0.59863946 0.12244898 0.27891156]

[0.1850301 0.61214342 0.20282649]

[0.00475059 0.00593824 0.98931116]]





precision recall f1-score support

hate 0.20 0.60 0.30 294

offensive 0.98 0.61 0.75 3821

neither 0.49 0.99 0.66 842

accuracy 0.68 4957

macro avg 0.56 0.73 0.57 4957

weighted avg 0.85 0.68 0.71 4957

average eval\_loss: 0.7095672156541578, accuracy: 0.6754085132136373

evaluating adversary...

Confusion matrix, without normalization

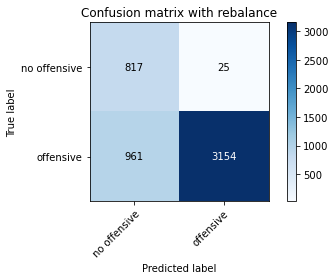
[[ 817 25]

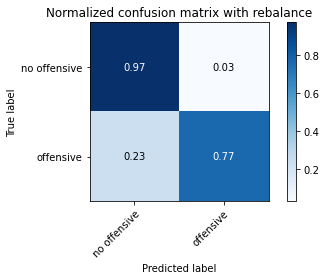
[ 961 3154]]

Normalized confusion matrix

[[0.97030879 0.02969121]

[0.23353584 0.76646416]]





precision recall f1-score support

no offensive 0.46 0.97 0.62 842

offensive 0.99 0.77 0.86 4115

accuracy 0.80 4957

macro avg 0.73 0.87 0.74 4957

weighted avg 0.90 0.80 0.82 4957

training 2 epoch...

45.011592388153076 seconds used......

2 training finished! average train loss: 0.33841936882824925

total clf loss: 96.65197432041168 total adv loss: 77.56452444940805

evaluating...

Confusion matrix, without normalization

[[ 239 41 14]

[ 845 2840 136]

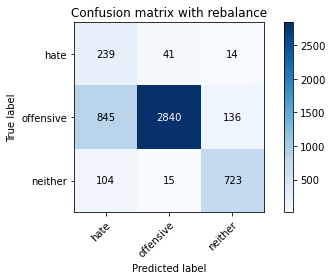
[ 104 15 723]]

Normalized confusion matrix

[[0.81292517 0.13945578 0.04761905]

[0.2211463 0.74326093 0.03559278]

[0.12351544 0.01781473 0.85866983]]





precision recall f1-score support

hate 0.20 0.81 0.32 294

offensive 0.98 0.74 0.85 3821

neither 0.83 0.86 0.84 842

accuracy 0.77 4957

macro avg 0.67 0.80 0.67 4957

weighted avg 0.91 0.77 0.81 4957

average eval\_loss: 0.6145746231439614, accuracy: 0.7669961670365141

evaluating adversary...

Confusion matrix, without normalization

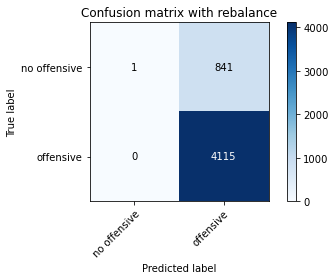
[[ 1 841]

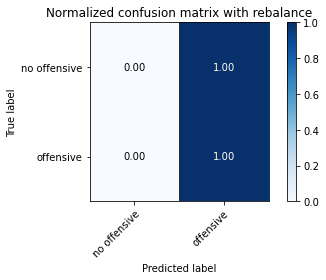
[ 0 4115]]

Normalized confusion matrix

[[0.00118765 0.99881235]

[0. 1. ]]





precision recall f1-score support

no offensive 1.00 0.00 0.00 842

offensive 0.83 1.00 0.91 4115

accuracy 0.83 4957

macro avg 0.92 0.50 0.45 4957

weighted avg 0.86 0.83 0.75 4957

training 3 epoch...

44.929988622665405 seconds used......

3 training finished! average train loss: 0.3312780056257694

total clf loss: 78.8750665448606 total adv loss: 44.45305564627051

evaluating...

Confusion matrix, without normalization

[[ 221 41 32]

[ 822 2846 153]

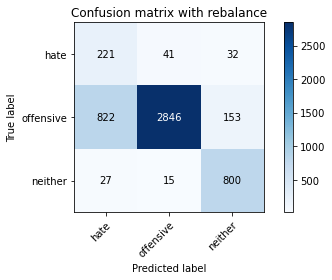
[ 27 15 800]]

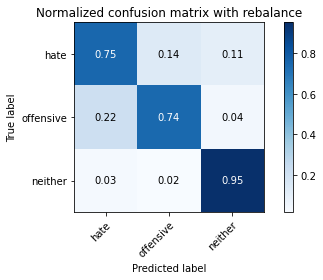
Normalized confusion matrix

[[0.75170068 0.13945578 0.10884354]

[0.21512693 0.7448312 0.04004187]

[0.03206651 0.01781473 0.95011876]]





precision recall f1-score support

hate 0.21 0.75 0.32 294

offensive 0.98 0.74 0.85 3821

neither 0.81 0.95 0.88 842

accuracy 0.78 4957

macro avg 0.67 0.82 0.68 4957

weighted avg 0.91 0.78 0.82 4957

average eval\_loss: 0.5957780366523131, accuracy: 0.7801089368569699

evaluating adversary...

Confusion matrix, without normalization

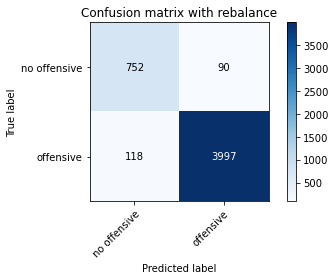
[[ 752 90]

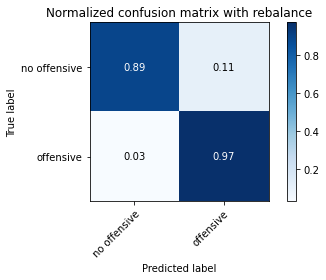
[ 118 3997]]

Normalized confusion matrix

[[0.89311164 0.10688836]

[0.02867558 0.97132442]]





precision recall f1-score support

no offensive 0.86 0.89 0.88 842

offensive 0.98 0.97 0.97 4115

accuracy 0.96 4957

macro avg 0.92 0.93 0.93 4957

weighted avg 0.96 0.96 0.96 4957

training 4 epoch...

44.9801926612854 seconds used......

4 training finished! average train loss: 0.24310034626757193

total clf loss: 53.67981066554785 total adv loss: 24.21930253179744

evaluating...

Confusion matrix, without normalization

[[ 253 29 12]

[1139 2578 104]

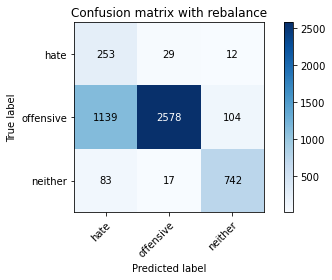
[ 83 17 742]]

Normalized confusion matrix

[[0.86054422 0.09863946 0.04081633]

[0.29808951 0.67469249 0.02721801]

[0.09857482 0.02019002 0.88123515]]





precision recall f1-score support

hate 0.17 0.86 0.29 294

offensive 0.98 0.67 0.80 3821

neither 0.86 0.88 0.87 842

accuracy 0.72 4957

macro avg 0.67 0.81 0.65 4957

weighted avg 0.91 0.72 0.78 4957

average eval\_loss: 0.6610797977735919, accuracy: 0.7207988702844462

evaluating adversary...

Confusion matrix, without normalization

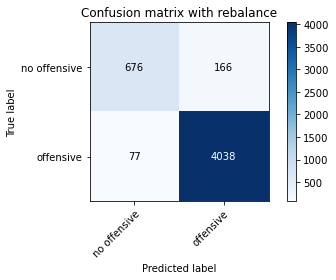
[[ 676 166]

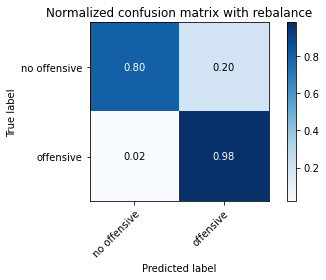
[ 77 4038]]

Normalized confusion matrix

[[0.80285036 0.19714964]

[0.01871203 0.98128797]]





precision recall f1-score support

no offensive 0.90 0.80 0.85 842

offensive 0.96 0.98 0.97 4115

accuracy 0.95 4957

macro avg 0.93 0.89 0.91 4957

weighted avg 0.95 0.95 0.95 4957

training 5 epoch...

44.910178899765015 seconds used......

5 training finished! average train loss: 0.16003250948300488

total clf loss: 35.81671594455838 total adv loss: 16.902313824743032

evaluating...

Confusion matrix, without normalization

[[ 198 69 27]

[ 727 2974 120]

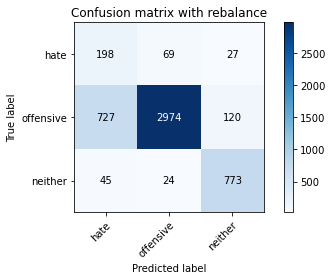
[ 45 24 773]]

Normalized confusion matrix

[[0.67346939 0.23469388 0.09183673]

[0.19026433 0.77833028 0.03140539]

[0.05344418 0.02850356 0.91805226]]





precision recall f1-score support

hate 0.20 0.67 0.31 294

offensive 0.97 0.78 0.86 3821

neither 0.84 0.92 0.88 842

accuracy 0.80 4957

macro avg 0.67 0.79 0.68 4957

weighted avg 0.90 0.80 0.83 4957

average eval\_loss: 0.544873054544892, accuracy: 0.795844260641517

evaluating adversary...

Confusion matrix, without normalization

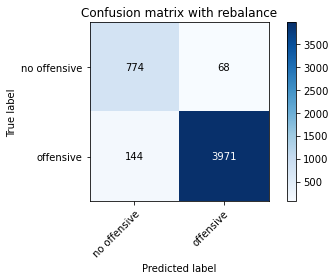
[[ 774 68]

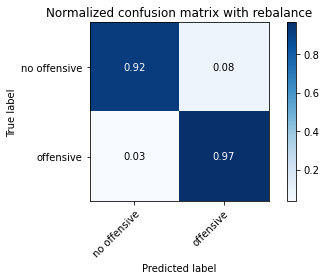
[ 144 3971]]

Normalized confusion matrix

[[0.9192399 0.0807601 ]

[0.03499392 0.96500608]]





precision recall f1-score support

no offensive 0.84 0.92 0.88 842

offensive 0.98 0.97 0.97 4115

accuracy 0.96 4957

macro avg 0.91 0.94 0.93 4957

weighted avg 0.96 0.96 0.96 4957

training 6 epoch...

45.0404109954834 seconds used......

6 training finished! average train loss: 0.09468215580332523

total clf loss: 21.503740382380784 total adv loss: 10.626183341257274

evaluating...

Confusion matrix, without normalization

[[ 214 58 22]

[ 763 2923 135]

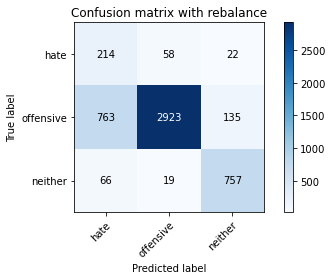
[ 66 19 757]]

Normalized confusion matrix

[[0.72789116 0.19727891 0.07482993]

[0.19968595 0.76498299 0.03533107]

[0.0783848 0.02256532 0.89904988]]





precision recall f1-score support

hate 0.21 0.73 0.32 294

offensive 0.97 0.76 0.86 3821

neither 0.83 0.90 0.86 842

accuracy 0.79 4957

macro avg 0.67 0.80 0.68 4957

weighted avg 0.90 0.79 0.83 4957

average eval\_loss: 0.7306152490012708, accuracy: 0.785555779705467

evaluating adversary...

Confusion matrix, without normalization

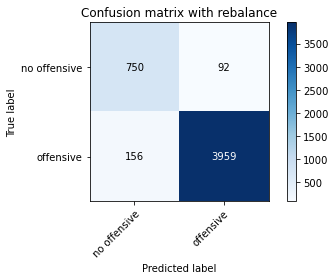
[[ 750 92]

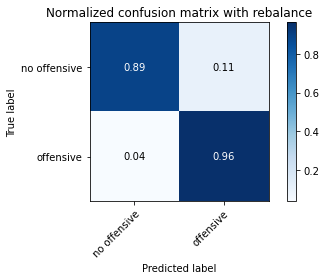
[ 156 3959]]

Normalized confusion matrix

[[0.89073634 0.10926366]

[0.03791009 0.96208991]]





precision recall f1-score support

no offensive 0.83 0.89 0.86 842

offensive 0.98 0.96 0.97 4115

accuracy 0.95 4957

macro avg 0.90 0.93 0.91 4957

weighted avg 0.95 0.95 0.95 4957

training 7 epoch...

44.9461874961853 seconds used......

7 training finished! average train loss: 0.05224609828951676

total clf loss: 12.484012172091752 total adv loss: 7.099858731802669

evaluating...

Confusion matrix, without normalization

[[ 173 67 54]

[ 558 3031 232]

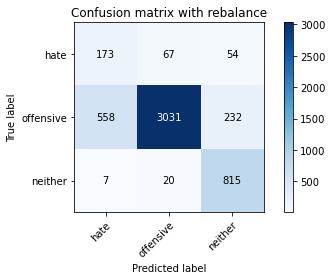
[ 7 20 815]]

Normalized confusion matrix

[[0.58843537 0.22789116 0.18367347]

[0.14603507 0.79324784 0.06071709]

[0.00831354 0.02375297 0.96793349]]





precision recall f1-score support

hate 0.23 0.59 0.34 294

offensive 0.97 0.79 0.87 3821

neither 0.74 0.97 0.84 842

accuracy 0.81 4957

macro avg 0.65 0.78 0.68 4957

weighted avg 0.89 0.81 0.84 4957

average eval\_loss: 0.7970391867603476, accuracy: 0.8107726447448054

evaluating adversary...

Confusion matrix, without normalization

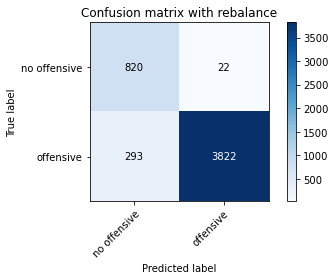
[[ 820 22]

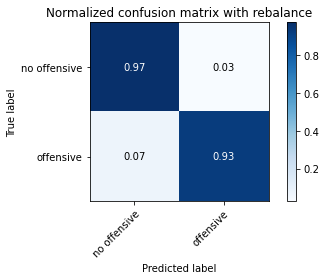
[ 293 3822]]

Normalized confusion matrix

[[0.97387173 0.02612827]

[0.07120292 0.92879708]]





precision recall f1-score support

no offensive 0.74 0.97 0.84 842

offensive 0.99 0.93 0.96 4115

accuracy 0.94 4957

macro avg 0.87 0.95 0.90 4957

weighted avg 0.95 0.94 0.94 4957

training 8 epoch...

44.94195890426636 seconds used......

8 training finished! average train loss: 0.03699167431413866

total clf loss: 13.437590537592769 total adv loss: 14.224028489901684

evaluating...

Confusion matrix, without normalization

[[ 223 56 15]

[ 845 2856 120]

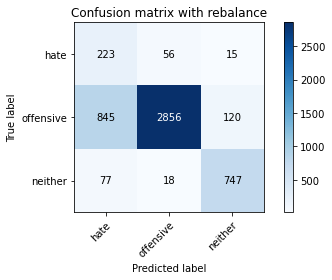
[ 77 18 747]]

Normalized confusion matrix

[[0.7585034 0.19047619 0.05102041]

[0.2211463 0.74744831 0.03140539]

[0.09144893 0.02137767 0.8871734 ]]





precision recall f1-score support

hate 0.19 0.76 0.31 294

offensive 0.97 0.75 0.85 3821

neither 0.85 0.89 0.87 842

accuracy 0.77 4957

macro avg 0.67 0.80 0.67 4957

weighted avg 0.91 0.77 0.82 4957

average eval\_loss: 1.1589986363335723, accuracy: 0.771837805124067

evaluating adversary...

Confusion matrix, without normalization

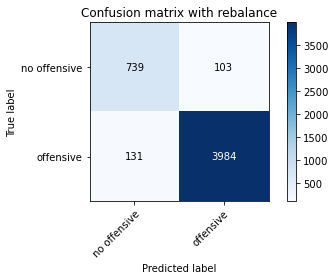
[[ 739 103]

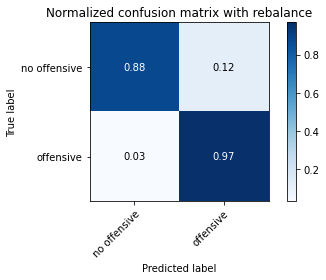
[ 131 3984]]

Normalized confusion matrix

[[0.87767221 0.12232779]

[0.03183475 0.96816525]]





precision recall f1-score support

no offensive 0.85 0.88 0.86 842

offensive 0.97 0.97 0.97 4115

accuracy 0.95 4957

macro avg 0.91 0.92 0.92 4957

weighted avg 0.95 0.95 0.95 4957

training 9 epoch...

45.022881269454956 seconds used......

9 training finished! average train loss: 0.03084140180130843

total clf loss: 8.518266088329256 total adv loss: 6.4887726765009575

evaluating...

Confusion matrix, without normalization

[[ 201 71 22]

[ 627 3063 131]

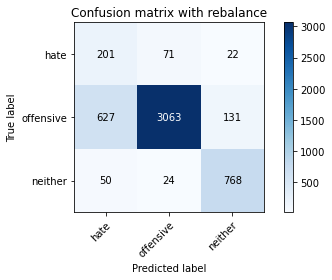
[ 50 24 768]]

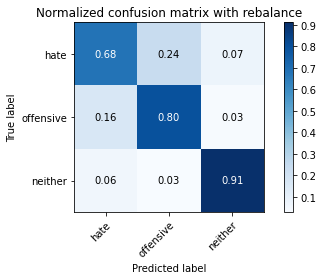
Normalized confusion matrix

[[0.68367347 0.2414966 0.07482993]

[0.16409317 0.80162261 0.03428422]

[0.05938242 0.02850356 0.91211401]]





precision recall f1-score support

hate 0.23 0.68 0.34 294

offensive 0.97 0.80 0.88 3821

neither 0.83 0.91 0.87 842

accuracy 0.81 4957

macro avg 0.68 0.80 0.70 4957

weighted avg 0.90 0.81 0.84 4957

average eval\_loss: 0.9614661299030206, accuracy: 0.8133951987088965

evaluating adversary...

Confusion matrix, without normalization

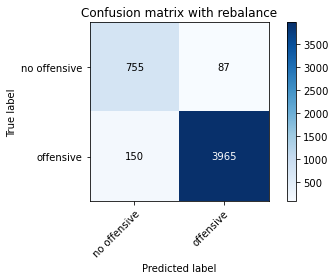
[[ 755 87]

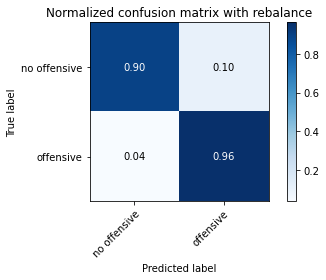
[ 150 3965]]

Normalized confusion matrix

[[0.89667458 0.10332542]

[0.036452 0.963548 ]]





precision recall f1-score support

no offensive 0.83 0.90 0.86 842

offensive 0.98 0.96 0.97 4115

accuracy 0.95 4957

macro avg 0.91 0.93 0.92 4957

weighted avg 0.95 0.95 0.95 4957

training 10 epoch...

44.883875608444214 seconds used......

10 training finished! average train loss: 0.009145317934126708

total clf loss: 27.413476226851344 total adv loss: 51.69925343513023

evaluating...

Confusion matrix, without normalization

[[ 201 74 19]

[ 593 3096 132]

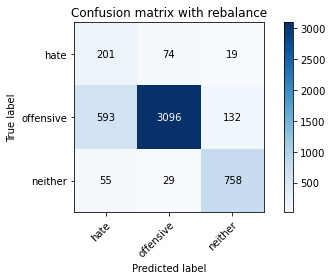
[ 55 29 758]]

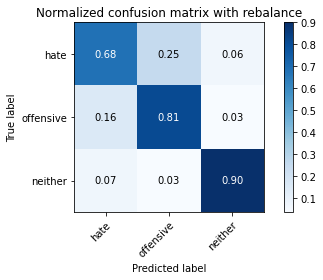
Normalized confusion matrix

[[0.68367347 0.25170068 0.06462585]

[0.15519498 0.81025909 0.03454593]

[0.06532067 0.03444181 0.90023753]]





precision recall f1-score support

hate 0.24 0.68 0.35 294

offensive 0.97 0.81 0.88 3821

neither 0.83 0.90 0.87 842

accuracy 0.82 4957

macro avg 0.68 0.80 0.70 4957

weighted avg 0.90 0.82 0.85 4957

average eval\_loss: 0.8408287794239098, accuracy: 0.8180351018761347

evaluating adversary...

Confusion matrix, without normalization

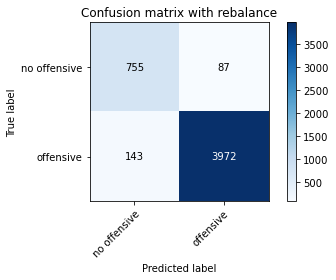
[[ 755 87]

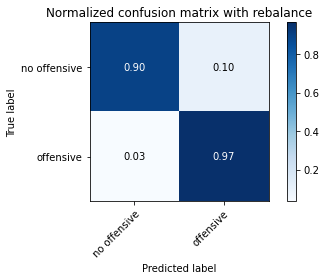
[ 143 3972]]

Normalized confusion matrix

[[0.89667458 0.10332542]

[0.03475091 0.96524909]]





precision recall f1-score support

no offensive 0.84 0.90 0.87 842

offensive 0.98 0.97 0.97 4115

accuracy 0.95 4957

macro avg 0.91 0.93 0.92 4957

weighted avg 0.96 0.95 0.95 4957

2 run

training 1 epoch...

45.01799535751343 seconds used......

1 training finished! average train loss: 0.5764407844048495

total clf loss: 145.51075053215027 total adv loss: 93.87875308096409

evaluating...

Confusion matrix, without normalization

[[ 233 56 5]

[ 723 3086 12]

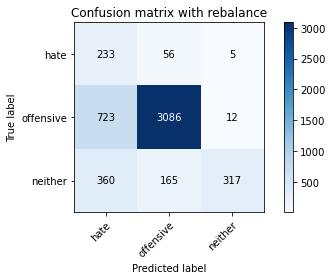
[ 360 165 317]]

Normalized confusion matrix

[[0.79251701 0.19047619 0.0170068 ]

[0.18921748 0.80764198 0.00314054]

[0.42755344 0.195962 0.37648456]]





precision recall f1-score support

hate 0.18 0.79 0.29 294

offensive 0.93 0.81 0.87 3821

neither 0.95 0.38 0.54 842

accuracy 0.73 4957

macro avg 0.69 0.66 0.56 4957

weighted avg 0.89 0.73 0.78 4957

average eval\_loss: 0.6194653026759624, accuracy: 0.7335081702642727

evaluating adversary...

Confusion matrix, without normalization

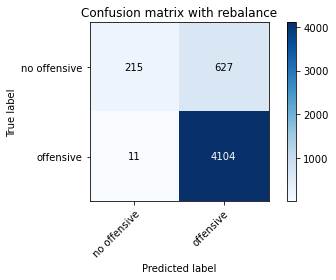
[[ 215 627]

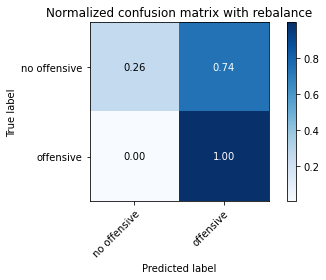
[ 11 4104]]

Normalized confusion matrix

[[0.25534442 0.74465558]

[0.00267315 0.99732685]]





precision recall f1-score support

no offensive 0.95 0.26 0.40 842

offensive 0.87 1.00 0.93 4115

accuracy 0.87 4957

macro avg 0.91 0.63 0.67 4957

weighted avg 0.88 0.87 0.84 4957

training 2 epoch...

45.088716983795166 seconds used......

2 training finished! average train loss: 0.32691399500383966

total clf loss: 96.11557479202747 total adv loss: 80.42656375095248

evaluating...

Confusion matrix, without normalization

[[ 221 54 19]

[ 730 2987 104]

[ 66 26 750]]

Normalized confusion matrix

[[0.75170068 0.18367347 0.06462585]

[0.19104946 0.78173253 0.02721801]

[0.0783848 0.03087886 0.89073634]]





precision recall f1-score support

hate 0.22 0.75 0.34 294

offensive 0.97 0.78 0.87 3821

neither 0.86 0.89 0.87 842

accuracy 0.80 4957

macro avg 0.68 0.81 0.69 4957

weighted avg 0.91 0.80 0.84 4957

average eval\_loss: 0.4960855595887669, accuracy: 0.7984668146056082

evaluating adversary...

Confusion matrix, without normalization

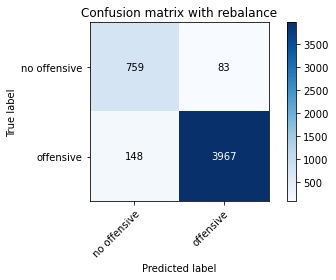
[[ 759 83]

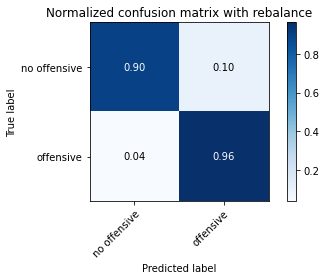
[ 148 3967]]

Normalized confusion matrix

[[0.90142518 0.09857482]

[0.03596598 0.96403402]]





precision recall f1-score support

no offensive 0.84 0.90 0.87 842

offensive 0.98 0.96 0.97 4115

accuracy 0.95 4957

macro avg 0.91 0.93 0.92 4957

weighted avg 0.96 0.95 0.95 4957

training 3 epoch...

44.952367305755615 seconds used......

3 training finished! average train loss: 0.29994562488288906

total clf loss: 66.96761800348759 total adv loss: 31.353831868618727

evaluating...

Confusion matrix, without normalization

[[ 247 39 8]

[ 930 2807 84]

[ 175 13 654]]

Normalized confusion matrix

[[0.84013605 0.13265306 0.02721088]

[0.24339178 0.73462444 0.02198377]

[0.20783848 0.01543943 0.77672209]]





precision recall f1-score support

hate 0.18 0.84 0.30 294

offensive 0.98 0.73 0.84 3821

neither 0.88 0.78 0.82 842

accuracy 0.75 4957

macro avg 0.68 0.78 0.65 4957

weighted avg 0.92 0.75 0.81 4957

average eval\_loss: 0.6393848087999129, accuracy: 0.7480330845269316

evaluating adversary...

Confusion matrix, without normalization

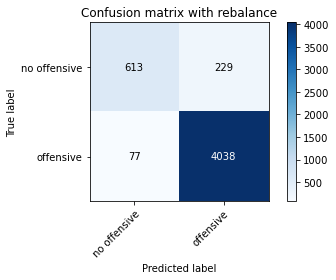
[[ 613 229]

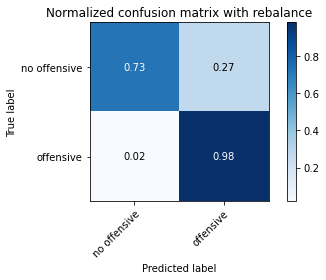
[ 77 4038]]

Normalized confusion matrix

[[0.7280285 0.2719715 ]

[0.01871203 0.98128797]]





precision recall f1-score support

no offensive 0.89 0.73 0.80 842

offensive 0.95 0.98 0.96 4115

accuracy 0.94 4957

macro avg 0.92 0.85 0.88 4957

weighted avg 0.94 0.94 0.94 4957

training 4 epoch...

44.99324560165405 seconds used......

4 training finished! average train loss: 0.23029845312499164

total clf loss: 54.20476224273443 total adv loss: 29.647453462705016

evaluating...

Confusion matrix, without normalization

[[ 186 57 51]

[ 555 2890 376]

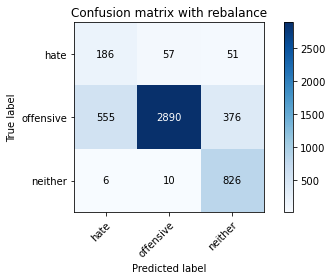
[ 6 10 826]]

Normalized confusion matrix

[[0.63265306 0.19387755 0.17346939]

[0.14524993 0.75634651 0.09840356]

[0.00712589 0.01187648 0.98099762]]





precision recall f1-score support

hate 0.25 0.63 0.36 294

offensive 0.98 0.76 0.85 3821

neither 0.66 0.98 0.79 842

accuracy 0.79 4957

macro avg 0.63 0.79 0.67 4957

weighted avg 0.88 0.79 0.81 4957

average eval\_loss: 0.5542438684271709, accuracy: 0.7871696590679846

evaluating adversary...

Confusion matrix, without normalization

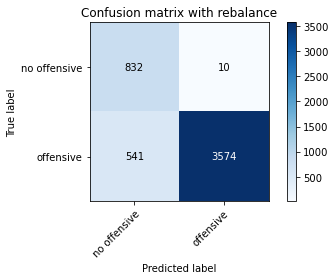
[[ 832 10]

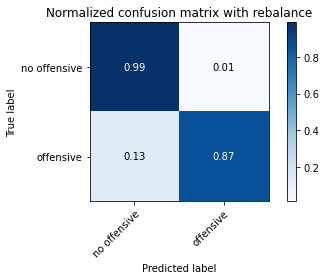
[ 541 3574]]

Normalized confusion matrix

[[0.98812352 0.01187648]

[0.13147023 0.86852977]]





precision recall f1-score support

no offensive 0.61 0.99 0.75 842

offensive 1.00 0.87 0.93 4115

accuracy 0.89 4957

macro avg 0.80 0.93 0.84 4957

weighted avg 0.93 0.89 0.90 4957

training 5 epoch...

44.939905881881714 seconds used......

5 training finished! average train loss: 0.20226541305320306

total clf loss: 77.7257781997323 total adv loss: 86.27678497321904

evaluating...

Confusion matrix, without normalization

[[ 214 37 43]

[ 583 2113 1125]

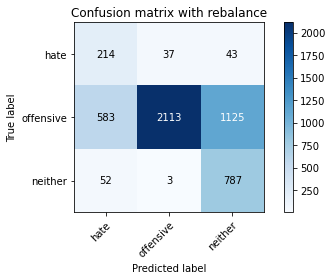
[ 52 3 787]]

Normalized confusion matrix

[[0.72789116 0.12585034 0.1462585 ]

[0.15257786 0.5529966 0.29442554]

[0.06175772 0.00356295 0.93467933]]





precision recall f1-score support

hate 0.25 0.73 0.37 294

offensive 0.98 0.55 0.71 3821

neither 0.40 0.93 0.56 842

accuracy 0.63 4957

macro avg 0.55 0.74 0.55 4957

weighted avg 0.84 0.63 0.66 4957

average eval\_loss: 0.9555576393200506, accuracy: 0.6282025418599959

evaluating adversary...

Confusion matrix, without normalization

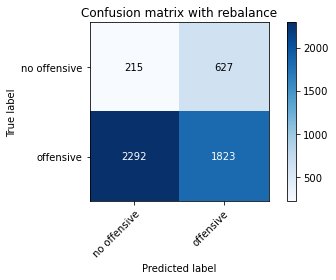
[[ 215 627]

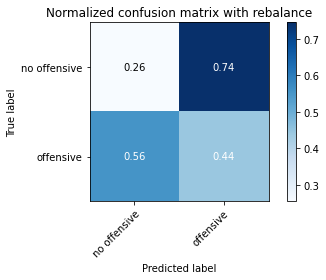
[2292 1823]]

Normalized confusion matrix

[[0.25534442 0.74465558]

[0.55698663 0.44301337]]





precision recall f1-score support

no offensive 0.09 0.26 0.13 842

offensive 0.74 0.44 0.56 4115

accuracy 0.41 4957

macro avg 0.41 0.35 0.34 4957

weighted avg 0.63 0.41 0.48 4957

training 6 epoch...

45.03216814994812 seconds used......

6 training finished! average train loss: 0.1443481395206256

total clf loss: 93.08722035586834 total adv loss: 136.80737673868134

evaluating...

Confusion matrix, without normalization

[[ 9 50 235]

[ 4 2843 974]

[ 1 103 738]]

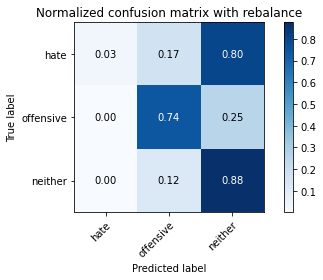
Normalized confusion matrix

[[0.03061224 0.17006803 0.79931973]

[0.00104685 0.74404606 0.25490709]

[0.00118765 0.12232779 0.87648456]]





precision recall f1-score support

hate 0.64 0.03 0.06 294

offensive 0.95 0.74 0.83 3821

neither 0.38 0.88 0.53 842

accuracy 0.72 4957

macro avg 0.66 0.55 0.47 4957

weighted avg 0.83 0.72 0.74 4957

average eval\_loss: 0.8318019290604899, accuracy: 0.7242283639297963

evaluating adversary...

Confusion matrix, without normalization

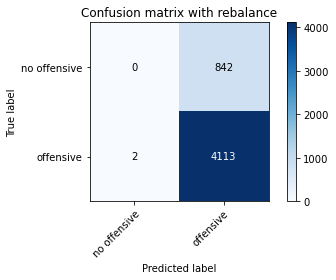
[[ 0 842]

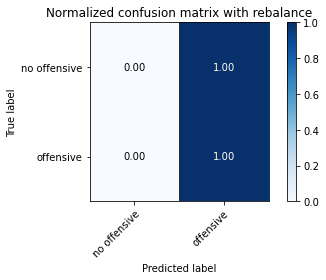
[ 2 4113]]

Normalized confusion matrix

[[0.00000000e+00 1.00000000e+00]

[4.86026731e-04 9.99513973e-01]]





precision recall f1-score support

no offensive 0.00 0.00 0.00 842

offensive 0.83 1.00 0.91 4115

accuracy 0.83 4957

macro avg 0.42 0.50 0.45 4957

weighted avg 0.69 0.83 0.75 4957

training 7 epoch...

45.07481265068054 seconds used......

7 training finished! average train loss: -0.274243939864008

total clf loss: 154.84403702616692 total adv loss: 403.47950115799904

evaluating...

Confusion matrix, without normalization

[[ 294 0 0]

[3801 20 0]

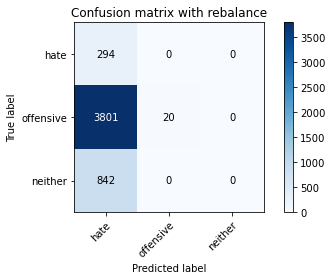
[ 842 0 0]]

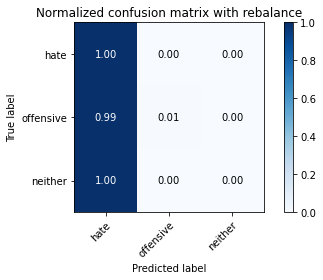
Normalized confusion matrix

[[1. 0. 0. ]

[0.99476577 0.00523423 0. ]

[1. 0. 0. ]]





/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

precision recall f1-score support

hate 0.06 1.00 0.11 294

offensive 1.00 0.01 0.01 3821

neither 0.00 0.00 0.00 842

accuracy 0.06 4957

macro avg 0.35 0.34 0.04 4957

weighted avg 0.77 0.06 0.01 4957

average eval\_loss: 1.7788607550244178, accuracy: 0.06334476497881783

evaluating adversary...

Confusion matrix, without normalization

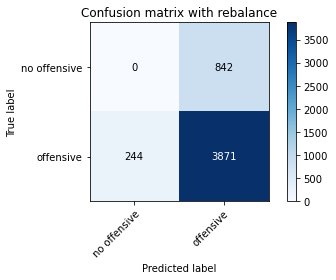
[[ 0 842]

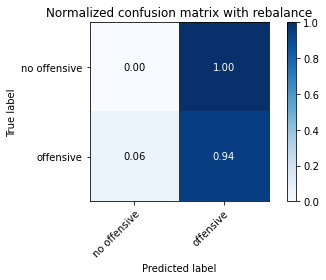
[ 244 3871]]

Normalized confusion matrix

[[0. 1. ]

[0.05929526 0.94070474]]





precision recall f1-score support

no offensive 0.00 0.00 0.00 842

offensive 0.82 0.94 0.88 4115

accuracy 0.78 4957

macro avg 0.41 0.47 0.44 4957

weighted avg 0.68 0.78 0.73 4957

training 8 epoch...

45.12533974647522 seconds used......

8 training finished! average train loss: -1.3212546221694055

total clf loss: 198.23231214284897 total adv loss: 848.3337054252625

evaluating...

Confusion matrix, without normalization

[[ 205 89 0]

[1481 2340 0]

[ 582 260 0]]

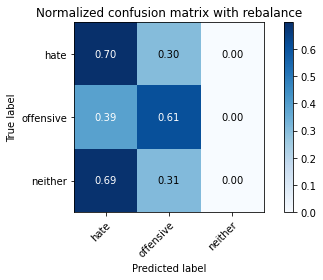
Normalized confusion matrix

[[0.69727891 0.30272109 0. ]

[0.38759487 0.61240513 0. ]

[0.6912114 0.3087886 0. ]]





/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

precision recall f1-score support

hate 0.09 0.70 0.16 294

offensive 0.87 0.61 0.72 3821

neither 0.00 0.00 0.00 842

accuracy 0.51 4957

macro avg 0.32 0.44 0.29 4957

weighted avg 0.68 0.51 0.56 4957

average eval\_loss: 0.9813240336314324, accuracy: 0.513415372200928

evaluating adversary...

Confusion matrix, without normalization

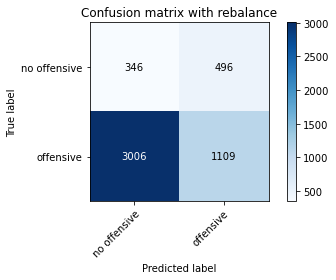
[[ 346 496]

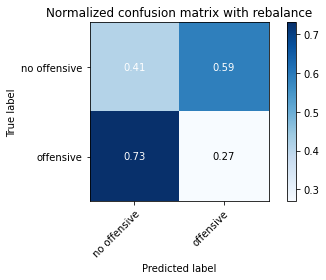
[3006 1109]]

Normalized confusion matrix

[[0.41092637 0.58907363]

[0.73049818 0.26950182]]





precision recall f1-score support

no offensive 0.10 0.41 0.16 842

offensive 0.69 0.27 0.39 4115

accuracy 0.29 4957

macro avg 0.40 0.34 0.28 4957

weighted avg 0.59 0.29 0.35 4957

training 9 epoch...

45.1252064704895 seconds used......

9 training finished! average train loss: -0.09556212899280571

total clf loss: 181.58685731887817 total adv loss: 395.85596349835396

evaluating...

Confusion matrix, without normalization

[[ 0 28 266]

[ 0 782 3039]

[ 0 2 840]]

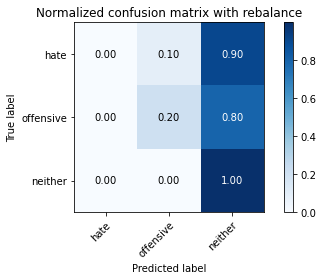
Normalized confusion matrix

[[0. 0.0952381 0.9047619 ]

[0. 0.20465847 0.79534153]

[0. 0.0023753 0.9976247 ]]





/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

precision recall f1-score support

hate 0.00 0.00 0.00 294

offensive 0.96 0.20 0.34 3821

neither 0.20 1.00 0.34 842

accuracy 0.33 4957

macro avg 0.39 0.40 0.22 4957

weighted avg 0.78 0.33 0.32 4957

average eval\_loss: 1.2286836927936924, accuracy: 0.3272140407504539

evaluating adversary...

Confusion matrix, without normalization

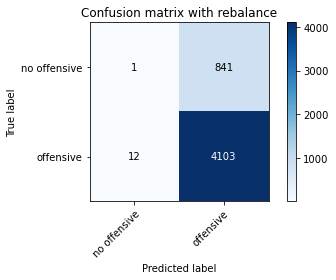
[[ 1 841]

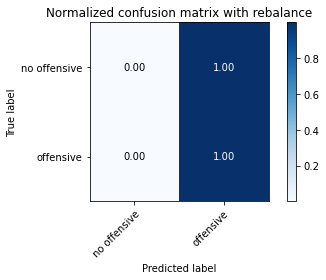
[ 12 4103]]

Normalized confusion matrix

[[0.00118765 0.99881235]

[0.00291616 0.99708384]]





precision recall f1-score support

no offensive 0.08 0.00 0.00 842

offensive 0.83 1.00 0.91 4115

accuracy 0.83 4957

macro avg 0.45 0.50 0.45 4957

weighted avg 0.70 0.83 0.75 4957

training 10 epoch...

45.057878732681274 seconds used......

10 training finished! average train loss: 0.5349902648326249

total clf loss: 168.2954169511795 total adv loss: 153.62416291236877

evaluating...

Confusion matrix, without normalization

[[ 168 55 71]

[1271 2102 448]

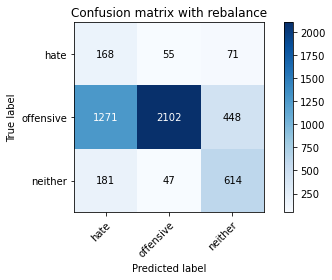
[ 181 47 614]]

Normalized confusion matrix

[[0.57142857 0.18707483 0.2414966 ]

[0.33263544 0.55011777 0.11724679]

[0.21496437 0.05581948 0.72921615]]





precision recall f1-score support

hate 0.10 0.57 0.18 294

offensive 0.95 0.55 0.70 3821

neither 0.54 0.73 0.62 842

accuracy 0.58 4957

macro avg 0.53 0.62 0.50 4957

weighted avg 0.83 0.58 0.65 4957

average eval\_loss: 0.7975963366127783, accuracy: 0.5818035101876134

evaluating adversary...

Confusion matrix, without normalization

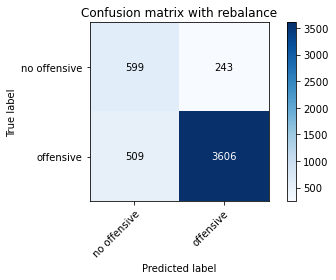
[[ 599 243]

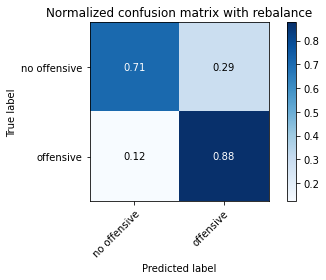
[ 509 3606]]

Normalized confusion matrix

[[0.71140143 0.28859857]

[0.1236938 0.8763062 ]]





precision recall f1-score support

no offensive 0.54 0.71 0.61 842

offensive 0.94 0.88 0.91 4115

accuracy 0.85 4957

macro avg 0.74 0.79 0.76 4957

weighted avg 0.87 0.85 0.86 4957