hate as offensive

alpha=0.05

input: mean of last 2 layers

1 run...

Downloading: 100%

411/411 [00:00<00:00, 5.33kB/s]

Downloading: 100%

263M/263M [00:12<00:00, 22.4MB/s]

training 1 epoch...

1319.6651628017426 seconds used......

1 training finished! average train loss: 0.3751245433576624

total clf loss: 528.3439631033689 total adv loss: 333.4817335046828

evaluating...

Confusion matrix, without normalization

[[ 381 310 331]

[ 191 4936 280]

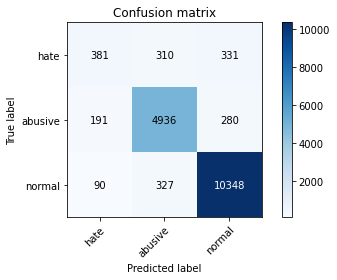
[ 90 327 10348]]

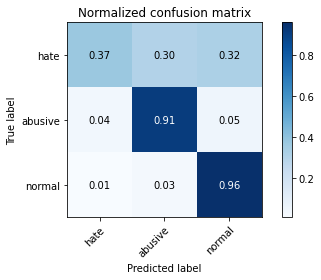
Normalized confusion matrix

[[0.37279843 0.30332681 0.32387476]

[0.03532458 0.9128907 0.05178472]

[0.00836043 0.03037622 0.96126335]]





precision recall f1-score support

hate 0.58 0.37 0.45 1022

abusive 0.89 0.91 0.90 5407

normal 0.94 0.96 0.95 10765

accuracy 0.91 17194

macro avg 0.80 0.75 0.77 17194

weighted avg 0.90 0.91 0.91 17194

average eval\_loss: 0.2929727533269067, accuracy: 0.9110736303361637

training 2 epoch...

1320.4241788387299 seconds used......

2 training finished! average train loss: 0.2867788195915935

total clf loss: 403.3680384736508 total adv loss: 244.0345723349601

evaluating...

Confusion matrix, without normalization

[[ 444 313 265]

[ 278 4920 209]

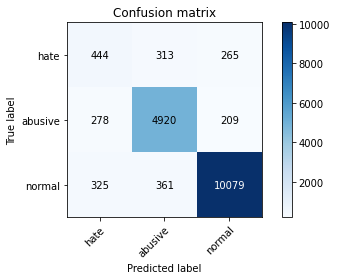
[ 325 361 10079]]

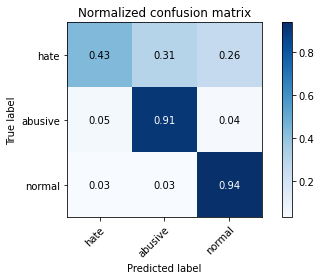
Normalized confusion matrix

[[0.43444227 0.30626223 0.2592955 ]

[0.05141483 0.90993157 0.0386536 ]

[0.03019043 0.0335346 0.93627497]]





precision recall f1-score support

hate 0.42 0.43 0.43 1022

abusive 0.88 0.91 0.89 5407

normal 0.96 0.94 0.95 10765

accuracy 0.90 17194

macro avg 0.75 0.76 0.76 17194

weighted avg 0.90 0.90 0.90 17194

average eval\_loss: 0.3102103885111594, accuracy: 0.8981621495870653

training 3 epoch...

1319.7269332408905 seconds used......

3 training finished! average train loss: 0.2105078315765179

total clf loss: 295.7226991346106 total adv loss: 171.80033542285673

evaluating...

Confusion matrix, without normalization

[[ 444 275 303]

[ 357 4764 286]

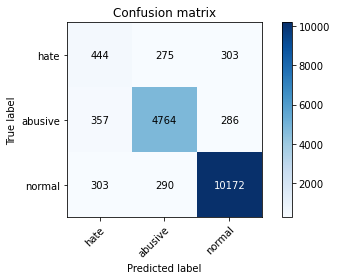
[ 303 290 10172]]

Normalized confusion matrix

[[0.43444227 0.26908023 0.2964775 ]

[0.06602552 0.88108008 0.0528944 ]

[0.02814677 0.02693915 0.94491407]]





precision recall f1-score support

hate 0.40 0.43 0.42 1022

abusive 0.89 0.88 0.89 5407

normal 0.95 0.94 0.95 10765

accuracy 0.89 17194

macro avg 0.75 0.75 0.75 17194

weighted avg 0.90 0.89 0.90 17194

average eval\_loss: 0.32924642304911517, accuracy: 0.8944980807258346

training 4 epoch...

1319.042095899582 seconds used......

4 training finished! average train loss: 0.13946330256055484

total clf loss: 195.32602235628292 total adv loss: 101.96155583602376

evaluating...

Confusion matrix, without normalization

[[ 441 331 250]

[ 301 4817 289]

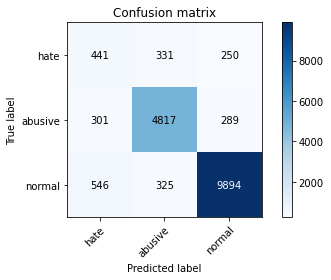
[ 546 325 9894]]

Normalized confusion matrix

[[0.43150685 0.32387476 0.2446184 ]

[0.05566858 0.89088219 0.05344923]

[0.05071993 0.03019043 0.91908964]]





precision recall f1-score support

hate 0.34 0.43 0.38 1022

abusive 0.88 0.89 0.89 5407

normal 0.95 0.92 0.93 10765

accuracy 0.88 17194

macro avg 0.72 0.75 0.73 17194

weighted avg 0.89 0.88 0.89 17194

average eval\_loss: 0.44221977945904484, accuracy: 0.8812376410375713

training 5 epoch...

1319.484138250351 seconds used......

5 training finished! average train loss: 0.09742345287909396

total clf loss: 136.35191017377656 total adv loss: 69.32640864397399

evaluating...

Confusion matrix, without normalization

[[ 402 323 297]

[ 262 4833 312]

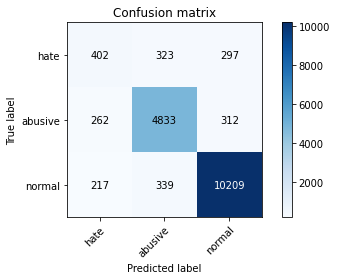
[ 217 339 10209]]

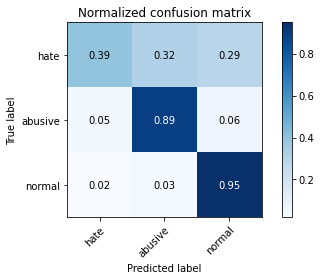
Normalized confusion matrix

[[0.39334638 0.31604697 0.29060665]

[0.04845571 0.89384132 0.05770298]

[0.02015792 0.03149094 0.94835114]]





precision recall f1-score support

hate 0.46 0.39 0.42 1022

abusive 0.88 0.89 0.89 5407

normal 0.94 0.95 0.95 10765

accuracy 0.90 17194

macro avg 0.76 0.75 0.75 17194

weighted avg 0.89 0.90 0.90 17194

average eval\_loss: 0.4781992619824704, accuracy: 0.8982203094102594

2 run...

training 1 epoch...

1320.0949096679688 seconds used......

1 training finished! average train loss: 0.37073405570134826

total clf loss: 522.3618360459805 total adv loss: 333.61166983284056

evaluating...

Confusion matrix, without normalization

[[ 468 271 283]

[ 338 4865 204]

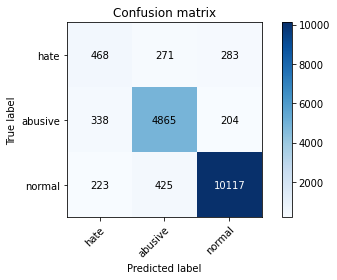
[ 223 425 10117]]

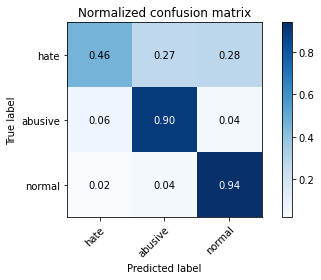
Normalized confusion matrix

[[0.45792564 0.26516634 0.27690802]

[0.06251156 0.89975957 0.03772887]

[0.02071528 0.0394798 0.93980492]]





precision recall f1-score support

hate 0.45 0.46 0.46 1022

abusive 0.87 0.90 0.89 5407

normal 0.95 0.94 0.95 10765

accuracy 0.90 17194

macro avg 0.76 0.77 0.76 17194

weighted avg 0.90 0.90 0.90 17194

average eval\_loss: 0.2970211916451537, accuracy: 0.8985692683494242

training 2 epoch...

1321.0536139011383 seconds used......

2 training finished! average train loss: 0.28862049790984157

total clf loss: 405.9478248041123 total adv loss: 245.38930828496814

evaluating...

Confusion matrix, without normalization

[[ 393 318 311]

[ 232 4923 252]

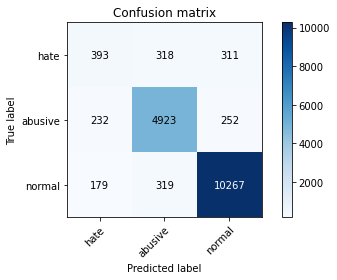
[ 179 319 10267]]

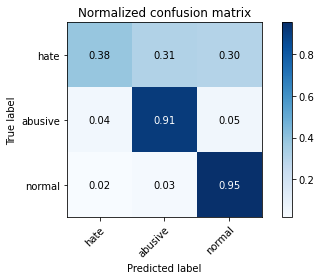
Normalized confusion matrix

[[0.38454012 0.3111546 0.30430528]

[0.04290734 0.91048641 0.04660625]

[0.01662796 0.02963307 0.95373897]]





precision recall f1-score support

hate 0.49 0.38 0.43 1022

abusive 0.89 0.91 0.90 5407

normal 0.95 0.95 0.95 10765

accuracy 0.91 17194

macro avg 0.77 0.75 0.76 17194

weighted avg 0.90 0.91 0.90 17194

average eval\_loss: 0.2798738498002464, accuracy: 0.9063045248342445

training 3 epoch...

1318.872077703476 seconds used......

3 training finished! average train loss: 0.21553284484180252

total clf loss: 302.6934915650636 total adv loss: 174.13382380921394

evaluating...

Confusion matrix, without normalization

[[ 456 316 250]

[ 294 4918 195]

[ 348 403 10014]]

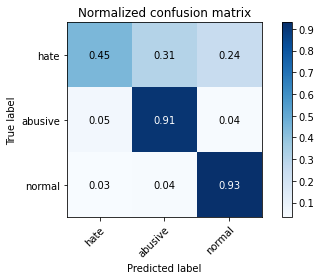
Normalized confusion matrix

[[0.44618395 0.30919765 0.2446184 ]

[0.05437396 0.90956168 0.03606436]

[0.03232699 0.03743614 0.93023688]]





precision recall f1-score support

hate 0.42 0.45 0.43 1022

abusive 0.87 0.91 0.89 5407

normal 0.96 0.93 0.94 10765

accuracy 0.89 17194

macro avg 0.75 0.76 0.75 17194

weighted avg 0.90 0.89 0.90 17194

average eval\_loss: 0.3311569435781864, accuracy: 0.8949633593113877

training 4 epoch...

1317.7540831565857 seconds used......

4 training finished! average train loss: 0.14207063862661998

total clf loss: 198.94145134883001 total adv loss: 103.14200442843139

evaluating...

Confusion matrix, without normalization

[[ 478 303 241]

[ 327 4909 171]

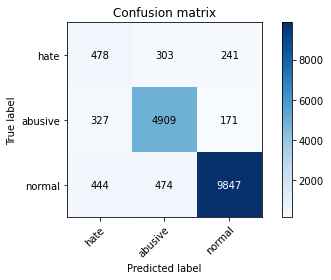
[ 444 474 9847]]

Normalized confusion matrix

[[0.46771037 0.2964775 0.23581213]

[0.06047716 0.90789717 0.03162567]

[0.04124477 0.04403158 0.91472364]]





precision recall f1-score support

hate 0.38 0.47 0.42 1022

abusive 0.86 0.91 0.89 5407

normal 0.96 0.91 0.94 10765

accuracy 0.89 17194

macro avg 0.74 0.76 0.75 17194

weighted avg 0.90 0.89 0.89 17194

average eval\_loss: 0.4414087401800464, accuracy: 0.8860067465394905

training 5 epoch...

1318.267418384552 seconds used......

5 training finished! average train loss: 0.0978294084149193

total clf loss: 136.9294534814544 total adv loss: 69.80280476796906

evaluating...

Confusion matrix, without normalization

[[ 416 274 332]

[ 379 4686 342]

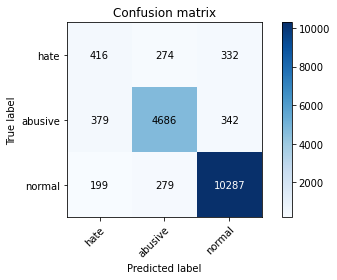
[ 199 279 10287]]

Normalized confusion matrix

[[0.40704501 0.26810176 0.32485323]

[0.07009432 0.86665434 0.06325134]

[0.01848583 0.02591732 0.95559684]]





precision recall f1-score support

hate 0.42 0.41 0.41 1022

abusive 0.89 0.87 0.88 5407

normal 0.94 0.96 0.95 10765

accuracy 0.90 17194

macro avg 0.75 0.74 0.75 17194

weighted avg 0.89 0.90 0.89 17194

average eval\_loss: 0.488419621697692, accuracy: 0.8950215191345818