hate as offensive

alpha=0.05

1 run

training 1 epoch...

810.3946263790131 seconds used......

1 training finished! average train loss: 0.640300583190037

total clf loss: 794.683562964201 total adv loss: 436.8151833638549

evaluating...

Confusion matrix, without normalization

[[ 615 177 154]

[ 651 4307 246]

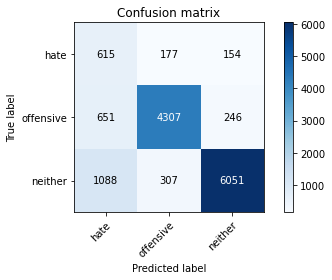
[1088 307 6051]]

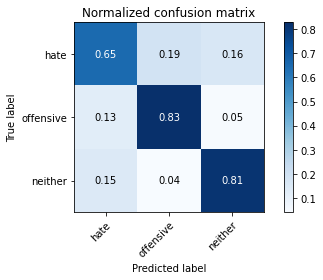
Normalized confusion matrix

[[0.65010571 0.18710359 0.1627907 ]

[0.12509608 0.82763259 0.04727133]

[0.14611872 0.04123019 0.81265109]]





precision recall f1-score support

hate 0.26 0.65 0.37 946

offensive 0.90 0.83 0.86 5204

neither 0.94 0.81 0.87 7446

accuracy 0.81 13596

macro avg 0.70 0.76 0.70 13596

weighted avg 0.88 0.81 0.83 13596

average eval\_loss: 0.49617595920229657, accuracy: 0.8070756104736687

training 2 epoch...

810.8135912418365 seconds used......

2 training finished! average train loss: 0.4917038623752827

total clf loss: 609.272312708199 total adv loss: 315.71502613835037

evaluating...

Confusion matrix, without normalization

[[ 654 152 140]

[ 775 4237 192]

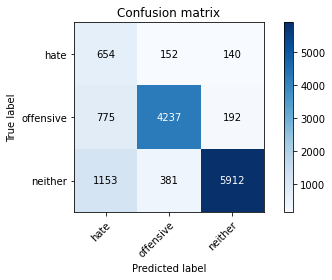
[1153 381 5912]]

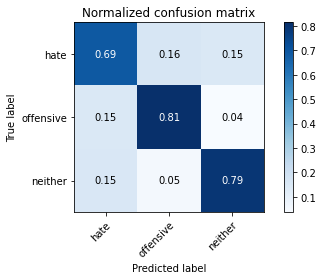
Normalized confusion matrix

[[0.69133192 0.16067653 0.14799154]

[0.1489239 0.8141814 0.0368947 ]

[0.15484824 0.05116841 0.79398335]]





precision recall f1-score support

hate 0.25 0.69 0.37 946

offensive 0.89 0.81 0.85 5204

neither 0.95 0.79 0.86 7446

accuracy 0.79 13596

macro avg 0.70 0.77 0.69 13596

weighted avg 0.88 0.79 0.82 13596

average eval\_loss: 0.515514539105489, accuracy: 0.7945719329214475

training 3 epoch...

811.2139692306519 seconds used......

3 training finished! average train loss: 0.373122142419503

total clf loss: 461.25502698868513 total adv loss: 217.9320118073374

evaluating...

Confusion matrix, without normalization

[[ 712 127 107]

[ 937 4102 165]

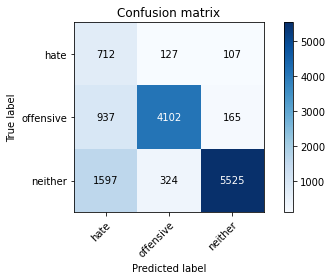
[1597 324 5525]]

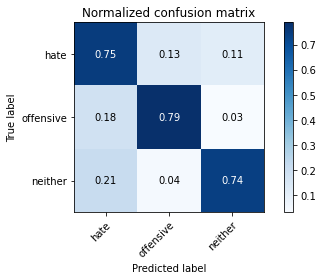
Normalized confusion matrix

[[0.75264271 0.13424947 0.11310782]

[0.1800538 0.78823982 0.03170638]

[0.21447757 0.0435133 0.74200913]]





precision recall f1-score support

hate 0.22 0.75 0.34 946

offensive 0.90 0.79 0.84 5204

neither 0.95 0.74 0.83 7446

accuracy 0.76 13596

macro avg 0.69 0.76 0.67 13596

weighted avg 0.88 0.76 0.80 13596

average eval\_loss: 0.653255897398819, accuracy: 0.760444248308326

Confusion matrix, without normalization

[[1445 165 225]

[1382 4233 310]

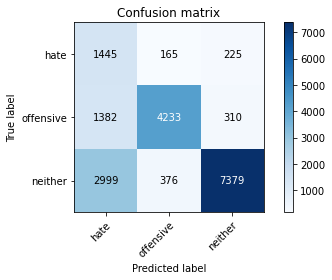
[2999 376 7379]]

Normalized confusion matrix

[[0.78746594 0.08991826 0.1226158 ]

[0.23324895 0.71443038 0.05232068]

[0.27887298 0.03496373 0.68616329]]





precision recall f1-score support

hate 0.25 0.79 0.38 1835

offensive 0.89 0.71 0.79 5925

neither 0.93 0.69 0.79 10754

accuracy 0.71 18514

macro avg 0.69 0.73 0.65 18514

weighted avg 0.85 0.71 0.75 18514

average eval\_loss: 0.8075072398775339, accuracy: 0.7052500810197688

0.7052500810197688

2 run

training 1 epoch...

811.4260900020599 seconds used......

1 training finished! average train loss: 0.6299808229062808

total clf loss: 781.8252285122871 total adv loss: 428.76749907806516

evaluating...

Confusion matrix, without normalization

[[ 649 169 128]

[ 780 4264 160]

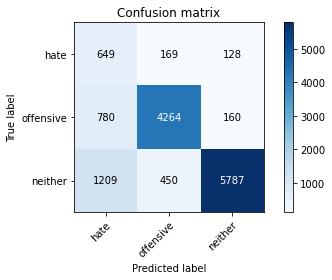
[1209 450 5787]]

Normalized confusion matrix

[[0.68604651 0.17864693 0.13530655]

[0.1498847 0.81936972 0.03074558]

[0.16236906 0.06043513 0.77719581]]





precision recall f1-score support

hate 0.25 0.69 0.36 946

offensive 0.87 0.82 0.85 5204

neither 0.95 0.78 0.86 7446

accuracy 0.79 13596

macro avg 0.69 0.76 0.69 13596

weighted avg 0.87 0.79 0.82 13596

average eval\_loss: 0.5143850639681605, accuracy: 0.7869961753456899

training 2 epoch...

812.1372072696686 seconds used......

2 training finished! average train loss: 0.4925707208537423

total clf loss: 610.4612724483013 total adv loss: 318.56823618896306

evaluating...

Confusion matrix, without normalization

[[ 575 170 201]

[ 594 4355 255]

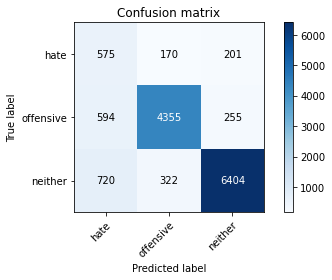
[ 720 322 6404]]

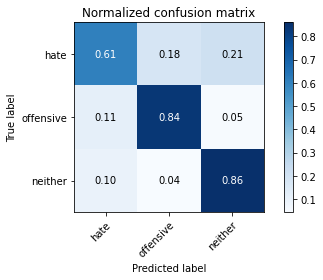
Normalized confusion matrix

[[0.60782241 0.17970402 0.21247357]

[0.11414297 0.83685626 0.04900077]

[0.09669621 0.0432447 0.86005909]]





precision recall f1-score support

hate 0.30 0.61 0.41 946

offensive 0.90 0.84 0.87 5204

neither 0.93 0.86 0.90 7446

accuracy 0.83 13596

macro avg 0.71 0.77 0.72 13596

weighted avg 0.88 0.83 0.85 13596

average eval\_loss: 0.4302794972105938, accuracy: 0.8336275375110327

training 3 epoch...

811.7072715759277 seconds used......

3 training finished! average train loss: 0.37101822108795746

total clf loss: 458.7792803272605 total adv loss: 219.20572878792882

evaluating...

Confusion matrix, without normalization

[[ 692 117 137]

[1015 3976 213]

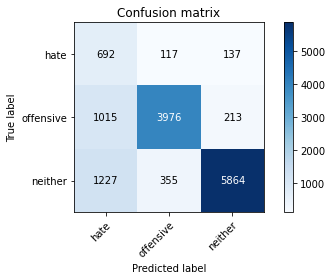
[1227 355 5864]]

Normalized confusion matrix

[[0.73150106 0.12367865 0.1448203 ]

[0.19504228 0.76402767 0.04093005]

[0.16478646 0.0476766 0.78753693]]





precision recall f1-score support

hate 0.24 0.73 0.36 946

offensive 0.89 0.76 0.82 5204

neither 0.94 0.79 0.86 7446

accuracy 0.77 13596

macro avg 0.69 0.76 0.68 13596

weighted avg 0.88 0.77 0.81 13596

average eval\_loss: 0.5745283308603307, accuracy: 0.7746395998823183

Confusion matrix, without normalization

[[1312 199 324]

[1320 4194 411]

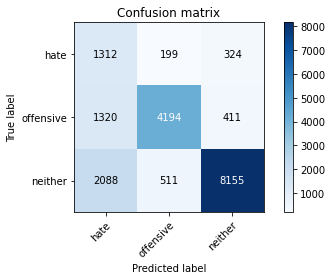
[2088 511 8155]]

Normalized confusion matrix

[[0.71498638 0.10844687 0.17656676]

[0.22278481 0.7078481 0.06936709]

[0.19416031 0.0475172 0.75832248]]





precision recall f1-score support

hate 0.28 0.71 0.40 1835

offensive 0.86 0.71 0.77 5925

neither 0.92 0.76 0.83 10754

accuracy 0.74 18514

macro avg 0.68 0.73 0.67 18514

weighted avg 0.83 0.74 0.77 18514

average eval\_loss: 0.696498114511081, accuracy: 0.7378740412660689

0.7378740412660689

3 run

training 1 epoch...

810.8723881244659 seconds used......

1 training finished! average train loss: 0.6340674286231386

total clf loss: 787.1393893808126 total adv loss: 436.40006510168314

evaluating...

Confusion matrix, without normalization

[[ 529 152 265]

[ 566 4110 528]

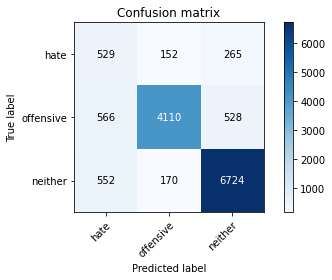
[ 552 170 6724]]

Normalized confusion matrix

[[0.55919662 0.16067653 0.28012685]

[0.10876249 0.78977709 0.10146042]

[0.07413376 0.02283105 0.90303519]]





precision recall f1-score support

hate 0.32 0.56 0.41 946

offensive 0.93 0.79 0.85 5204

neither 0.89 0.90 0.90 7446

accuracy 0.84 13596

macro avg 0.71 0.75 0.72 13596

weighted avg 0.87 0.84 0.85 13596

average eval\_loss: 0.4428593167816015, accuracy: 0.8357605177993528

training 2 epoch...

811.0630061626434 seconds used......

2 training finished! average train loss: 0.49003406998631177

total clf loss: 607.3194625973701 total adv loss: 316.96679739654064

evaluating...

Confusion matrix, without normalization

[[ 638 162 146]

[ 725 4261 218]

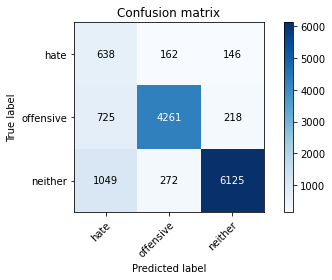
[1049 272 6125]]

Normalized confusion matrix

[[0.6744186 0.17124736 0.15433404]

[0.13931591 0.81879324 0.04189085]

[0.14088101 0.03652968 0.82258931]]





precision recall f1-score support

hate 0.26 0.67 0.38 946

offensive 0.91 0.82 0.86 5204

neither 0.94 0.82 0.88 7446

accuracy 0.81 13596

macro avg 0.71 0.77 0.71 13596

weighted avg 0.88 0.81 0.84 13596

average eval\_loss: 0.4716684648219277, accuracy: 0.8108267137393351

training 3 epoch...

811.1294844150543 seconds used......

3 training finished! average train loss: 0.36999322499619997

total clf loss: 457.49979957193136 total adv loss: 218.35953614255413

evaluating...

Confusion matrix, without normalization

[[ 690 95 161]

[1102 3867 235]

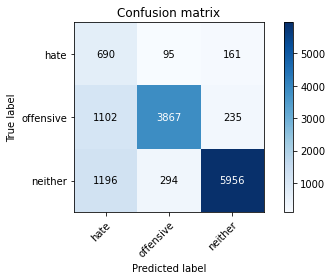
[1196 294 5956]]

Normalized confusion matrix

[[0.72938689 0.10042283 0.17019027]

[0.21176018 0.74308224 0.04515757]

[0.16062315 0.03948429 0.79989256]]





precision recall f1-score support

hate 0.23 0.73 0.35 946

offensive 0.91 0.74 0.82 5204

neither 0.94 0.80 0.86 7446

accuracy 0.77 13596

macro avg 0.69 0.76 0.68 13596

weighted avg 0.88 0.77 0.81 13596

average eval\_loss: 0.6035326223382178, accuracy: 0.7732421300382465

Confusion matrix, without normalization

[[1380 150 305]

[1499 4003 423]

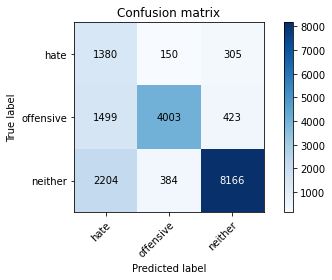
[2204 384 8166]]

Normalized confusion matrix

[[0.7520436 0.08174387 0.16621253]

[0.25299578 0.67561181 0.07139241]

[0.204947 0.03570764 0.75934536]]





precision recall f1-score support

hate 0.27 0.75 0.40 1835

offensive 0.88 0.68 0.77 5925

neither 0.92 0.76 0.83 10754

accuracy 0.73 18514

macro avg 0.69 0.73 0.67 18514

weighted avg 0.84 0.73 0.77 18514

average eval\_loss: 0.7322338784262377, accuracy: 0.7318245651939073

0.7318245651939073