1 run

hate as no-offensive

alpha=2

training 1 epoch...

107.6885232925415 seconds used......

1 training finished! average train loss: 1.475420523265546

total clf loss: 118.62265515327454 total adv loss: 70.5256778113544

evaluating...

Confusion matrix, without normalization

[[ 106 13 15]

[ 451 1440 46]

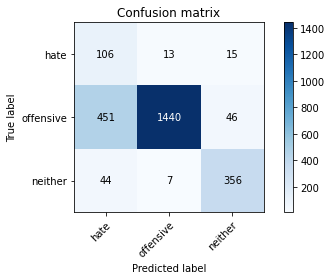
[ 44 7 356]]

Normalized confusion matrix

[[0.79104478 0.09701493 0.1119403 ]

[0.23283428 0.74341766 0.02374806]

[0.10810811 0.01719902 0.87469287]]





precision recall f1-score support

hate 0.18 0.79 0.29 134

offensive 0.99 0.74 0.85 1937

neither 0.85 0.87 0.86 407

accuracy 0.77 2478

macro avg 0.67 0.80 0.67 2478

weighted avg 0.92 0.77 0.82 2478

average eval\_loss: 0.5209590040387646, accuracy: 0.7675544794188862

training 2 epoch...

106.77542352676392 seconds used......

2 training finished! average train loss: 0.9655564549294385

total clf loss: 72.2141400501132 total adv loss: 48.86189803108573

evaluating...

Confusion matrix, without normalization

[[ 101 17 16]

[ 360 1517 60]

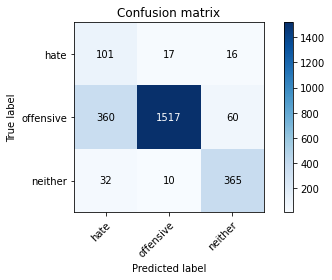
[ 32 10 365]]

Normalized confusion matrix

[[0.75373134 0.12686567 0.11940299]

[0.18585441 0.78316985 0.03097574]

[0.07862408 0.02457002 0.8968059 ]]





precision recall f1-score support

hate 0.20 0.75 0.32 134

offensive 0.98 0.78 0.87 1937

neither 0.83 0.90 0.86 407

accuracy 0.80 2478

macro avg 0.67 0.81 0.68 2478

weighted avg 0.92 0.80 0.84 2478

average eval\_loss: 0.4470233919639741, accuracy: 0.8002421307506054

training 3 epoch...

106.74873185157776 seconds used......

3 training finished! average train loss: 0.6620229037309234

total clf loss: 50.46455197036266 total adv loss: 33.02573946677148

evaluating...

Confusion matrix, without normalization

[[ 87 28 19]

[ 265 1599 73]

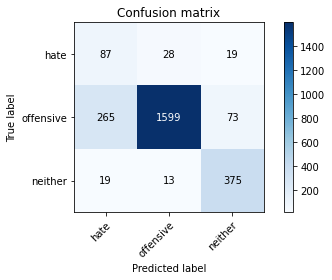
[ 19 13 375]]

Normalized confusion matrix

[[0.64925373 0.20895522 0.14179104]

[0.1368095 0.82550336 0.03768715]

[0.04668305 0.03194103 0.92137592]]





precision recall f1-score support

hate 0.23 0.65 0.34 134

offensive 0.97 0.83 0.89 1937

neither 0.80 0.92 0.86 407

accuracy 0.83 2478

macro avg 0.67 0.80 0.70 2478

weighted avg 0.91 0.83 0.86 2478

average eval\_loss: 0.5016408336739386, accuracy: 0.8317191283292978

training 4 epoch...

106.87634325027466 seconds used......

4 training finished! average train loss: 0.37974258309060877

total clf loss: 29.325000582262874 total adv loss: 18.754847193369642

evaluating...

Confusion matrix, without normalization

[[ 94 22 18]

[ 294 1573 70]

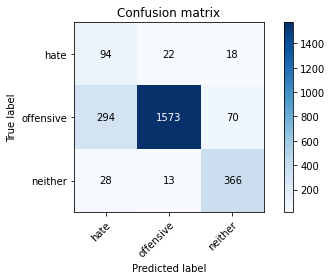
[ 28 13 366]]

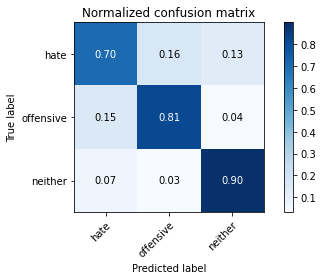
Normalized confusion matrix

[[0.70149254 0.1641791 0.13432836]

[0.1517811 0.81208054 0.03613836]

[0.06879607 0.03194103 0.8992629 ]]





precision recall f1-score support

hate 0.23 0.70 0.34 134

offensive 0.98 0.81 0.89 1937

neither 0.81 0.90 0.85 407

accuracy 0.82 2478

macro avg 0.67 0.80 0.69 2478

weighted avg 0.91 0.82 0.85 2478

average eval\_loss: 0.5552974749176253, accuracy: 0.8204196933010492

training 5 epoch...

106.77687478065491 seconds used......

5 training finished! average train loss: 0.23413138444514267

total clf loss: 17.1394905555062 total adv loss: 12.033816701296018

evaluating...

Confusion matrix, without normalization

[[ 103 14 17]

[ 482 1367 88]

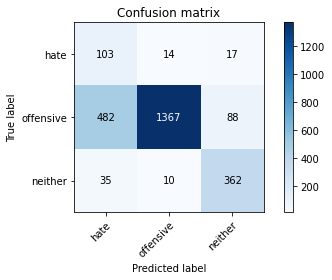
[ 35 10 362]]

Normalized confusion matrix

[[0.76865672 0.10447761 0.12686567]

[0.24883841 0.70573051 0.04543108]

[0.08599509 0.02457002 0.88943489]]





precision recall f1-score support

hate 0.17 0.77 0.27 134

offensive 0.98 0.71 0.82 1937

neither 0.78 0.89 0.83 407

accuracy 0.74 2478

macro avg 0.64 0.79 0.64 2478

weighted avg 0.90 0.74 0.79 2478

average eval\_loss: 0.9855633670043561, accuracy: 0.7393058918482648

model of 1 epoch

Confusion matrix, without normalization

[[ 94 23 7]

[ 425 1462 55]

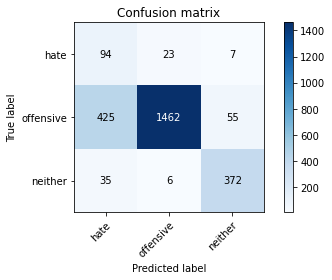
[ 35 6 372]]

Normalized confusion matrix

[[0.75806452 0.18548387 0.05645161]

[0.21884655 0.75283213 0.02832132]

[0.08474576 0.01452785 0.90072639]]





precision recall f1-score support

hate 0.17 0.76 0.28 124

offensive 0.98 0.75 0.85 1942

neither 0.86 0.90 0.88 413

accuracy 0.78 2479

macro avg 0.67 0.80 0.67 2479

weighted avg 0.92 0.78 0.83 2479

average eval\_loss: 0.5368707510492494, accuracy: 0.7777329568374345

[18]:

0.7777329568374345

2 run

training 1 epoch...

106.67369961738586 seconds used......

1 training finished! average train loss: 1.471313931386579

total clf loss: 119.13852368295193 total adv loss: 69.90636368095875

evaluating...

Confusion matrix, without normalization

[[ 63 44 27]

[ 156 1699 82]

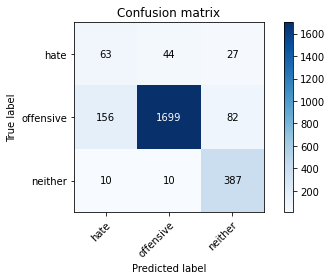
[ 10 10 387]]

Normalized confusion matrix

[[0.47014925 0.32835821 0.20149254]

[0.08053691 0.87712958 0.04233351]

[0.02457002 0.02457002 0.95085995]]





precision recall f1-score support

hate 0.28 0.47 0.35 134

offensive 0.97 0.88 0.92 1937

neither 0.78 0.95 0.86 407

accuracy 0.87 2478

macro avg 0.67 0.77 0.71 2478

weighted avg 0.90 0.87 0.88 2478

average eval\_loss: 0.3959970881621684, accuracy: 0.867231638418079

training 2 epoch...

106.68524670600891 seconds used......

2 training finished! average train loss: 0.9729205397042361

total clf loss: 72.83006901293993 total adv loss: 49.20197331905365

evaluating...

Confusion matrix, without normalization

[[ 106 11 17]

[ 442 1431 64]

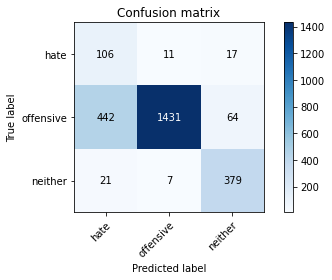
[ 21 7 379]]

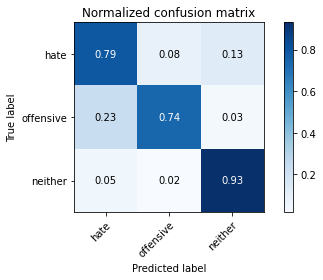
Normalized confusion matrix

[[0.79104478 0.08208955 0.12686567]

[0.22818792 0.7387713 0.03304078]

[0.05159705 0.01719902 0.93120393]]





precision recall f1-score support

hate 0.19 0.79 0.30 134

offensive 0.99 0.74 0.85 1937

neither 0.82 0.93 0.87 407

accuracy 0.77 2478

macro avg 0.67 0.82 0.67 2478

weighted avg 0.92 0.77 0.82 2478

average eval\_loss: 0.5458313203867404, accuracy: 0.7732041969330105

training 3 epoch...

106.71107149124146 seconds used......

3 training finished! average train loss: 0.714224742598493

total clf loss: 52.6131141372025 total adv loss: 36.54522046633065

evaluating...

Confusion matrix, without normalization

[[ 101 19 14]

[ 334 1537 66]

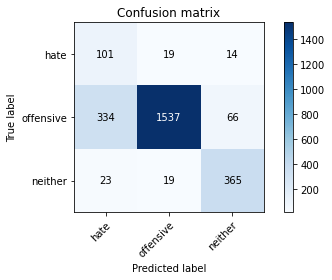
[ 23 19 365]]

Normalized confusion matrix

[[0.75373134 0.14179104 0.10447761]

[0.1724316 0.7934951 0.03407331]

[0.05651106 0.04668305 0.8968059 ]]





precision recall f1-score support

hate 0.22 0.75 0.34 134

offensive 0.98 0.79 0.88 1937

neither 0.82 0.90 0.86 407

accuracy 0.81 2478

macro avg 0.67 0.81 0.69 2478

weighted avg 0.91 0.81 0.84 2478

average eval\_loss: 0.4867505493663972, accuracy: 0.8083131557707829

training 4 epoch...

106.65097451210022 seconds used......

4 training finished! average train loss: 0.438775188755244

total clf loss: 33.24585161358118 total adv loss: 21.989290863275528

evaluating...

Confusion matrix, without normalization

[[ 83 32 19]

[ 270 1549 118]

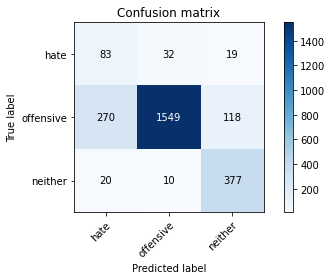
[ 20 10 377]]

Normalized confusion matrix

[[0.61940299 0.23880597 0.14179104]

[0.13939081 0.79969024 0.06091895]

[0.04914005 0.02457002 0.92628993]]





precision recall f1-score support

hate 0.22 0.62 0.33 134

offensive 0.97 0.80 0.88 1937

neither 0.73 0.93 0.82 407

accuracy 0.81 2478

macro avg 0.64 0.78 0.67 2478

weighted avg 0.89 0.81 0.84 2478

average eval\_loss: 0.5910683938812825, accuracy: 0.8107344632768362

training 5 epoch...

106.61352443695068 seconds used......

5 training finished! average train loss: 0.25734941104680975

total clf loss: 19.25615531951189 total adv loss: 13.018670444260351

evaluating...

Confusion matrix, without normalization

[[ 100 21 13]

[ 378 1484 75]

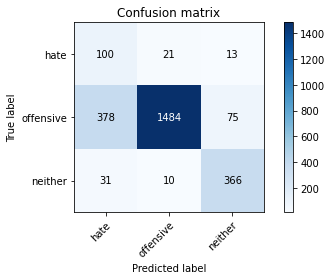
[ 31 10 366]]

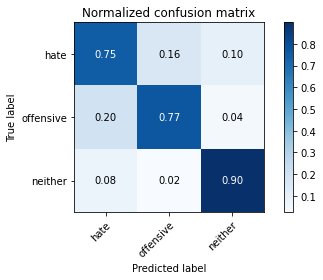
Normalized confusion matrix

[[0.74626866 0.15671642 0.09701493]

[0.19514713 0.7661332 0.03871967]

[0.07616708 0.02457002 0.8992629 ]]





precision recall f1-score support

hate 0.20 0.75 0.31 134

offensive 0.98 0.77 0.86 1937

neither 0.81 0.90 0.85 407

accuracy 0.79 2478

macro avg 0.66 0.80 0.67 2478

weighted avg 0.91 0.79 0.83 2478

average eval\_loss: 0.799176387308586, accuracy: 0.7869249394673123

Confusion matrix, without normalization

[[ 92 23 9]

[ 457 1411 74]

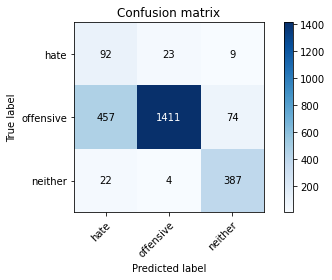
[ 22 4 387]]

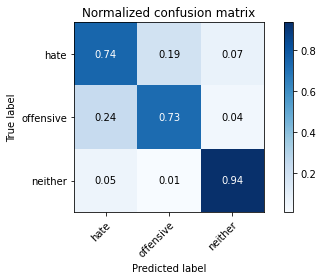
Normalized confusion matrix

[[0.74193548 0.18548387 0.07258065]

[0.23532441 0.72657055 0.03810505]

[0.05326877 0.00968523 0.937046 ]]





precision recall f1-score support

hate 0.16 0.74 0.26 124

offensive 0.98 0.73 0.83 1942

neither 0.82 0.94 0.88 413

accuracy 0.76 2479

macro avg 0.66 0.80 0.66 2479

weighted avg 0.91 0.76 0.81 2479

average eval\_loss: 0.5854892270459283, accuracy: 0.7624041952400161

[31]:

0.7624041952400161

3 run

training 1 epoch...

106.57259106636047 seconds used......

1 training finished! average train loss: 1.4419106326319955

total clf loss: 116.74500209838152 total adv loss: 68.51563476771116

evaluating...

Confusion matrix, without normalization

[[ 76 41 17]

[ 155 1740 42]

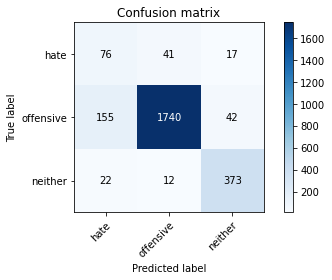
[ 22 12 373]]

Normalized confusion matrix

[[0.56716418 0.30597015 0.12686567]

[0.08002065 0.89829633 0.02168301]

[0.05405405 0.02948403 0.91646192]]





precision recall f1-score support

hate 0.30 0.57 0.39 134

offensive 0.97 0.90 0.93 1937

neither 0.86 0.92 0.89 407

accuracy 0.88 2478

macro avg 0.71 0.79 0.74 2478

weighted avg 0.92 0.88 0.90 2478

average eval\_loss: 0.34148448417263644, accuracy: 0.8833736884584342

training 2 epoch...

106.58604693412781 seconds used......

2 training finished! average train loss: 0.9392201047898694

total clf loss: 69.9055258333683 total adv loss: 47.69860624521971

evaluating...

Confusion matrix, without normalization

[[ 84 29 21]

[ 203 1664 70]

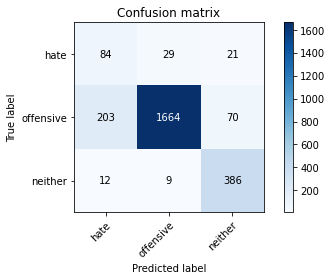
[ 12 9 386]]

Normalized confusion matrix

[[0.62686567 0.21641791 0.15671642]

[0.10480124 0.8590604 0.03613836]

[0.02948403 0.02211302 0.94840295]]





precision recall f1-score support

hate 0.28 0.63 0.39 134

offensive 0.98 0.86 0.91 1937

neither 0.81 0.95 0.87 407

accuracy 0.86 2478

macro avg 0.69 0.81 0.73 2478

weighted avg 0.91 0.86 0.88 2478

average eval\_loss: 0.3875430725034206, accuracy: 0.8611783696529459

training 3 epoch...

106.57074093818665 seconds used......

3 training finished! average train loss: 0.6487953272986818

total clf loss: 47.24722243100405 total adv loss: 33.47037754114717

evaluating...

Confusion matrix, without normalization

[[ 98 19 17]

[ 451 1408 78]

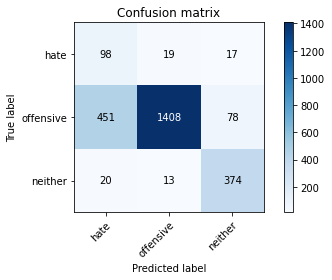
[ 20 13 374]]

Normalized confusion matrix

[[0.73134328 0.14179104 0.12686567]

[0.23283428 0.72689726 0.04026846]

[0.04914005 0.03194103 0.91891892]]





precision recall f1-score support

hate 0.17 0.73 0.28 134

offensive 0.98 0.73 0.83 1937

neither 0.80 0.92 0.85 407

accuracy 0.76 2478

macro avg 0.65 0.79 0.66 2478

weighted avg 0.90 0.76 0.81 2478

average eval\_loss: 0.5781077970900843, accuracy: 0.7586763518966909

training 4 epoch...

106.63913774490356 seconds used......

4 training finished! average train loss: 0.3420089565915987

total clf loss: 26.099847267381847 total adv loss: 17.04686469375156

evaluating...

Confusion matrix, without normalization

[[ 95 25 14]

[ 294 1601 42]

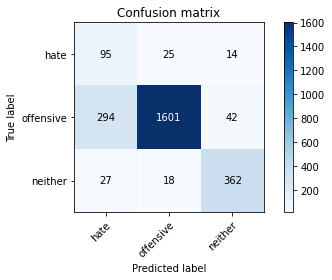
[ 27 18 362]]

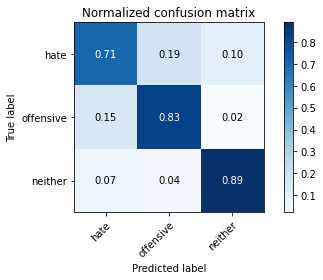
Normalized confusion matrix

[[0.70895522 0.18656716 0.10447761]

[0.1517811 0.82653588 0.02168301]

[0.06633907 0.04422604 0.88943489]]





precision recall f1-score support

hate 0.23 0.71 0.35 134

offensive 0.97 0.83 0.89 1937

neither 0.87 0.89 0.88 407

accuracy 0.83 2478

macro avg 0.69 0.81 0.71 2478

weighted avg 0.92 0.83 0.86 2478

average eval\_loss: 0.5399512004768175, accuracy: 0.8305084745762712

training 5 epoch...

106.64064407348633 seconds used......

5 training finished! average train loss: 0.2106794919130731

total clf loss: 16.081229587085545 total adv loss: 10.499180432700086

evaluating...

Confusion matrix, without normalization

[[ 94 27 13]

[ 284 1589 64]

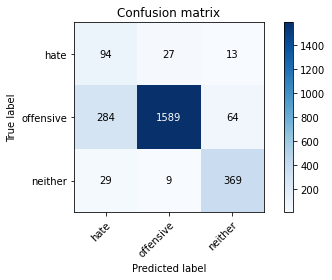
[ 29 9 369]]

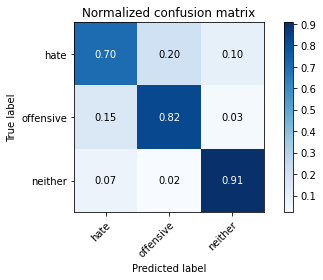
Normalized confusion matrix

[[0.70149254 0.20149254 0.09701493]

[0.14661848 0.82034073 0.03304078]

[0.07125307 0.02211302 0.90663391]]





precision recall f1-score support

hate 0.23 0.70 0.35 134

offensive 0.98 0.82 0.89 1937

neither 0.83 0.91 0.87 407

accuracy 0.83 2478

macro avg 0.68 0.81 0.70 2478

weighted avg 0.91 0.83 0.86 2478

average eval\_loss: 0.7221252065031759, accuracy: 0.8280871670702179

Confusion matrix, without normalization

[[ 90 26 8]

[ 469 1395 78]

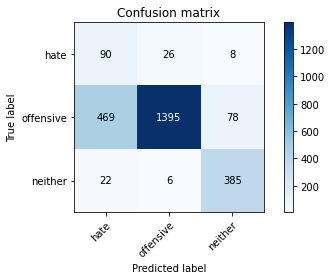
[ 22 6 385]]

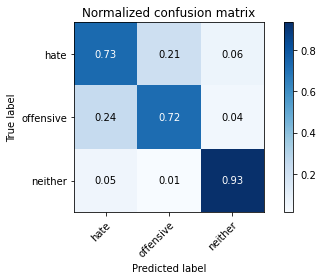
Normalized confusion matrix

[[0.72580645 0.20967742 0.06451613]

[0.2415036 0.71833162 0.04016478]

[0.05326877 0.01452785 0.93220339]]





precision recall f1-score support

hate 0.15 0.73 0.26 124

offensive 0.98 0.72 0.83 1942

neither 0.82 0.93 0.87 413

accuracy 0.75 2479

macro avg 0.65 0.79 0.65 2479

weighted avg 0.91 0.75 0.81 2479

average eval\_loss: 0.5845868718479911, accuracy: 0.754336425978217

[33]:

0.754336425978217

4 run

training 1 epoch...

107.48105669021606 seconds used......

1 training finished! average train loss: 1.4653157905082812

total clf loss: 119.98334963619709 total adv loss: 68.95611403882504

evaluating...

Confusion matrix, without normalization

[[ 85 30 19]

[ 286 1585 66]

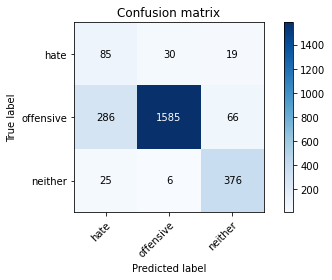
[ 25 6 376]]

Normalized confusion matrix

[[0.63432836 0.2238806 0.14179104]

[0.14765101 0.81827568 0.03407331]

[0.06142506 0.01474201 0.92383292]]





precision recall f1-score support

hate 0.21 0.63 0.32 134

offensive 0.98 0.82 0.89 1937

neither 0.82 0.92 0.87 407

accuracy 0.83 2478

macro avg 0.67 0.79 0.69 2478

weighted avg 0.91 0.83 0.86 2478

average eval\_loss: 0.4549936625745989, accuracy: 0.8256658595641646

training 2 epoch...

106.75632953643799 seconds used......

2 training finished! average train loss: 0.9837645358321342

total clf loss: 74.08707826584578 total adv loss: 49.52773992717266

evaluating...

Confusion matrix, without normalization

[[ 87 31 16]

[ 222 1660 55]

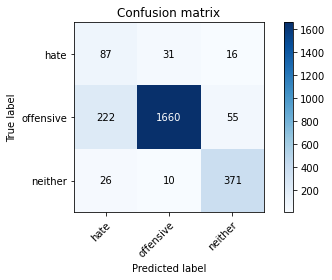
[ 26 10 371]]

Normalized confusion matrix

[[0.64925373 0.23134328 0.11940299]

[0.11461022 0.85699535 0.02839442]

[0.06388206 0.02457002 0.91154791]]





precision recall f1-score support

hate 0.26 0.65 0.37 134

offensive 0.98 0.86 0.91 1937

neither 0.84 0.91 0.87 407

accuracy 0.85 2478

macro avg 0.69 0.81 0.72 2478

weighted avg 0.91 0.85 0.88 2478

average eval\_loss: 0.4119875520587929, accuracy: 0.8547215496368039

training 3 epoch...

106.64335560798645 seconds used......

3 training finished! average train loss: 0.73438297288323

total clf loss: 54.14193505421281 total adv loss: 37.55473418533802

evaluating...

Confusion matrix, without normalization

[[ 86 25 23]

[ 234 1609 94]

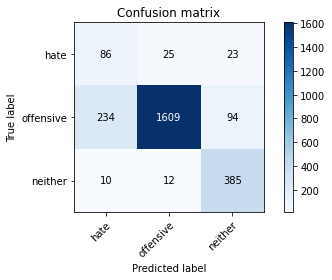
[ 10 12 385]]

Normalized confusion matrix

[[0.64179104 0.18656716 0.17164179]

[0.12080537 0.83066598 0.04852865]

[0.02457002 0.02948403 0.94594595]]





precision recall f1-score support

hate 0.26 0.64 0.37 134

offensive 0.98 0.83 0.90 1937

neither 0.77 0.95 0.85 407

accuracy 0.84 2478

macro avg 0.67 0.81 0.71 2478

weighted avg 0.90 0.84 0.86 2478

average eval\_loss: 0.4606177949256474, accuracy: 0.8393866020984665

training 4 epoch...

106.66119456291199 seconds used......

4 training finished! average train loss: 0.4619512820235369

total clf loss: 34.88153553009033 total adv loss: 23.21094509214163

evaluating...

Confusion matrix, without normalization

[[ 88 27 19]

[ 232 1633 72]

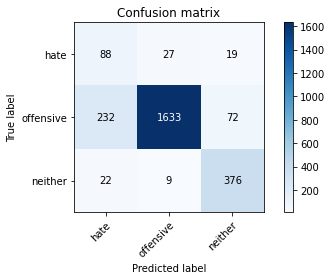
[ 22 9 376]]

Normalized confusion matrix

[[0.65671642 0.20149254 0.14179104]

[0.11977284 0.84305627 0.03717088]

[0.05405405 0.02211302 0.92383292]]





precision recall f1-score support

hate 0.26 0.66 0.37 134

offensive 0.98 0.84 0.91 1937

neither 0.81 0.92 0.86 407

accuracy 0.85 2478

macro avg 0.68 0.81 0.71 2478

weighted avg 0.91 0.85 0.87 2478

average eval\_loss: 0.4823011563670251, accuracy: 0.8462469733656174

training 5 epoch...

106.67514634132385 seconds used......

5 training finished! average train loss: 0.26257600052684377

total clf loss: 19.35468725953251 total adv loss: 13.42934451677138

evaluating...

Confusion matrix, without normalization

[[ 108 15 11]

[ 441 1462 34]

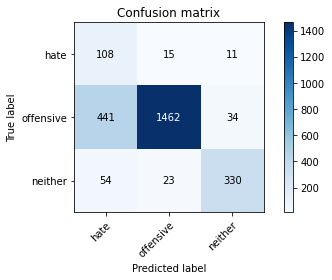
[ 54 23 330]]

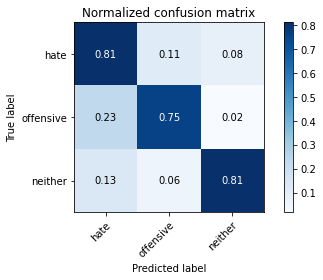
Normalized confusion matrix

[[0.80597015 0.1119403 0.08208955]

[0.22767166 0.75477543 0.01755292]

[0.13267813 0.05651106 0.81081081]]





precision recall f1-score support

hate 0.18 0.81 0.29 134

offensive 0.97 0.75 0.85 1937

neither 0.88 0.81 0.84 407

accuracy 0.77 2478

macro avg 0.68 0.79 0.66 2478

weighted avg 0.92 0.77 0.82 2478

average eval\_loss: 0.7368757937614235, accuracy: 0.7667473769168685

Confusion matrix, without normalization

[[ 100 21 3]

[ 434 1466 42]

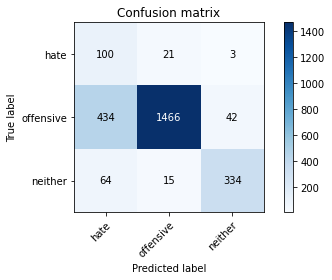
[ 64 15 334]]

Normalized confusion matrix

[[0.80645161 0.16935484 0.02419355]

[0.22348095 0.75489186 0.02162719]

[0.15496368 0.03631961 0.80871671]]





precision recall f1-score support

hate 0.17 0.81 0.28 124

offensive 0.98 0.75 0.85 1942

neither 0.88 0.81 0.84 413

accuracy 0.77 2479

macro avg 0.67 0.79 0.66 2479

weighted avg 0.92 0.77 0.82 2479

average eval\_loss: 0.8156250166075845, accuracy: 0.7664380798709157

[18]:

0.7664380798709157

5 run

training 1 epoch...

106.70979762077332 seconds used......

1 training finished! average train loss: 1.4625852361998775

total clf loss: 118.86088556051254 total adv loss: 69.27705800533295

evaluating...

Confusion matrix, without normalization

[[ 85 28 21]

[ 302 1549 86]

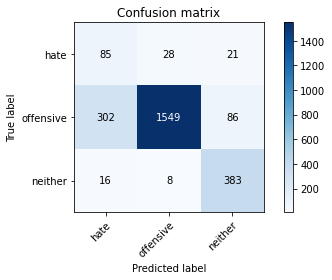
[ 16 8 383]]

Normalized confusion matrix

[[0.63432836 0.20895522 0.15671642]

[0.1559112 0.79969024 0.04439855]

[0.03931204 0.01965602 0.94103194]]





precision recall f1-score support

hate 0.21 0.63 0.32 134

offensive 0.98 0.80 0.88 1937

neither 0.78 0.94 0.85 407

accuracy 0.81 2478

macro avg 0.66 0.79 0.68 2478

weighted avg 0.90 0.81 0.84 2478

average eval\_loss: 0.4886254163519029, accuracy: 0.8139628732849071

training 2 epoch...

106.6011381149292 seconds used......

2 training finished! average train loss: 0.9789400383491408

total clf loss: 73.0968332439661 total adv loss: 49.5983069576323

evaluating...

Confusion matrix, without normalization

[[ 90 25 19]

[ 283 1601 53]

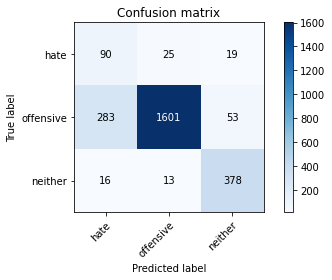
[ 16 13 378]]

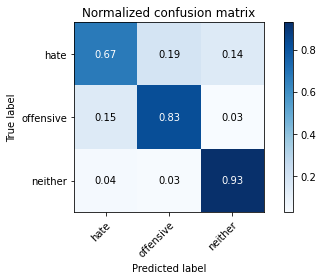
Normalized confusion matrix

[[0.67164179 0.18656716 0.14179104]

[0.14610222 0.82653588 0.0273619 ]

[0.03931204 0.03194103 0.92874693]]





precision recall f1-score support

hate 0.23 0.67 0.34 134

offensive 0.98 0.83 0.90 1937

neither 0.84 0.93 0.88 407

accuracy 0.83 2478

macro avg 0.68 0.81 0.71 2478

weighted avg 0.91 0.83 0.86 2478

average eval\_loss: 0.427713745063351, accuracy: 0.8349475383373689

training 3 epoch...

106.60847878456116 seconds used......

3 training finished! average train loss: 0.7122213637286966

total clf loss: 51.703914284706116 total adv loss: 36.82352289929986

evaluating...

Confusion matrix, without normalization

[[ 96 20 18]

[ 385 1466 86]

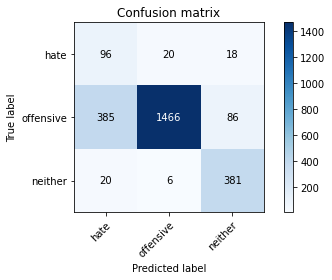
[ 20 6 381]]

Normalized confusion matrix

[[0.71641791 0.14925373 0.13432836]

[0.19876097 0.75684047 0.04439855]

[0.04914005 0.01474201 0.93611794]]





precision recall f1-score support

hate 0.19 0.72 0.30 134

offensive 0.98 0.76 0.86 1937

neither 0.79 0.94 0.85 407

accuracy 0.78 2478

macro avg 0.65 0.80 0.67 2478

weighted avg 0.91 0.78 0.83 2478

average eval\_loss: 0.5832766927778721, accuracy: 0.7841000807102502

training 4 epoch...

106.54098129272461 seconds used......

4 training finished! average train loss: 0.4438891096082939

total clf loss: 33.00282375700772 total adv loss: 22.560829835943878

evaluating...

Confusion matrix, without normalization

[[ 108 15 11]

[ 467 1414 56]

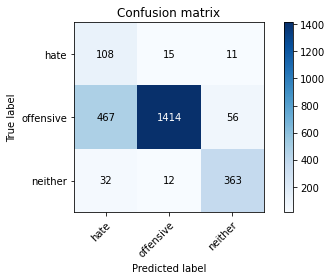
[ 32 12 363]]

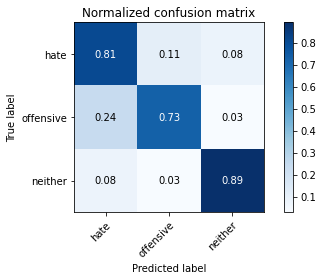
Normalized confusion matrix

[[0.80597015 0.1119403 0.08208955]

[0.24109448 0.72999484 0.02891069]

[0.07862408 0.02948403 0.89189189]]





precision recall f1-score support

hate 0.18 0.81 0.29 134

offensive 0.98 0.73 0.84 1937

neither 0.84 0.89 0.87 407

accuracy 0.76 2478

macro avg 0.67 0.81 0.67 2478

weighted avg 0.92 0.76 0.81 2478

average eval\_loss: 0.7380589217308068, accuracy: 0.7606941081517352

training 5 epoch...

106.5233862400055 seconds used......

5 training finished! average train loss: 0.25359140999551694

total clf loss: 18.791415308136493 total adv loss: 12.920336428214796

evaluating...

Confusion matrix, without normalization

[[ 96 25 13]

[ 315 1563 59]

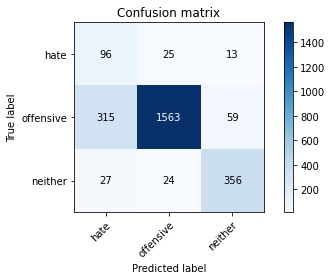
[ 27 24 356]]

Normalized confusion matrix

[[0.71641791 0.18656716 0.09701493]

[0.16262261 0.80691791 0.03045947]

[0.06633907 0.05896806 0.87469287]]





precision recall f1-score support

hate 0.22 0.72 0.34 134

offensive 0.97 0.81 0.88 1937

neither 0.83 0.87 0.85 407

accuracy 0.81 2478

macro avg 0.67 0.80 0.69 2478

weighted avg 0.91 0.81 0.85 2478

average eval\_loss: 0.6294484626982482, accuracy: 0.8131557707828895

Confusion matrix, without normalization

[[ 96 23 5]

[ 499 1379 64]

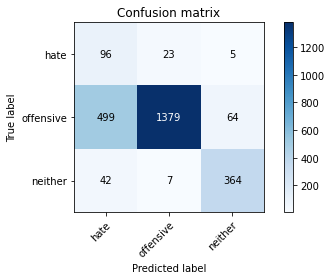
[ 42 7 364]]

Normalized confusion matrix

[[0.77419355 0.18548387 0.04032258]

[0.2569516 0.71009269 0.03295572]

[0.10169492 0.01694915 0.88135593]]





precision recall f1-score support

hate 0.15 0.77 0.25 124

offensive 0.98 0.71 0.82 1942

neither 0.84 0.88 0.86 413

accuracy 0.74 2479

macro avg 0.66 0.79 0.65 2479

weighted avg 0.91 0.74 0.80 2479

average eval\_loss: 0.797616075664278, accuracy: 0.7418313836224284

[20]:

0.7418313836224284

5 run

training 1 epoch...

106.62323451042175 seconds used......

1 training finished! average train loss: 1.4479425579986789

total clf loss: 117.48694701492786 total adv loss: 68.67547146230936

evaluating...

Confusion matrix, without normalization

[[ 96 20 18]

[ 340 1540 57]

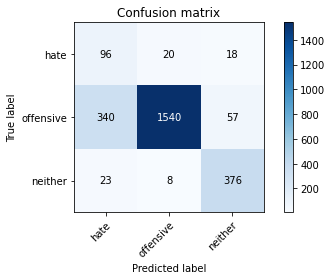
[ 23 8 376]]

Normalized confusion matrix

[[0.71641791 0.14925373 0.13432836]

[0.17552917 0.79504388 0.02942695]

[0.05651106 0.01965602 0.92383292]]





precision recall f1-score support

hate 0.21 0.72 0.32 134

offensive 0.98 0.80 0.88 1937

neither 0.83 0.92 0.88 407

accuracy 0.81 2478

macro avg 0.67 0.81 0.69 2478

weighted avg 0.92 0.81 0.85 2478

average eval\_loss: 0.5060796617500244, accuracy: 0.8119451170298628

training 2 epoch...

106.55185151100159 seconds used......

2 training finished! average train loss: 0.9651318165388975

total clf loss: 70.50350444763899 total adv loss: 49.679847694933414

evaluating...

Confusion matrix, without normalization

[[ 102 20 12]

[ 362 1515 60]

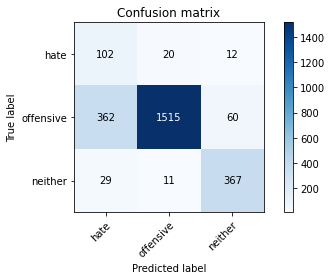
[ 29 11 367]]

Normalized confusion matrix

[[0.76119403 0.14925373 0.08955224]

[0.18688694 0.78213733 0.03097574]

[0.07125307 0.02702703 0.9017199 ]]





precision recall f1-score support

hate 0.21 0.76 0.33 134

offensive 0.98 0.78 0.87 1937

neither 0.84 0.90 0.87 407

accuracy 0.80 2478

macro avg 0.67 0.82 0.69 2478

weighted avg 0.91 0.80 0.84 2478

average eval\_loss: 0.48906717365307195, accuracy: 0.8006456820016142

training 3 epoch...

106.59650611877441 seconds used......

3 training finished! average train loss: 0.6962087478658016

total clf loss: 51.059938199818134 total adv loss: 35.73640073277056

evaluating...

Confusion matrix, without normalization

[[ 97 22 15]

[ 320 1549 68]

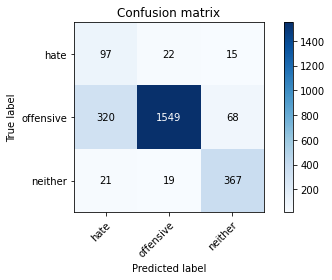
[ 21 19 367]]

Normalized confusion matrix

[[0.7238806 0.1641791 0.1119403 ]

[0.16520392 0.79969024 0.03510583]

[0.05159705 0.04668305 0.9017199 ]]





precision recall f1-score support

hate 0.22 0.72 0.34 134

offensive 0.97 0.80 0.88 1937

neither 0.82 0.90 0.86 407

accuracy 0.81 2478

macro avg 0.67 0.81 0.69 2478

weighted avg 0.91 0.81 0.85 2478

average eval\_loss: 0.47437662793503654, accuracy: 0.8123486682808717

training 4 epoch...

106.6593918800354 seconds used......

4 training finished! average train loss: 0.4450809453741055

total clf loss: 32.64411250501871 total adv loss: 22.845066979527473

evaluating...

Confusion matrix, without normalization

[[ 105 19 10]

[ 330 1555 52]

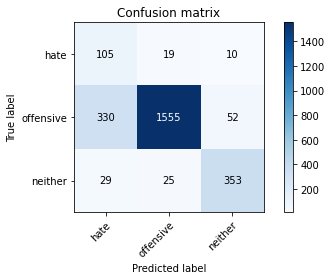
[ 29 25 353]]

Normalized confusion matrix

[[0.78358209 0.14179104 0.07462687]

[0.17036655 0.80278782 0.02684564]

[0.07125307 0.06142506 0.86732187]]





precision recall f1-score support

hate 0.23 0.78 0.35 134

offensive 0.97 0.80 0.88 1937

neither 0.85 0.87 0.86 407

accuracy 0.81 2478

macro avg 0.68 0.82 0.70 2478

weighted avg 0.91 0.81 0.85 2478

average eval\_loss: 0.5596416319810575, accuracy: 0.8123486682808717

training 5 epoch...

106.6315221786499 seconds used......

5 training finished! average train loss: 0.20618512680944565

total clf loss: 15.977365533821285 total adv loss: 10.15560839814134

evaluating...

Confusion matrix, without normalization

[[ 106 20 8]

[ 437 1442 58]

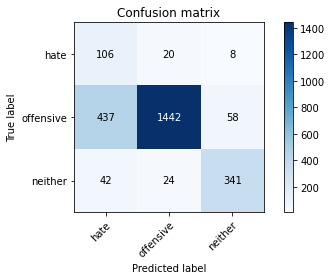
[ 42 24 341]]

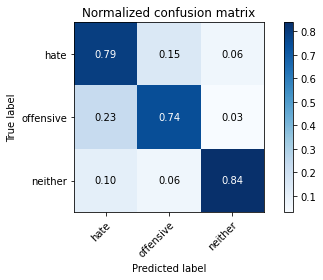
Normalized confusion matrix

[[0.79104478 0.14925373 0.05970149]

[0.22560661 0.74445018 0.02994321]

[0.1031941 0.05896806 0.83783784]]





precision recall f1-score support

hate 0.18 0.79 0.29 134

offensive 0.97 0.74 0.84 1937

neither 0.84 0.84 0.84 407

accuracy 0.76 2478

macro avg 0.66 0.79 0.66 2478

weighted avg 0.91 0.76 0.81 2478

average eval\_loss: 0.8687997196891135, accuracy: 0.7623083131557707

Confusion matrix, without normalization

[[ 93 26 5]

[ 462 1421 59]

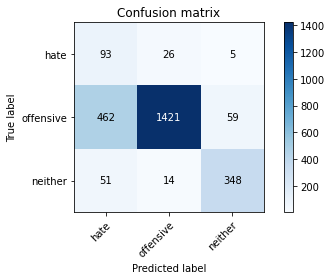
[ 51 14 348]]

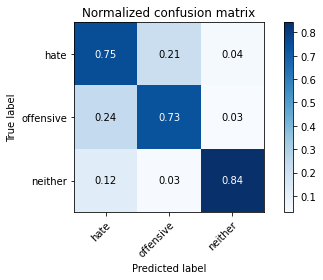
Normalized confusion matrix

[[0.75 0.20967742 0.04032258]

[0.23789907 0.73171988 0.03038105]

[0.12348668 0.03389831 0.84261501]]





precision recall f1-score support

hate 0.15 0.75 0.25 124

offensive 0.97 0.73 0.84 1942

neither 0.84 0.84 0.84 413

accuracy 0.75 2479

macro avg 0.66 0.77 0.64 2479

weighted avg 0.91 0.75 0.81 2479

average eval\_loss: 0.96072506712329, accuracy: 0.7511093182734974

[23]:

0.7511093182734974

6 run

training 1 epoch...

107.4319953918457 seconds used......

1 training finished! average train loss: 1.4596122675998644

total clf loss: 118.1554882824421 total adv loss: 69.36813519895077

evaluating...

Confusion matrix, without normalization

[[ 92 15 27]

[ 371 1471 95]

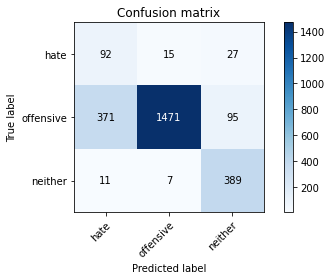
[ 11 7 389]]

Normalized confusion matrix

[[0.68656716 0.1119403 0.20149254]

[0.1915333 0.75942179 0.04904491]

[0.02702703 0.01719902 0.95577396]]





precision recall f1-score support

hate 0.19 0.69 0.30 134

offensive 0.99 0.76 0.86 1937

neither 0.76 0.96 0.85 407

accuracy 0.79 2478

macro avg 0.65 0.80 0.67 2478

weighted avg 0.91 0.79 0.83 2478

average eval\_loss: 0.49043737213698124, accuracy: 0.7877320419693301

training 2 epoch...

106.5961766242981 seconds used......

2 training finished! average train loss: 0.9317321138964458

total clf loss: 69.85775028169155 total adv loss: 47.06355077587068

evaluating...

Confusion matrix, without normalization

[[ 94 24 16]

[ 292 1598 47]

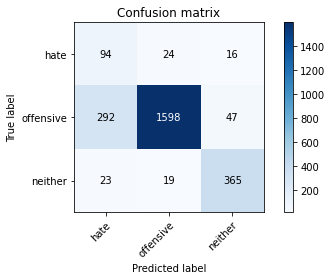
[ 23 19 365]]

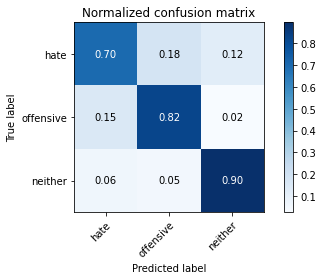
Normalized confusion matrix

[[0.70149254 0.17910448 0.11940299]

[0.15074858 0.82498709 0.02426433]

[0.05651106 0.04668305 0.8968059 ]]





precision recall f1-score support

hate 0.23 0.70 0.35 134

offensive 0.97 0.82 0.89 1937

neither 0.85 0.90 0.87 407

accuracy 0.83 2478

macro avg 0.69 0.81 0.70 2478

weighted avg 0.91 0.83 0.86 2478

average eval\_loss: 0.3864129697603564, accuracy: 0.8301049233252623

training 3 epoch...

106.70591497421265 seconds used......

3 training finished! average train loss: 0.6689489843662489

total clf loss: 49.04362466931343 total adv loss: 34.34569817222655

evaluating...

Confusion matrix, without normalization

[[ 99 15 20]

[ 428 1393 116]

[ 18 7 382]]

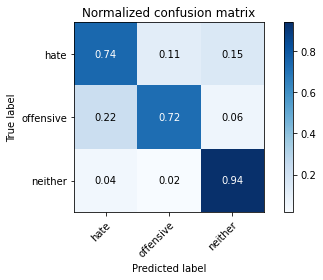
Normalized confusion matrix

[[0.73880597 0.1119403 0.14925373]

[0.22096025 0.71915333 0.05988642]

[0.04422604 0.01719902 0.93857494]]





precision recall f1-score support

hate 0.18 0.74 0.29 134

offensive 0.98 0.72 0.83 1937

neither 0.74 0.94 0.83 407

accuracy 0.76 2478

macro avg 0.63 0.80 0.65 2478

weighted avg 0.90 0.76 0.80 2478

average eval\_loss: 0.6774860684189105, accuracy: 0.7562550443906376

training 4 epoch...

106.71993589401245 seconds used......

4 training finished! average train loss: 0.39311031736856833

total clf loss: 28.85855976305902 total adv loss: 20.16442811768502

evaluating...

Confusion matrix, without normalization

[[ 94 27 13]

[ 294 1578 65]

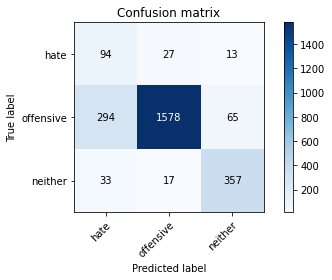
[ 33 17 357]]

Normalized confusion matrix

[[0.70149254 0.20149254 0.09701493]

[0.1517811 0.81466185 0.03355705]

[0.08108108 0.04176904 0.87714988]]





precision recall f1-score support

hate 0.22 0.70 0.34 134

offensive 0.97 0.81 0.89 1937

neither 0.82 0.88 0.85 407

accuracy 0.82 2478

macro avg 0.67 0.80 0.69 2478

weighted avg 0.91 0.82 0.85 2478

average eval\_loss: 0.5395440221313508, accuracy: 0.8188054882970137

training 5 epoch...

106.73828840255737 seconds used......

5 training finished! average train loss: 0.2571502498202873

total clf loss: 18.502725150901824 total adv loss: 13.377859355634428

evaluating...

Confusion matrix, without normalization

[[ 98 23 13]

[ 344 1556 37]

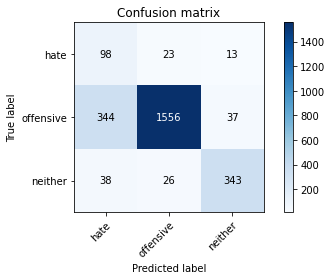
[ 38 26 343]]

Normalized confusion matrix

[[0.73134328 0.17164179 0.09701493]

[0.17759422 0.80330408 0.0191017 ]

[0.09336609 0.06388206 0.84275184]]





precision recall f1-score support

hate 0.20 0.73 0.32 134

offensive 0.97 0.80 0.88 1937

neither 0.87 0.84 0.86 407

accuracy 0.81 2478

macro avg 0.68 0.79 0.69 2478

weighted avg 0.91 0.81 0.84 2478

average eval\_loss: 0.6257685970875525, accuracy: 0.8058918482647296

Confusion matrix, without normalization

[[ 94 22 8]

[ 418 1418 106]

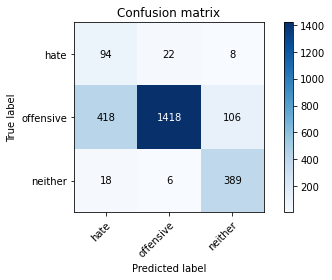
[ 18 6 389]]

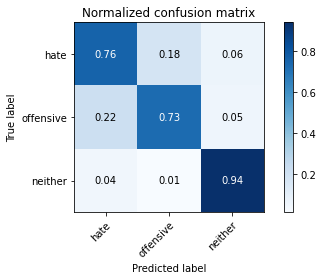
Normalized confusion matrix

[[0.75806452 0.17741935 0.06451613]

[0.21524202 0.73017508 0.0545829 ]

[0.04358354 0.01452785 0.94188862]]





precision recall f1-score support

hate 0.18 0.76 0.29 124

offensive 0.98 0.73 0.84 1942

neither 0.77 0.94 0.85 413

accuracy 0.77 2479

macro avg 0.64 0.81 0.66 2479

weighted avg 0.91 0.77 0.81 2479

average eval\_loss: 0.6924109623557136, accuracy: 0.7668414683340057

[19]:

0.7668414683340057

alpha = 0.05

1 run

training 1 epoch...

106.7902250289917 seconds used......

1 training finished! average train loss: 0.6713428319516507

total clf loss: 114.48142738640308 total adv loss: 73.4982153326273

evaluating...

Confusion matrix, without normalization

[[ 99 13 22]

[ 432 1431 74]

[ 20 6 381]]

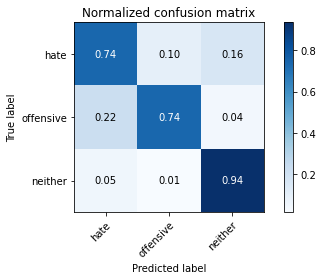
Normalized confusion matrix

[[0.73880597 0.09701493 0.1641791 ]

[0.2230253 0.7387713 0.03820341]

[0.04914005 0.01474201 0.93611794]]





precision recall f1-score support

hate 0.18 0.74 0.29 134

offensive 0.99 0.74 0.84 1937

neither 0.80 0.94 0.86 407

accuracy 0.77 2478

macro avg 0.66 0.80 0.67 2478

weighted avg 0.91 0.77 0.82 2478

average eval\_loss: 0.5583295347469468, accuracy: 0.7711864406779662

training 2 epoch...

106.63755679130554 seconds used......

2 training finished! average train loss: 0.40788620147346094

total clf loss: 69.18068335205317 total adv loss: 52.14576071500778

evaluating...

Confusion matrix, without normalization

[[ 106 13 15]

[ 452 1446 39]

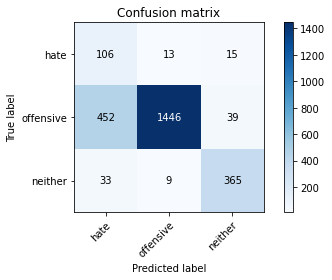
[ 33 9 365]]

Normalized confusion matrix

[[0.79104478 0.09701493 0.1119403 ]

[0.23335054 0.74651523 0.02013423]

[0.08108108 0.02211302 0.8968059 ]]





precision recall f1-score support

hate 0.18 0.79 0.29 134

offensive 0.99 0.75 0.85 1937

neither 0.87 0.90 0.88 407

accuracy 0.77 2478

macro avg 0.68 0.81 0.68 2478

weighted avg 0.92 0.77 0.82 2478

average eval\_loss: 0.5341592898532268, accuracy: 0.7736077481840193

training 3 epoch...

106.64973187446594 seconds used......

3 training finished! average train loss: 0.2822538458064876

total clf loss: 47.65975833311677 total adv loss: 40.33837401308119

evaluating...

Confusion matrix, without normalization

[[ 101 19 14]

[ 382 1515 40]

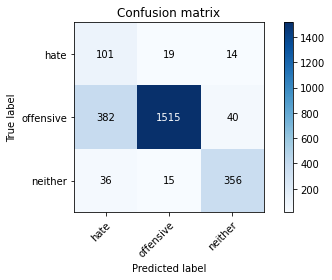
[ 36 15 356]]

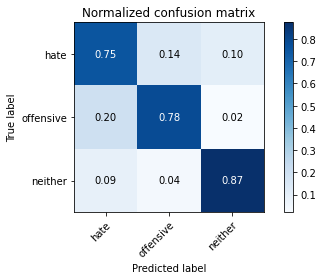
Normalized confusion matrix

[[0.75373134 0.14179104 0.10447761]

[0.19721218 0.78213733 0.02065049]

[0.08845209 0.03685504 0.87469287]]





precision recall f1-score support

hate 0.19 0.75 0.31 134

offensive 0.98 0.78 0.87 1937

neither 0.87 0.87 0.87 407

accuracy 0.80 2478

macro avg 0.68 0.80 0.68 2478

weighted avg 0.92 0.80 0.84 2478

average eval\_loss: 0.5440574555147079, accuracy: 0.7958030669895076

training 4 epoch...

106.68238234519958 seconds used......

4 training finished! average train loss: 0.17957311582920904

total clf loss: 30.219991205260158 total adv loss: 27.697542816400528

evaluating...

Confusion matrix, without normalization

[[ 102 19 13]

[ 378 1518 41]

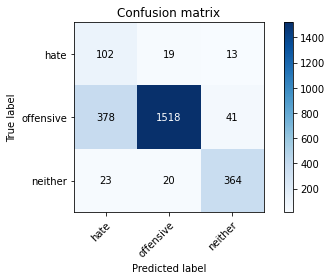
[ 23 20 364]]

Normalized confusion matrix

[[0.76119403 0.14179104 0.09701493]

[0.19514713 0.78368611 0.02116675]

[0.05651106 0.04914005 0.89434889]]





precision recall f1-score support

hate 0.20 0.76 0.32 134

offensive 0.97 0.78 0.87 1937

neither 0.87 0.89 0.88 407

accuracy 0.80 2478

macro avg 0.68 0.81 0.69 2478

weighted avg 0.92 0.80 0.84 2478

average eval\_loss: 0.5516557032542844, accuracy: 0.8006456820016142

training 5 epoch...

106.6414487361908 seconds used......

5 training finished! average train loss: 0.10307169912001965

total clf loss: 17.293526860885322 total adv loss: 16.94184376206249

evaluating...

Confusion matrix, without normalization

[[ 101 22 11]

[ 335 1557 45]

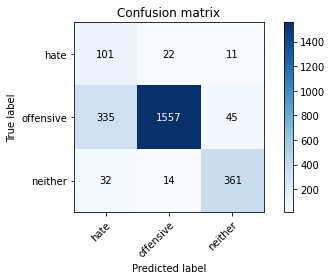
[ 32 14 361]]

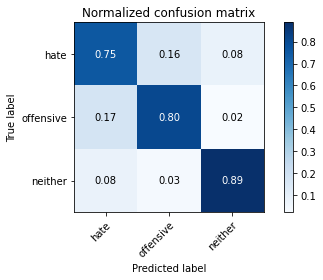
Normalized confusion matrix

[[0.75373134 0.1641791 0.08208955]

[0.17294786 0.80382034 0.0232318 ]

[0.07862408 0.03439803 0.88697789]]





precision recall f1-score support

hate 0.22 0.75 0.34 134

offensive 0.98 0.80 0.88 1937

neither 0.87 0.89 0.88 407

accuracy 0.81 2478

macro avg 0.69 0.81 0.70 2478

weighted avg 0.92 0.81 0.85 2478

average eval\_loss: 0.6653842523453697, accuracy: 0.8147699757869249

Confusion matrix, without normalization

[[ 94 25 5]

[ 437 1453 52]

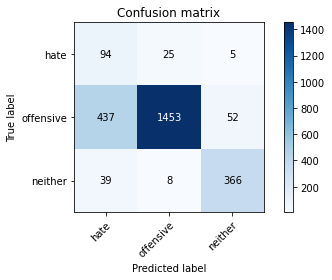
[ 39 8 366]]

Normalized confusion matrix

[[0.75806452 0.2016129 0.04032258]

[0.22502575 0.74819773 0.02677652]

[0.09443099 0.01937046 0.88619855]]





precision recall f1-score support

hate 0.16 0.76 0.27 124

offensive 0.98 0.75 0.85 1942

neither 0.87 0.89 0.88 413

accuracy 0.77 2479

macro avg 0.67 0.80 0.66 2479

weighted avg 0.92 0.77 0.82 2479

average eval\_loss: 0.5773481944756161, accuracy: 0.7716821298910851

[27]:

0.7716821298910851

2 run

training 1 epoch...

106.68588304519653 seconds used......

1 training finished! average train loss: 0.664372198368338

total clf loss: 113.18472105264664 total adv loss: 74.89571671932936

evaluating...

Confusion matrix, without normalization

[[ 78 41 15]

[ 155 1744 38]

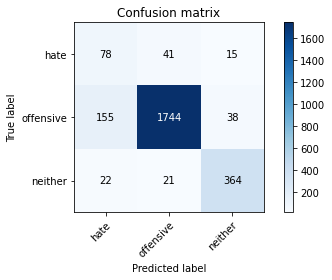
[ 22 21 364]]

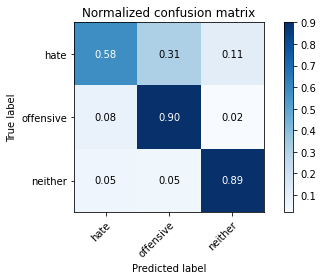
Normalized confusion matrix

[[0.58208955 0.30597015 0.1119403 ]

[0.08002065 0.90036138 0.01961797]

[0.05405405 0.05159705 0.89434889]]





precision recall f1-score support

hate 0.31 0.58 0.40 134

offensive 0.97 0.90 0.93 1937

neither 0.87 0.89 0.88 407

accuracy 0.88 2478

macro avg 0.71 0.79 0.74 2478

weighted avg 0.91 0.88 0.90 2478

average eval\_loss: 0.3298983436919028, accuracy: 0.8821630347054076

training 2 epoch...

106.66306376457214 seconds used......

2 training finished! average train loss: 0.3984389476986094

total clf loss: 67.51966727524996 total adv loss: 52.11174834519625

evaluating...

Confusion matrix, without normalization

[[ 86 23 25]

[ 327 1521 89]

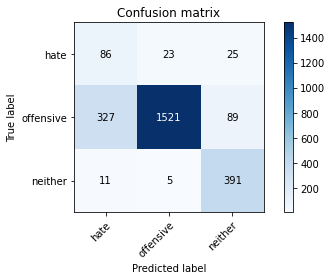
[ 11 5 391]]

Normalized confusion matrix

[[0.64179104 0.17164179 0.18656716]

[0.16881776 0.7852349 0.04594734]

[0.02702703 0.01228501 0.96068796]]





precision recall f1-score support

hate 0.20 0.64 0.31 134

offensive 0.98 0.79 0.87 1937

neither 0.77 0.96 0.86 407

accuracy 0.81 2478

macro avg 0.65 0.80 0.68 2478

weighted avg 0.91 0.81 0.84 2478

average eval\_loss: 0.48562866689697387, accuracy: 0.8062953995157385

training 3 epoch...

106.63338732719421 seconds used......

3 training finished! average train loss: 0.27213343025439163

total clf loss: 45.96434595063329 total adv loss: 38.62275755405426

evaluating...

Confusion matrix, without normalization

[[ 87 33 14]

[ 235 1670 32]

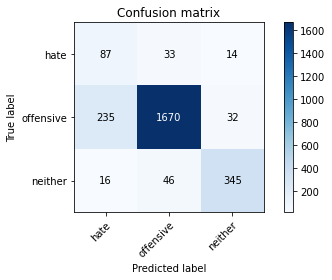
[ 16 46 345]]

Normalized confusion matrix

[[0.64925373 0.24626866 0.10447761]

[0.12132163 0.86215798 0.01652039]

[0.03931204 0.11302211 0.84766585]]





precision recall f1-score support

hate 0.26 0.65 0.37 134

offensive 0.95 0.86 0.91 1937

neither 0.88 0.85 0.86 407

accuracy 0.85 2478

macro avg 0.70 0.79 0.71 2478

weighted avg 0.91 0.85 0.87 2478

average eval\_loss: 0.36770141782659677, accuracy: 0.8482647296206618

training 4 epoch...

106.61737823486328 seconds used......

4 training finished! average train loss: 0.15942440666681665

total clf loss: 26.791553433984518 total adv loss: 25.342843062244356

evaluating...

Confusion matrix, without normalization

[[ 107 13 14]

[ 543 1353 41]

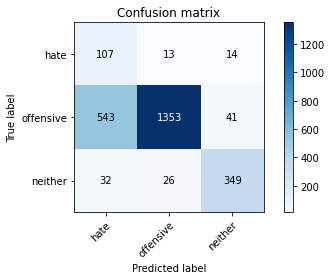
[ 32 26 349]]

Normalized confusion matrix

[[0.79850746 0.09701493 0.10447761]

[0.28033041 0.69850284 0.02116675]

[0.07862408 0.06388206 0.85749386]]





precision recall f1-score support

hate 0.16 0.80 0.26 134

offensive 0.97 0.70 0.81 1937

neither 0.86 0.86 0.86 407

accuracy 0.73 2478

macro avg 0.66 0.78 0.65 2478

weighted avg 0.91 0.73 0.79 2478

average eval\_loss: 0.698904060067669, accuracy: 0.7300242130750605

training 5 epoch...

106.6334171295166 seconds used......

5 training finished! average train loss: 0.08679465288878419

total clf loss: 14.477948290295899 total adv loss: 15.958213014062494

evaluating...

Confusion matrix, without normalization

[[ 95 21 18]

[ 306 1587 44]

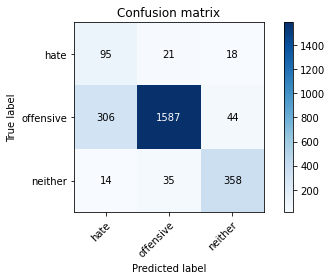
[ 14 35 358]]

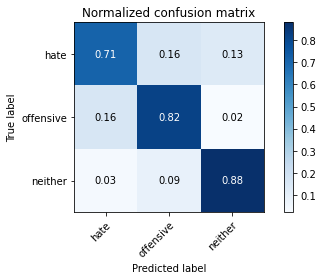
Normalized confusion matrix

[[0.70895522 0.15671642 0.13432836]

[0.15797625 0.81930821 0.02271554]

[0.03439803 0.08599509 0.87960688]]





precision recall f1-score support

hate 0.23 0.71 0.35 134

offensive 0.97 0.82 0.89 1937

neither 0.85 0.88 0.87 407

accuracy 0.82 2478

macro avg 0.68 0.80 0.70 2478

weighted avg 0.91 0.82 0.85 2478

average eval\_loss: 0.5946593117569724, accuracy: 0.8232445520581114

Confusion matrix, without normalization

[[ 94 23 7]

[ 526 1359 57]

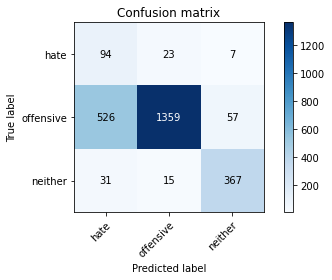
[ 31 15 367]]

Normalized confusion matrix

[[0.75806452 0.18548387 0.05645161]

[0.27085479 0.69979403 0.02935118]

[0.07506053 0.03631961 0.88861985]]





precision recall f1-score support

hate 0.14 0.76 0.24 124

offensive 0.97 0.70 0.81 1942

neither 0.85 0.89 0.87 413

accuracy 0.73 2479

macro avg 0.66 0.78 0.64 2479

weighted avg 0.91 0.73 0.79 2479

average eval\_loss: 0.7600568163418963, accuracy: 0.7341670028237193

[30]:

0.7341670028237193

alpha=0.8

training 1 epoch...

106.75828170776367 seconds used......

1 training finished! average train loss: 0.9679547110572457

total clf loss: 115.07668390870094 total adv loss: 69.10418012738228

evaluating...

Confusion matrix, without normalization

[[ 105 14 15]

[ 421 1470 46]

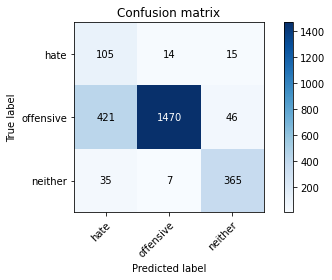
[ 35 7 365]]

Normalized confusion matrix

[[0.78358209 0.10447761 0.1119403 ]

[0.21734641 0.75890552 0.02374806]

[0.08599509 0.01719902 0.8968059 ]]





precision recall f1-score support

hate 0.19 0.78 0.30 134

offensive 0.99 0.76 0.86 1937

neither 0.86 0.90 0.88 407

accuracy 0.78 2478

macro avg 0.68 0.81 0.68 2478

weighted avg 0.92 0.78 0.83 2478

average eval\_loss: 0.5408123766462649, accuracy: 0.7828894269572235

training 2 epoch...

106.66502118110657 seconds used......

2 training finished! average train loss: 0.6173097183081236

total clf loss: 68.89511422812939 total adv loss: 49.68924441561103

evaluating...

Confusion matrix, without normalization

[[ 103 17 14]

[ 415 1477 45]

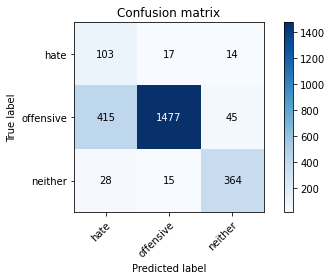
[ 28 15 364]]

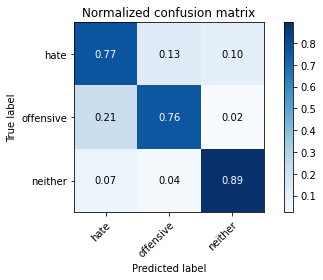
Normalized confusion matrix

[[0.76865672 0.12686567 0.10447761]

[0.21424884 0.76251936 0.0232318 ]

[0.06879607 0.03685504 0.89434889]]





precision recall f1-score support

hate 0.19 0.77 0.30 134

offensive 0.98 0.76 0.86 1937

neither 0.86 0.89 0.88 407

accuracy 0.78 2478

macro avg 0.68 0.81 0.68 2478

weighted avg 0.92 0.78 0.83 2478

average eval\_loss: 0.48178985102042077, accuracy: 0.784503631961259

training 3 epoch...

106.72989559173584 seconds used......

3 training finished! average train loss: 0.4301361450078813

total clf loss: 47.10932318866253 total adv loss: 35.743297036737204

evaluating...

Confusion matrix, without normalization

[[ 78 34 22]

[ 211 1651 75]

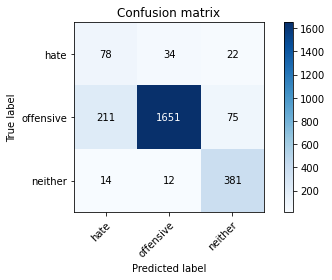
[ 14 12 381]]

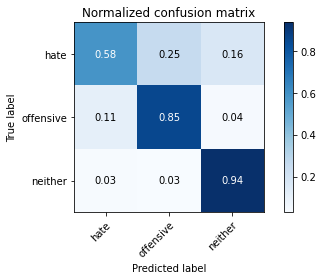
Normalized confusion matrix

[[0.58208955 0.25373134 0.1641791 ]

[0.10893134 0.85234899 0.03871967]

[0.03439803 0.02948403 0.93611794]]





precision recall f1-score support

hate 0.26 0.58 0.36 134

offensive 0.97 0.85 0.91 1937

neither 0.80 0.94 0.86 407

accuracy 0.85 2478

macro avg 0.68 0.79 0.71 2478

weighted avg 0.91 0.85 0.87 2478

average eval\_loss: 0.4570691152025134, accuracy: 0.8514931396287329

training 4 epoch...

106.69065856933594 seconds used......

4 training finished! average train loss: 0.25486031134443526

total clf loss: 27.101963536813855 total adv loss: 22.19181373436004

evaluating...

Confusion matrix, without normalization

[[ 110 14 10]

[ 450 1442 45]

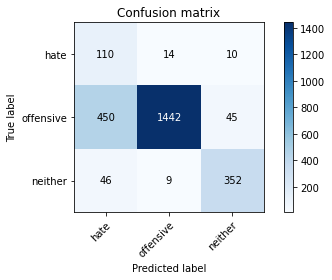
[ 46 9 352]]

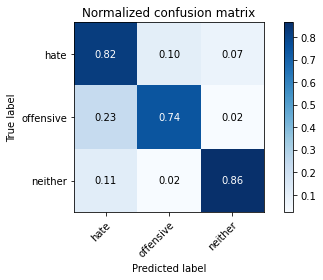
Normalized confusion matrix

[[0.82089552 0.10447761 0.07462687]

[0.23231802 0.74445018 0.0232318 ]

[0.11302211 0.02211302 0.86486486]]





precision recall f1-score support

hate 0.18 0.82 0.30 134

offensive 0.98 0.74 0.85 1937

neither 0.86 0.86 0.86 407

accuracy 0.77 2478

macro avg 0.68 0.81 0.67 2478

weighted avg 0.92 0.77 0.82 2478

average eval\_loss: 0.729209604282533, accuracy: 0.768361581920904

training 5 epoch...

106.69792914390564 seconds used......

5 training finished! average train loss: 0.10931571665357544

total clf loss: 11.90251542115584 total adv loss: 9.171313180326251

evaluating...

Confusion matrix, without normalization

[[ 111 11 12]

[ 530 1321 86]

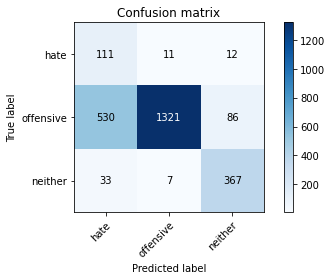
[ 33 7 367]]

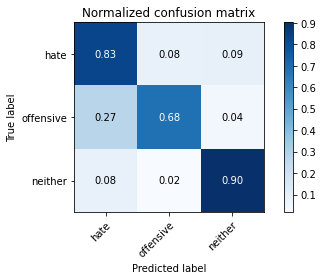
Normalized confusion matrix

[[0.82835821 0.08208955 0.08955224]

[0.273619 0.68198245 0.04439855]

[0.08108108 0.01719902 0.9017199 ]]





precision recall f1-score support

hate 0.16 0.83 0.27 134

offensive 0.99 0.68 0.81 1937

neither 0.79 0.90 0.84 407

accuracy 0.73 2478

macro avg 0.65 0.80 0.64 2478

weighted avg 0.91 0.73 0.78 2478

average eval\_loss: 1.0090256719098938, accuracy: 0.7259887005649718

Confusion matrix, without normalization

[[ 96 21 7]

[ 551 1307 84]

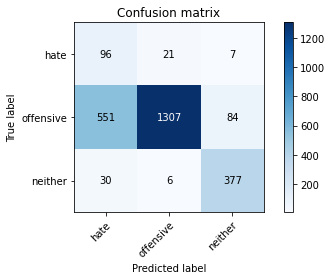
[ 30 6 377]]

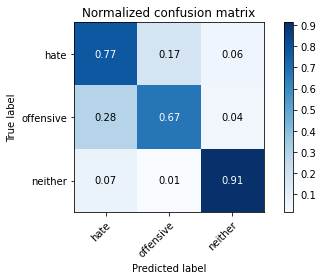
Normalized confusion matrix

[[0.77419355 0.16935484 0.05645161]

[0.28372812 0.67301751 0.04325438]

[0.07263923 0.01452785 0.91283293]]





precision recall f1-score support

hate 0.14 0.77 0.24 124

offensive 0.98 0.67 0.80 1942

neither 0.81 0.91 0.86 413

accuracy 0.72 2479

macro avg 0.64 0.79 0.63 2479

weighted avg 0.91 0.72 0.78 2479

average eval\_loss: 1.066469636295111, accuracy: 0.718031464300121

[34]:

0.718031464300121

alpha = 1.2

training 1 epoch...

108.34121751785278 seconds used......

1 training finished! average train loss: 1.1795054730027914

total clf loss: 121.15726724267006 total adv loss: 72.02974366396666

evaluating...

Confusion matrix, without normalization

[[ 84 29 21]

[ 284 1590 63]

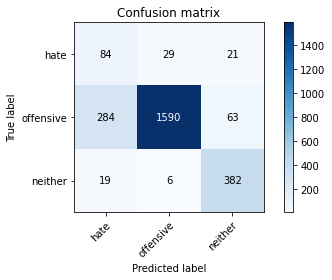
[ 19 6 382]]

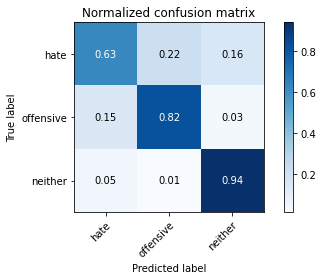
Normalized confusion matrix

[[0.62686567 0.21641791 0.15671642]

[0.14661848 0.820857 0.03252452]

[0.04668305 0.01474201 0.93857494]]





precision recall f1-score support

hate 0.22 0.63 0.32 134

offensive 0.98 0.82 0.89 1937

neither 0.82 0.94 0.88 407

accuracy 0.83 2478

macro avg 0.67 0.80 0.70 2478

weighted avg 0.91 0.83 0.86 2478

average eval\_loss: 0.4622981519468369, accuracy: 0.8297013720742534

training 2 epoch...

107.44312644004822 seconds used......

2 training finished! average train loss: 0.7331838131966916

total clf loss: 70.63636042177677 total adv loss: 48.669990569353104

evaluating...

Confusion matrix, without normalization

[[ 101 17 16]

[ 433 1465 39]

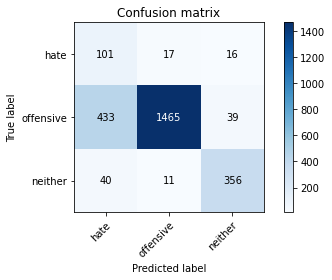
[ 40 11 356]]

Normalized confusion matrix

[[0.75373134 0.12686567 0.11940299]

[0.22354156 0.75632421 0.02013423]

[0.0982801 0.02702703 0.87469287]]





precision recall f1-score support

hate 0.18 0.75 0.29 134

offensive 0.98 0.76 0.85 1937

neither 0.87 0.87 0.87 407

accuracy 0.78 2478

macro avg 0.67 0.79 0.67 2478

weighted avg 0.92 0.78 0.83 2478

average eval\_loss: 0.5250332300220767, accuracy: 0.7756255044390638

training 3 epoch...

107.55324792861938 seconds used......

3 training finished! average train loss: 0.529318431138315

total clf loss: 49.57836990058422 total adv loss: 36.318059887737036

evaluating...

Confusion matrix, without normalization

[[ 113 12 9]

[ 498 1387 52]

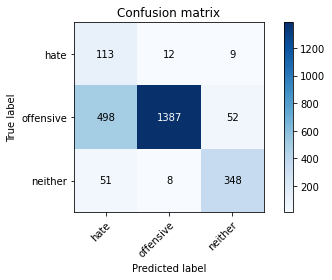
[ 51 8 348]]

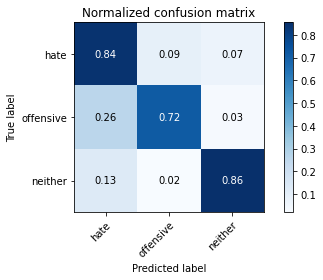
Normalized confusion matrix

[[0.84328358 0.08955224 0.06716418]

[0.25709861 0.71605576 0.02684564]

[0.12530713 0.01965602 0.85503686]]





precision recall f1-score support

hate 0.17 0.84 0.28 134

offensive 0.99 0.72 0.83 1937

neither 0.85 0.86 0.85 407

accuracy 0.75 2478

macro avg 0.67 0.80 0.66 2478

weighted avg 0.92 0.75 0.80 2478

average eval\_loss: 0.6636188742854903, accuracy: 0.7457627118644068

training 4 epoch...

107.60751509666443 seconds used......

4 training finished! average train loss: 0.31826430912637577

total clf loss: 28.676040241494775 total adv loss: 22.78206441923976

evaluating...

Confusion matrix, without normalization

[[ 111 11 12]

[ 628 1250 59]

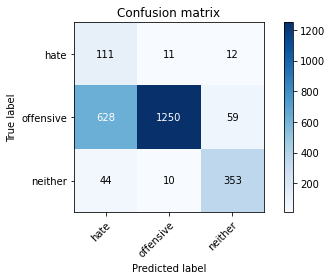
[ 44 10 353]]

Normalized confusion matrix

[[0.82835821 0.08208955 0.08955224]

[0.3242127 0.64532783 0.03045947]

[0.10810811 0.02457002 0.86732187]]





precision recall f1-score support

hate 0.14 0.83 0.24 134

offensive 0.98 0.65 0.78 1937

neither 0.83 0.87 0.85 407

accuracy 0.69 2478

macro avg 0.65 0.78 0.62 2478

weighted avg 0.91 0.69 0.76 2478

average eval\_loss: 0.8585992462692722, accuracy: 0.6916868442292171

training 5 epoch...

107.46945595741272 seconds used......

5 training finished! average train loss: 0.20114104701629418

total clf loss: 17.985526049509645 total adv loss: 14.512747886881698

evaluating...

Confusion matrix, without normalization

[[ 94 25 15]

[ 274 1604 59]

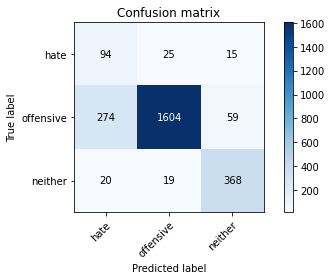
[ 20 19 368]]

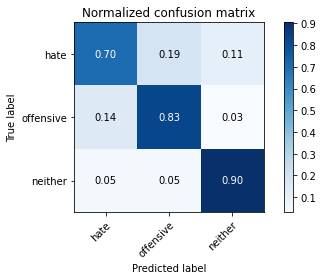
Normalized confusion matrix

[[0.70149254 0.18656716 0.1119403 ]

[0.14145586 0.82808467 0.03045947]

[0.04914005 0.04668305 0.9041769 ]]





precision recall f1-score support

hate 0.24 0.70 0.36 134

offensive 0.97 0.83 0.89 1937

neither 0.83 0.90 0.87 407

accuracy 0.83 2478

macro avg 0.68 0.81 0.71 2478

weighted avg 0.91 0.83 0.86 2478

average eval\_loss: 0.582085655100884, accuracy: 0.8337368845843423

Confusion matrix, without normalization

[[ 98 21 5]

[ 529 1365 48]

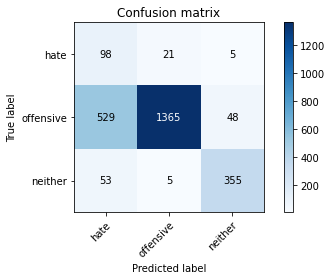
[ 53 5 355]]

Normalized confusion matrix

[[0.79032258 0.16935484 0.04032258]

[0.27239959 0.70288363 0.02471679]

[0.1283293 0.01210654 0.85956416]]





precision recall f1-score support

hate 0.14 0.79 0.24 124

offensive 0.98 0.70 0.82 1942

neither 0.87 0.86 0.86 413

accuracy 0.73 2479

macro avg 0.67 0.78 0.64 2479

weighted avg 0.92 0.73 0.80 2479

average eval\_loss: 0.7046092332851502, accuracy: 0.7333602258975394

[20]:

0.7333602258975394

alpha=1.2

training 1 epoch...

107.49295926094055 seconds used......

1 training finished! average train loss: 1.1450812508436767

total clf loss: 116.98682977259159 total adv loss: 70.45622286945581

evaluating...

Confusion matrix, without normalization

[[ 80 33 21]

[ 232 1630 75]

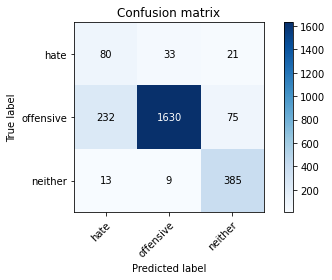
[ 13 9 385]]

Normalized confusion matrix

[[0.59701493 0.24626866 0.15671642]

[0.11977284 0.84150749 0.03871967]

[0.03194103 0.02211302 0.94594595]]





precision recall f1-score support

hate 0.25 0.60 0.35 134

offensive 0.97 0.84 0.90 1937

neither 0.80 0.95 0.87 407

accuracy 0.85 2478

macro avg 0.67 0.79 0.71 2478

weighted avg 0.91 0.85 0.87 2478

average eval\_loss: 0.4508960639997836, accuracy: 0.8454398708635997

training 2 epoch...

107.50501441955566 seconds used......

2 training finished! average train loss: 0.732239244505763

total clf loss: 70.05434714257717 total adv loss: 49.016464691609144

evaluating...

Confusion matrix, without normalization

[[ 90 22 22]

[ 344 1538 55]

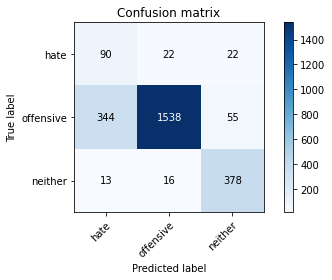
[ 13 16 378]]

Normalized confusion matrix

[[0.67164179 0.1641791 0.1641791 ]

[0.17759422 0.79401136 0.02839442]

[0.03194103 0.03931204 0.92874693]]





precision recall f1-score support

hate 0.20 0.67 0.31 134

offensive 0.98 0.79 0.88 1937

neither 0.83 0.93 0.88 407

accuracy 0.81 2478

macro avg 0.67 0.80 0.69 2478

weighted avg 0.91 0.81 0.85 2478

average eval\_loss: 0.462767755432475, accuracy: 0.8095238095238095

training 3 epoch...

107.52208828926086 seconds used......

3 training finished! average train loss: 0.53621010850607

total clf loss: 49.497152239084244 total adv loss: 37.396521179005504

evaluating...

Confusion matrix, without normalization

[[ 71 46 17]

[ 170 1708 59]

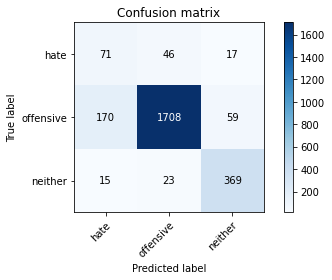
[ 15 23 369]]

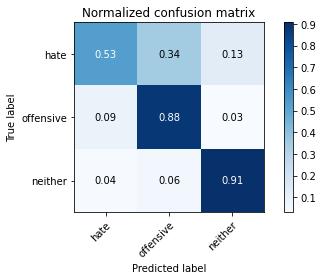
Normalized confusion matrix

[[0.52985075 0.34328358 0.12686567]

[0.08776458 0.88177594 0.03045947]

[0.03685504 0.05651106 0.90663391]]





precision recall f1-score support

hate 0.28 0.53 0.36 134

offensive 0.96 0.88 0.92 1937

neither 0.83 0.91 0.87 407

accuracy 0.87 2478

macro avg 0.69 0.77 0.72 2478

weighted avg 0.90 0.87 0.88 2478

average eval\_loss: 0.3899929702972933, accuracy: 0.8668280871670703

training 4 epoch...

107.48565459251404 seconds used......

4 training finished! average train loss: 0.32877963527359744

total clf loss: 30.44766502827406 total adv loss: 22.847958193626255

evaluating...

Confusion matrix, without normalization

[[ 101 20 13]

[ 334 1565 38]

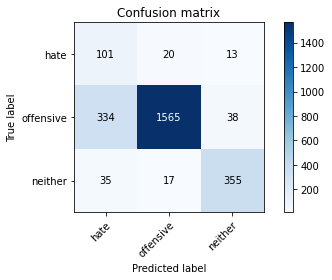
[ 35 17 355]]

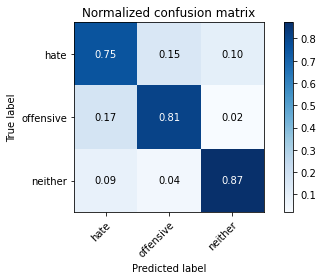
Normalized confusion matrix

[[0.75373134 0.14925373 0.09701493]

[0.1724316 0.80795044 0.01961797]

[0.08599509 0.04176904 0.87223587]]





precision recall f1-score support

hate 0.21 0.75 0.33 134

offensive 0.98 0.81 0.88 1937

neither 0.87 0.87 0.87 407

accuracy 0.82 2478

macro avg 0.69 0.81 0.70 2478

weighted avg 0.92 0.82 0.85 2478

average eval\_loss: 0.5073243255817121, accuracy: 0.8155770782889427

training 5 epoch...

107.5222179889679 seconds used......

5 training finished! average train loss: 0.18684331787517294

total clf loss: 16.615485776681453 total adv loss: 13.55744791374309

evaluating...

Confusion matrix, without normalization

[[ 105 18 11]

[ 366 1516 55]

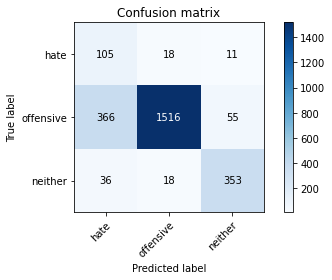
[ 36 18 353]]

Normalized confusion matrix

[[0.78358209 0.13432836 0.08208955]

[0.18895199 0.78265359 0.02839442]

[0.08845209 0.04422604 0.86732187]]





precision recall f1-score support

hate 0.21 0.78 0.33 134

offensive 0.98 0.78 0.87 1937

neither 0.84 0.87 0.85 407

accuracy 0.80 2478

macro avg 0.68 0.81 0.68 2478

weighted avg 0.91 0.80 0.84 2478

average eval\_loss: 0.7432159732186026, accuracy: 0.7966101694915254

Confusion matrix, without normalization

[[ 92 28 4]

[ 411 1476 55]

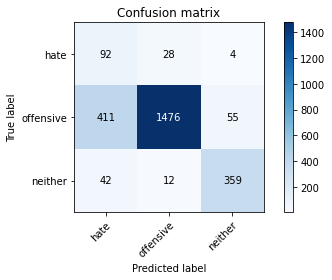
[ 42 12 359]]

Normalized confusion matrix

[[0.74193548 0.22580645 0.03225806]

[0.21163749 0.76004119 0.02832132]

[0.10169492 0.02905569 0.86924939]]





precision recall f1-score support

hate 0.17 0.74 0.28 124

offensive 0.97 0.76 0.85 1942

neither 0.86 0.87 0.86 413

accuracy 0.78 2479

macro avg 0.67 0.79 0.66 2479

weighted avg 0.91 0.78 0.83 2479

average eval\_loss: 0.8214578017472259, accuracy: 0.7773295683743445

[25]:

0.7773295683743445

alpha =0.2

training 1 epoch...

107.06914186477661 seconds used......

1 training finished! average train loss: 0.7454944236881353

total clf loss: 116.410052344203 total adv loss: 73.9848303757608

evaluating...

Confusion matrix, without normalization

[[ 109 9 16]

[ 539 1343 55]

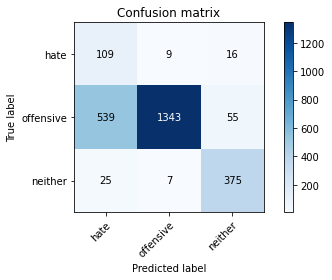
[ 25 7 375]]

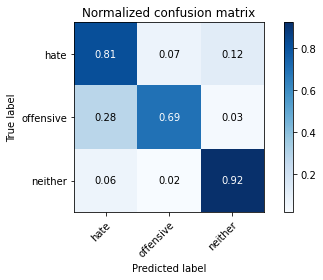
Normalized confusion matrix

[[0.81343284 0.06716418 0.11940299]

[0.27826536 0.69334022 0.02839442]

[0.06142506 0.01719902 0.92137592]]





precision recall f1-score support

hate 0.16 0.81 0.27 134

offensive 0.99 0.69 0.81 1937

neither 0.84 0.92 0.88 407

accuracy 0.74 2478

macro avg 0.66 0.81 0.65 2478

weighted avg 0.92 0.74 0.80 2478

average eval\_loss: 0.6269497659898573, accuracy: 0.7372881355932204

training 2 epoch...

106.99750351905823 seconds used......

2 training finished! average train loss: 0.4421884395097467

total clf loss: 67.7539640814066 total adv loss: 50.35600603930652

evaluating...

Confusion matrix, without normalization

[[ 74 35 25]

[ 178 1692 67]

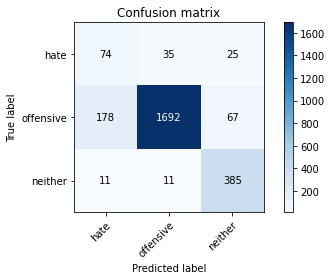
[ 11 11 385]]

Normalized confusion matrix

[[0.55223881 0.26119403 0.18656716]

[0.09189468 0.87351575 0.03458957]

[0.02702703 0.02702703 0.94594595]]





precision recall f1-score support

hate 0.28 0.55 0.37 134

offensive 0.97 0.87 0.92 1937

neither 0.81 0.95 0.87 407

accuracy 0.87 2478

macro avg 0.69 0.79 0.72 2478

weighted avg 0.91 0.87 0.88 2478

average eval\_loss: 0.34712111475246565, accuracy: 0.8680387409200968

training 3 epoch...

106.97371029853821 seconds used......

3 training finished! average train loss: 0.3188462092287161

total clf loss: 48.071486704051495 total adv loss: 40.22723028063774

evaluating...

Confusion matrix, without normalization

[[ 82 32 20]

[ 161 1672 104]

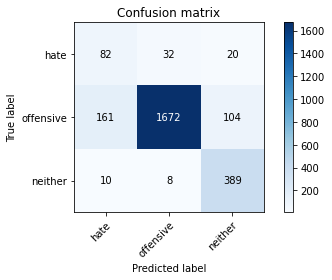
[ 10 8 389]]

Normalized confusion matrix

[[0.6119403 0.23880597 0.14925373]

[0.08311822 0.8631905 0.05369128]

[0.02457002 0.01965602 0.95577396]]





precision recall f1-score support

hate 0.32 0.61 0.42 134

offensive 0.98 0.86 0.92 1937

neither 0.76 0.96 0.85 407

accuracy 0.86 2478

macro avg 0.69 0.81 0.73 2478

weighted avg 0.91 0.86 0.88 2478

average eval\_loss: 0.4301011778654591, accuracy: 0.8648103309120259

training 4 epoch...

107.06610560417175 seconds used......

4 training finished! average train loss: 0.18672844600884922

total clf loss: 27.83509961795062 total adv loss: 25.14553426206112

evaluating...

Confusion matrix, without normalization

[[ 84 41 9]

[ 187 1723 27]

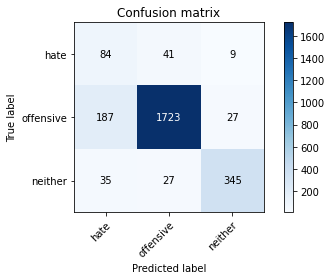
[ 35 27 345]]

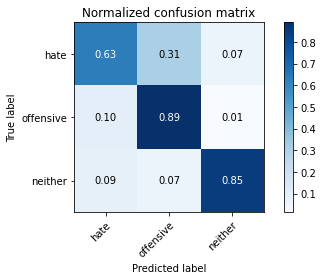
Normalized confusion matrix

[[0.62686567 0.30597015 0.06716418]

[0.09654104 0.88951988 0.01393908]

[0.08599509 0.06633907 0.84766585]]





precision recall f1-score support

hate 0.27 0.63 0.38 134

offensive 0.96 0.89 0.92 1937

neither 0.91 0.85 0.88 407

accuracy 0.87 2478

macro avg 0.71 0.79 0.73 2478

weighted avg 0.92 0.87 0.89 2478

average eval\_loss: 0.4040321245969784, accuracy: 0.8684422921711057

training 5 epoch...

107.08911275863647 seconds used......

5 training finished! average train loss: 0.08610311536573466

total clf loss: 12.613770632538944 total adv loss: 12.701887778006494

evaluating...

Confusion matrix, without normalization

[[ 92 32 10]

[ 196 1707 34]

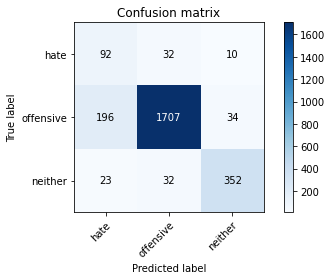
[ 23 32 352]]

Normalized confusion matrix

[[0.68656716 0.23880597 0.07462687]

[0.1011874 0.88125968 0.01755292]

[0.05651106 0.07862408 0.86486486]]





precision recall f1-score support

hate 0.30 0.69 0.41 134

offensive 0.96 0.88 0.92 1937

neither 0.89 0.86 0.88 407

accuracy 0.87 2478

macro avg 0.72 0.81 0.74 2478

weighted avg 0.92 0.87 0.89 2478

average eval\_loss: 0.48978857215373744, accuracy: 0.8680387409200968

Confusion matrix, without normalization

[[ 97 20 7]

[ 543 1345 54]

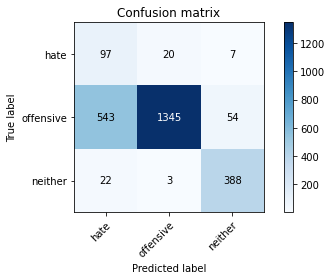
[ 22 3 388]]

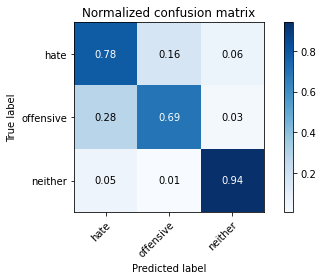
Normalized confusion matrix

[[0.78225806 0.16129032 0.05645161]

[0.27960865 0.69258496 0.02780639]

[0.05326877 0.00726392 0.93946731]]





precision recall f1-score support

hate 0.15 0.78 0.25 124

offensive 0.98 0.69 0.81 1942

neither 0.86 0.94 0.90 413

accuracy 0.74 2479

macro avg 0.66 0.80 0.65 2479

weighted avg 0.92 0.74 0.80 2479

average eval\_loss: 0.6575899431782384, accuracy: 0.7382008874546188

[24]:

0.7382008874546188

alpha=0.2

training 1 epoch...

106.99204540252686 seconds used......

1 training finished! average train loss: 0.7282103675586934

total clf loss: 113.6239121556282 total adv loss: 72.70556196570396

evaluating...

Confusion matrix, without normalization

[[ 101 12 21]

[ 437 1411 89]

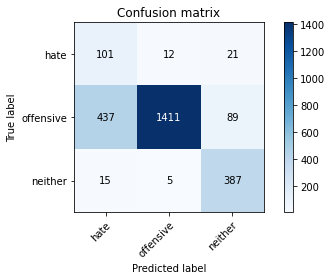
[ 15 5 387]]

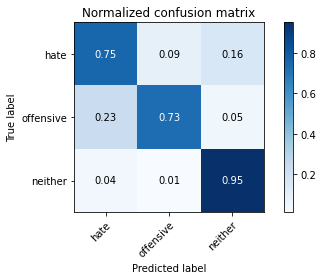
Normalized confusion matrix

[[0.75373134 0.08955224 0.15671642]

[0.22560661 0.72844605 0.04594734]

[0.03685504 0.01228501 0.95085995]]





precision recall f1-score support

hate 0.18 0.75 0.29 134

offensive 0.99 0.73 0.84 1937

neither 0.78 0.95 0.86 407

accuracy 0.77 2478

macro avg 0.65 0.81 0.66 2478

weighted avg 0.91 0.77 0.81 2478

average eval\_loss: 0.5796124935150146, accuracy: 0.7663438256658596

training 2 epoch...

107.00434255599976 seconds used......

2 training finished! average train loss: 0.44158962491730397

total clf loss: 67.69677241146564 total adv loss: 50.115006828680634

evaluating...

Confusion matrix, without normalization

[[ 105 20 9]

[ 298 1607 32]

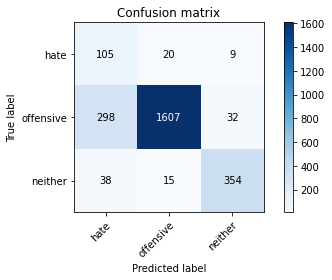
[ 38 15 354]]

Normalized confusion matrix

[[0.78358209 0.14925373 0.06716418]

[0.15384615 0.82963345 0.01652039]

[0.09336609 0.03685504 0.86977887]]





precision recall f1-score support

hate 0.24 0.78 0.37 134

offensive 0.98 0.83 0.90 1937

neither 0.90 0.87 0.88 407

accuracy 0.83 2478

macro avg 0.70 0.83 0.72 2478

weighted avg 0.93 0.83 0.87 2478

average eval\_loss: 0.3966764703152641, accuracy: 0.8337368845843423

training 3 epoch...

107.04097771644592 seconds used......

3 training finished! average train loss: 0.3046281303236769

total clf loss: 45.98026207461953 total adv loss: 38.171443324536085

evaluating...

Confusion matrix, without normalization

[[ 102 16 16]

[ 381 1487 69]

[ 27 9 371]]

Normalized confusion matrix

[[0.76119403 0.11940299 0.11940299]

[0.19669592 0.76768198 0.0356221 ]

[0.06633907 0.02211302 0.91154791]]





precision recall f1-score support

hate 0.20 0.76 0.32 134

offensive 0.98 0.77 0.86 1937

neither 0.81 0.91 0.86 407

accuracy 0.79 2478

macro avg 0.67 0.81 0.68 2478

weighted avg 0.91 0.79 0.83 2478

average eval\_loss: 0.5336936582480708, accuracy: 0.7909604519774012

training 4 epoch...

107.03426671028137 seconds used......

4 training finished! average train loss: 0.17877117758193475

total clf loss: 26.453560680150986 total adv loss: 25.050831764005125

evaluating...

Confusion matrix, without normalization

[[ 100 18 16]

[ 307 1566 64]

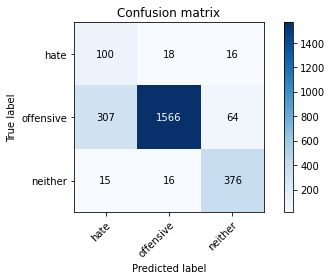
[ 15 16 376]]

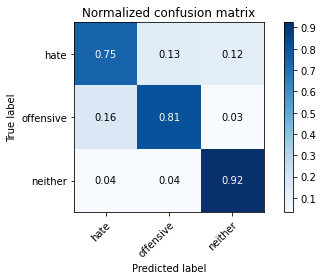
Normalized confusion matrix

[[0.74626866 0.13432836 0.11940299]

[0.15849251 0.8084667 0.03304078]

[0.03685504 0.03931204 0.92383292]]





precision recall f1-score support

hate 0.24 0.75 0.36 134

offensive 0.98 0.81 0.89 1937

neither 0.82 0.92 0.87 407

accuracy 0.82 2478

macro avg 0.68 0.83 0.71 2478

weighted avg 0.91 0.82 0.85 2478

average eval\_loss: 0.5333526631396625, accuracy: 0.8240516545601292

training 5 epoch...

107.05308055877686 seconds used......

5 training finished! average train loss: 0.08740301233906807

total clf loss: 12.902434842195362 total adv loss: 12.402475980110466

evaluating...

Confusion matrix, without normalization

[[ 97 21 16]

[ 291 1595 51]

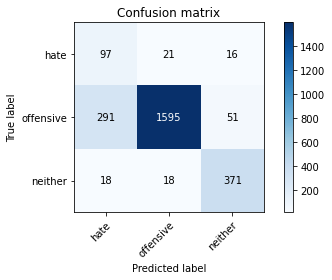
[ 18 18 371]]

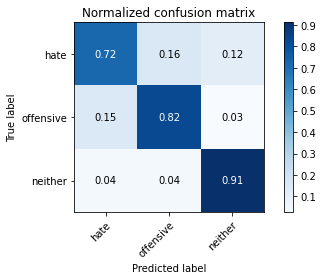
Normalized confusion matrix

[[0.7238806 0.15671642 0.11940299]

[0.15023232 0.82343831 0.02632938]

[0.04422604 0.04422604 0.91154791]]





precision recall f1-score support

hate 0.24 0.72 0.36 134

offensive 0.98 0.82 0.89 1937

neither 0.85 0.91 0.88 407

accuracy 0.83 2478

macro avg 0.69 0.82 0.71 2478

weighted avg 0.92 0.83 0.86 2478

average eval\_loss: 0.5781543588551182, accuracy: 0.8325262308313156

Confusion matrix, without normalization

[[ 94 28 2]

[ 337 1566 39]

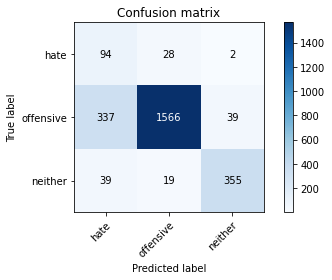
[ 39 19 355]]

Normalized confusion matrix

[[0.75806452 0.22580645 0.01612903]

[0.17353244 0.80638517 0.02008239]

[0.09443099 0.04600484 0.85956416]]





precision recall f1-score support

hate 0.20 0.76 0.32 124

offensive 0.97 0.81 0.88 1942

neither 0.90 0.86 0.88 413

accuracy 0.81 2479

macro avg 0.69 0.81 0.69 2479

weighted avg 0.92 0.81 0.85 2479

average eval\_loss: 0.4468144163970024, accuracy: 0.8128277531262605

[26]:

alpha=0.2

training 1 epoch...

106.9900438785553 seconds used......

1 training finished! average train loss: 0.7400016525590961

total clf loss: 115.5007728934288 total adv loss: 73.69758582860231

evaluating...

Confusion matrix, without normalization

[[ 98 18 18]

[ 375 1499 63]

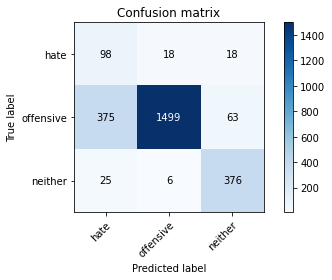
[ 25 6 376]]

Normalized confusion matrix

[[0.73134328 0.13432836 0.13432836]

[0.19359835 0.77387713 0.03252452]

[0.06142506 0.01474201 0.92383292]]





precision recall f1-score support

hate 0.20 0.73 0.31 134

offensive 0.98 0.77 0.87 1937

neither 0.82 0.92 0.87 407

accuracy 0.80 2478

macro avg 0.67 0.81 0.68 2478

weighted avg 0.92 0.80 0.84 2478

average eval\_loss: 0.5395318024581478, accuracy: 0.7962066182405165

training 2 epoch...

107.09231114387512 seconds used......

2 training finished! average train loss: 0.4353028280152516

total clf loss: 66.60779722779989 total adv loss: 50.02750267833471

evaluating...

Confusion matrix, without normalization

[[ 93 21 20]

[ 277 1604 56]

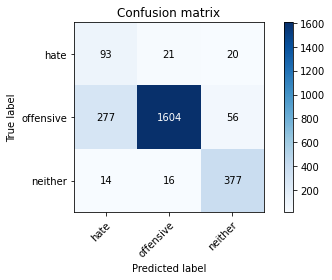
[ 14 16 377]]

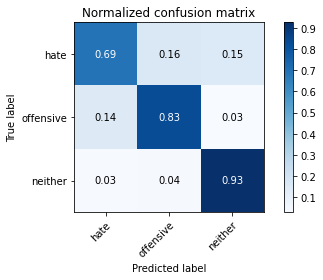
Normalized confusion matrix

[[0.69402985 0.15671642 0.14925373]

[0.14300465 0.82808467 0.02891069]

[0.03439803 0.03931204 0.92628993]]





precision recall f1-score support

hate 0.24 0.69 0.36 134

offensive 0.98 0.83 0.90 1937

neither 0.83 0.93 0.88 407

accuracy 0.84 2478

macro avg 0.68 0.82 0.71 2478

weighted avg 0.91 0.84 0.86 2478

average eval\_loss: 0.4159763976931572, accuracy: 0.8369652945924132

training 3 epoch...

106.93765425682068 seconds used......

3 training finished! average train loss: 0.2973518843804909

total clf loss: 44.780519746243954 total adv loss: 37.76705936715007

evaluating...

Confusion matrix, without normalization

[[ 104 16 14]

[ 416 1462 59]

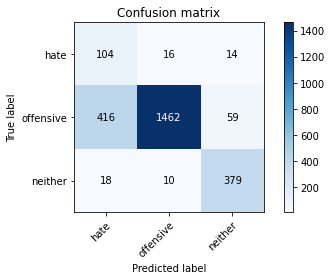
[ 18 10 379]]

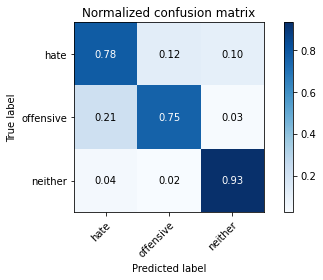
Normalized confusion matrix

[[0.7761194 0.11940299 0.10447761]

[0.2147651 0.75477543 0.03045947]

[0.04422604 0.02457002 0.93120393]]





precision recall f1-score support

hate 0.19 0.78 0.31 134

offensive 0.98 0.75 0.85 1937

neither 0.84 0.93 0.88 407

accuracy 0.78 2478

macro avg 0.67 0.82 0.68 2478

weighted avg 0.92 0.78 0.83 2478

average eval\_loss: 0.5641586207334073, accuracy: 0.7849071832122679

training 4 epoch...

106.95657515525818 seconds used......

4 training finished! average train loss: 0.18139319769530135

total clf loss: 26.78712230361998 total adv loss: 25.690402201376855

evaluating...

Confusion matrix, without normalization

[[ 110 16 8]

[ 447 1465 25]

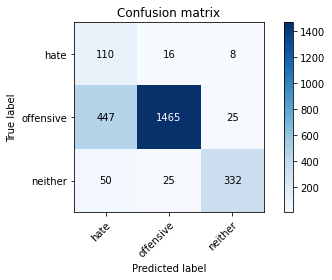
[ 50 25 332]]

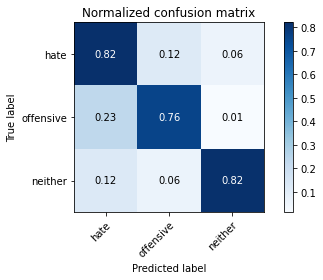
Normalized confusion matrix

[[0.82089552 0.11940299 0.05970149]

[0.23076923 0.75632421 0.01290656]

[0.12285012 0.06142506 0.81572482]]





precision recall f1-score support

hate 0.18 0.82 0.30 134

offensive 0.97 0.76 0.85 1937

neither 0.91 0.82 0.86 407

accuracy 0.77 2478

macro avg 0.69 0.80 0.67 2478

weighted avg 0.92 0.77 0.82 2478

average eval\_loss: 0.7256636040345315, accuracy: 0.7695722356739306

training 5 epoch...

107.0853021144867 seconds used......

5 training finished! average train loss: 0.11906608659244905

total clf loss: 17.501484506763518 total adv loss: 17.270733218290843

evaluating...

Confusion matrix, without normalization

[[ 98 23 13]

[ 295 1604 38]

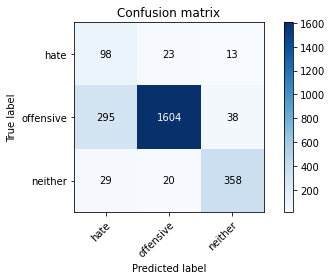
[ 29 20 358]]

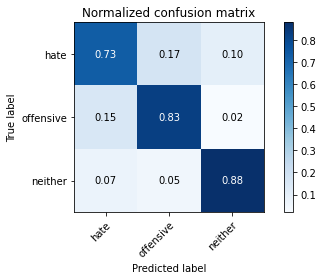
Normalized confusion matrix

[[0.73134328 0.17164179 0.09701493]

[0.15229737 0.82808467 0.01961797]

[0.07125307 0.04914005 0.87960688]]





precision recall f1-score support

hate 0.23 0.73 0.35 134

offensive 0.97 0.83 0.90 1937

neither 0.88 0.88 0.88 407

accuracy 0.83 2478

macro avg 0.69 0.81 0.71 2478

weighted avg 0.92 0.83 0.86 2478

average eval\_loss: 0.6074743557481035, accuracy: 0.831315577078289

Confusion matrix, without normalization

[[ 92 28 4]

[ 469 1434 39]

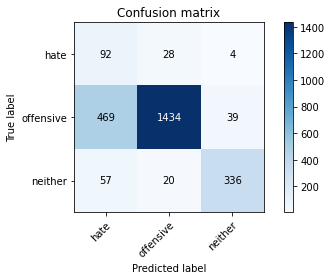
[ 57 20 336]]

Normalized confusion matrix

[[0.74193548 0.22580645 0.03225806]

[0.2415036 0.73841401 0.02008239]

[0.13801453 0.04842615 0.81355932]]





precision recall f1-score support

hate 0.15 0.74 0.25 124

offensive 0.97 0.74 0.84 1942

neither 0.89 0.81 0.85 413

accuracy 0.75 2479

macro avg 0.67 0.76 0.64 2479

weighted avg 0.91 0.75 0.81 2479

average eval\_loss: 0.8326451348200921, accuracy: 0.7511093182734974

[29]:

0.7511093182734974