1 run

hate as offensive alpha=0.05

training 1 epoch...

107.3185224533081 seconds used......

1 training finished! average train loss: 0.6399346016855403

total clf loss: 115.52645055949688 total adv loss: 57.95920928567648

evaluating...

Confusion matrix, without normalization

[[ 95 22 17]

[ 354 1523 60]

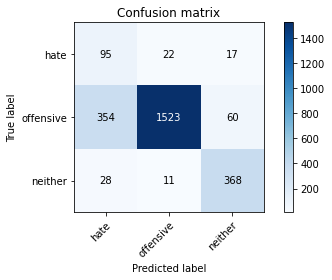
[ 28 11 368]]

Normalized confusion matrix

[[0.70895522 0.1641791 0.12686567]

[0.18275684 0.78626742 0.03097574]

[0.06879607 0.02702703 0.9041769 ]]





precision recall f1-score support

hate 0.20 0.71 0.31 134

offensive 0.98 0.79 0.87 1937

neither 0.83 0.90 0.86 407

accuracy 0.80 2478

macro avg 0.67 0.80 0.68 2478

weighted avg 0.91 0.80 0.84 2478

average eval\_loss: 0.4864188460332732, accuracy: 0.801452784503632

training 2 epoch...

106.59346079826355 seconds used......

2 training finished! average train loss: 0.38244483717293903

total clf loss: 68.69447787851095 total adv loss: 27.683729426003993

evaluating...

Confusion matrix, without normalization

[[ 110 13 11]

[ 475 1432 30]

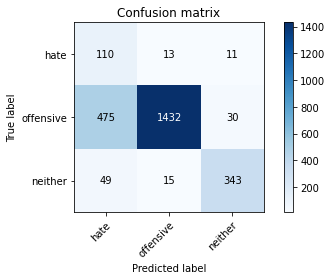
[ 49 15 343]]

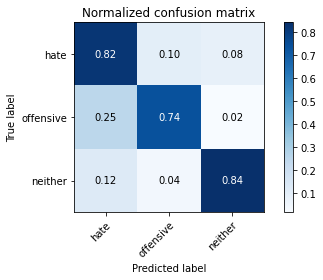
Normalized confusion matrix

[[0.82089552 0.09701493 0.08208955]

[0.24522457 0.73928756 0.01548787]

[0.12039312 0.03685504 0.84275184]]





precision recall f1-score support

hate 0.17 0.82 0.29 134

offensive 0.98 0.74 0.84 1937

neither 0.89 0.84 0.87 407

accuracy 0.76 2478

macro avg 0.68 0.80 0.67 2478

weighted avg 0.92 0.76 0.82 2478

average eval\_loss: 0.5432009310010941, accuracy: 0.7606941081517352

training 3 epoch...

106.71440577507019 seconds used......

3 training finished! average train loss: 0.2581874185919084

total clf loss: 46.1933504268527 total adv loss: 15.04729157476686

evaluating...

Confusion matrix, without normalization

[[ 99 18 17]

[ 408 1483 46]

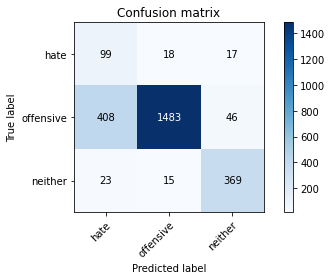
[ 23 15 369]]

Normalized confusion matrix

[[0.73880597 0.13432836 0.12686567]

[0.210635 0.76561693 0.02374806]

[0.05651106 0.03685504 0.90663391]]





precision recall f1-score support

hate 0.19 0.74 0.30 134

offensive 0.98 0.77 0.86 1937

neither 0.85 0.91 0.88 407

accuracy 0.79 2478

macro avg 0.67 0.80 0.68 2478

weighted avg 0.92 0.79 0.83 2478

average eval\_loss: 0.5277010141721656, accuracy: 0.7873284907183212

training 4 epoch...

106.67679476737976 seconds used......

4 training finished! average train loss: 0.15613869443239475

total clf loss: 27.79887641593814 total adv loss: 6.369322533078957

evaluating...

Confusion matrix, without normalization

[[ 88 25 21]

[ 232 1609 96]

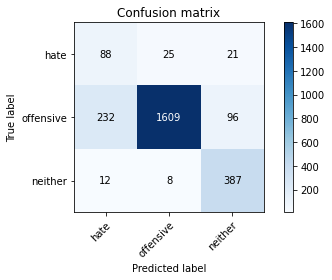
[ 12 8 387]]

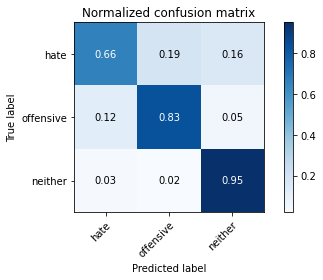
Normalized confusion matrix

[[0.65671642 0.18656716 0.15671642]

[0.11977284 0.83066598 0.04956118]

[0.02948403 0.01965602 0.95085995]]





precision recall f1-score support

hate 0.27 0.66 0.38 134

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 0.91 0.84 0.86 2478

average eval\_loss: 0.5731069597926352, accuracy: 0.841000807102502

training 5 epoch...

106.57078289985657 seconds used......

5 training finished! average train loss: 0.08917058702420698

total clf loss: 15.87949653249234 total adv loss: 3.7094634177410626

evaluating...

Confusion matrix, without normalization

[[ 91 27 16]

[ 348 1549 40]

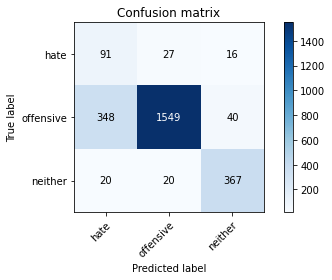
[ 20 20 367]]

Normalized confusion matrix

[[0.67910448 0.20149254 0.11940299]

[0.17965927 0.79969024 0.02065049]

[0.04914005 0.04914005 0.9017199 ]]





precision recall f1-score support

hate 0.20 0.68 0.31 134

offensive 0.97 0.80 0.88 1937

neither 0.87 0.90 0.88 407

accuracy 0.81 2478

macro avg 0.68 0.79 0.69 2478

weighted avg 0.91 0.81 0.85 2478

average eval\_loss: 0.6112767064403142, accuracy: 0.8099273607748184

Confusion matrix, without normalization

[[ 96 25 3]

[ 471 1432 39]

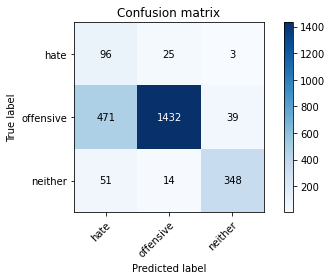
[ 51 14 348]]

Normalized confusion matrix

[[0.77419355 0.2016129 0.02419355]

[0.24253347 0.73738414 0.02008239]

[0.12348668 0.03389831 0.84261501]]





precision recall f1-score support

hate 0.16 0.77 0.26 124

offensive 0.97 0.74 0.84 1942

neither 0.89 0.84 0.87 413

accuracy 0.76 2479

macro avg 0.67 0.78 0.65 2479

weighted avg 0.92 0.76 0.81 2479

average eval\_loss: 0.5883548581792463, accuracy: 0.7567567567567568

[18]:

0.7567567567567568

2 run

training 1 epoch...

106.6114227771759 seconds used......

1 training finished! average train loss: 0.6315586253238673

total clf loss: 114.0428679510951 total adv loss: 57.77100170776248

evaluating...

Confusion matrix, without normalization

[[ 81 33 20]

[ 230 1651 56]

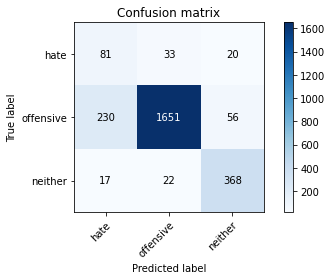
[ 17 22 368]]

Normalized confusion matrix

[[0.60447761 0.24626866 0.14925373]

[0.11874032 0.85234899 0.02891069]

[0.04176904 0.05405405 0.9041769 ]]





precision recall f1-score support

hate 0.25 0.60 0.35 134

offensive 0.97 0.85 0.91 1937

neither 0.83 0.90 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: 0.39916819910849294, accuracy: 0.847457627118644

training 2 epoch...

106.61712574958801 seconds used......

2 training finished! average train loss: 0.3673171108927239

total clf loss: 65.9577082619071 total adv loss: 26.197937555145472

evaluating...

Confusion matrix, without normalization

[[ 99 16 19]

[ 362 1506 69]

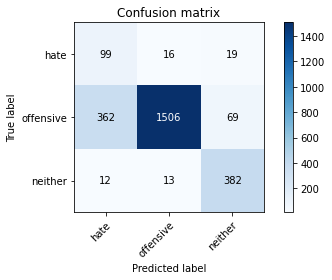
[ 12 13 382]]

Normalized confusion matrix

[[0.73880597 0.11940299 0.14179104]

[0.18688694 0.77749097 0.0356221 ]

[0.02948403 0.03194103 0.93857494]]





precision recall f1-score support

hate 0.21 0.74 0.33 134

offensive 0.98 0.78 0.87 1937

neither 0.81 0.94 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: 0.4615205433219671, accuracy: 0.8018563357546409

training 3 epoch...

106.69355249404907 seconds used......

3 training finished! average train loss: 0.25322923256846314

total clf loss: 45.28623527288437 total adv loss: 14.357805907959118

evaluating...

Confusion matrix, without normalization

[[ 101 23 10]

[ 294 1584 59]

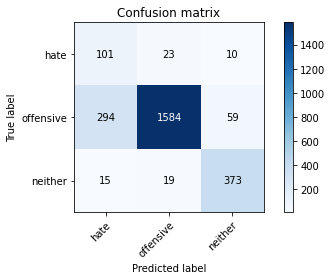
[ 15 19 373]]

Normalized confusion matrix

[[0.75373134 0.17164179 0.07462687]

[0.1517811 0.81775942 0.03045947]

[0.03685504 0.04668305 0.91646192]]





precision recall f1-score support

hate 0.25 0.75 0.37 134

offensive 0.97 0.82 0.89 1937

neither 0.84 0.92 0.88 407

accuracy 0.83 2478

macro avg 0.69 0.83 0.71 2478

weighted avg 0.91 0.83 0.86 2478

average eval\_loss: 0.4536690799399249, accuracy: 0.8305084745762712

training 4 epoch...

106.70135402679443 seconds used......

4 training finished! average train loss: 0.1390631206873381

total clf loss: 24.764556505717337 total adv loss: 5.788942882383708

evaluating...

Confusion matrix, without normalization

[[ 94 26 14]

[ 299 1579 59]

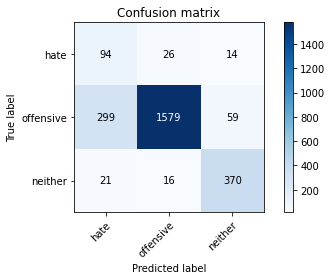
[ 21 16 370]]

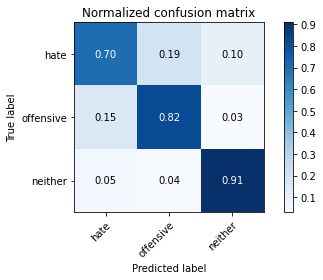
Normalized confusion matrix

[[0.70149254 0.19402985 0.10447761]

[0.15436242 0.81517811 0.03045947]

[0.05159705 0.03931204 0.90909091]]





precision recall f1-score support

hate 0.23 0.70 0.34 134

offensive 0.97 0.82 0.89 1937

neither 0.84 0.91 0.87 407

accuracy 0.82 2478

macro avg 0.68 0.81 0.70 2478

weighted avg 0.91 0.82 0.86 2478

average eval\_loss: 0.5469943736830065, accuracy: 0.8244552058111381

training 5 epoch...

106.67122554779053 seconds used......

5 training finished! average train loss: 0.07311606170630759

total clf loss: 12.992955232970417 total adv loss: 2.4905668711489852

evaluating...

Confusion matrix, without normalization

[[ 99 18 17]

[ 335 1523 79]

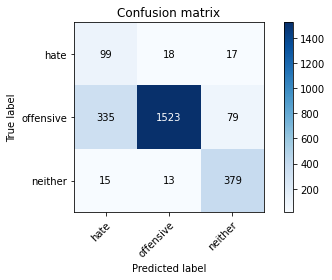
[ 15 13 379]]

Normalized confusion matrix

[[0.73880597 0.13432836 0.12686567]

[0.17294786 0.78626742 0.04078472]

[0.03685504 0.03194103 0.93120393]]





precision recall f1-score support

hate 0.22 0.74 0.34 134

offensive 0.98 0.79 0.87 1937

neither 0.80 0.93 0.86 407

accuracy 0.81 2478

macro avg 0.67 0.82 0.69 2478

weighted avg 0.91 0.81 0.84 2478

average eval\_loss: 0.6802248331688104, accuracy: 0.8075060532687651

Confusion matrix, without normalization

[[ 83 33 8]

[ 333 1543 66]

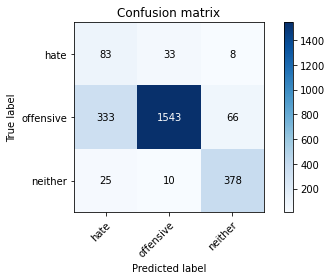
[ 25 10 378]]

Normalized confusion matrix

[[0.66935484 0.26612903 0.06451613]

[0.17147271 0.79454171 0.03398558]

[0.06053269 0.02421308 0.91525424]]





precision recall f1-score support

hate 0.19 0.67 0.29 124

offensive 0.97 0.79 0.87 1942

neither 0.84 0.92 0.87 413

accuracy 0.81 2479

macro avg 0.67 0.79 0.68 2479

weighted avg 0.91 0.81 0.85 2479

average eval\_loss: 0.518293627267403, accuracy: 0.8083904800322711

[20]:

0.8083904800322711