1 run

hate as no-off

alpha=2

training 1 epoch...

518.154595375061 seconds used......

1 training finished! average train loss: 1.2629019072833094

total clf loss: 349.6439098119736 total adv loss: 198.9970094896853

evaluating...

Confusion matrix, without normalization

[[ 273 118 99]

[ 239 2373 85]

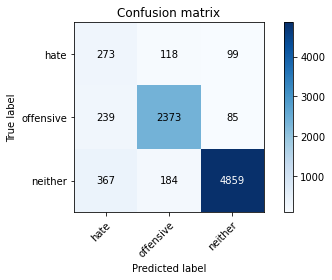
[ 367 184 4859]]

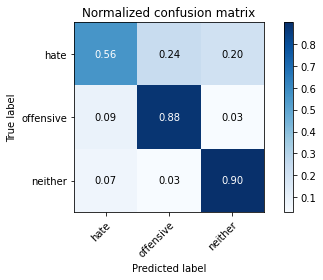
Normalized confusion matrix

[[0.55714286 0.24081633 0.20204082]

[0.08861698 0.87986652 0.0315165 ]

[0.06783734 0.03401109 0.89815157]]





precision recall f1-score support

hate 0.31 0.56 0.40 490

offensive 0.89 0.88 0.88 2697

neither 0.96 0.90 0.93 5410

accuracy 0.87 8597

macro avg 0.72 0.78 0.74 8597

weighted avg 0.90 0.87 0.88 8597

average eval\_loss: 0.3749695762472097, accuracy: 0.8729789461440037

training 2 epoch...

518.9343659877777 seconds used......

2 training finished! average train loss: 0.9677598921431078

total clf loss: 267.87848914414644 total adv loss: 152.51768307574093

evaluating...

Confusion matrix, without normalization

[[ 291 93 106]

[ 330 2243 124]

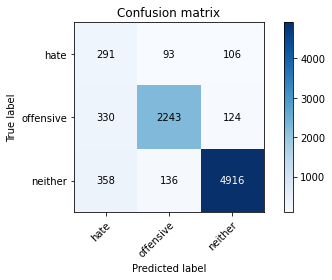
[ 358 136 4916]]

Normalized confusion matrix

[[0.59387755 0.18979592 0.21632653]

[0.12235818 0.83166481 0.04597701]

[0.06617375 0.02513863 0.90868762]]





precision recall f1-score support

hate 0.30 0.59 0.40 490

offensive 0.91 0.83 0.87 2697

neither 0.96 0.91 0.93 5410

accuracy 0.87 8597

macro avg 0.72 0.78 0.73 8597

weighted avg 0.90 0.87 0.88 8597

average eval\_loss: 0.3826631617234197, accuracy: 0.8665813655926486

training 3 epoch...

519.219966173172 seconds used......

3 training finished! average train loss: 0.7526904379294531

total clf loss: 207.90088015794754 total adv loss: 118.84593029692769

evaluating...

Confusion matrix, without normalization

[[ 306 117 67]

[ 333 2299 65]

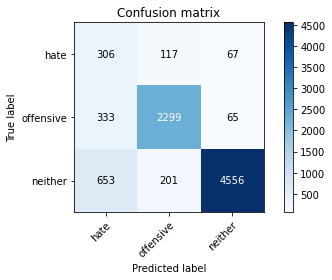
[ 653 201 4556]]

Normalized confusion matrix

[[0.6244898 0.23877551 0.13673469]

[0.12347052 0.85242862 0.02410085]

[0.1207024 0.03715342 0.84214418]]





precision recall f1-score support

hate 0.24 0.62 0.34 490

offensive 0.88 0.85 0.87 2697

neither 0.97 0.84 0.90 5410

accuracy 0.83 8597

macro avg 0.70 0.77 0.70 8597

weighted avg 0.90 0.83 0.86 8597

average eval\_loss: 0.48786980662234997, accuracy: 0.8329649877864371

training 4 epoch...

519.1631343364716 seconds used......

4 training finished! average train loss: 0.5499944566321131

total clf loss: 150.66023406386375 total adv loss: 87.46824196795933

evaluating...

Confusion matrix, without normalization

[[ 301 99 90]

[ 430 2181 86]

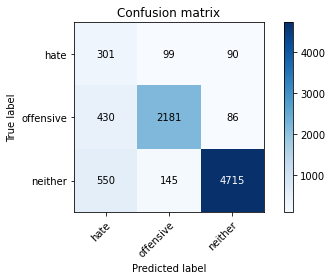
[ 550 145 4715]]

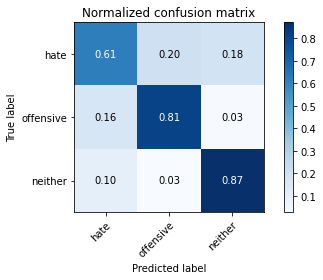
Normalized confusion matrix

[[0.61428571 0.20204082 0.18367347]

[0.15943641 0.80867631 0.03188728]

[0.10166359 0.02680222 0.8715342 ]]





precision recall f1-score support

hate 0.23 0.61 0.34 490

offensive 0.90 0.81 0.85 2697

neither 0.96 0.87 0.92 5410

accuracy 0.84 8597

macro avg 0.70 0.76 0.70 8597

weighted avg 0.90 0.84 0.86 8597

average eval\_loss: 0.5150090045360631, accuracy: 0.837152495056415

training 5 epoch...

518.7880721092224 seconds used......

5 training finished! average train loss: 0.3999192299812126

total clf loss: 108.64018574915826 total adv loss: 64.05599936182261

evaluating...

Confusion matrix, without normalization

[[ 311 98 81]

[ 470 2146 81]

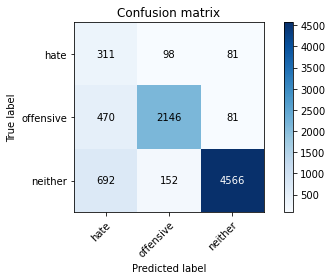
[ 692 152 4566]]

Normalized confusion matrix

[[0.63469388 0.2 0.16530612]

[0.1742677 0.79569892 0.03003337]

[0.12791128 0.02809612 0.84399261]]





precision recall f1-score support

hate 0.21 0.63 0.32 490

offensive 0.90 0.80 0.84 2697

neither 0.97 0.84 0.90 5410

accuracy 0.82 8597

macro avg 0.69 0.76 0.69 8597

weighted avg 0.90 0.82 0.85 8597

average eval\_loss: 0.6641345571769878, accuracy: 0.8169128765848552

model of 5 epoch

Confusion matrix, without normalization

[[ 350 101 81]

[ 439 2196 75]

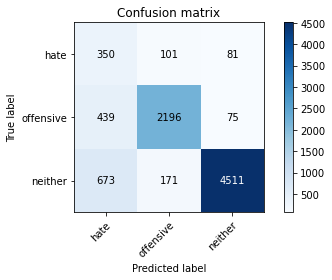
[ 673 171 4511]]

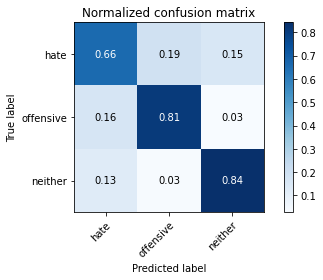
Normalized confusion matrix

[[0.65789474 0.18984962 0.15225564]

[0.16199262 0.8103321 0.02767528]

[0.12567694 0.03193277 0.84239029]]





precision recall f1-score support

hate 0.24 0.66 0.35 532

offensive 0.89 0.81 0.85 2710

neither 0.97 0.84 0.90 5355

accuracy 0.82 8597

macro avg 0.70 0.77 0.70 8597

weighted avg 0.90 0.82 0.85 8597

average eval\_loss: 0.663271612071887, accuracy: 0.8208677445620566

[20]:

0.8208677445620566

2 run

training 1 epoch...

518.9745042324066 seconds used......

1 training finished! average train loss: 1.2422417221037116

total clf loss: 347.64145812392235 total adv loss: 193.882820636034

evaluating...

Confusion matrix, without normalization

[[ 333 70 87]

[ 475 2130 92]

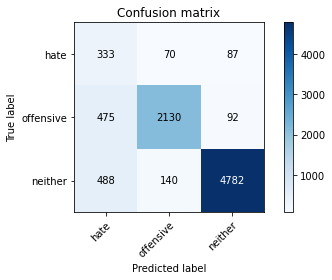
[ 488 140 4782]]

Normalized confusion matrix

[[0.67959184 0.14285714 0.17755102]

[0.17612162 0.78976641 0.03411198]

[0.09020333 0.025878 0.88391867]]





precision recall f1-score support

hate 0.26 0.68 0.37 490

offensive 0.91 0.79 0.85 2697

neither 0.96 0.88 0.92 5410

accuracy 0.84 8597

macro avg 0.71 0.78 0.71 8597

weighted avg 0.91 0.84 0.87 8597

average eval\_loss: 0.4554126424844875, accuracy: 0.8427358380830522

training 2 epoch...

519.1080183982849 seconds used......

2 training finished! average train loss: 0.9617123863584286

total clf loss: 266.2191577255726 total adv loss: 151.55728725716472

evaluating...

Confusion matrix, without normalization

[[ 300 102 88]

[ 328 2278 91]

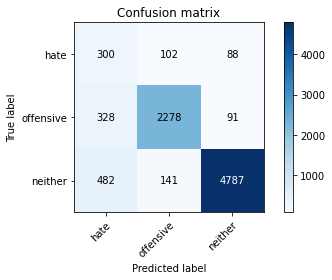
[ 482 141 4787]]

Normalized confusion matrix

[[0.6122449 0.20816327 0.17959184]

[0.12161661 0.8446422 0.03374119]

[0.08909427 0.02606285 0.88484288]]





precision recall f1-score support

hate 0.27 0.61 0.38 490

offensive 0.90 0.84 0.87 2697

neither 0.96 0.88 0.92 5410

accuracy 0.86 8597

macro avg 0.71 0.78 0.72 8597

weighted avg 0.91 0.86 0.88 8597

average eval\_loss: 0.40927249023387596, accuracy: 0.8566941956496452

training 3 epoch...

518.9869477748871 seconds used......

3 training finished! average train loss: 0.7534278227972824

total clf loss: 210.34773695468903 total adv loss: 117.84076652862132

evaluating...

Confusion matrix, without normalization

[[ 284 118 88]

[ 272 2346 79]

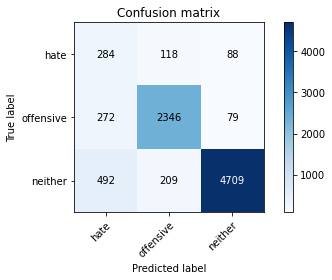
[ 492 209 4709]]

Normalized confusion matrix

[[0.57959184 0.24081633 0.17959184]

[0.1008528 0.86985539 0.02929181]

[0.0909427 0.03863216 0.87042514]]





precision recall f1-score support

hate 0.27 0.58 0.37 490

offensive 0.88 0.87 0.87 2697

neither 0.97 0.87 0.92 5410

accuracy 0.85 8597

macro avg 0.70 0.77 0.72 8597

weighted avg 0.90 0.85 0.87 8597

average eval\_loss: 0.450802528936156, accuracy: 0.8536698848435501

training 4 epoch...

518.9877264499664 seconds used......

4 training finished! average train loss: 0.5491767514849434

total clf loss: 150.5525097027421 total adv loss: 87.28006394580007

evaluating...

Confusion matrix, without normalization

[[ 311 80 99]

[ 506 2062 129]

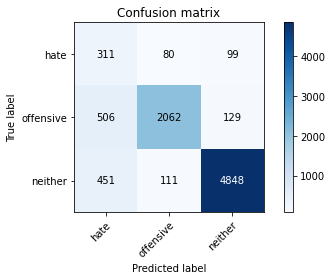
[ 451 111 4848]]

Normalized confusion matrix

[[0.63469388 0.16326531 0.20204082]

[0.18761587 0.76455321 0.04783092]

[0.08336414 0.02051756 0.8961183 ]]





precision recall f1-score support

hate 0.25 0.63 0.35 490

offensive 0.92 0.76 0.83 2697

neither 0.96 0.90 0.92 5410

accuracy 0.84 8597

macro avg 0.71 0.77 0.70 8597

weighted avg 0.90 0.84 0.86 8597

average eval\_loss: 0.508204358538916, accuracy: 0.8399441665697336

training 5 epoch...

519.0987253189087 seconds used......

5 training finished! average train loss: 0.38450128663759175

total clf loss: 105.11772003956139 total adv loss: 61.25352091179229

evaluating...

Confusion matrix, without normalization

[[ 342 83 65]

[ 493 2148 56]

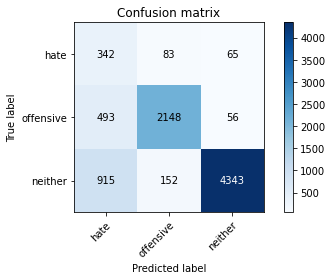
[ 915 152 4343]]

Normalized confusion matrix

[[0.69795918 0.16938776 0.13265306]

[0.1827957 0.79644049 0.02076381]

[0.16913124 0.02809612 0.80277264]]





precision recall f1-score support

hate 0.20 0.70 0.31 490

offensive 0.90 0.80 0.85 2697

neither 0.97 0.80 0.88 5410

accuracy 0.79 8597

macro avg 0.69 0.77 0.68 8597

weighted avg 0.91 0.79 0.84 8597

average eval\_loss: 0.7281716783119495, accuracy: 0.7948121437710829

Confusion matrix, without normalization

[[ 362 101 69]

[ 457 2194 59]

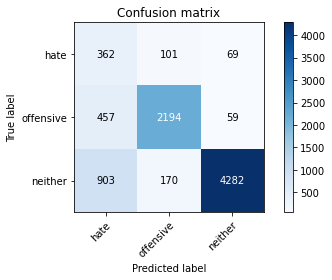
[ 903 170 4282]]

Normalized confusion matrix

[[0.68045113 0.18984962 0.12969925]

[0.16863469 0.8095941 0.02177122]

[0.16862745 0.03174603 0.79962652]]





precision recall f1-score support

hate 0.21 0.68 0.32 532

offensive 0.89 0.81 0.85 2710

neither 0.97 0.80 0.88 5355

accuracy 0.80 8597

macro avg 0.69 0.76 0.68 8597

weighted avg 0.90 0.80 0.83 8597

average eval\_loss: 0.7386458080478533, accuracy: 0.7953937420030243

[30]:

0.7953937420030243

3 run

training 1 epoch...

526.4603109359741 seconds used......

1 training finished! average train loss: 1.2470652632717345

total clf loss: 349.80069221556187 total adv loss: 194.23097119480371

evaluating...

Confusion matrix, without normalization

[[ 334 78 78]

[ 410 2216 71]

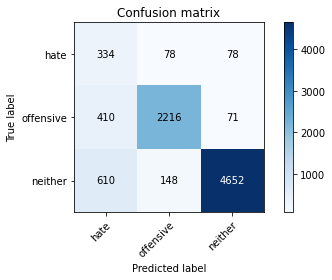
[ 610 148 4652]]

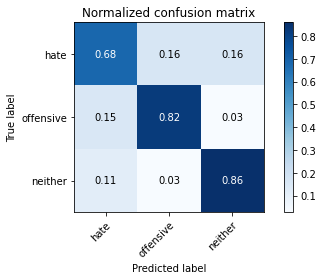
Normalized confusion matrix

[[0.68163265 0.15918367 0.15918367]

[0.15202076 0.82165369 0.02632555]

[0.11275416 0.02735675 0.85988909]]





precision recall f1-score support

hate 0.25 0.68 0.36 490

offensive 0.91 0.82 0.86 2697

neither 0.97 0.86 0.91 5410

accuracy 0.84 8597

macro avg 0.71 0.79 0.71 8597

weighted avg 0.91 0.84 0.86 8597

average eval\_loss: 0.4611202320387197, accuracy: 0.8377340932883564

training 2 epoch...

525.504105091095 seconds used......

2 training finished! average train loss: 0.9610139557940734

total clf loss: 264.308432854712 total adv loss: 152.30591443181038

evaluating...

Confusion matrix, without normalization

[[ 316 76 98]

[ 469 2129 99]

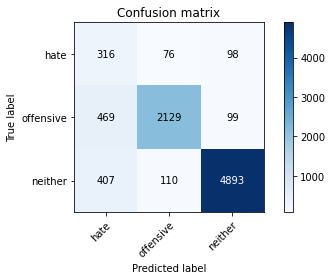
[ 407 110 4893]]

Normalized confusion matrix

[[0.64489796 0.15510204 0.2 ]

[0.17389692 0.78939562 0.03670745]

[0.07523105 0.02033272 0.90443623]]





precision recall f1-score support

hate 0.27 0.64 0.38 490

offensive 0.92 0.79 0.85 2697

neither 0.96 0.90 0.93 5410

accuracy 0.85 8597

macro avg 0.72 0.78 0.72 8597

weighted avg 0.91 0.85 0.87 8597

average eval\_loss: 0.39350874259374863, accuracy: 0.8535535651971617

training 3 epoch...

526.08669090271 seconds used......

3 training finished! average train loss: 0.7570305495095011

total clf loss: 210.8518292531371 total adv loss: 118.6551278559491

evaluating...

Confusion matrix, without normalization

[[ 352 73 65]

[ 490 2151 56]

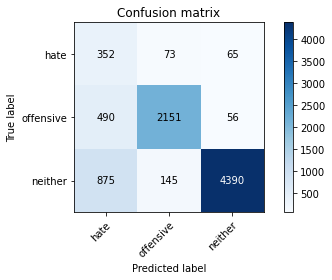
[ 875 145 4390]]

Normalized confusion matrix

[[0.71836735 0.14897959 0.13265306]

[0.18168335 0.79755284 0.02076381]

[0.16173752 0.02680222 0.81146026]]





precision recall f1-score support

hate 0.21 0.72 0.32 490

offensive 0.91 0.80 0.85 2697

neither 0.97 0.81 0.88 5410

accuracy 0.80 8597

macro avg 0.70 0.78 0.68 8597

weighted avg 0.91 0.80 0.84 8597

average eval\_loss: 0.5750295840203762, accuracy: 0.8017913225543795

training 4 epoch...

525.6065788269043 seconds used......

4 training finished! average train loss: 0.5424748592431078

total clf loss: 150.20810917392373 total adv loss: 85.4685035077855

evaluating...

Confusion matrix, without normalization

[[ 342 67 81]

[ 601 2012 84]

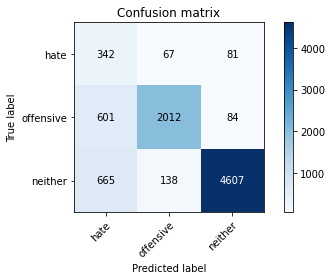
[ 665 138 4607]]

Normalized confusion matrix

[[0.69795918 0.13673469 0.16530612]

[0.22284019 0.74601409 0.03114572]

[0.12292052 0.02550832 0.85157116]]





precision recall f1-score support

hate 0.21 0.70 0.33 490

offensive 0.91 0.75 0.82 2697

neither 0.97 0.85 0.90 5410

accuracy 0.81 8597

macro avg 0.70 0.77 0.68 8597

weighted avg 0.90 0.81 0.84 8597

average eval\_loss: 0.5794213418523932, accuracy: 0.8097010585087822

training 5 epoch...

525.0519721508026 seconds used......

5 training finished! average train loss: 0.37972519130483773

total clf loss: 104.1938804499805 total adv loss: 60.30171628901735

evaluating...

Confusion matrix, without normalization

[[ 299 103 88]

[ 453 2179 65]

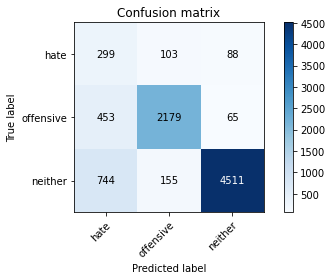
[ 744 155 4511]]

Normalized confusion matrix

[[0.61020408 0.21020408 0.17959184]

[0.1679644 0.80793474 0.02410085]

[0.13752311 0.02865065 0.83382625]]





precision recall f1-score support

hate 0.20 0.61 0.30 490

offensive 0.89 0.81 0.85 2697

neither 0.97 0.83 0.90 5410

accuracy 0.81 8597

macro avg 0.69 0.75 0.68 8597

weighted avg 0.90 0.81 0.85 8597

average eval\_loss: 0.6212594184418058, accuracy: 0.8129580086076539

Confusion matrix, without normalization

[[ 385 82 65]

[ 477 2191 42]

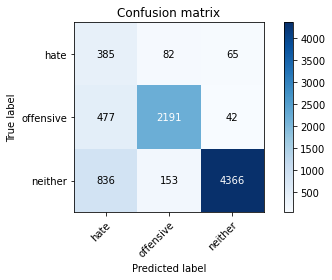
[ 836 153 4366]]

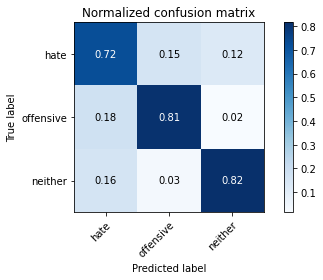
Normalized confusion matrix

[[0.72368421 0.15413534 0.12218045]

[0.17601476 0.80848708 0.01549815]

[0.15611578 0.02857143 0.81531279]]





precision recall f1-score support

hate 0.23 0.72 0.35 532

offensive 0.90 0.81 0.85 2710

neither 0.98 0.82 0.89 5355

accuracy 0.81 8597

macro avg 0.70 0.78 0.70 8597

weighted avg 0.91 0.81 0.84 8597

average eval\_loss: 0.565979355868212, accuracy: 0.8074909852274049

[18]:

0.8074909852274049