0.0

training 1 epoch...

44.90124535560608 seconds used......

1 training finished! average train loss: 0.6751120390250669

total clf loss: 115.44415867328644 total adv loss: 74.4114280231297

evaluating...

Confusion matrix, without normalization

[[ 219 43 32]

[ 747 2918 156]

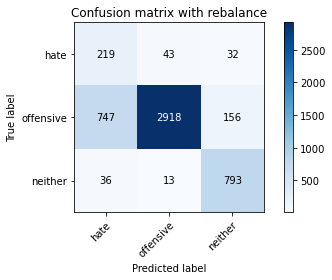
[ 36 13 793]]

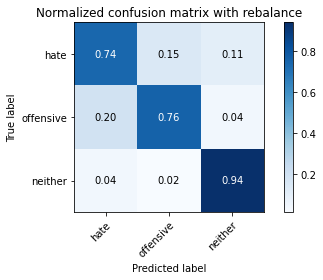
Normalized confusion matrix

[[0.74489796 0.1462585 0.10884354]

[0.19549856 0.76367443 0.04082701]

[0.04275534 0.01543943 0.94180523]]





precision recall f1-score support

hate 0.22 0.74 0.34 294

offensive 0.98 0.76 0.86 3821

neither 0.81 0.94 0.87 842

accuracy 0.79 4957

macro avg 0.67 0.82 0.69 4957

weighted avg 0.91 0.79 0.83 4957

average eval\_loss: 0.46986491746839976, accuracy: 0.7928182368367964

evaluating adversary...

0.1

training 1 epoch...

44.877102851867676 seconds used......

1 training finished! average train loss: 0.6915939269010086

total clf loss: 111.16345500946045 total adv loss: 70.99106374382973

evaluating...

Confusion matrix, without normalization

[[ 233 34 27]

[ 843 2849 129]

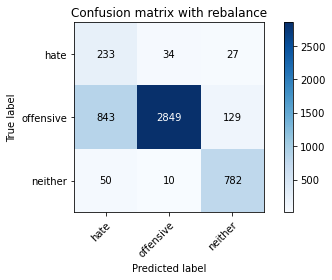
[ 50 10 782]]

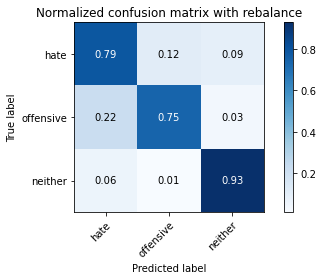
Normalized confusion matrix

[[0.79251701 0.11564626 0.09183673]

[0.22062287 0.74561633 0.0337608 ]

[0.05938242 0.01187648 0.92874109]]





precision recall f1-score support

hate 0.21 0.79 0.33 294

offensive 0.98 0.75 0.85 3821

neither 0.83 0.93 0.88 842

accuracy 0.78 4957

macro avg 0.68 0.82 0.69 4957

weighted avg 0.91 0.78 0.82 4957

average eval\_loss: 0.5643114311199996, accuracy: 0.7795037320960259

evaluating adversary...

0.2

training 1 epoch...

44.86839699745178 seconds used......

1 training finished! average train loss: 0.7384659221820664

total clf loss: 111.89416246116161 total adv loss: 71.91755194962025

evaluating...

Confusion matrix, without normalization

[[ 189 72 33]

[ 448 3191 182]

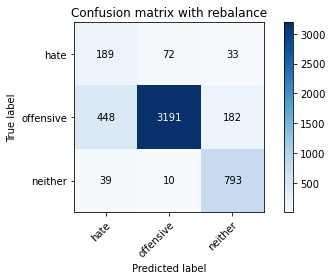
[ 39 10 793]]

Normalized confusion matrix

[[0.64285714 0.24489796 0.1122449 ]

[0.11724679 0.8351217 0.04763151]

[0.04631829 0.01187648 0.94180523]]





precision recall f1-score support

hate 0.28 0.64 0.39 294

offensive 0.97 0.84 0.90 3821

neither 0.79 0.94 0.86 842

accuracy 0.84 4957

macro avg 0.68 0.81 0.72 4957

weighted avg 0.90 0.84 0.86 4957

average eval\_loss: 0.4055572144927517, accuracy: 0.8418398224732702

evaluating adversary...

0.30000000000000004

training 1 epoch...

45.04794931411743 seconds used......

1 training finished! average train loss: 0.7907140476313251

total clf loss: 113.62639981508255 total adv loss: 71.95233839005232

evaluating...

Confusion matrix, without normalization

[[ 226 53 15]

[ 725 3005 91]

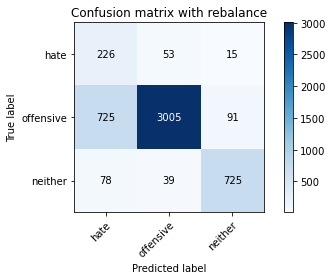
[ 78 39 725]]

Normalized confusion matrix

[[0.76870748 0.18027211 0.05102041]

[0.18974091 0.78644334 0.02381576]

[0.09263658 0.04631829 0.86104513]]





precision recall f1-score support

hate 0.22 0.77 0.34 294

offensive 0.97 0.79 0.87 3821

neither 0.87 0.86 0.87 842

accuracy 0.80 4957

macro avg 0.69 0.81 0.69 4957

weighted avg 0.91 0.80 0.84 4957

average eval\_loss: 0.44285051861116964, accuracy: 0.7980633447649789

evaluating adversary...

0.4

training 1 epoch...

44.92207050323486 seconds used......

1 training finished! average train loss: 0.856749071625241

total clf loss: 117.24102665483952 total adv loss: 73.15765894949436

evaluating...

Confusion matrix, without normalization

[[ 221 39 34]

[ 885 2775 161]

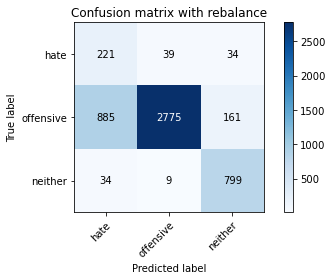
[ 34 9 799]]

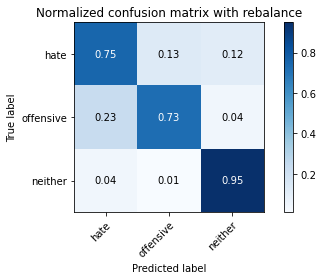
Normalized confusion matrix

[[0.75170068 0.13265306 0.11564626]

[0.23161476 0.72624967 0.04213557]

[0.04038005 0.01068884 0.94893112]]





precision recall f1-score support

hate 0.19 0.75 0.31 294

offensive 0.98 0.73 0.84 3821

neither 0.80 0.95 0.87 842

accuracy 0.77 4957

macro avg 0.66 0.81 0.67 4957

weighted avg 0.91 0.77 0.81 4957

average eval\_loss: 0.5975794280248303, accuracy: 0.765584022594311

evaluating adversary...

0.5

training 1 epoch...

44.990957498550415 seconds used......

1 training finished! average train loss: 0.8417199656628725

total clf loss: 109.74452693760395 total adv loss: 68.37917505204678

evaluating...

Confusion matrix, without normalization

[[ 216 44 34]

[ 757 2895 169]

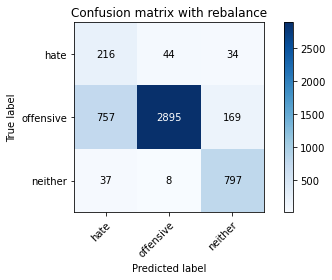
[ 37 8 797]]

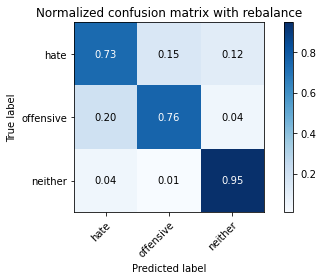
Normalized confusion matrix

[[0.73469388 0.14965986 0.11564626]

[0.19811568 0.75765506 0.04422926]

[0.04394299 0.00950119 0.94655582]]





precision recall f1-score support

hate 0.21 0.73 0.33 294

offensive 0.98 0.76 0.86 3821

neither 0.80 0.95 0.87 842

accuracy 0.79 4957

macro avg 0.66 0.81 0.68 4957

weighted avg 0.91 0.79 0.83 4957

average eval\_loss: 0.5003707727357265, accuracy: 0.7883800685898729

evaluating adversary...

0.6000000000000001

training 1 epoch...

44.90506291389465 seconds used......

1 training finished! average train loss: 0.8951162371719092

total clf loss: 110.8904395699501 total adv loss: 70.29072579741478

evaluating...

Confusion matrix, without normalization

[[ 232 40 22]

[ 823 2899 99]

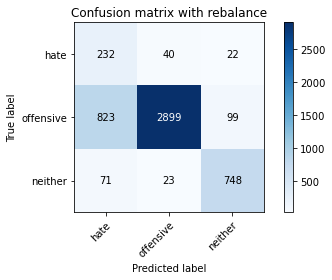
[ 71 23 748]]

Normalized confusion matrix

[[0.78911565 0.13605442 0.07482993]

[0.21538864 0.75870191 0.02590945]

[0.08432304 0.02731591 0.88836105]]





precision recall f1-score support

hate 0.21 0.79 0.33 294

offensive 0.98 0.76 0.85 3821

neither 0.86 0.89 0.87 842

accuracy 0.78 4957

macro avg 0.68 0.81 0.69 4957

weighted avg 0.91 0.78 0.83 4957

average eval\_loss: 0.48442621745409503, accuracy: 0.7825297559007465

evaluating adversary...

0.7000000000000001

training 1 epoch...

44.99803423881531 seconds used......

1 training finished! average train loss: 0.9326643438367118

total clf loss: 110.93433827161789 total adv loss: 69.35895095020533

evaluating...

Confusion matrix, without normalization

[[ 194 68 32]

[ 514 3182 125]

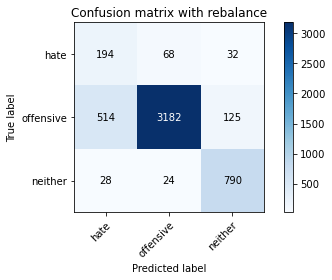
[ 28 24 790]]

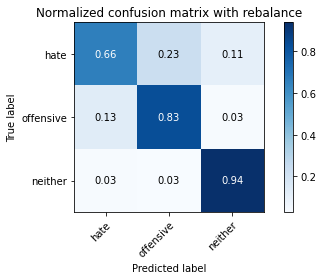
Normalized confusion matrix

[[0.65986395 0.23129252 0.10884354]

[0.13451976 0.83276629 0.03271395]

[0.03325416 0.02850356 0.93824228]]





precision recall f1-score support

hate 0.26 0.66 0.38 294

offensive 0.97 0.83 0.90 3821

neither 0.83 0.94 0.88 842

accuracy 0.84 4957

macro avg 0.69 0.81 0.72 4957

weighted avg 0.91 0.84 0.86 4957

average eval\_loss: 0.3748401150167469, accuracy: 0.8404276780310672

evaluating adversary...

0.8

training 1 epoch...

44.8702118396759 seconds used......

1 training finished! average train loss: 1.0028993572756562

total clf loss: 114.30615510046482 total adv loss: 71.48704186826944

evaluating...

Confusion matrix, without normalization

[[ 224 37 33]

[ 799 2869 153]

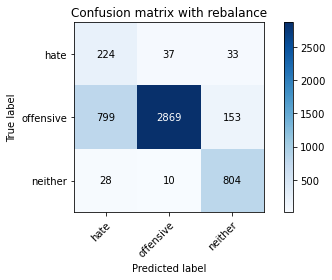
[ 28 10 804]]

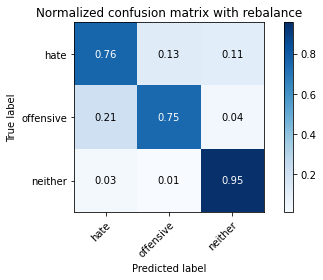
Normalized confusion matrix

[[0.76190476 0.12585034 0.1122449 ]

[0.20910756 0.75085056 0.04004187]

[0.03325416 0.01187648 0.95486936]]





precision recall f1-score support

hate 0.21 0.76 0.33 294

offensive 0.98 0.75 0.85 3821

neither 0.81 0.95 0.88 842

accuracy 0.79 4957

macro avg 0.67 0.82 0.69 4957

weighted avg 0.91 0.79 0.83 4957

average eval\_loss: 0.5248051851747497, accuracy: 0.7861609844664111

evaluating adversary...

0.9

training 1 epoch...

45.06028985977173 seconds used......

1 training finished! average train loss: 1.0084223585170613

total clf loss: 110.17170049250126 total adv loss: 69.1872486397624

evaluating...

Confusion matrix, without normalization

[[ 214 45 35]

[ 750 2916 155]

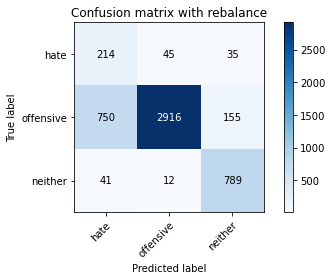
[ 41 12 789]]

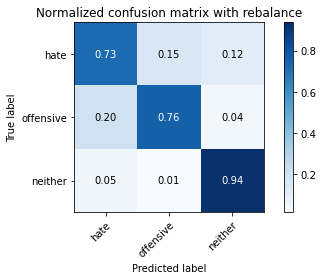
Normalized confusion matrix

[[0.72789116 0.15306122 0.11904762]

[0.1962837 0.76315101 0.0405653 ]

[0.04869359 0.01425178 0.93705463]]





precision recall f1-score support

hate 0.21 0.73 0.33 294

offensive 0.98 0.76 0.86 3821

neither 0.81 0.94 0.87 842

accuracy 0.79 4957

macro avg 0.67 0.81 0.68 4957

weighted avg 0.91 0.79 0.83 4957

average eval\_loss: 0.4503089730898219, accuracy: 0.7905991527133347

evaluating adversary...

1.0

training 1 epoch...

45.05418848991394 seconds used......

1 training finished! average train loss: 1.10404594930989

total clf loss: 118.07286316156387 total adv loss: 70.71899365633726

evaluating...

Confusion matrix, without normalization

[[ 203 53 38]

[ 608 3068 145]

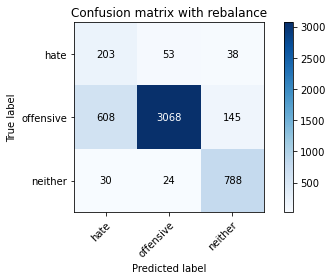
[ 30 24 788]]

Normalized confusion matrix

[[0.69047619 0.18027211 0.1292517 ]

[0.15912065 0.80293117 0.03794818]

[0.03562945 0.02850356 0.93586698]]





precision recall f1-score support

hate 0.24 0.69 0.36 294

offensive 0.98 0.80 0.88 3821

neither 0.81 0.94 0.87 842

accuracy 0.82 4957

macro avg 0.68 0.81 0.70 4957

weighted avg 0.90 0.82 0.85 4957

average eval\_loss: 0.41925681839066165, accuracy: 0.8188420415573936

evaluating adversary...

1.1

training 1 epoch...

44.993661642074585 seconds used......

1 training finished! average train loss: 1.1703093344705147

total clf loss: 119.02334377169609 total adv loss: 73.72686459869146

evaluating...

Confusion matrix, without normalization

[[ 194 68 32]

[ 514 3191 116]

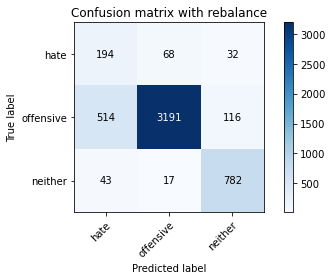
[ 43 17 782]]

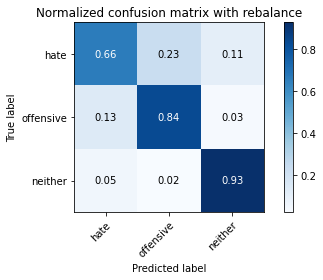
Normalized confusion matrix

[[0.65986395 0.23129252 0.10884354]

[0.13451976 0.8351217 0.03035854]

[0.05106888 0.02019002 0.92874109]]





precision recall f1-score support

hate 0.26 0.66 0.37 294

offensive 0.97 0.84 0.90 3821

neither 0.84 0.93 0.88 842

accuracy 0.84 4957

macro avg 0.69 0.81 0.72 4957

weighted avg 0.91 0.84 0.87 4957

average eval\_loss: 0.38902170730814817, accuracy: 0.8406294129513819

evaluating adversary...

1.2000000000000002

training 1 epoch...

44.92350125312805 seconds used......

1 training finished! average train loss: 1.1866891410267144

total clf loss: 116.29667890071869 total adv loss: 72.18930097669363

evaluating...

Confusion matrix, without normalization

[[ 216 42 36]

[ 766 2923 132]

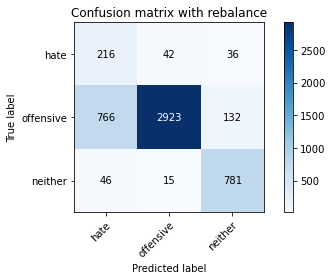
[ 46 15 781]]

Normalized confusion matrix

[[0.73469388 0.14285714 0.12244898]

[0.20047108 0.76498299 0.03454593]

[0.05463183 0.01781473 0.92755344]]





precision recall f1-score support

hate 0.21 0.73 0.33 294

offensive 0.98 0.76 0.86 3821

neither 0.82 0.93 0.87 842

accuracy 0.79 4957

macro avg 0.67 0.81 0.69 4957

weighted avg 0.91 0.79 0.83 4957

average eval\_loss: 0.5284906537422249, accuracy: 0.7908008876336494

evaluating adversary...

1.3

training 1 epoch...

44.912662744522095 seconds used......

1 training finished! average train loss: 1.2286379351253398

total clf loss: 116.84404085576534 total adv loss: 71.73311492055655

evaluating...

Confusion matrix, without normalization

[[ 217 41 36]

[ 782 2891 148]

[ 36 9 797]]

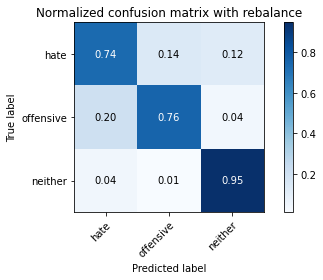
Normalized confusion matrix

[[0.73809524 0.13945578 0.12244898]

[0.20465847 0.75660822 0.03873332]

[0.04275534 0.01068884 0.94655582]]





precision recall f1-score support

hate 0.21 0.74 0.33 294

offensive 0.98 0.76 0.86 3821

neither 0.81 0.95 0.87 842

accuracy 0.79 4957

macro avg 0.67 0.81 0.69 4957

weighted avg 0.91 0.79 0.83 4957

average eval\_loss: 0.46932620543145365, accuracy: 0.7877748638289288

evaluating adversary...

1.4000000000000001

training 1 epoch...

44.97136211395264 seconds used......

1 training finished! average train loss: 1.2508488015124672

total clf loss: 116.3555543422699 total adv loss: 69.67113684862852

evaluating...

Confusion matrix, without normalization

[[ 240 31 23]

[ 999 2700 122]

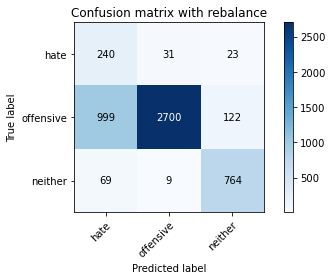
[ 69 9 764]]

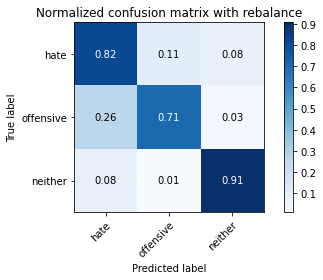
Normalized confusion matrix

[[0.81632653 0.10544218 0.07823129]

[0.26144988 0.7066213 0.03192881]

[0.08194774 0.01068884 0.90736342]]





precision recall f1-score support

hate 0.18 0.82 0.30 294

offensive 0.99 0.71 0.82 3821

neither 0.84 0.91 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: 0.6186091103741238, accuracy: 0.7472261448456727

evaluating adversary...

1.5

training 1 epoch...

44.86948275566101 seconds used......

1 training finished! average train loss: 1.330164281771197

total clf loss: 118.49133463203907 total adv loss: 72.64450445026159

evaluating...

Confusion matrix, without normalization

[[ 189 68 37]

[ 466 3212 143]

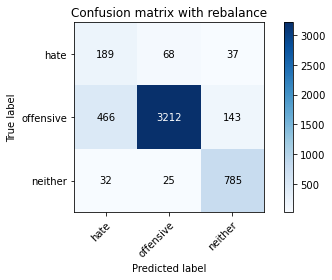
[ 32 25 785]]

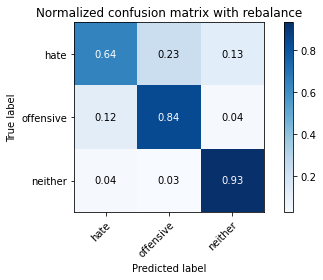
Normalized confusion matrix

[[0.64285714 0.23129252 0.12585034]

[0.1219576 0.84061764 0.03742476]

[0.03800475 0.02969121 0.93230404]]





precision recall f1-score support

hate 0.28 0.64 0.39 294

offensive 0.97 0.84 0.90 3821

neither 0.81 0.93 0.87 842

accuracy 0.84 4957

macro avg 0.69 0.81 0.72 4957

weighted avg 0.90 0.84 0.87 4957

average eval\_loss: 0.45188820091707094, accuracy: 0.8444623764373613

evaluating adversary...

1.6

training 1 epoch...

44.91610026359558 seconds used......

1 training finished! average train loss: 1.372889682563425

total clf loss: 119.08833876252174 total adv loss: 72.29737222194672

evaluating...

Confusion matrix, without normalization

[[ 241 31 22]

[1034 2710 77]

[ 122 12 708]]

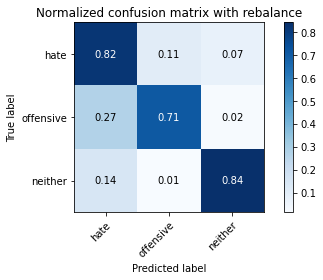
Normalized confusion matrix

[[0.81972789 0.10544218 0.07482993]

[0.27060979 0.70923842 0.02015179]

[0.14489311 0.01425178 0.84085511]]





precision recall f1-score support

hate 0.17 0.82 0.29 294

offensive 0.98 0.71 0.82 3821

neither 0.88 0.84 0.86 842

accuracy 0.74 4957

macro avg 0.68 0.79 0.66 4957

weighted avg 0.92 0.74 0.80 4957

average eval\_loss: 0.6207005064333638, accuracy: 0.738148073431511

evaluating adversary...

1.7000000000000002

training 1 epoch...

44.806339263916016 seconds used......

1 training finished! average train loss: 1.3838657144217463

total clf loss: 116.9421549141407 total adv loss: 70.41110567003489

evaluating...

Confusion matrix, without normalization

[[ 245 29 20]

[1206 2520 95]

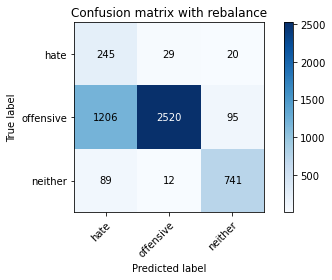
[ 89 12 741]]

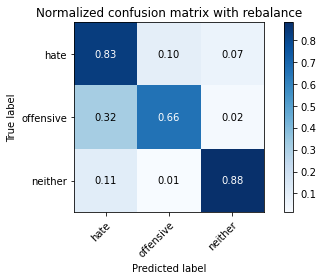
Normalized confusion matrix

[[0.83333333 0.09863946 0.06802721]

[0.31562418 0.65951322 0.0248626 ]

[0.10570071 0.01425178 0.88004751]]





precision recall f1-score support

hate 0.16 0.83 0.27 294

offensive 0.98 0.66 0.79 3821

neither 0.87 0.88 0.87 842

accuracy 0.71 4957

macro avg 0.67 0.79 0.64 4957

weighted avg 0.91 0.71 0.77 4957

average eval\_loss: 0.7262627155550064, accuracy: 0.7072826306233609

evaluating adversary...

1.8

training 1 epoch...

44.93340015411377 seconds used......

1 training finished! average train loss: 1.4520840962030734

total clf loss: 117.40905480086803 total adv loss: 72.72073844075203

evaluating...

Confusion matrix, without normalization

[[ 233 36 25]

[ 858 2846 117]

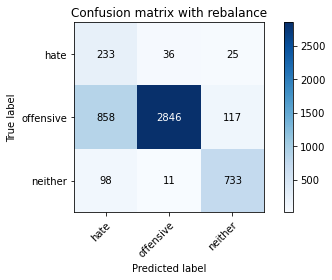
[ 98 11 733]]

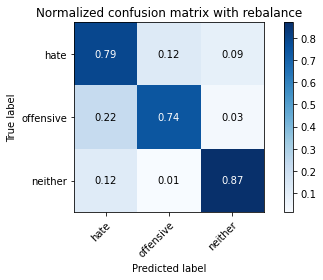
Normalized confusion matrix

[[0.79251701 0.12244898 0.08503401]

[0.22454855 0.7448312 0.03062026]

[0.11638955 0.01306413 0.87054632]]





precision recall f1-score support

hate 0.20 0.79 0.31 294

offensive 0.98 0.74 0.85 3821

neither 0.84 0.87 0.85 842

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: 0.5535300858318806, accuracy: 0.7690135162396611

evaluating adversary...

1.9000000000000001

training 1 epoch...

44.96239709854126 seconds used......

1 training finished! average train loss: 1.4732174815838797

total clf loss: 117.77795495092869 total adv loss: 70.60117656737566

evaluating...

Confusion matrix, without normalization

[[ 181 73 40]

[ 399 3269 153]

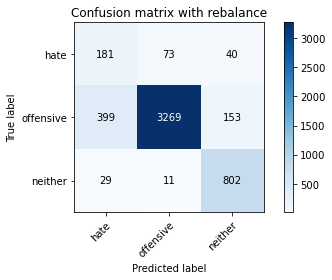
[ 29 11 802]]

Normalized confusion matrix

[[0.61564626 0.24829932 0.13605442]

[0.10442293 0.8555352 0.04004187]

[0.03444181 0.01306413 0.95249406]]





precision recall f1-score support

hate 0.30 0.62 0.40 294

offensive 0.97 0.86 0.91 3821

neither 0.81 0.95 0.87 842

accuracy 0.86 4957

macro avg 0.69 0.81 0.73 4957

weighted avg 0.91 0.86 0.87 4957

average eval\_loss: 0.42025790862258405, accuracy: 0.8577768811781319

evaluating adversary...

2.0

training 1 epoch...

44.94124889373779 seconds used......

1 training finished! average train loss: 1.5052392308474982

total clf loss: 117.0698324739933 total adv loss: 70.16303783655167

evaluating...

Confusion matrix, without normalization

[[ 207 49 38]

[ 677 2990 154]

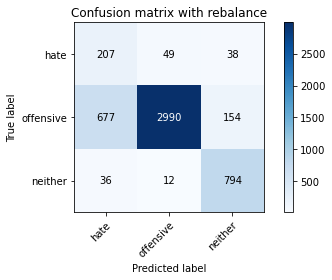
[ 36 12 794]]

Normalized confusion matrix

[[0.70408163 0.16666667 0.1292517 ]

[0.17717875 0.78251767 0.04030359]

[0.04275534 0.01425178 0.94299287]]





precision recall f1-score support

hate 0.23 0.70 0.34 294

offensive 0.98 0.78 0.87 3821

neither 0.81 0.94 0.87 842

accuracy 0.81 4957

macro avg 0.67 0.81 0.69 4957

weighted avg 0.91 0.81 0.84 4957

average eval\_loss: 0.4813600376008018, accuracy: 0.8051240669759936

evaluating adversary...

2.1

training 1 epoch...

44.9605610370636 seconds used......

1 training finished! average train loss: 1.556324470287178

total clf loss: 117.80319851636887 total adv loss: 70.63252053409815

evaluating...

Confusion matrix, without normalization

[[ 219 40 35]

[ 805 2896 120]

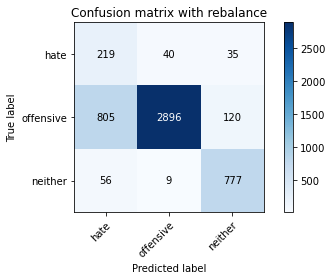
[ 56 9 777]]

Normalized confusion matrix

[[0.74489796 0.13605442 0.11904762]

[0.21067783 0.75791678 0.03140539]

[0.06650831 0.01068884 0.92280285]]





precision recall f1-score support

hate 0.20 0.74 0.32 294

offensive 0.98 0.76 0.86 3821

neither 0.83 0.92 0.88 842

accuracy 0.79 4957

macro avg 0.67 0.81 0.68 4957

weighted avg 0.91 0.79 0.83 4957

average eval\_loss: 0.5164903335032924, accuracy: 0.7851523098648376

evaluating adversary...

2.2

training 1 epoch...

45.019880056381226 seconds used......

1 training finished! average train loss: 1.6440076524751228

total clf loss: 124.00129280984402 total adv loss: 71.42000576481223

evaluating...

Confusion matrix, without normalization

[[ 200 67 27]

[ 529 3200 92]

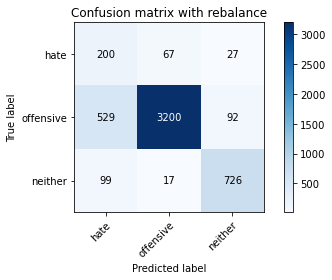
[ 99 17 726]]

Normalized confusion matrix

[[0.68027211 0.22789116 0.09183673]

[0.13844543 0.8374771 0.02407747]

[0.1175772 0.02019002 0.86223278]]





precision recall f1-score support

hate 0.24 0.68 0.36 294

offensive 0.97 0.84 0.90 3821

neither 0.86 0.86 0.86 842

accuracy 0.83 4957

macro avg 0.69 0.79 0.71 4957

weighted avg 0.91 0.83 0.86 4957

average eval\_loss: 0.4418493865598594, accuracy: 0.8323582812184789

evaluating adversary...

2.3000000000000003

training 1 epoch...

44.99794101715088 seconds used......

1 training finished! average train loss: 1.6185215450169748

total clf loss: 117.21952264010906 total adv loss: 69.36855005472898

evaluating...

Confusion matrix, without normalization

[[ 220 40 34]

[ 762 2910 149]

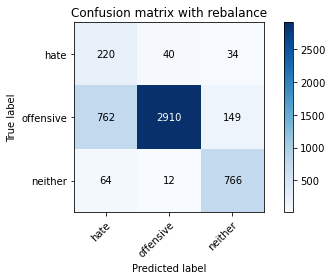
[ 64 12 766]]

Normalized confusion matrix

[[0.74829932 0.13605442 0.11564626]

[0.19942423 0.76158074 0.03899503]

[0.0760095 0.01425178 0.90973872]]





precision recall f1-score support

hate 0.21 0.75 0.33 294

offensive 0.98 0.76 0.86 3821

neither 0.81 0.91 0.86 842

accuracy 0.79 4957

macro avg 0.67 0.81 0.68 4957

weighted avg 0.91 0.79 0.83 4957

average eval\_loss: 0.5024529455770408, accuracy: 0.7859592495460964

evaluating adversary...

2.4000000000000004

training 1 epoch...

45.039764165878296 seconds used......

1 training finished! average train loss: 1.7674111312592935

total clf loss: 124.36855378746986 total adv loss: 74.10780879110098

evaluating...

Confusion matrix, without normalization

[[ 222 45 27]

[ 761 2946 114]

[ 108 13 721]]

Normalized confusion matrix

[[0.75510204 0.15306122 0.09183673]

[0.19916252 0.77100236 0.02983512]

[0.12826603 0.01543943 0.85629454]]





precision recall f1-score support

hate 0.20 0.76 0.32 294

offensive 0.98 0.77 0.86 3821

neither 0.84 0.86 0.85 842

accuracy 0.78 4957

macro avg 0.67 0.79 0.68 4957

weighted avg 0.91 0.78 0.83 4957

average eval\_loss: 0.4979179607884538, accuracy: 0.7845471051038935

evaluating adversary...

2.5

training 1 epoch...

44.90902352333069 seconds used......

1 training finished! average train loss: 1.7265440839075903

total clf loss: 121.16485054790974 total adv loss: 69.62967497855425

evaluating...

Confusion matrix, without normalization

[[ 224 36 34]

[ 845 2860 116]

[ 69 12 761]]

Normalized confusion matrix

[[0.76190476 0.12244898 0.11564626]

[0.2211463 0.74849516 0.03035854]

[0.08194774 0.01425178 0.90380048]]





precision recall f1-score support

hate 0.20 0.76 0.31 294

offensive 0.98 0.75 0.85 3821

neither 0.84 0.90 0.87 842

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: 0.5051771077297388, accuracy: 0.7756707686100464

evaluating adversary...

2.6

training 1 epoch...

44.92484927177429 seconds used......

1 training finished! average train loss: 1.768993946195346

total clf loss: 119.16650441288948 total adv loss: 70.51210306584835

evaluating...

Confusion matrix, without normalization

[[ 204 56 34]

[ 610 3060 151]

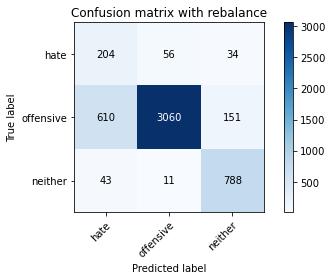
[ 43 11 788]]

Normalized confusion matrix

[[0.69387755 0.19047619 0.11564626]

[0.15964407 0.80083748 0.03951845]

[0.05106888 0.01306413 0.93586698]]





precision recall f1-score support

hate 0.24 0.69 0.35 294

offensive 0.98 0.80 0.88 3821

neither 0.81 0.94 0.87 842

accuracy 0.82 4957

macro avg 0.68 0.81 0.70 4957

weighted avg 0.91 0.82 0.85 4957

average eval\_loss: 0.4560034534623546, accuracy: 0.8174298971151907

evaluating adversary...

2.7

training 1 epoch...

44.90429759025574 seconds used......

1 training finished! average train loss: 1.828648268827918

total clf loss: 122.7169280052185 total adv loss: 70.36367528513074

evaluating...

Confusion matrix, without normalization

[[ 151 85 58]

[ 245 3367 209]

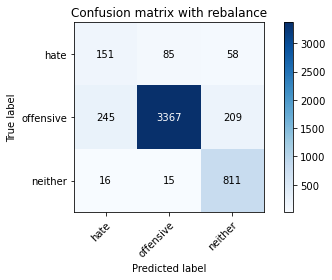
[ 16 15 811]]

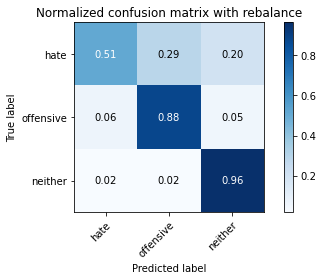
Normalized confusion matrix

[[0.51360544 0.28911565 0.19727891]

[0.06411934 0.88118294 0.05469772]

[0.01900238 0.01781473 0.9631829 ]]





precision recall f1-score support

hate 0.37 0.51 0.43 294

offensive 0.97 0.88 0.92 3821

neither 0.75 0.96 0.84 842

accuracy 0.87 4957

macro avg 0.70 0.79 0.73 4957

weighted avg 0.90 0.87 0.88 4957

average eval\_loss: 0.3673440700456981, accuracy: 0.8733104700423643

evaluating adversary...

2.8000000000000003

training 1 epoch...

44.934221506118774 seconds used......

1 training finished! average train loss: 1.821565546138942

total clf loss: 117.49830581247807 total adv loss: 69.28193003684282

evaluating...

Confusion matrix, without normalization

[[ 231 37 26]

[ 879 2828 114]

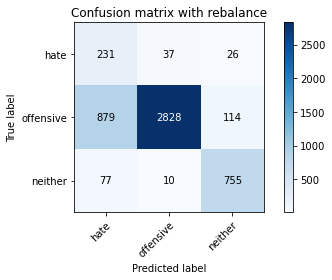
[ 77 10 755]]

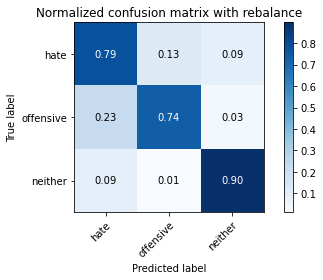
Normalized confusion matrix

[[0.78571429 0.12585034 0.08843537]

[0.23004449 0.74012039 0.02983512]

[0.09144893 0.01187648 0.89667458]]





precision recall f1-score support

hate 0.19 0.79 0.31 294

offensive 0.98 0.74 0.84 3821

neither 0.84 0.90 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: 0.5174441545480682, accuracy: 0.7694169860802905

evaluating adversary...

2.9000000000000004

training 1 epoch...

44.93741583824158 seconds used......

1 training finished! average train loss: 1.935004712893949

total clf loss: 121.26740658283234 total adv loss: 72.28220350295305

evaluating...

Confusion matrix, without normalization

[[ 213 39 42]

[ 766 2915 140]

[ 31 13 798]]

Normalized confusion matrix

[[0.7244898 0.13265306 0.14285714]

[0.20047108 0.7628893 0.03663962]

[0.0368171 0.01543943 0.94774347]]





precision recall f1-score support

hate 0.21 0.72 0.33 294

offensive 0.98 0.76 0.86 3821

neither 0.81 0.95 0.88 842

accuracy 0.79 4957

macro avg 0.67 0.81 0.69 4957

weighted avg 0.91 0.79 0.83 4957

average eval\_loss: 0.4790528680408193, accuracy: 0.7920112971555376

evaluating adversary…