with adv

normal examples = abusive examples

training 1 epoch...

1563.4894840717316 seconds used......

1 training finished! average train loss: 0.8187477691606281

total clf loss: 629.7506555654109 total adv loss: 414.21956035122275

evaluating...

Confusion matrix, without normalization

[[ 334 411 247]

[ 106 5160 143]

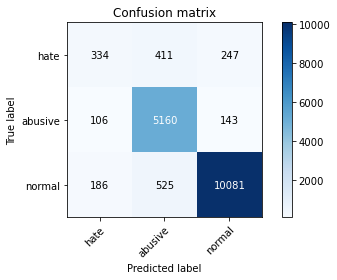
[ 186 525 10081]]

Normalized confusion matrix

[[0.33669355 0.41431452 0.24899194]

[0.01959697 0.95396561 0.02643742]

[0.01723499 0.04864715 0.93411787]]





precision recall f1-score support

hate 0.53 0.34 0.41 992

abusive 0.85 0.95 0.90 5409

normal 0.96 0.93 0.95 10792

accuracy 0.91 17193

macro avg 0.78 0.74 0.75 17193

weighted avg 0.90 0.91 0.90 17193

average eval\_loss: 0.2976298939783213, accuracy: 0.9058919327633339

training 2 epoch...

1564.0033478736877 seconds used......

2 training finished! average train loss: 0.6605138359429259

total clf loss: 512.8450668863952 total adv loss: 331.7650383124128

evaluating...

Confusion matrix, without normalization

[[ 359 361 272]

[ 126 5110 173]

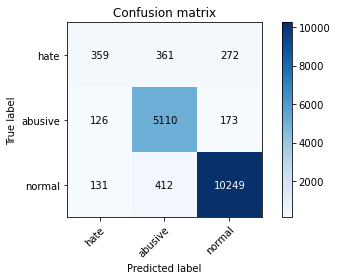
[ 131 412 10249]]

Normalized confusion matrix

[[0.36189516 0.36391129 0.27419355]

[0.02329451 0.94472176 0.03198373]

[0.01213862 0.03817643 0.94968495]]





precision recall f1-score support

hate 0.58 0.36 0.45 992

abusive 0.87 0.94 0.91 5409

normal 0.96 0.95 0.95 10792

accuracy 0.91 17193

macro avg 0.80 0.75 0.77 17193

weighted avg 0.91 0.91 0.91 17193

average eval\_loss: 0.2706131179037309, accuracy: 0.9142092712150294

training 3 epoch...

1564.217046737671 seconds used......

3 training finished! average train loss: 0.5282841285733545

total clf loss: 415.08555131964386 total adv loss: 262.8942405786365

evaluating...

Confusion matrix, without normalization

[[ 331 450 211]

[ 104 5168 137]

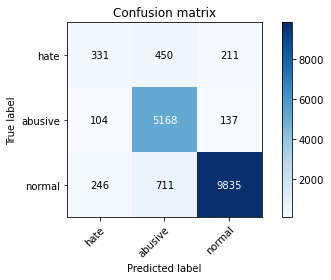
[ 246 711 9835]]

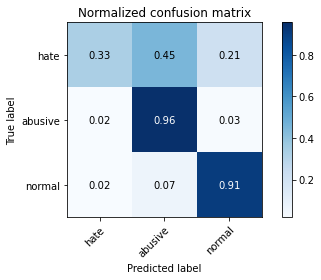
Normalized confusion matrix

[[0.33366935 0.45362903 0.21270161]

[0.01922721 0.95544463 0.02532816]

[0.02279466 0.06588213 0.9113232 ]]





precision recall f1-score support

hate 0.49 0.33 0.40 992

abusive 0.82 0.96 0.88 5409

normal 0.97 0.91 0.94 10792

accuracy 0.89 17193

macro avg 0.76 0.73 0.74 17193

weighted avg 0.89 0.89 0.89 17193

average eval\_loss: 0.3339818447438437, accuracy: 0.891874600127959

training 4 epoch...

1564.7633891105652 seconds used......

4 training finished! average train loss: 0.4009682066395569

total clf loss: 319.29970167484134 total adv loss: 197.41233795601875

evaluating...

Confusion matrix, without normalization

[[ 374 354 264]

[ 194 4999 216]

[ 208 368 10216]]

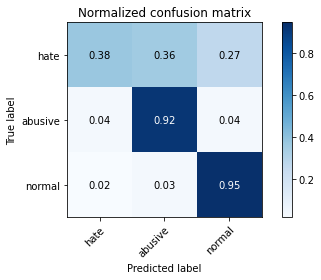
Normalized confusion matrix

[[0.37701613 0.35685484 0.26612903]

[0.03586615 0.92420041 0.03993344]

[0.01927354 0.03409933 0.94662713]]





precision recall f1-score support

hate 0.48 0.38 0.42 992

abusive 0.87 0.92 0.90 5409

normal 0.96 0.95 0.95 10792

accuracy 0.91 17193

macro avg 0.77 0.75 0.76 17193

weighted avg 0.90 0.91 0.90 17193

average eval\_loss: 0.3315751517224017, accuracy: 0.9067062176467167

training 5 epoch...

1564.8422508239746 seconds used......

5 training finished! average train loss: 0.3077561464945063

total clf loss: 241.22568044019863 total adv loss: 153.44400798596325

evaluating...

Confusion matrix, without normalization

[[ 433 374 185]

[ 219 5086 104]

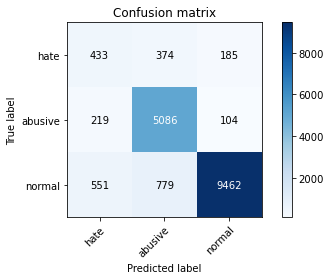
[ 551 779 9462]]

Normalized confusion matrix

[[0.43649194 0.37701613 0.18649194]

[0.04048808 0.94028471 0.01922721]

[0.05105634 0.0721831 0.87676056]]





precision recall f1-score support

hate 0.36 0.44 0.39 992

abusive 0.82 0.94 0.87 5409

normal 0.97 0.88 0.92 10792

accuracy 0.87 17193

macro avg 0.72 0.75 0.73 17193

weighted avg 0.89 0.87 0.88 17193

average eval\_loss: 0.45390312063702665, accuracy: 0.871342988425522

with adv

normal examples = hate speech examples

training 1 epoch...

984.68133020401 seconds used......

1 training finished! average train loss: 0.9630528763740611

total clf loss: 433.49577996879816 total adv loss: 318.7095100991428

evaluating...

Confusion matrix, without normalization

[[ 300 554 138]

[ 60 5292 57]

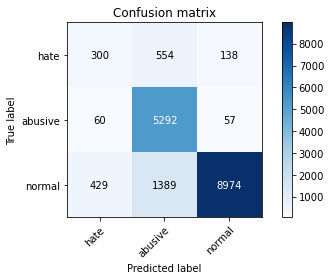
[ 429 1389 8974]]

Normalized confusion matrix

[[0.30241935 0.55846774 0.1391129 ]

[0.01109262 0.97836938 0.01053799]

[0.03975167 0.12870645 0.83154188]]





precision recall f1-score support

hate 0.38 0.30 0.34 992

abusive 0.73 0.98 0.84 5409

normal 0.98 0.83 0.90 10792

accuracy 0.85 17193

macro avg 0.70 0.70 0.69 17193

weighted avg 0.87 0.85 0.85 17193

average eval\_loss: 0.5022185951991137, accuracy: 0.8472052579538184

training 2 epoch...

983.1054022312164 seconds used......

2 training finished! average train loss: 0.7699557992884313

total clf loss: 345.13647773116827 total adv loss: 255.52718534879386

evaluating...

Confusion matrix, without normalization

[[ 498 366 128]

[ 213 5120 76]

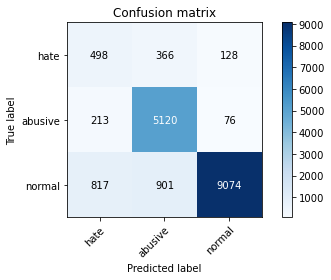
[ 817 901 9074]]

Normalized confusion matrix

[[0.50201613 0.36895161 0.12903226]

[0.03937881 0.94657053 0.01405066]

[0.07570423 0.08348777 0.84080801]]





precision recall f1-score support

hate 0.33 0.50 0.40 992

abusive 0.80 0.95 0.87 5409

normal 0.98 0.84 0.90 10792

accuracy 0.85 17193

macro avg 0.70 0.76 0.72 17193

weighted avg 0.88 0.85 0.86 17193

average eval\_loss: 0.465868100939795, accuracy: 0.8545338219042634

training 3 epoch...

982.1637961864471 seconds used......

3 training finished! average train loss: 0.6134758782874552

total clf loss: 276.81842491962016 total adv loss: 202.6833760542795

evaluating...

Confusion matrix, without normalization

[[ 430 388 174]

[ 171 5144 94]

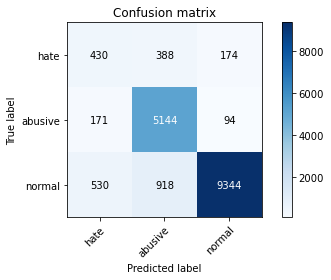
[ 530 918 9344]]

Normalized confusion matrix

[[0.43346774 0.39112903 0.17540323]

[0.03161398 0.95100758 0.01737844]

[0.04911045 0.08506301 0.86582654]]





precision recall f1-score support

hate 0.38 0.43 0.41 992

abusive 0.80 0.95 0.87 5409

normal 0.97 0.87 0.92 10792

accuracy 0.87 17193

macro avg 0.72 0.75 0.73 17193

weighted avg 0.88 0.87 0.87 17193

average eval\_loss: 0.4318348436941241, accuracy: 0.8676787064502995

training 4 epoch...

981.3576276302338 seconds used......

4 training finished! average train loss: 0.4595772669811418

total clf loss: 208.92732724361122 total adv loss: 151.06129675963894

evaluating...

Confusion matrix, without normalization

[[ 522 388 82]

[ 248 5121 40]

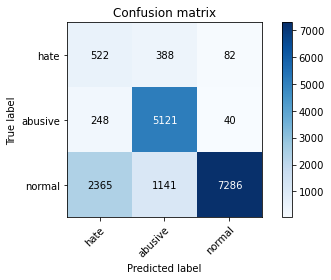
[2365 1141 7286]]

Normalized confusion matrix

[[0.52620968 0.39112903 0.08266129]

[0.04584951 0.94675541 0.00739508]

[0.21914381 0.10572646 0.67512973]]





precision recall f1-score support

hate 0.17 0.53 0.25 992

abusive 0.77 0.95 0.85 5409

normal 0.98 0.68 0.80 10792

accuracy 0.75 17193

macro avg 0.64 0.72 0.63 17193

weighted avg 0.87 0.75 0.78 17193

average eval\_loss: 0.7567726745854977, accuracy: 0.75199208980399

training 5 epoch...

980.9276835918427 seconds used......

5 training finished! average train loss: 0.33746729932328445

total clf loss: 153.4502929467708 total adv loss: 110.9066714826622

evaluating...

Confusion matrix, without normalization

[[ 458 404 130]

[ 208 5120 81]

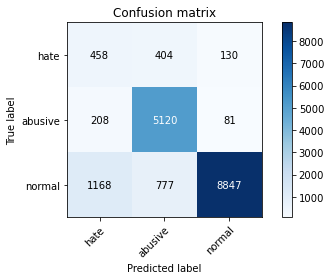
[1168 777 8847]]

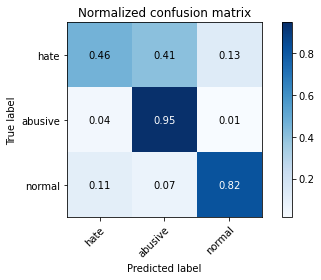
Normalized confusion matrix

[[0.46169355 0.40725806 0.13104839]

[0.03845443 0.94657053 0.01497504]

[0.10822832 0.07199778 0.81977391]]





precision recall f1-score support

hate 0.25 0.46 0.32 992

abusive 0.81 0.95 0.87 5409

normal 0.98 0.82 0.89 10792

accuracy 0.84 17193

macro avg 0.68 0.74 0.70 17193

weighted avg 0.88 0.84 0.85 17193

average eval\_loss: 0.6195938056778856, accuracy: 0.8390042459140348