alpha = 0.45

training 1 epoch...

46.801477909088135 seconds used......

1 training finished! average train loss: 0.543278969955026

total clf loss: 139.79655307531357 total adv loss: 104.2130012884736

evaluating...

Confusion matrix, without normalization

[[ 203 42 49]

[ 744 2879 198]

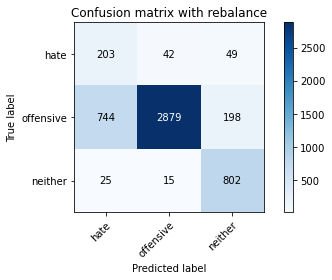
[ 25 15 802]]

Normalized confusion matrix

[[0.69047619 0.14285714 0.16666667]

[0.19471343 0.75346768 0.0518189 ]

[0.02969121 0.01781473 0.95249406]]





precision recall f1-score support

hate 0.21 0.69 0.32 294

offensive 0.98 0.75 0.85 3821

neither 0.76 0.95 0.85 842

accuracy 0.78 4957

macro avg 0.65 0.80 0.67 4957

weighted avg 0.90 0.78 0.82 4957

average eval\_loss: 0.5865956288311751, accuracy: 0.78353843050232

evaluating adversary...

Confusion matrix, without normalization

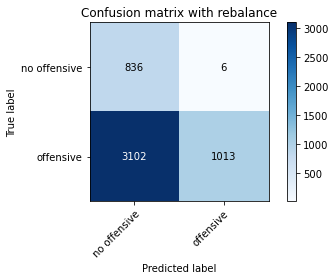
[[ 836 6]

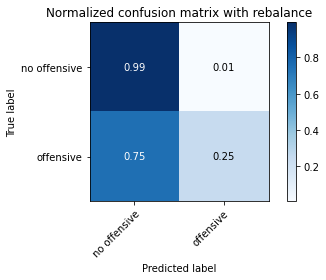
[3102 1013]]

Normalized confusion matrix

[[0.99287411 0.00712589]

[0.75382746 0.24617254]]





precision recall f1-score support

no offensive 0.21 0.99 0.35 842

offensive 0.99 0.25 0.39 4115

accuracy 0.37 4957

macro avg 0.60 0.62 0.37 4957

weighted avg 0.86 0.37 0.39 4957

training 2 epoch...

46.73608756065369 seconds used......

2 training finished! average train loss: 0.35239944532949324

total clf loss: 86.8523505628109 total adv loss: 59.093435417860746

evaluating...

Confusion matrix, without normalization

[[ 181 77 36]

[ 286 3390 145]

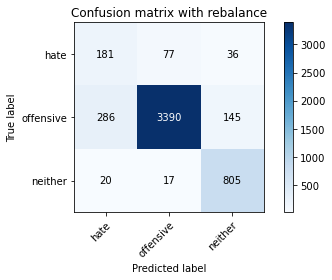
[ 20 17 805]]

Normalized confusion matrix

[[0.61564626 0.26190476 0.12244898]

[0.07484952 0.8872023 0.03794818]

[0.02375297 0.02019002 0.95605701]]





precision recall f1-score support

hate 0.37 0.62 0.46 294

offensive 0.97 0.89 0.93 3821

neither 0.82 0.96 0.88 842

accuracy 0.88 4957

macro avg 0.72 0.82 0.76 4957

weighted avg 0.91 0.88 0.89 4957

average eval\_loss: 0.3760022836166524, accuracy: 0.8827920112971556

evaluating adversary...

Confusion matrix, without normalization

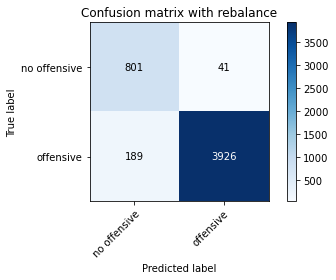
[[ 801 41]

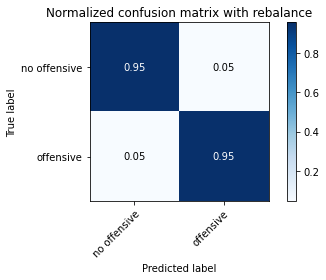
[ 189 3926]]

Normalized confusion matrix

[[0.95130641 0.04869359]

[0.04592953 0.95407047]]





precision recall f1-score support

no offensive 0.81 0.95 0.87 842

offensive 0.99 0.95 0.97 4115

accuracy 0.95 4957

macro avg 0.90 0.95 0.92 4957

weighted avg 0.96 0.95 0.96 4957

training 3 epoch...

46.77133536338806 seconds used......

3 training finished! average train loss: 0.3116796236289175

total clf loss: 66.4592345803976 total adv loss: 29.248932091519237

evaluating...

Confusion matrix, without normalization

[[ 255 28 11]

[1287 2439 95]

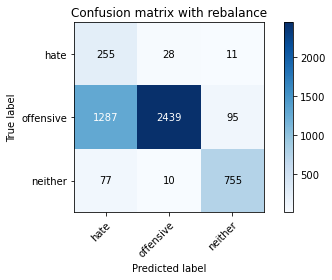
[ 77 10 755]]

Normalized confusion matrix

[[0.86734694 0.0952381 0.03741497]

[0.33682282 0.63831458 0.0248626 ]

[0.09144893 0.01187648 0.89667458]]





precision recall f1-score support

hate 0.16 0.87 0.27 294

offensive 0.98 0.64 0.77 3821

neither 0.88 0.90 0.89 842

accuracy 0.70 4957

macro avg 0.67 0.80 0.64 4957

weighted avg 0.92 0.70 0.76 4957

average eval\_loss: 0.7366269938046893, accuracy: 0.6957837401654227

evaluating adversary...

Confusion matrix, without normalization

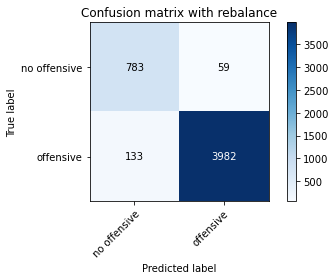
[[ 783 59]

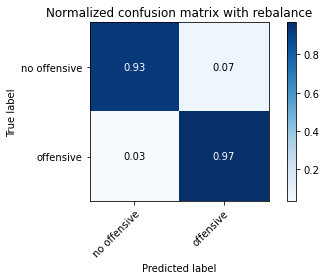
[ 133 3982]]

Normalized confusion matrix

[[0.92992874 0.07007126]

[0.03232078 0.96767922]]





precision recall f1-score support

no offensive 0.85 0.93 0.89 842

offensive 0.99 0.97 0.98 4115

accuracy 0.96 4957

macro avg 0.92 0.95 0.93 4957

weighted avg 0.96 0.96 0.96 4957

training 4 epoch...

46.88863754272461 seconds used......

4 training finished! average train loss: 0.22502259029979593

total clf loss: 47.979114059358835 total adv loss: 21.111669609323144

evaluating...

Confusion matrix, without normalization

[[ 241 43 10]

[ 802 2923 96]

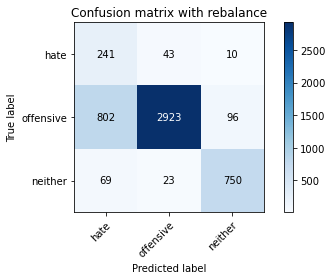
[ 69 23 750]]

Normalized confusion matrix

[[0.81972789 0.1462585 0.03401361]

[0.2098927 0.76498299 0.02512431]

[0.08194774 0.02731591 0.89073634]]





precision recall f1-score support

hate 0.22 0.82 0.34 294

offensive 0.98 0.76 0.86 3821

neither 0.88 0.89 0.88 842

accuracy 0.79 4957

macro avg 0.69 0.83 0.69 4957

weighted avg 0.92 0.79 0.83 4957

average eval\_loss: 0.5091838793258273, accuracy: 0.7895904781117612

evaluating adversary...

Confusion matrix, without normalization

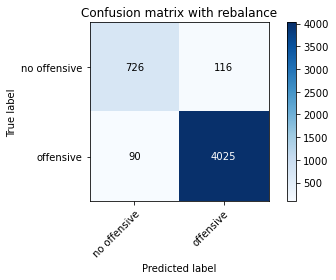
[[ 726 116]

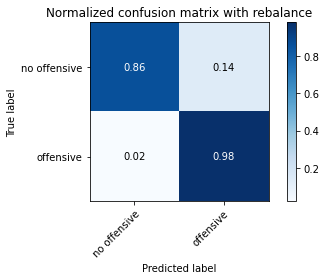
[ 90 4025]]

Normalized confusion matrix

[[0.86223278 0.13776722]

[0.0218712 0.9781288 ]]





precision recall f1-score support

no offensive 0.89 0.86 0.88 842

offensive 0.97 0.98 0.98 4115

accuracy 0.96 4957

macro avg 0.93 0.92 0.93 4957

weighted avg 0.96 0.96 0.96 4957

training 5 epoch...

46.91153812408447 seconds used......

5 training finished! average train loss: 0.14187966269163185

total clf loss: 30.74671147763729 total adv loss: 14.411753965076059

evaluating...

Confusion matrix, without normalization

[[ 214 72 8]

[ 559 3185 77]

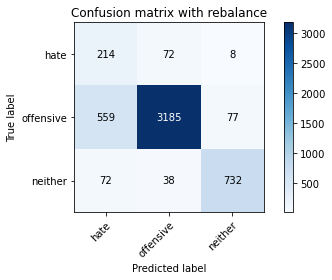
[ 72 38 732]]

Normalized confusion matrix

[[0.72789116 0.24489796 0.02721088]

[0.14629678 0.83355143 0.02015179]

[0.08551069 0.04513064 0.86935867]]





precision recall f1-score support

hate 0.25 0.73 0.38 294

offensive 0.97 0.83 0.90 3821

neither 0.90 0.87 0.88 842

accuracy 0.83 4957

macro avg 0.71 0.81 0.72 4957

weighted avg 0.91 0.83 0.86 4957

average eval\_loss: 0.5545673720809, accuracy: 0.8333669558200525

evaluating adversary...

Confusion matrix, without normalization

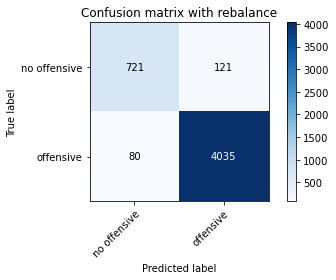
[[ 721 121]

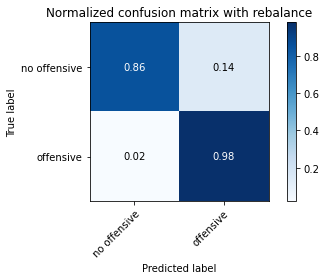
[ 80 4035]]

Normalized confusion matrix

[[0.85629454 0.14370546]

[0.01944107 0.98055893]]





precision recall f1-score support

no offensive 0.90 0.86 0.88 842

offensive 0.97 0.98 0.98 4115

accuracy 0.96 4957

macro avg 0.94 0.92 0.93 4957

weighted avg 0.96 0.96 0.96 4957

training 6 epoch...

46.93400025367737 seconds used......

6 training finished! average train loss: 0.08112157693906137

total clf loss: 18.136511572636664 total adv loss: 9.477159939240664

evaluating...

Confusion matrix, without normalization

[[ 241 38 15]

[ 992 2710 119]

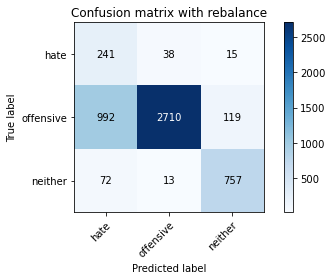
[ 72 13 757]]

Normalized confusion matrix

[[0.81972789 0.1292517 0.05102041]

[0.2596179 0.70923842 0.03114368]

[0.08551069 0.01543943 0.89904988]]





precision recall f1-score support

hate 0.18 0.82 0.30 294

offensive 0.98 0.71 0.82 3821

neither 0.85 0.90 0.87 842

accuracy 0.75 4957

macro avg 0.67 0.81 0.67 4957

weighted avg 0.91 0.75 0.80 4957

average eval\_loss: 1.0146696808252242, accuracy: 0.7480330845269316

evaluating adversary...

Confusion matrix, without normalization

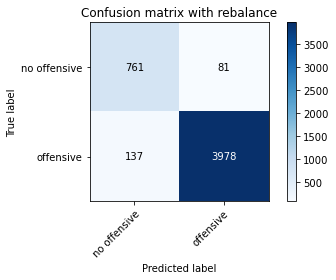
[[ 761 81]

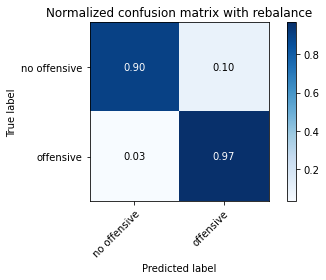
[ 137 3978]]

Normalized confusion matrix

[[0.90380048 0.09619952]

[0.03329283 0.96670717]]





precision recall f1-score support

no offensive 0.85 0.90 0.87 842

offensive 0.98 0.97 0.97 4115

accuracy 0.96 4957

macro avg 0.91 0.94 0.92 4957

weighted avg 0.96 0.96 0.96 4957

training 7 epoch...

46.692848920822144 seconds used......

7 training finished! average train loss: 0.04476366612559173

total clf loss: 9.592891945503652 total adv loss: 4.307344645058038

evaluating...

Confusion matrix, without normalization

[[ 234 49 11]

[ 913 2813 95]

[ 81 22 739]]

Normalized confusion matrix

[[0.79591837 0.16666667 0.03741497]

[0.23894269 0.73619471 0.0248626 ]

[0.09619952 0.02612827 0.87767221]]





precision recall f1-score support

hate 0.19 0.80 0.31 294

offensive 0.98 0.74 0.84 3821

neither 0.87 0.88 0.88 842

accuracy 0.76 4957

macro avg 0.68 0.80 0.67 4957

weighted avg 0.91 0.76 0.81 4957

average eval\_loss: 1.0925151663118853, accuracy: 0.7637684083114787

evaluating adversary...

Confusion matrix, without normalization

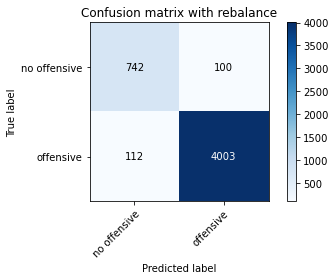
[[ 742 100]

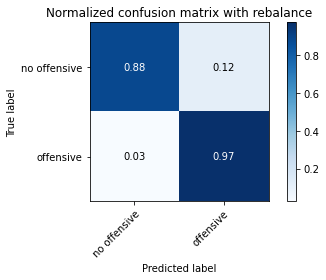
[ 112 4003]]

Normalized confusion matrix

[[0.88123515 0.11876485]

[0.0272175 0.9727825 ]]





precision recall f1-score support

no offensive 0.87 0.88 0.88 842

offensive 0.98 0.97 0.97 4115

accuracy 0.96 4957

macro avg 0.92 0.93 0.92 4957

weighted avg 0.96 0.96 0.96 4957

training 8 epoch...

46.54600191116333 seconds used......

8 training finished! average train loss: 0.03662753840145922

total clf loss: 7.978485728497617 total adv loss: 3.811503673943662

evaluating...

Confusion matrix, without normalization

[[ 211 67 16]

[ 692 3019 110]

[ 47 34 761]]

Normalized confusion matrix

[[0.71768707 0.22789116 0.05442177]

[0.18110442 0.7901073 0.02878828]

[0.05581948 0.04038005 0.90380048]]





precision recall f1-score support

hate 0.22 0.72 0.34 294

offensive 0.97 0.79 0.87 3821

neither 0.86 0.90 0.88 842

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: 0.8263468290106468, accuracy: 0.8051240669759936

evaluating adversary...

Confusion matrix, without normalization

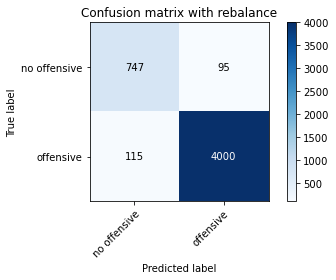
[[ 747 95]

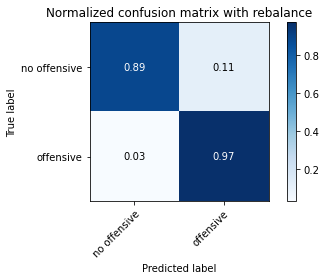
[ 115 4000]]

Normalized confusion matrix

[[0.8871734 0.1128266 ]

[0.02794654 0.97205346]]





precision recall f1-score support

no offensive 0.87 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 9 epoch...

46.736021757125854 seconds used......

9 training finished! average train loss: 0.01683245793091166

total clf loss: 3.7430178048089147 total adv loss: 1.9214833630248904

evaluating...

Confusion matrix, without normalization

[[ 229 54 11]

[ 772 2953 96]

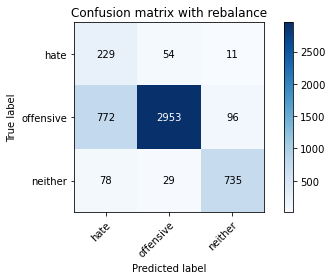
[ 78 29 735]]

Normalized confusion matrix

[[0.77891156 0.18367347 0.03741497]

[0.20204135 0.77283434 0.02512431]

[0.09263658 0.03444181 0.87292162]]





precision recall f1-score support

hate 0.21 0.78 0.33 294

offensive 0.97 0.77 0.86 3821

neither 0.87 0.87 0.87 842

accuracy 0.79 4957

macro avg 0.69 0.81 0.69 4957

weighted avg 0.91 0.79 0.83 4957

average eval\_loss: 1.0838366431905138, accuracy: 0.7901956828727053

evaluating adversary...

Confusion matrix, without normalization

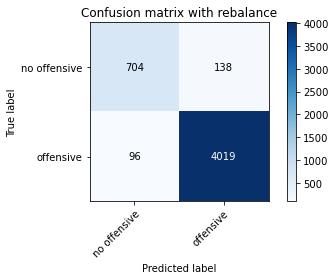
[[ 704 138]

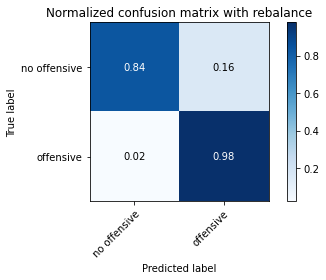
[ 96 4019]]

Normalized confusion matrix

[[0.83610451 0.16389549]

[0.02332928 0.97667072]]





precision recall f1-score support

no offensive 0.88 0.84 0.86 842

offensive 0.97 0.98 0.97 4115

accuracy 0.95 4957

macro avg 0.92 0.91 0.91 4957

weighted avg 0.95 0.95 0.95 4957

training 10 epoch...

46.64911866188049 seconds used......

10 training finished! average train loss: 0.013807384382691017

total clf loss: 3.109728091978468 total adv loss: 1.6637008597326712

evaluating...

Confusion matrix, without normalization

[[ 201 73 20]

[ 652 3047 122]

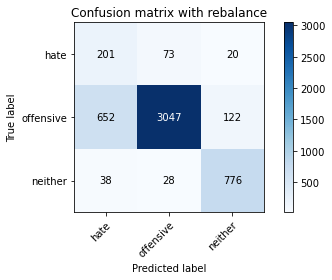
[ 38 28 776]]

Normalized confusion matrix

[[0.68367347 0.24829932 0.06802721]

[0.17063596 0.79743523 0.03192881]

[0.04513064 0.03325416 0.9216152 ]]





precision recall f1-score support

hate 0.23 0.68 0.34 294

offensive 0.97 0.80 0.87 3821

neither 0.85 0.92 0.88 842

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: 1.0642632683147601, accuracy: 0.8117813193463789

evaluating adversary...

Confusion matrix, without normalization

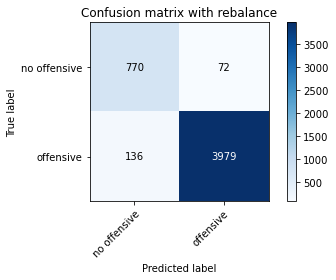
[[ 770 72]

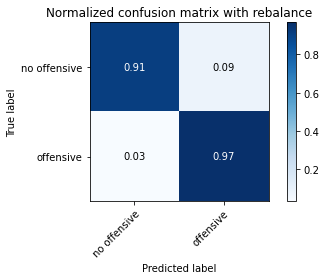
[ 136 3979]]

Normalized confusion matrix

[[0.91448931 0.08551069]

[0.03304982 0.96695018]]





precision recall f1-score support

no offensive 0.85 0.91 0.88 842

offensive 0.98 0.97 0.97 4115

accuracy 0.96 4957

macro avg 0.92 0.94 0.93 4957

weighted avg 0.96 0.96 0.96 4957

2 run

training 1 epoch...

46.64477825164795 seconds used......

1 training finished! average train loss: 0.48069811625438824

total clf loss: 134.22101190686226 total adv loss: 115.60363358259201

evaluating...

Confusion matrix, without normalization

[[ 172 70 52]

[ 380 3198 243]

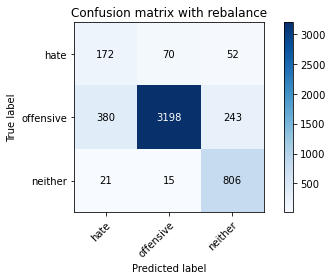
[ 21 15 806]]

Normalized confusion matrix

[[0.58503401 0.23809524 0.17687075]

[0.09945041 0.83695368 0.06359592]

[0.02494062 0.01781473 0.95724466]]





precision recall f1-score support

hate 0.30 0.59 0.40 294

offensive 0.97 0.84 0.90 3821

neither 0.73 0.96 0.83 842

accuracy 0.84 4957

macro avg 0.67 0.79 0.71 4957

weighted avg 0.89 0.84 0.86 4957

average eval\_loss: 0.4428900758765878, accuracy: 0.8424450272342142

evaluating adversary...

Confusion matrix, without normalization

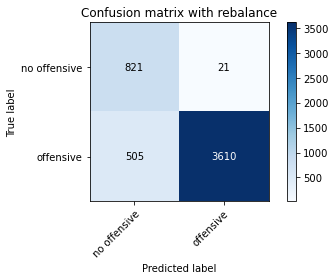
[[ 821 21]

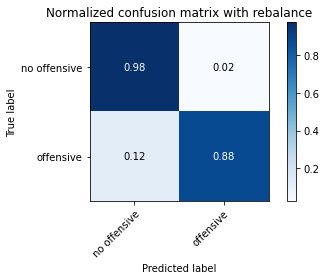
[ 505 3610]]

Normalized confusion matrix

[[0.97505938 0.02494062]

[0.12272175 0.87727825]]





precision recall f1-score support

no offensive 0.62 0.98 0.76 842

offensive 0.99 0.88 0.93 4115

accuracy 0.89 4957

macro avg 0.81 0.93 0.84 4957

weighted avg 0.93 0.89 0.90 4957

training 2 epoch...

46.592897176742554 seconds used......

2 training finished! average train loss: 0.36491937059582324

total clf loss: 80.69224298000336 total adv loss: 40.64673610217869

evaluating...

Confusion matrix, without normalization

[[ 201 83 10]

[ 310 3412 99]

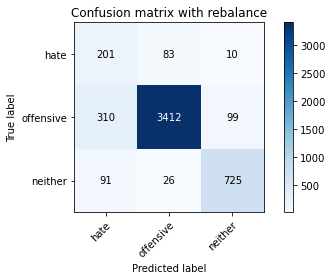
[ 91 26 725]]

Normalized confusion matrix

[[0.68367347 0.28231293 0.03401361]

[0.08113059 0.89295996 0.02590945]

[0.10807601 0.03087886 0.86104513]]





precision recall f1-score support

hate 0.33 0.68 0.45 294

offensive 0.97 0.89 0.93 3821

neither 0.87 0.86 0.87 842

accuracy 0.88 4957

macro avg 0.72 0.81 0.75 4957

weighted avg 0.91 0.88 0.89 4957

average eval\_loss: 0.34878553143672403, accuracy: 0.8751260843251967

evaluating adversary...

Confusion matrix, without normalization

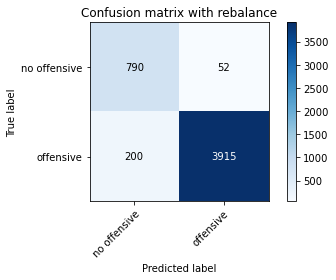
[[ 790 52]

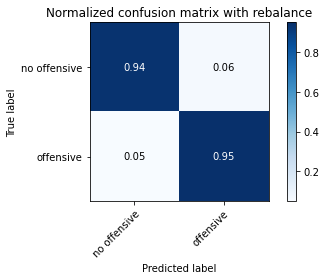
[ 200 3915]]

Normalized confusion matrix

[[0.93824228 0.06175772]

[0.04860267 0.95139733]]





precision recall f1-score support

no offensive 0.80 0.94 0.86 842

offensive 0.99 0.95 0.97 4115

accuracy 0.95 4957

macro avg 0.89 0.94 0.92 4957

weighted avg 0.95 0.95 0.95 4957

training 3 epoch...

46.59215474128723 seconds used......

3 training finished! average train loss: 0.2977282117986888

total clf loss: 66.11057634651661 total adv loss: 33.77567263599485

evaluating...

Confusion matrix, without normalization

[[ 232 49 13]

[ 791 2911 119]

[ 87 12 743]]

Normalized confusion matrix

[[0.78911565 0.16666667 0.04421769]

[0.20701387 0.76184245 0.03114368]

[0.10332542 0.01425178 0.8824228 ]]





precision recall f1-score support

hate 0.21 0.79 0.33 294

offensive 0.98 0.76 0.86 3821

neither 0.85 0.88 0.87 842

accuracy 0.78 4957

macro avg 0.68 0.81 0.68 4957

weighted avg 0.91 0.78 0.83 4957

average eval\_loss: 0.5602074053378836, accuracy: 0.7839419003429494

evaluating adversary...

Confusion matrix, without normalization

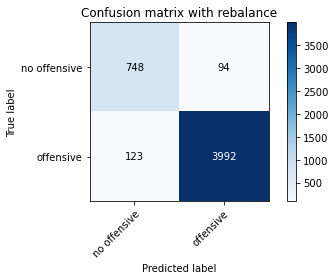
[[ 748 94]

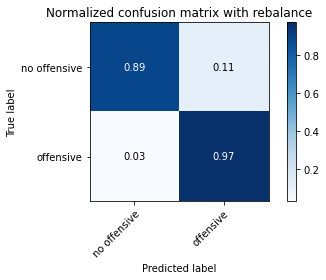
[ 123 3992]]

Normalized confusion matrix

[[0.88836105 0.11163895]

[0.02989064 0.97010936]]





precision recall f1-score support

no offensive 0.86 0.89 0.87 842

offensive 0.98 0.97 0.97 4115

accuracy 0.96 4957

macro avg 0.92 0.93 0.92 4957

weighted avg 0.96 0.96 0.96 4957

training 4 epoch...

46.61317849159241 seconds used......

4 training finished! average train loss: 0.22136564101827771

total clf loss: 46.35134110599756 total adv loss: 18.88403701223433

evaluating...

Confusion matrix, without normalization

[[ 231 49 14]

[ 828 2873 120]

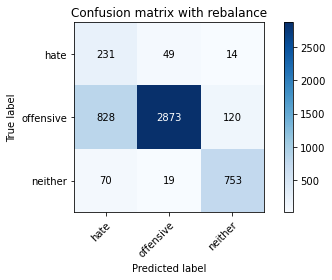
[ 70 19 753]]

Normalized confusion matrix

[[0.78571429 0.16666667 0.04761905]

[0.2166972 0.75189741 0.03140539]

[0.08313539 0.02256532 0.89429929]]





precision recall f1-score support

hate 0.20 0.79 0.32 294

offensive 0.98 0.75 0.85 3821

neither 0.85 0.89 0.87 842

accuracy 0.78 4957

macro avg 0.68 0.81 0.68 4957

weighted avg 0.91 0.78 0.82 4957

average eval\_loss: 0.5610411205599385, accuracy: 0.7780915876538229

evaluating adversary...

Confusion matrix, without normalization

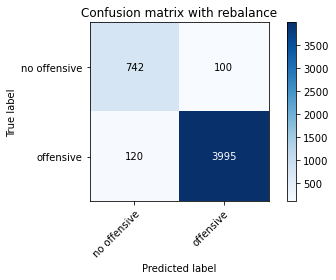
[[ 742 100]

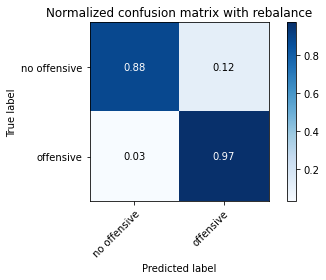
[ 120 3995]]

Normalized confusion matrix

[[0.88123515 0.11876485]

[0.0291616 0.9708384 ]]





precision recall f1-score support

no offensive 0.86 0.88 0.87 842

offensive 0.98 0.97 0.97 4115

accuracy 0.96 4957

macro avg 0.92 0.93 0.92 4957

weighted avg 0.96 0.96 0.96 4957

training 5 epoch...

46.691121339797974 seconds used......

5 training finished! average train loss: 0.13807975184390245

total clf loss: 29.361193856224418 total adv loss: 12.776792009361088

evaluating...

Confusion matrix, without normalization

[[ 222 57 15]

[ 718 2990 113]

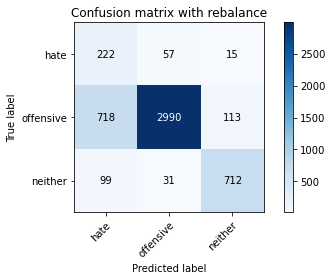
[ 99 31 712]]

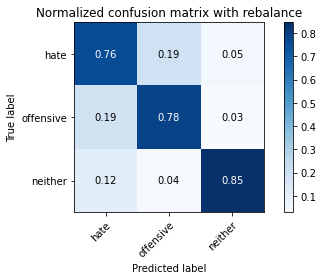
Normalized confusion matrix

[[0.75510204 0.19387755 0.05102041]

[0.18790892 0.78251767 0.02957341]

[0.1175772 0.0368171 0.8456057 ]]





precision recall f1-score support

hate 0.21 0.76 0.33 294

offensive 0.97 0.78 0.87 3821

neither 0.85 0.85 0.85 842

accuracy 0.79 4957

macro avg 0.68 0.79 0.68 4957

weighted avg 0.91 0.79 0.83 4957

average eval\_loss: 0.6411037143777996, accuracy: 0.7916078273149082

evaluating adversary...

Confusion matrix, without normalization

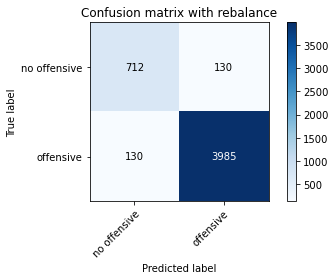
[[ 712 130]

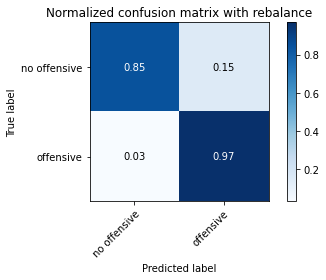
[ 130 3985]]

Normalized confusion matrix

[[0.8456057 0.1543943 ]

[0.03159174 0.96840826]]





precision recall f1-score support

no offensive 0.85 0.85 0.85 842

offensive 0.97 0.97 0.97 4115

accuracy 0.95 4957

macro avg 0.91 0.91 0.91 4957

weighted avg 0.95 0.95 0.95 4957

training 6 epoch...

46.52571487426758 seconds used......

6 training finished! average train loss: 0.07775193052529766

total clf loss: 17.645402452908456 total adv loss: 9.666272032947745

evaluating...

Confusion matrix, without normalization

[[ 221 53 20]

[ 809 2868 144]

[ 44 33 765]]

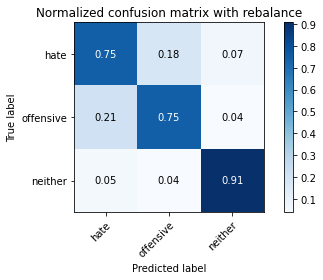
Normalized confusion matrix

[[0.75170068 0.18027211 0.06802721]

[0.21172468 0.75058885 0.03768647]

[0.05225653 0.0391924 0.90855107]]





precision recall f1-score support

hate 0.21 0.75 0.32 294

offensive 0.97 0.75 0.85 3821

neither 0.82 0.91 0.86 842

accuracy 0.78 4957

macro avg 0.67 0.80 0.68 4957

weighted avg 0.90 0.78 0.82 4957

average eval\_loss: 0.8252451859917792, accuracy: 0.7774863828928787

evaluating adversary...

Confusion matrix, without normalization

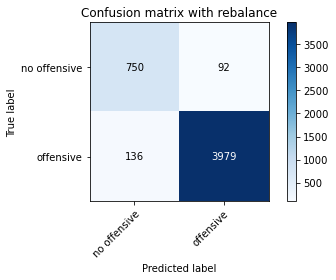
[[ 750 92]

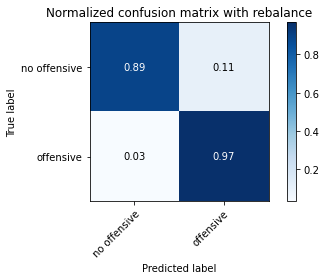
[ 136 3979]]

Normalized confusion matrix

[[0.89073634 0.10926366]

[0.03304982 0.96695018]]





precision recall f1-score support

no offensive 0.85 0.89 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

training 7 epoch...

46.41549754142761 seconds used......

7 training finished! average train loss: 0.04365427666398212

total clf loss: 9.725480638910085 total adv loss: 5.023554160812637

evaluating...

Confusion matrix, without normalization

[[ 207 69 18]

[ 628 3045 148]

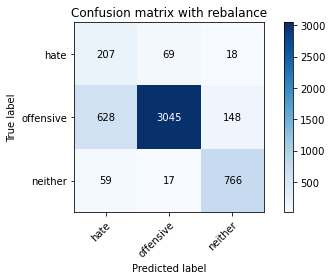
[ 59 17 766]]

Normalized confusion matrix

[[0.70408163 0.23469388 0.06122449]

[0.16435488 0.7969118 0.03873332]

[0.07007126 0.02019002 0.90973872]]





precision recall f1-score support

hate 0.23 0.70 0.35 294

offensive 0.97 0.80 0.88 3821

neither 0.82 0.91 0.86 842

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: 0.7715315052417047, accuracy: 0.8105709098244906

evaluating adversary...

Confusion matrix, without normalization

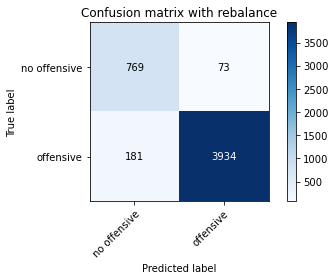
[[ 769 73]

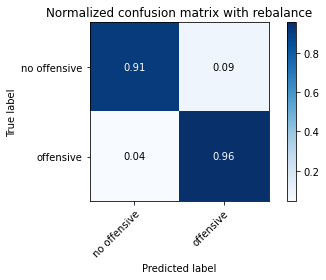
[ 181 3934]]

Normalized confusion matrix

[[0.91330166 0.08669834]

[0.04398542 0.95601458]]





precision recall f1-score support

no offensive 0.81 0.91 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 8 epoch...

46.64174938201904 seconds used......

8 training finished! average train loss: 0.026213861172692032

total clf loss: 5.670527271460742 total adv loss: 2.6399045862053754

evaluating...

Confusion matrix, without normalization

[[ 202 72 20]

[ 584 3105 132]

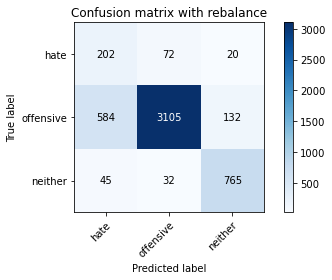
[ 45 32 765]]

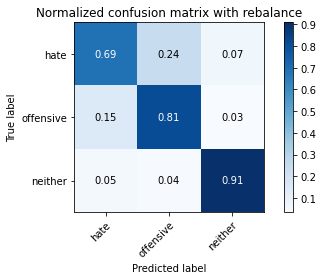
Normalized confusion matrix

[[0.68707483 0.24489796 0.06802721]

[0.15283957 0.8126145 0.03454593]

[0.05344418 0.03800475 0.90855107]]





precision recall f1-score support

hate 0.24 0.69 0.36 294

offensive 0.97 0.81 0.88 3821

neither 0.83 0.91 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: 0.85780503705281, accuracy: 0.8214645955214848

evaluating adversary...

Confusion matrix, without normalization

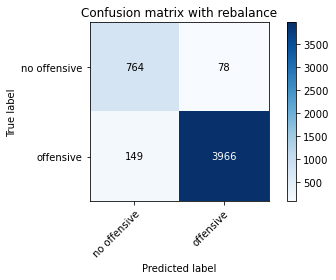
[[ 764 78]

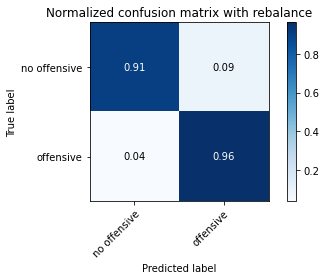
[ 149 3966]]

Normalized confusion matrix

[[0.90736342 0.09263658]

[0.03620899 0.96379101]]





precision recall f1-score support

no offensive 0.84 0.91 0.87 842

offensive 0.98 0.96 0.97 4115

accuracy 0.95 4957

macro avg 0.91 0.94 0.92 4957

weighted avg 0.96 0.95 0.95 4957

training 9 epoch...

46.542394161224365 seconds used......

9 training finished! average train loss: 0.023623464590929753

total clf loss: 4.976150431786664 total adv loss: 2.0811955843819305

evaluating...

Confusion matrix, without normalization

[[ 226 51 17]

[ 906 2778 137]

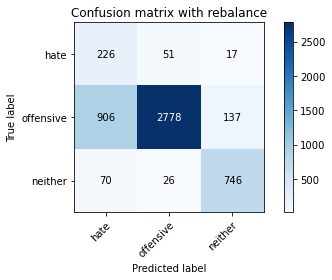
[ 70 26 746]]

Normalized confusion matrix

[[0.76870748 0.17346939 0.05782313]

[0.2371107 0.72703481 0.03585449]

[0.08313539 0.03087886 0.88598575]]





precision recall f1-score support

hate 0.19 0.77 0.30 294

offensive 0.97 0.73 0.83 3821

neither 0.83 0.89 0.86 842

accuracy 0.76 4957

macro avg 0.66 0.79 0.66 4957

weighted avg 0.90 0.76 0.80 4957

average eval\_loss: 1.261145712197533, accuracy: 0.7565059511801493

evaluating adversary...

Confusion matrix, without normalization

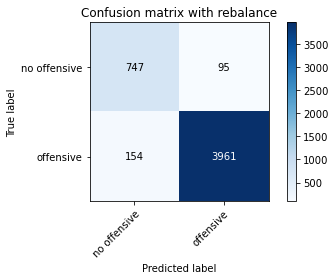
[[ 747 95]

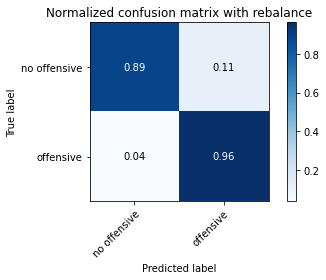
[ 154 3961]]

Normalized confusion matrix

[[0.8871734 0.1128266 ]

[0.03742406 0.96257594]]





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 0.92 0.91 4957

weighted avg 0.95 0.95 0.95 4957

training 10 epoch...

46.97283387184143 seconds used......

10 training finished! average train loss: 0.015869455121757806

total clf loss: 3.261357896059053 total adv loss: 1.2170690493021539

evaluating...

Confusion matrix, without normalization

[[ 209 66 19]

[ 579 3107 135]

[ 59 25 758]]

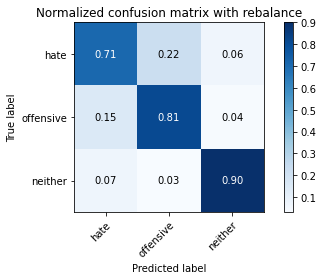
Normalized confusion matrix

[[0.71088435 0.2244898 0.06462585]

[0.15153101 0.81313792 0.03533107]

[0.07007126 0.02969121 0.90023753]]





precision recall f1-score support

hate 0.25 0.71 0.37 294

offensive 0.97 0.81 0.89 3821

neither 0.83 0.90 0.86 842

accuracy 0.82 4957

macro avg 0.68 0.81 0.71 4957

weighted avg 0.90 0.82 0.85 4957

average eval\_loss: 1.0616533065753644, accuracy: 0.8218680653621142

evaluating adversary...

Confusion matrix, without normalization

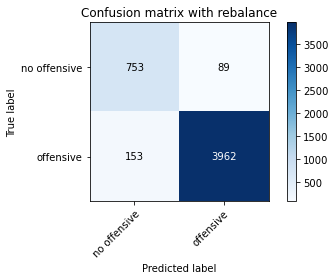
[[ 753 89]

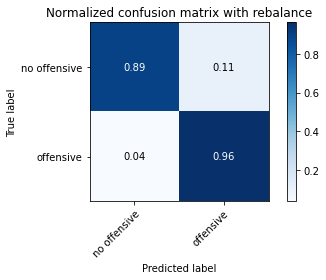
[ 153 3962]]

Normalized confusion matrix

[[0.89429929 0.10570071]

[0.03718104 0.96281896]]





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 0.92 4957

weighted avg 0.95 0.95 0.95 4957

3 run

training 1 epoch...

46.664013385772705 seconds used......

1 training finished! average train loss: 0.5345707351004171

total clf loss: 137.4601612985134 total adv loss: 102.33014786243439

evaluating...

Confusion matrix, without normalization

[[ 210 42 42]

[ 823 2854 144]

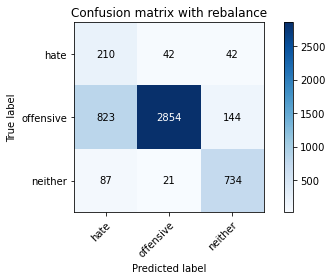
[ 87 21 734]]

Normalized confusion matrix

[[0.71428571 0.14285714 0.14285714]

[0.21538864 0.74692489 0.03768647]

[0.10332542 0.02494062 0.87173397]]





precision recall f1-score support

hate 0.19 0.71 0.30 294

offensive 0.98 0.75 0.85 3821

neither 0.80 0.87 0.83 842

accuracy 0.77 4957

macro avg 0.65 0.78 0.66 4957

weighted avg 0.90 0.77 0.81 4957

average eval\_loss: 0.6456995690541882, accuracy: 0.7661892273552552

evaluating adversary...

Confusion matrix, without normalization

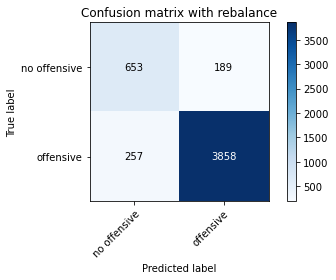
[[ 653 189]

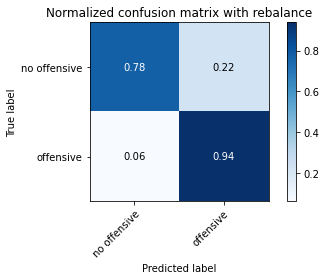
[ 257 3858]]

Normalized confusion matrix

[[0.77553444 0.22446556]

[0.06245443 0.93754557]]





precision recall f1-score support

no offensive 0.72 0.78 0.75 842

offensive 0.95 0.94 0.95 4115

accuracy 0.91 4957

macro avg 0.84 0.86 0.85 4957

weighted avg 0.91 0.91 0.91 4957

training 2 epoch...

46.60363531112671 seconds used......

2 training finished! average train loss: 0.3748715617503339

total clf loss: 84.91783273220062 total adv loss: 46.25510246120393

evaluating...

Confusion matrix, without normalization

[[ 236 45 13]

[ 820 2912 89]

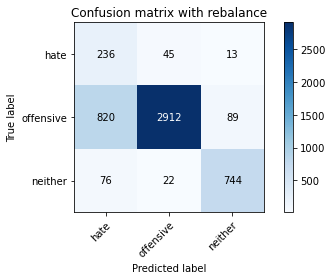
[ 76 22 744]]

Normalized confusion matrix

[[0.80272109 0.15306122 0.04421769]

[0.21460351 0.76210416 0.02329233]

[0.09026128 0.02612827 0.88361045]]





precision recall f1-score support

hate 0.21 0.80 0.33 294

offensive 0.98 0.76 0.86 3821

neither 0.88 0.88 0.88 842

accuracy 0.79 4957

macro avg 0.69 0.82 0.69 4957

weighted avg 0.92 0.79 0.83 4957

average eval\_loss: 0.580899213290503, accuracy: 0.7851523098648376

evaluating adversary...

Confusion matrix, without normalization

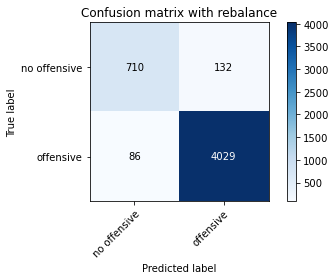
[[ 710 132]

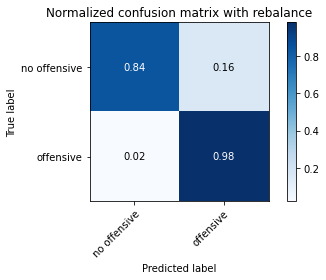
[ 86 4029]]

Normalized confusion matrix

[[0.8432304 0.1567696 ]

[0.02089915 0.97910085]]





precision recall f1-score support

no offensive 0.89 0.84 0.87 842

offensive 0.97 0.98 0.97 4115

accuracy 0.96 4957

macro avg 0.93 0.91 0.92 4957

weighted avg 0.96 0.96 0.96 4957

training 3 epoch...

46.58311653137207 seconds used......

3 training finished! average train loss: 0.2853526702226951

total clf loss: 60.80181911587715 total adv loss: 26.681139819789678

evaluating...

Confusion matrix, without normalization

[[ 201 64 29]

[ 515 3156 150]

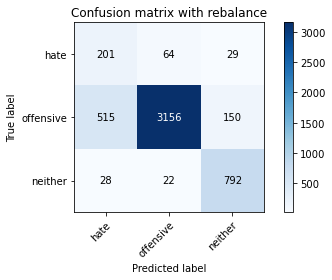
[ 28 22 792]]

Normalized confusion matrix

[[0.68367347 0.21768707 0.09863946]

[0.13478147 0.82596179 0.03925674]

[0.03325416 0.02612827 0.94061758]]





precision recall f1-score support

hate 0.27 0.68 0.39 294

offensive 0.97 0.83 0.89 3821

neither 0.82 0.94 0.87 842

accuracy 0.84 4957

macro avg 0.69 0.82 0.72 4957

weighted avg 0.90 0.84 0.86 4957

average eval\_loss: 0.474482073996877, accuracy: 0.8369981843857172

evaluating adversary...

Confusion matrix, without normalization

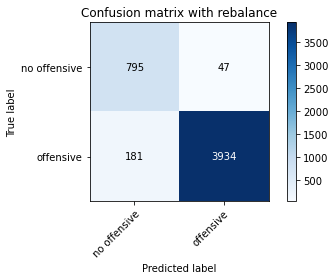
[[ 795 47]

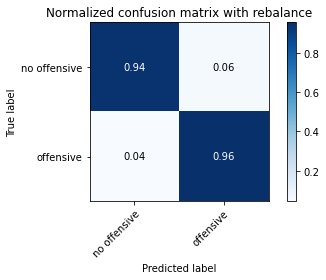
[ 181 3934]]

Normalized confusion matrix

[[0.94418052 0.05581948]

[0.04398542 0.95601458]]





precision recall f1-score support

no offensive 0.81 0.94 0.87 842

offensive 0.99 0.96 0.97 4115

accuracy 0.95 4957

macro avg 0.90 0.95 0.92 4957

weighted avg 0.96 0.95 0.96 4957

training 4 epoch...

46.78208589553833 seconds used......

4 training finished! average train loss: 0.1884458561520479

total clf loss: 40.05144656449556 total adv loss: 17.393789728172123

evaluating...

Confusion matrix, without normalization

[[ 208 55 31]

[ 685 2966 170]

[ 31 14 797]]

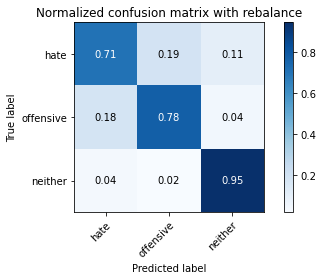
Normalized confusion matrix

[[0.70748299 0.18707483 0.10544218]

[0.17927244 0.77623659 0.04449097]

[0.0368171 0.01662708 0.94655582]]





precision recall f1-score support

hate 0.23 0.71 0.34 294

offensive 0.98 0.78 0.87 3821

neither 0.80 0.95 0.87 842

accuracy 0.80 4957

macro avg 0.67 0.81 0.69 4957

weighted avg 0.90 0.80 0.83 4957

average eval\_loss: 0.5963108355194451, accuracy: 0.8010893685696994

evaluating adversary...

Confusion matrix, without normalization

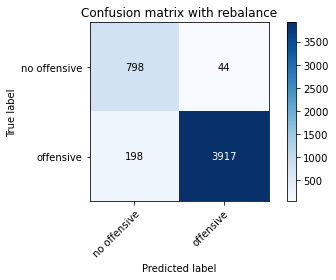
[[ 798 44]

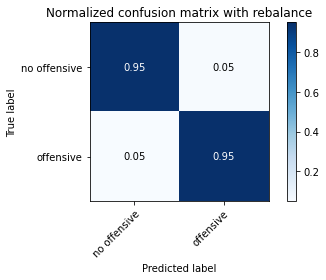
[ 198 3917]]

Normalized confusion matrix

[[0.94774347 0.05225653]

[0.04811665 0.95188335]]





precision recall f1-score support

no offensive 0.80 0.95 0.87 842

offensive 0.99 0.95 0.97 4115

accuracy 0.95 4957

macro avg 0.90 0.95 0.92 4957

weighted avg 0.96 0.95 0.95 4957

training 5 epoch...

46.47878813743591 seconds used......

5 training finished! average train loss: 0.11399999095333947

total clf loss: 25.406304840929806 total adv loss: 13.138458999572322

evaluating...

Confusion matrix, without normalization

[[ 218 42 34]

[ 890 2726 205]

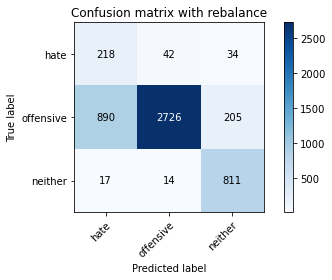
[ 17 14 811]]

Normalized confusion matrix

[[0.7414966 0.14285714 0.11564626]

[0.23292332 0.7134258 0.05365088]

[0.02019002 0.01662708 0.9631829 ]]





precision recall f1-score support

hate 0.19 0.74 0.31 294

offensive 0.98 0.71 0.83 3821

neither 0.77 0.96 0.86 842

accuracy 0.76 4957

macro avg 0.65 0.81 0.66 4957

weighted avg 0.90 0.76 0.80 4957

average eval\_loss: 0.8996455105877812, accuracy: 0.7575146257817228

evaluating adversary...

Confusion matrix, without normalization

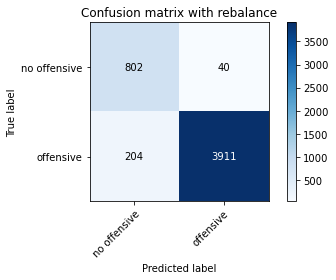
[[ 802 40]

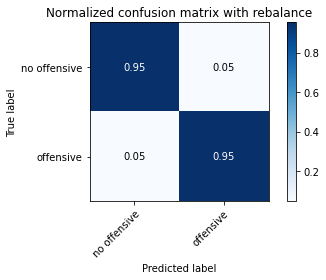
[ 204 3911]]

Normalized confusion matrix

[[0.95249406 0.04750594]

[0.04957473 0.95042527]]





precision recall f1-score support

no offensive 0.80 0.95 0.87 842

offensive 0.99 0.95 0.97 4115

accuracy 0.95 4957

macro avg 0.89 0.95 0.92 4957

weighted avg 0.96 0.95 0.95 4957

training 6 epoch...

46.714585304260254 seconds used......

6 training finished! average train loss: 0.07463538793476614

total clf loss: 24.588814626447856 total adv loss: 26.2803636752069

evaluating...

Confusion matrix, without normalization

[[ 205 80 9]

[ 513 3241 67]

[ 93 99 650]]

Normalized confusion matrix

[[0.69727891 0.27210884 0.03061224]

[0.13425805 0.84820728 0.01753468]

[0.11045131 0.1175772 0.7719715 ]]





precision recall f1-score support

hate 0.25 0.70 0.37 294

offensive 0.95 0.85 0.90 3821

neither 0.90 0.77 0.83 842

accuracy 0.83 4957

macro avg 0.70 0.77 0.70 4957

weighted avg 0.90 0.83 0.85 4957

average eval\_loss: 0.6371250778526789, accuracy: 0.8263062336090378

evaluating adversary...

Confusion matrix, without normalization

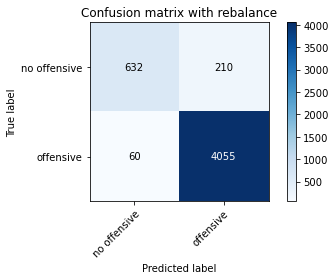
[[ 632 210]

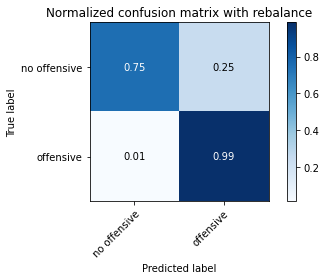
[ 60 4055]]

Normalized confusion matrix

[[0.75059382 0.24940618]

[0.0145808 0.9854192 ]]





precision recall f1-score support

no offensive 0.91 0.75 0.82 842

offensive 0.95 0.99 0.97 4115

accuracy 0.95 4957

macro avg 0.93 0.87 0.90 4957

weighted avg 0.94 0.95 0.94 4957

training 7 epoch...

46.65276908874512 seconds used......

7 training finished! average train loss: 0.04417694490797796

total clf loss: 12.551279639592394 total adv loss: 11.104493723833002

evaluating...

Confusion matrix, without normalization

[[ 214 47 33]

[ 651 2952 218]

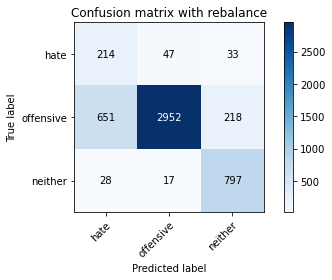
[ 28 17 797]]

Normalized confusion matrix

[[0.72789116 0.15986395 0.1122449 ]

[0.17037425 0.77257262 0.05705313]

[0.03325416 0.02019002 0.94655582]]





precision recall f1-score support

hate 0.24 0.73 0.36 294

offensive 0.98 0.77 0.86 3821

neither 0.76 0.95 0.84 842

accuracy 0.80 4957

macro avg 0.66 0.82 0.69 4957

weighted avg 0.90 0.80 0.83 4957

average eval\_loss: 0.9203458775488299, accuracy: 0.7994754892071818

evaluating adversary...

Confusion matrix, without normalization

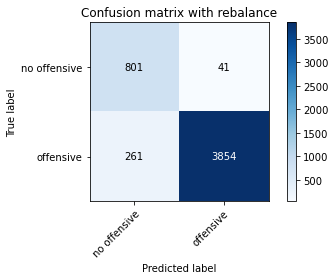
[[ 801 41]

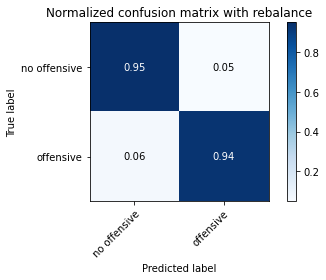
[ 261 3854]]

Normalized confusion matrix

[[0.95130641 0.04869359]

[0.06342649 0.93657351]]





precision recall f1-score support

no offensive 0.75 0.95 0.84 842

offensive 0.99 0.94 0.96 4115

accuracy 0.94 4957

macro avg 0.87 0.94 0.90 4957

weighted avg 0.95 0.94 0.94 4957

training 8 epoch...

46.64030933380127 seconds used......

8 training finished! average train loss: 0.03658790615651408

total clf loss: 11.898527628509328 total adv loss: 12.537768448761199

evaluating...

Confusion matrix, without normalization

[[ 236 44 14]

[ 952 2747 122]

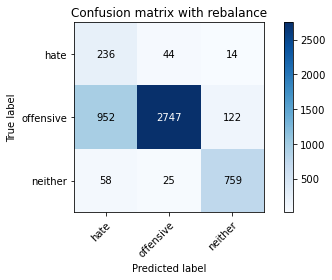
[ 58 25 759]]

Normalized confusion matrix

[[0.80272109 0.14965986 0.04761905]

[0.24914944 0.71892175 0.03192881]

[0.06888361 0.02969121 0.90142518]]





precision recall f1-score support

hate 0.19 0.80 0.31 294

offensive 0.98 0.72 0.83 3821

neither 0.85 0.90 0.87 842

accuracy 0.75 4957

macro avg 0.67 0.81 0.67 4957

weighted avg 0.91 0.75 0.80 4957

average eval\_loss: 1.1383815450113146, accuracy: 0.7548920718176316

evaluating adversary...

Confusion matrix, without normalization

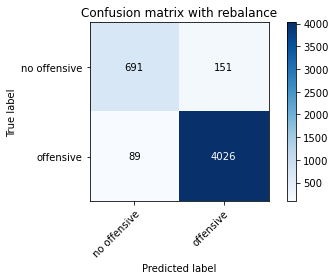
[[ 691 151]

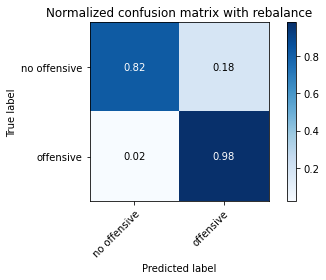
[ 89 4026]]

Normalized confusion matrix

[[0.82066508 0.17933492]

[0.02162819 0.97837181]]





precision recall f1-score support

no offensive 0.89 0.82 0.85 842

offensive 0.96 0.98 0.97 4115

accuracy 0.95 4957

macro avg 0.92 0.90 0.91 4957

weighted avg 0.95 0.95 0.95 4957

training 9 epoch...

46.82638144493103 seconds used......

9 training finished! average train loss: 0.028835667178786377

total clf loss: 10.537100633140653 total adv loss: 12.458225938491523

evaluating...

Confusion matrix, without normalization

[[ 229 47 18]

[ 849 2837 135]

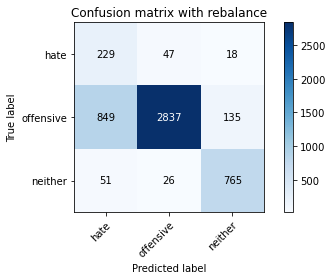
[ 51 26 765]]

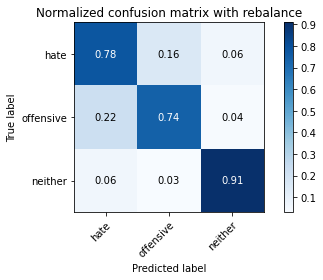
Normalized confusion matrix

[[0.77891156 0.15986395 0.06122449]

[0.22219314 0.74247579 0.03533107]

[0.06057007 0.03087886 0.90855107]]





precision recall f1-score support

hate 0.20 0.78 0.32 294

offensive 0.97 0.74 0.84 3821

neither 0.83 0.91 0.87 842

accuracy 0.77 4957

macro avg 0.67 0.81 0.68 4957

weighted avg 0.91 0.77 0.82 4957

average eval\_loss: 1.1545793234686097, accuracy: 0.7728464797256405

evaluating adversary...

Confusion matrix, without normalization

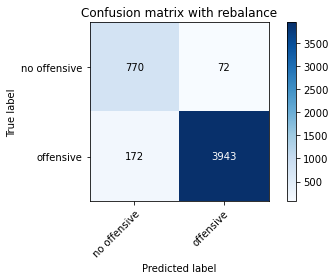
[[ 770 72]

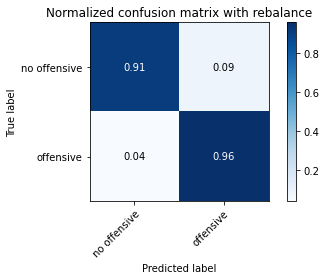
[ 172 3943]]

Normalized confusion matrix

[[0.91448931 0.08551069]

[0.0417983 0.9582017 ]]





precision recall f1-score support

no offensive 0.82 0.91 0.86 842

offensive 0.98 0.96 0.97 4115

accuracy 0.95 4957

macro avg 0.90 0.94 0.92 4957

weighted avg 0.95 0.95 0.95 4957

training 10 epoch...

46.622427225112915 seconds used......

10 training finished! average train loss: 0.0127371242239238

total clf loss: 3.6342569511616603 total adv loss: 3.2360194393404527

evaluating...

Confusion matrix, without normalization

[[ 211 69 14]

[ 597 3105 119]

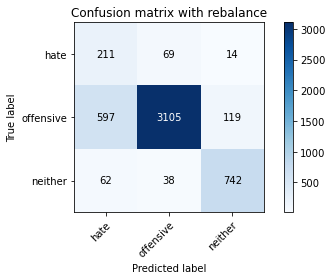
[ 62 38 742]]

Normalized confusion matrix

[[0.71768707 0.23469388 0.04761905]

[0.15624182 0.8126145 0.03114368]

[0.0736342 0.04513064 0.88123515]]





precision recall f1-score support

hate 0.24 0.72 0.36 294

offensive 0.97 0.81 0.88 3821

neither 0.85 0.88 0.86 842

accuracy 0.82 4957

macro avg 0.69 0.80 0.70 4957

weighted avg 0.90 0.82 0.85 4957

average eval\_loss: 0.9921020459521734, accuracy: 0.8186403066370789

evaluating adversary...

Confusion matrix, without normalization

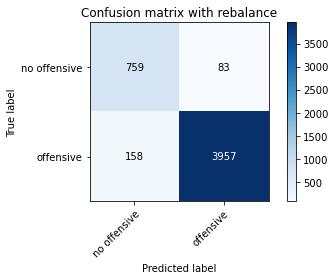
[[ 759 83]

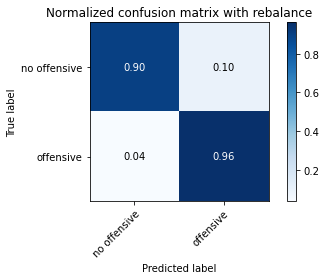
[ 158 3957]]

Normalized confusion matrix

[[0.90142518 0.09857482]

[0.03839611 0.96160389]]





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 0.90 0.93 0.92 4957

weighted avg 0.95 0.95 0.95 4957

4 run

training 1 epoch...

46.723856925964355 seconds used......

1 training finished! average train loss: 0.5986209378866424

total clf loss: 145.50306530296803 total adv loss: 95.86419135332108

evaluating...

Confusion matrix, without normalization

[[ 216 68 10]

[ 561 3232 28]

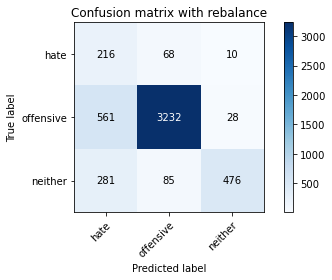
[ 281 85 476]]

Normalized confusion matrix

[[0.73469388 0.23129252 0.03401361]

[0.1468202 0.84585187 0.00732792]

[0.33372922 0.10095012 0.56532067]]





precision recall f1-score support

hate 0.20 0.73 0.32 294

offensive 0.95 0.85 0.90 3821

neither 0.93 0.57 0.70 842

accuracy 0.79 4957

macro avg 0.70 0.72 0.64 4957

weighted avg 0.91 0.79 0.83 4957

average eval\_loss: 0.5064508517182642, accuracy: 0.7916078273149082

evaluating adversary...

Confusion matrix, without normalization

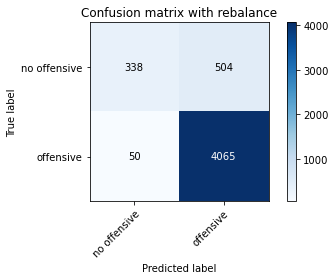
[[ 338 504]

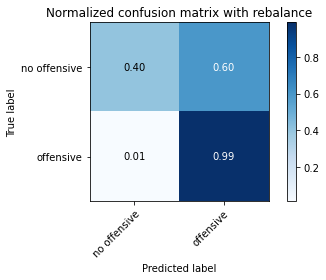
[ 50 4065]]

Normalized confusion matrix

[[0.40142518 0.59857482]

[0.01215067 0.98784933]]





precision recall f1-score support

no offensive 0.87 0.40 0.55 842

offensive 0.89 0.99 0.94 4115

accuracy 0.89 4957

macro avg 0.88 0.69 0.74 4957

weighted avg 0.89 0.89 0.87 4957

training 2 epoch...

46.63413047790527 seconds used......

2 training finished! average train loss: 0.3654525198957376

total clf loss: 92.76632256805897 total adv loss: 67.27542736381292

evaluating...

Confusion matrix, without normalization

[[ 217 45 32]

[ 667 2971 183]

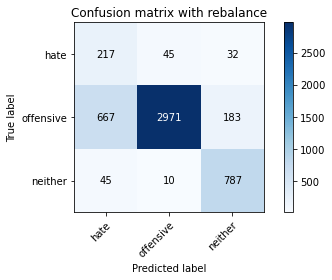
[ 45 10 787]]

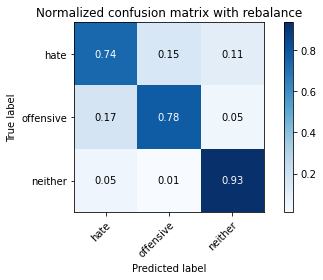
Normalized confusion matrix

[[0.73809524 0.15306122 0.10884354]

[0.17456163 0.77754515 0.04789322]

[0.05344418 0.01187648 0.93467933]]





precision recall f1-score support

hate 0.23 0.74 0.35 294

offensive 0.98 0.78 0.87 3821

neither 0.79 0.93 0.85 842

accuracy 0.80 4957

macro avg 0.67 0.82 0.69 4957

weighted avg 0.90 0.80 0.83 4957

average eval\_loss: 0.4731116633742086, accuracy: 0.8018963082509583

evaluating adversary...

Confusion matrix, without normalization

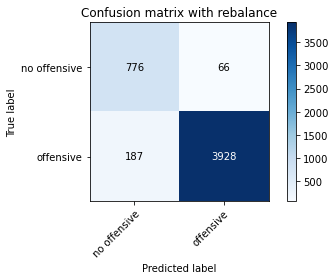
[[ 776 66]

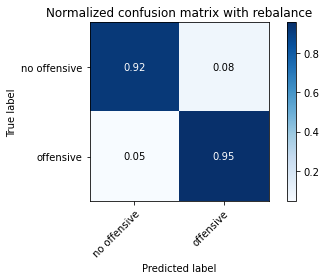
[ 187 3928]]

Normalized confusion matrix

[[0.9216152 0.0783848]

[0.0454435 0.9545565]]





precision recall f1-score support

no offensive 0.81 0.92 0.86 842

offensive 0.98 0.95 0.97 4115

accuracy 0.95 4957

macro avg 0.89 0.94 0.91 4957

weighted avg 0.95 0.95 0.95 4957

training 3 epoch...

46.4853630065918 seconds used......

3 training finished! average train loss: 0.31109125001562965

total clf loss: 69.37703073769808 total adv loss: 35.956505118869245

evaluating...

Confusion matrix, without normalization

[[ 197 74 23]

[ 417 3292 112]

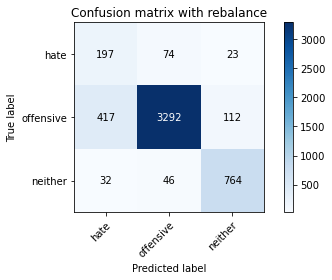
[ 32 46 764]]

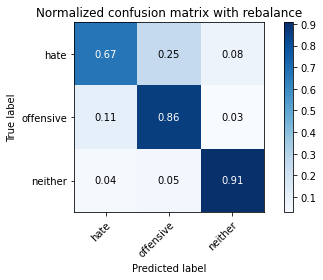
Normalized confusion matrix

[[0.67006803 0.25170068 0.07823129]

[0.10913373 0.86155457 0.0293117 ]

[0.03800475 0.05463183 0.90736342]]





precision recall f1-score support

hate 0.30 0.67 0.42 294

offensive 0.96 0.86 0.91 3821

neither 0.85 0.91 0.88 842

accuracy 0.86 4957

macro avg 0.71 0.81 0.74 4957

weighted avg 0.91 0.86 0.88 4957

average eval\_loss: 0.3733568117322941, accuracy: 0.8579786160984466

evaluating adversary...

Confusion matrix, without normalization

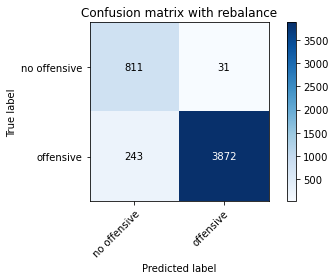
[[ 811 31]

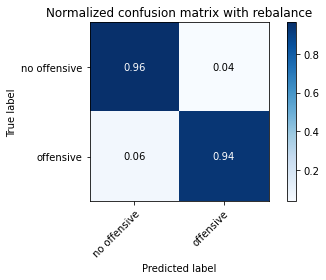
[ 243 3872]]

Normalized confusion matrix

[[0.9631829 0.0368171 ]

[0.05905225 0.94094775]]





precision recall f1-score support

no offensive 0.77 0.96 0.86 842

offensive 0.99 0.94 0.97 4115

accuracy 0.94 4957

macro avg 0.88 0.95 0.91 4957

weighted avg 0.95 0.94 0.95 4957

training 4 epoch...

46.59550619125366 seconds used......

4 training finished! average train loss: 0.22794056975702096

total clf loss: 50.14122624695301 total adv loss: 24.80753143131733

evaluating...

Confusion matrix, without normalization

[[ 214 64 16]

[ 546 3165 110]

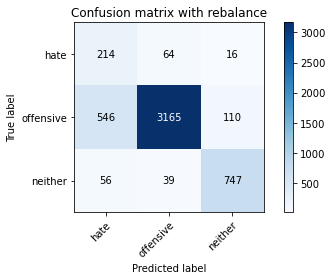
[ 56 39 747]]

Normalized confusion matrix

[[0.72789116 0.21768707 0.05442177]

[0.14289453 0.82831719 0.02878828]

[0.06650831 0.04631829 0.8871734 ]]





precision recall f1-score support

hate 0.26 0.73 0.39 294

offensive 0.97 0.83 0.89 3821

neither 0.86 0.89 0.87 842

accuracy 0.83 4957

macro avg 0.70 0.81 0.72 4957

weighted avg 0.91 0.83 0.86 4957

average eval\_loss: 0.43338846391247166, accuracy: 0.8323582812184789

evaluating adversary...

Confusion matrix, without normalization

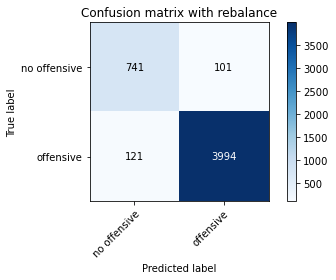
[[ 741 101]

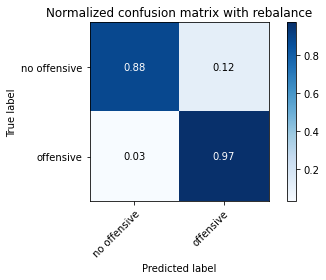
[ 121 3994]]

Normalized confusion matrix

[[0.88004751 0.11995249]

[0.02940462 0.97059538]]





precision recall f1-score support

no offensive 0.86 0.88 0.87 842

offensive 0.98 0.97 0.97 4115

accuracy 0.96 4957

macro avg 0.92 0.93 0.92 4957

weighted avg 0.96 0.96 0.96 4957

training 5 epoch...

46.68880295753479 seconds used......

5 training finished! average train loss: 0.1522796546593744

total clf loss: 32.91467622667551 total adv loss: 15.277456477982923

evaluating...

Confusion matrix, without normalization

[[ 209 60 25]

[ 638 3027 156]

[ 50 18 774]]

Normalized confusion matrix

[[0.71088435 0.20408163 0.08503401]

[0.166972 0.79220099 0.04082701]

[0.05938242 0.02137767 0.9192399 ]]





precision recall f1-score support

hate 0.23 0.71 0.35 294

offensive 0.97 0.79 0.87 3821

neither 0.81 0.92 0.86 842

accuracy 0.81 4957

macro avg 0.67 0.81 0.70 4957

weighted avg 0.90 0.81 0.84 4957

average eval\_loss: 0.5447701801876387, accuracy: 0.808957030461973

evaluating adversary...

Confusion matrix, without normalization

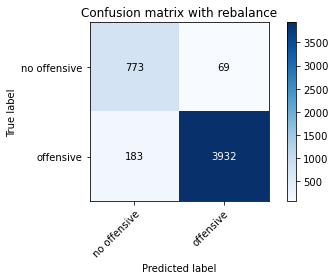
[[ 773 69]

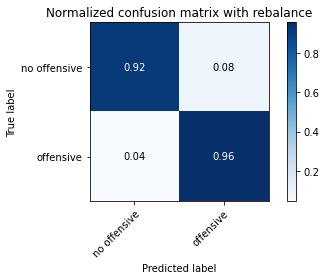
[ 183 3932]]

Normalized confusion matrix

[[0.91805226 0.08194774]

[0.04447145 0.95552855]]





precision recall f1-score support

no offensive 0.81 0.92 0.86 842

offensive 0.98 0.96 0.97 4115

accuracy 0.95 4957

macro avg 0.90 0.94 0.91 4957

weighted avg 0.95 0.95 0.95 4957

training 6 epoch...

46.66539120674133 seconds used......

6 training finished! average train loss: 0.08279646264269339

total clf loss: 18.033915404230356 total adv loss: 8.612712066154927

evaluating...

Confusion matrix, without normalization

[[ 209 65 20]

[ 700 2983 138]

[ 66 33 743]]

Normalized confusion matrix

[[0.71088435 0.22108844 0.06802721]

[0.18319812 0.78068568 0.0361162 ]

[0.0783848 0.0391924 0.8824228 ]]





precision recall f1-score support

hate 0.21 0.71 0.33 294

offensive 0.97 0.78 0.86 3821

neither 0.82 0.88 0.85 842

accuracy 0.79 4957

macro avg 0.67 0.79 0.68 4957

weighted avg 0.90 0.79 0.83 4957

average eval\_loss: 0.7458177611321932, accuracy: 0.79382691143837

evaluating adversary...

Confusion matrix, without normalization

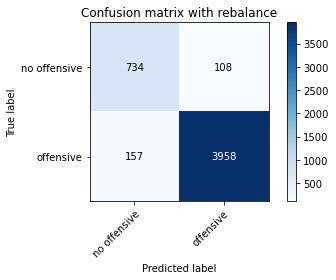
[[ 734 108]

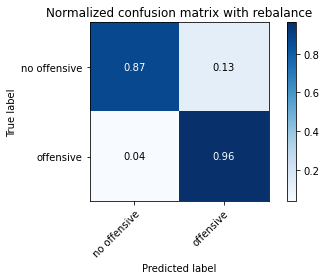
[ 157 3958]]

Normalized confusion matrix

[[0.87173397 0.12826603]

[0.0381531 0.9618469 ]]





precision recall f1-score support

no offensive 0.82 0.87 0.85 842

offensive 0.97 0.96 0.97 4115

accuracy 0.95 4957

macro avg 0.90 0.92 0.91 4957

weighted avg 0.95 0.95 0.95 4957

training 7 epoch...

46.74950313568115 seconds used......

7 training finished! average train loss: 0.06140094118503233

total clf loss: 28.365009179105982 total adv loss: 39.70099698344711

evaluating...

Confusion matrix, without normalization

[[ 188 102 4]

[ 354 3446 21]

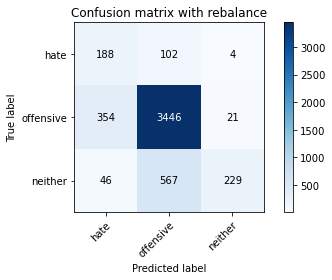
[ 46 567 229]]

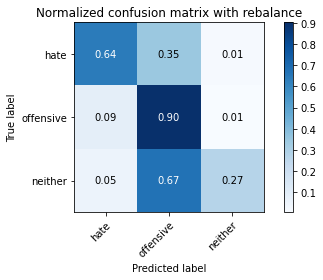
Normalized confusion matrix

[[0.63945578 0.34693878 0.01360544]

[0.0926459 0.90185815 0.00549594]

[0.05463183 0.67339667 0.2719715 ]]





precision recall f1-score support

hate 0.32 0.64 0.43 294

offensive 0.84 0.90 0.87 3821

neither 0.90 0.27 0.42 842

accuracy 0.78 4957

macro avg 0.69 0.60 0.57 4957

weighted avg 0.82 0.78 0.77 4957

average eval\_loss: 0.5711197836039167, accuracy: 0.7793019971757111

evaluating adversary...

Confusion matrix, without normalization

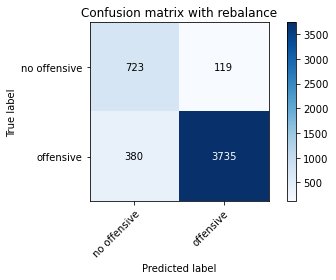
[[ 723 119]

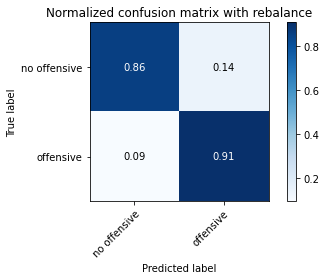
[ 380 3735]]

Normalized confusion matrix

[[0.85866983 0.14133017]

[0.09234508 0.90765492]]





precision recall f1-score support

no offensive 0.66 0.86 0.74 842

offensive 0.97 0.91 0.94 4115

accuracy 0.90 4957

macro avg 0.81 0.88 0.84 4957

weighted avg 0.92 0.90 0.90 4957

training 8 epoch...

46.66118597984314 seconds used......

8 training finished! average train loss: 0.13414564996696354

total clf loss: 62.294215861707926 total adv loss: 87.45624564460013

evaluating...

Confusion matrix, without normalization

[[ 221 64 9]

[ 653 2659 509]

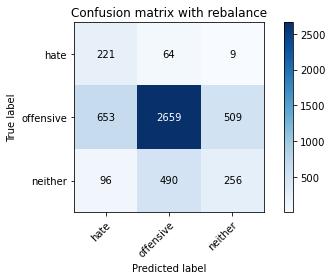
[ 96 490 256]]

Normalized confusion matrix

[[0.75170068 0.21768707 0.03061224]

[0.17089767 0.69589113 0.1332112 ]

[0.11401425 0.58194774 0.304038 ]]





precision recall f1-score support

hate 0.23 0.75 0.35 294

offensive 0.83 0.70 0.76 3821

neither 0.33 0.30 0.32 842

accuracy 0.63 4957

macro avg 0.46 0.58 0.47 4957

weighted avg 0.71 0.63 0.66 4957

average eval\_loss: 0.9253013827387364, accuracy: 0.6326407101069195

evaluating adversary...

Confusion matrix, without normalization

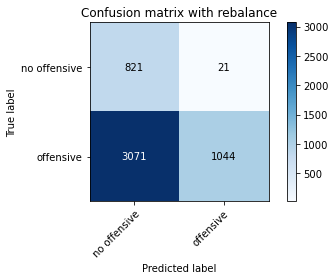
[[ 821 21]

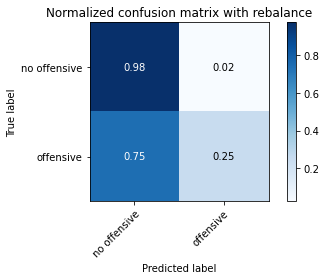
[3071 1044]]

Normalized confusion matrix

[[0.97505938 0.02494062]

[0.74629405 0.25370595]]





precision recall f1-score support

no offensive 0.21 0.98 0.35 842

offensive 0.98 0.25 0.40 4115

accuracy 0.38 4957

macro avg 0.60 0.61 0.37 4957

weighted avg 0.85 0.38 0.39 4957

training 9 epoch...

46.633822441101074 seconds used......

9 training finished! average train loss: 0.06100530509853311

total clf loss: 27.76932759443298 total adv loss: 38.52760201867204

evaluating...

Confusion matrix, without normalization

[[ 232 46 16]

[ 815 2877 129]

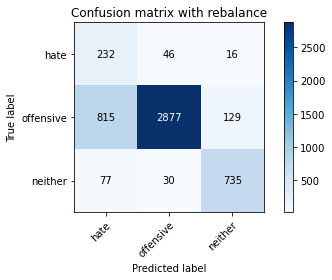
[ 77 30 735]]

Normalized confusion matrix

[[0.78911565 0.15646259 0.05442177]

[0.21329495 0.75294426 0.0337608 ]

[0.09144893 0.03562945 0.87292162]]





precision recall f1-score support

hate 0.21 0.79 0.33 294

offensive 0.97 0.75 0.85 3821

neither 0.84 0.87 0.85 842

accuracy 0.78 4957

macro avg 0.67 0.80 0.68 4957

weighted avg 0.91 0.78 0.82 4957

average eval\_loss: 1.133571438239737, accuracy: 0.7754690336897317

evaluating adversary...

Confusion matrix, without normalization

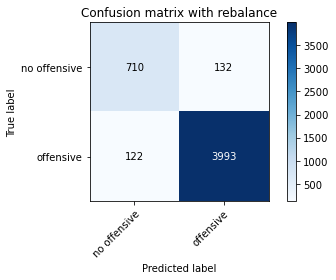
[[ 710 132]

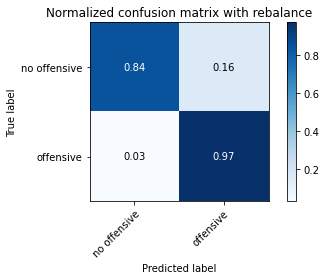
[ 122 3993]]

Normalized confusion matrix

[[0.8432304 0.1567696 ]

[0.02964763 0.97035237]]





precision recall f1-score support

no offensive 0.85 0.84 0.85 842

offensive 0.97 0.97 0.97 4115

accuracy 0.95 4957

macro avg 0.91 0.91 0.91 4957

weighted avg 0.95 0.95 0.95 4957

training 10 epoch...

46.55795240402222 seconds used......

10 training finished! average train loss: 0.035528077481162046

total clf loss: 10.191691228887066 total adv loss: 9.147533542709425

evaluating...

Confusion matrix, without normalization

[[ 207 63 24]

[ 688 2977 156]

[ 50 24 768]]

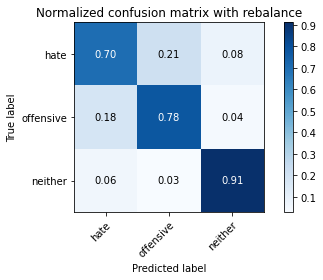
Normalized confusion matrix

[[0.70408163 0.21428571 0.08163265]

[0.18005758 0.77911541 0.04082701]

[0.05938242 0.02850356 0.91211401]]





precision recall f1-score support

hate 0.22 0.70 0.33 294

offensive 0.97 0.78 0.86 3821

neither 0.81 0.91 0.86 842

accuracy 0.80 4957

macro avg 0.67 0.80 0.69 4957

weighted avg 0.90 0.80 0.83 4957

average eval\_loss: 1.0298618786112075, accuracy: 0.79725640508372

evaluating adversary...

Confusion matrix, without normalization

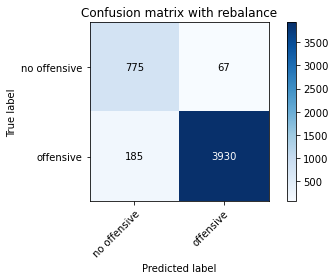
[[ 775 67]

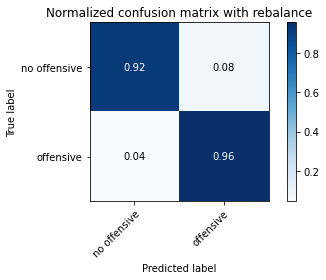
[ 185 3930]]

Normalized confusion matrix

[[0.92042755 0.07957245]

[0.04495747 0.95504253]]





precision recall f1-score support

no offensive 0.81 0.92 0.86 842

offensive 0.98 0.96 0.97 4115

accuracy 0.95 4957

macro avg 0.90 0.94 0.91 4957

weighted avg 0.95 0.95 0.95 4957

alpha = 0.46

1 run

training 1 epoch...

46.72237944602966 seconds used......

1 training finished! average train loss: 0.5373418380816778

total clf loss: 135.32334902882576 total adv loss: 94.43020385503769

evaluating...

Confusion matrix, without normalization

[[ 211 49 34]

[ 759 2914 148]

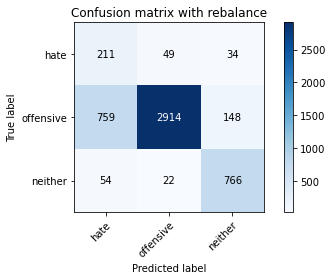
[ 54 22 766]]

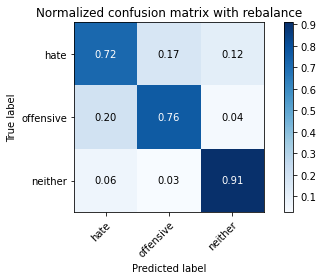
Normalized confusion matrix

[[0.71768707 0.16666667 0.11564626]

[0.1986391 0.76262758 0.03873332]

[0.06413302 0.02612827 0.90973872]]





precision recall f1-score support

hate 0.21 0.72 0.32 294

offensive 0.98 0.76 0.86 3821

neither 0.81 0.91 0.86 842

accuracy 0.78 4957

macro avg 0.66 0.80 0.68 4957

weighted avg 0.90 0.78 0.82 4957

average eval\_loss: 0.5692260849379724, accuracy: 0.7849505749445229

evaluating adversary...

Confusion matrix, without normalization

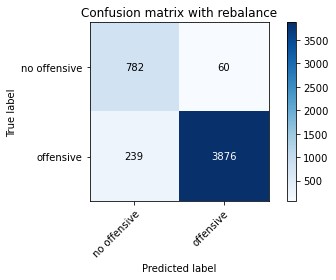
[[ 782 60]

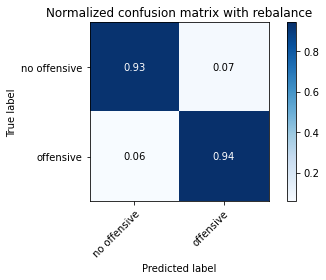
[ 239 3876]]

Normalized confusion matrix

[[0.92874109 0.07125891]

[0.05808019 0.94191981]]





precision recall f1-score support

no offensive 0.77 0.93 0.84 842

offensive 0.98 0.94 0.96 4115

accuracy 0.94 4957

macro avg 0.88 0.94 0.90 4957

weighted avg 0.95 0.94 0.94 4957

training 2 epoch...

46.5129337310791 seconds used......

2 training finished! average train loss: 0.35711062350991174

total clf loss: 81.23953023552895 total adv loss: 43.855680545791984

evaluating...

Confusion matrix, without normalization

[[ 252 34 8]

[1026 2773 22]

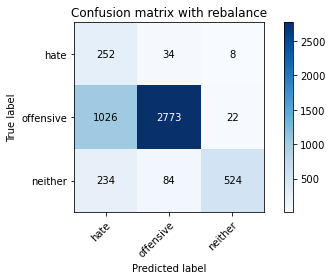
[ 234 84 524]]

Normalized confusion matrix

[[0.85714286 0.11564626 0.02721088]

[0.2685161 0.72572625 0.00575766]

[0.27790974 0.09976247 0.62232779]]





precision recall f1-score support

hate 0.17 0.86 0.28 294

offensive 0.96 0.73 0.83 3821

neither 0.95 0.62 0.75 842

accuracy 0.72 4957

macro avg 0.69 0.74 0.62 4957

weighted avg 0.91 0.72 0.78 4957

average eval\_loss: 0.6390899622272099, accuracy: 0.7159572321968933

evaluating adversary...

Confusion matrix, without normalization

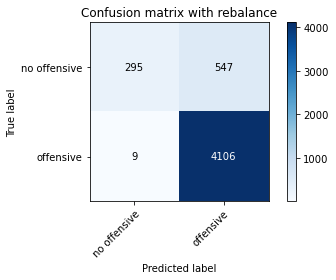
[[ 295 547]

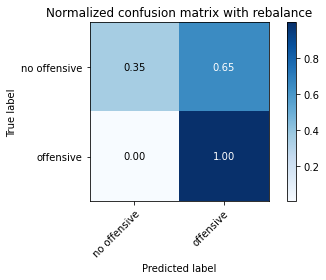
[ 9 4106]]

Normalized confusion matrix

[[0.35035629 0.64964371]

[0.00218712 0.99781288]]





precision recall f1-score support

no offensive 0.97 0.35 0.51 842

offensive 0.88 1.00 0.94 4115

accuracy 0.89 4957

macro avg 0.93 0.67 0.73 4957

weighted avg 0.90 0.89 0.86 4957

training 3 epoch...

46.69325280189514 seconds used......

3 training finished! average train loss: 0.29070726198236846

total clf loss: 66.88353614509106 total adv loss: 37.33172616735101

evaluating...

Confusion matrix, without normalization

[[ 220 69 5]

[ 541 3234 46]

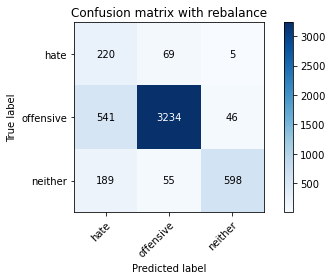
[ 189 55 598]]

Normalized confusion matrix

[[0.74829932 0.23469388 0.0170068 ]

[0.14158597 0.84637529 0.01203873]

[0.22446556 0.06532067 0.71021378]]





precision recall f1-score support

hate 0.23 0.75 0.35 294

offensive 0.96 0.85 0.90 3821

neither 0.92 0.71 0.80 842

accuracy 0.82 4957

macro avg 0.71 0.77 0.69 4957

weighted avg 0.91 0.82 0.85 4957

average eval\_loss: 0.5114324281232492, accuracy: 0.8174298971151907

evaluating adversary...

Confusion matrix, without normalization

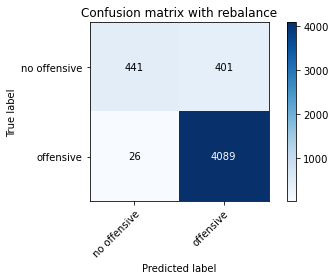
[[ 441 401]

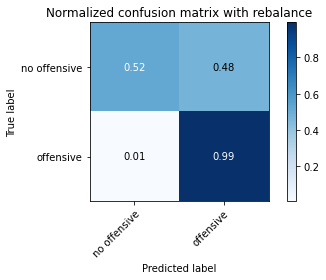
[ 26 4089]]

Normalized confusion matrix

[[0.52375297 0.47624703]

[0.00631835 0.99368165]]





precision recall f1-score support

no offensive 0.94 0.52 0.67 842

offensive 0.91 0.99 0.95 4115

accuracy 0.91 4957

macro avg 0.93 0.76 0.81 4957

weighted avg 0.92 0.91 0.90 4957

training 4 epoch...

46.62454080581665 seconds used......

4 training finished! average train loss: 0.20824814246891185

total clf loss: 50.234601728618145 total adv loss: 31.791671561077237

evaluating...

Confusion matrix, without normalization

[[ 237 38 19]

[1090 2615 116]

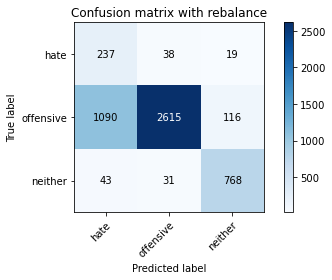
[ 43 31 768]]

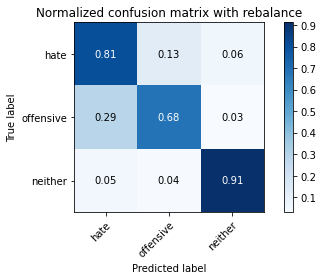
Normalized confusion matrix

[[0.80612245 0.1292517 0.06462585]

[0.28526564 0.68437582 0.03035854]

[0.05106888 0.0368171 0.91211401]]





precision recall f1-score support

hate 0.17 0.81 0.28 294

offensive 0.97 0.68 0.80 3821

neither 0.85 0.91 0.88 842

accuracy 0.73 4957

macro avg 0.67 0.80 0.66 4957

weighted avg 0.91 0.73 0.79 4957

average eval\_loss: 0.6767478779919686, accuracy: 0.7302804115392374

evaluating adversary...

Confusion matrix, without normalization

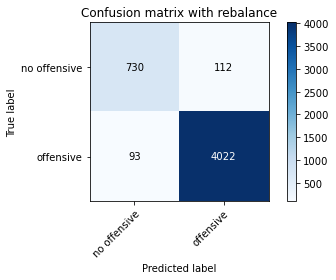
[[ 730 112]

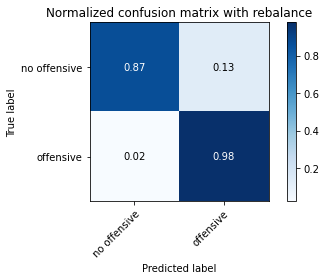
[ 93 4022]]

Normalized confusion matrix

[[0.86698337 0.13301663]

[0.02260024 0.97739976]]





precision recall f1-score support

no offensive 0.89 0.87 0.88 842

offensive 0.97 0.98 0.98 4115

accuracy 0.96 4957

macro avg 0.93 0.92 0.93 4957

weighted avg 0.96 0.96 0.96 4957

training 5 epoch...

46.8075008392334 seconds used......

5 training finished! average train loss: 0.12477437731924287

total clf loss: 26.83229641057551 total adv loss: 11.947560478933156

evaluating...

Confusion matrix, without normalization

[[ 191 71 32]

[ 518 3154 149]

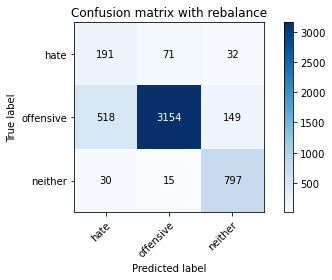
[ 30 15 797]]

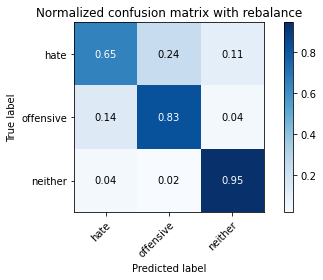
Normalized confusion matrix

[[0.64965986 0.2414966 0.10884354]

[0.13556661 0.82543837 0.03899503]

[0.03562945 0.01781473 0.94655582]]





precision recall f1-score support

hate 0.26 0.65 0.37 294

offensive 0.97 0.83 0.89 3821

neither 0.81 0.95 0.88 842

accuracy 0.84 4957

macro avg 0.68 0.81 0.71 4957

weighted avg 0.90 0.84 0.86 4957

average eval\_loss: 0.5369540765626176, accuracy: 0.8355860399435142

evaluating adversary...

Confusion matrix, without normalization

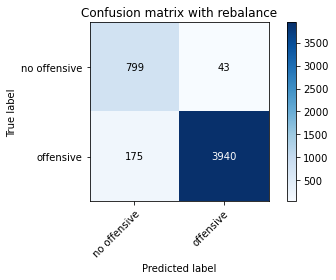
[[ 799 43]

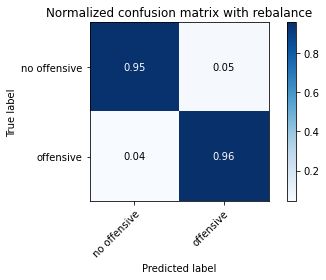
[ 175 3940]]

Normalized confusion matrix

[[0.94893112 0.05106888]

[0.04252734 0.95747266]]





precision recall f1-score support

no offensive 0.82 0.95 0.88 842

offensive 0.99 0.96 0.97 4115

accuracy 0.96 4957

macro avg 0.90 0.95 0.93 4957

weighted avg 0.96 0.96 0.96 4957

training 6 epoch...

46.87358570098877 seconds used......

6 training finished! average train loss: 0.05450672767130516

total clf loss: 13.53015054576099 total adv loss: 9.151087114820257

evaluating...

Confusion matrix, without normalization

[[ 207 64 23]

[ 552 3141 128]

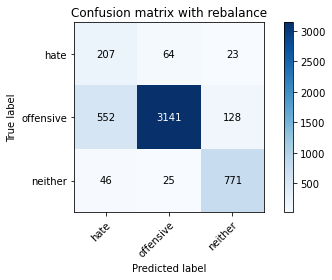
[ 46 25 771]]

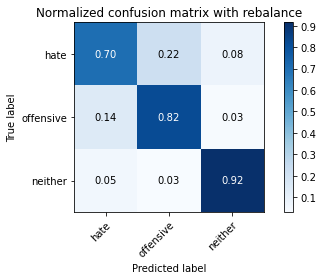
Normalized confusion matrix

[[0.70408163 0.21768707 0.07823129]

[0.1444648 0.82203612 0.03349908]

[0.05463183 0.02969121 0.91567696]]





precision recall f1-score support

hate 0.26 0.70 0.38 294

offensive 0.97 0.82 0.89 3821

neither 0.84 0.92 0.87 842

accuracy 0.83 4957

macro avg 0.69 0.81 0.71 4957

weighted avg 0.91 0.83 0.86 4957

average eval\_loss: 0.6970388385271954, accuracy: 0.8309461367762759

evaluating adversary...

Confusion matrix, without normalization

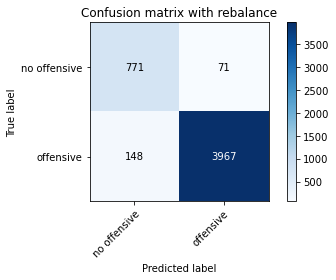
[[ 771 71]

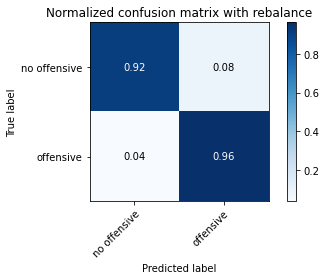
[ 148 3967]]

Normalized confusion matrix

[[0.91567696 0.08432304]

[0.03596598 0.96403402]]





precision recall f1-score support

no offensive 0.84 0.92 0.88 842

offensive 0.98 0.96 0.97 4115

accuracy 0.96 4957

macro avg 0.91 0.94 0.92 4957

weighted avg 0.96 0.96 0.96 4957

training 7 epoch...

46.74081110954285 seconds used......

7 training finished! average train loss: 0.027330778850302405

total clf loss: 6.91829412127845 total adv loss: 4.879849757417105

evaluating...

Confusion matrix, without normalization

[[ 201 69 24]

[ 538 3153 130]

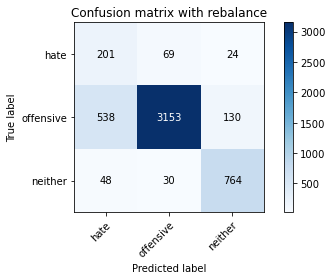
[ 48 30 764]]

Normalized confusion matrix

[[0.68367347 0.23469388 0.08163265]

[0.14080084 0.82517666 0.03402251]

[0.05700713 0.03562945 0.90736342]]





precision recall f1-score support

hate 0.26 0.68 0.37 294

offensive 0.97 0.83 0.89 3821

neither 0.83 0.91 0.87 842

accuracy 0.83 4957

macro avg 0.69 0.81 0.71 4957

weighted avg 0.90 0.83 0.86 4957

average eval\_loss: 0.7529280492406895, accuracy: 0.8307444018559612

evaluating adversary...

Confusion matrix, without normalization

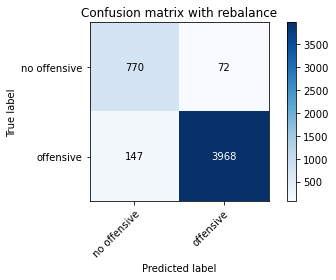
[[ 770 72]

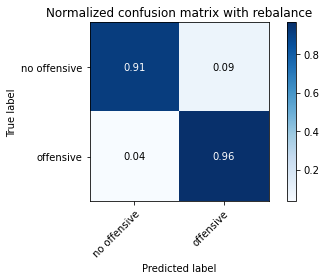
[ 147 3968]]

Normalized confusion matrix

[[0.91448931 0.08551069]

[0.03572296 0.96427704]]





precision recall f1-score support

no offensive 0.84 0.91 0.88 842

offensive 0.98 0.96 0.97 4115

accuracy 0.96 4957

macro avg 0.91 0.94 0.92 4957

weighted avg 0.96 0.96 0.96 4957

training 8 epoch...

46.926268339157104 seconds used......

8 training finished! average train loss: 0.021515650819709294

total clf loss: 5.48970743722748 total adv loss: 3.9359371672180714

evaluating...

Confusion matrix, without normalization

[[ 213 59 22]

[ 700 2958 163]

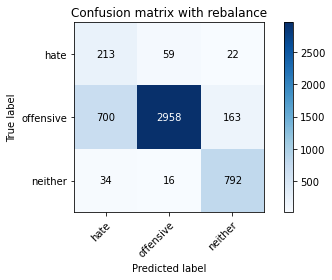
[ 34 16 792]]

Normalized confusion matrix

[[0.7244898 0.20068027 0.07482993]

[0.18319812 0.77414289 0.04265899]

[0.04038005 0.01900238 0.94061758]]





precision recall f1-score support

hate 0.22 0.72 0.34 294

offensive 0.98 0.77 0.86 3821

neither 0.81 0.94 0.87 842

accuracy 0.80 4957

macro avg 0.67 0.81 0.69 4957

weighted avg 0.90 0.80 0.83 4957

average eval\_loss: 1.0426644818902375, accuracy: 0.7994754892071818

evaluating adversary...

Confusion matrix, without normalization

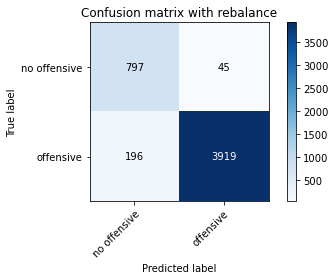
[[ 797 45]

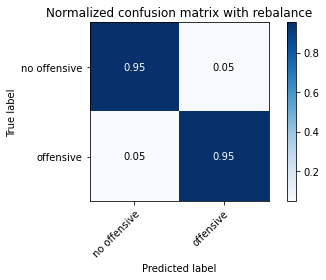
[ 196 3919]]

Normalized confusion matrix

[[0.94655582 0.05344418]

[0.04763062 0.95236938]]





precision recall f1-score support

no offensive 0.80 0.95 0.87 842

offensive 0.99 0.95 0.97 4115

accuracy 0.95 4957

macro avg 0.90 0.95 0.92 4957

weighted avg 0.96 0.95 0.95 4957

training 9 epoch...

46.718873023986816 seconds used......

9 training finished! average train loss: 0.02665301241907863

total clf loss: 6.153474254417233 total adv loss: 3.469150239663577

evaluating...

Confusion matrix, without normalization

[[ 218 55 21]

[ 665 3003 153]

[ 41 18 783]]

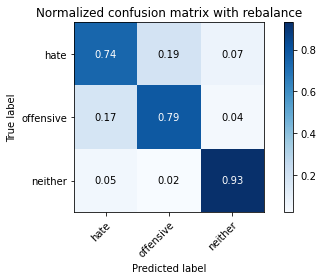
Normalized confusion matrix

[[0.7414966 0.18707483 0.07142857]

[0.17403821 0.78591992 0.04004187]

[0.04869359 0.02137767 0.92992874]]





precision recall f1-score support

hate 0.24 0.74 0.36 294

offensive 0.98 0.79 0.87 3821

neither 0.82 0.93 0.87 842

accuracy 0.81 4957

macro avg 0.68 0.82 0.70 4957

weighted avg 0.91 0.81 0.84 4957

average eval\_loss: 1.0015155979574868, accuracy: 0.8077466209400848

evaluating adversary...

Confusion matrix, without normalization

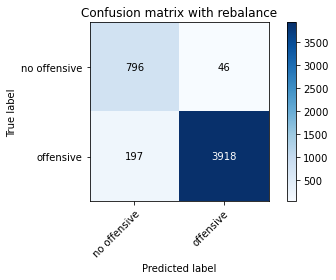
[[ 796 46]

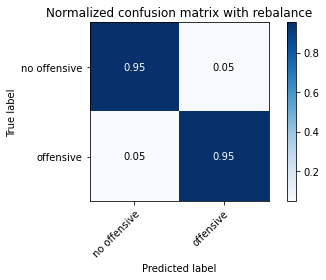
[ 197 3918]]

Normalized confusion matrix

[[0.94536817 0.05463183]

[0.04787363 0.95212637]]





precision recall f1-score support

no offensive 0.80 0.95 0.87 842

offensive 0.99 0.95 0.97 4115

accuracy 0.95 4957

macro avg 0.90 0.95 0.92 4957

weighted avg 0.96 0.95 0.95 4957

training 10 epoch...

46.684271574020386 seconds used......

10 training finished! average train loss: 0.020252765966810787

total clf loss: 6.714231296442449 total adv loss: 7.067409350253001

evaluating...

Confusion matrix, without normalization

[[ 243 34 17]

[1108 2554 159]

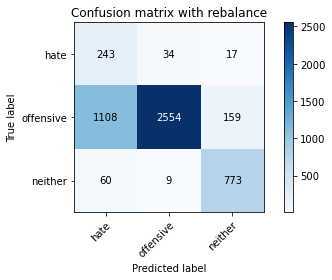
[ 60 9 773]]

Normalized confusion matrix

[[0.82653061 0.11564626 0.05782313]

[0.28997645 0.66841141 0.04161214]

[0.07125891 0.01068884 0.91805226]]





precision recall f1-score support

hate 0.17 0.83 0.29 294

offensive 0.98 0.67 0.80 3821

neither 0.81 0.92 0.86 842

accuracy 0.72 4957

macro avg 0.66 0.80 0.65 4957

weighted avg 0.91 0.72 0.78 4957

average eval\_loss: 1.7590403646475263, accuracy: 0.7201936655235022

evaluating adversary...

Confusion matrix, without normalization

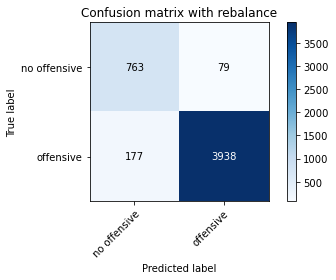
[[ 763 79]

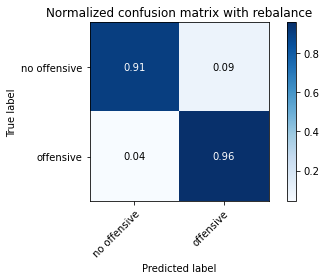
[ 177 3938]]

Normalized confusion matrix

[[0.90617577 0.09382423]

[0.04301337 0.95698663]]





precision recall f1-score support

no offensive 0.81 0.91 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

alpha = 0.46

2 run