with adv

normal examples = hate speech examples

1 run

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]

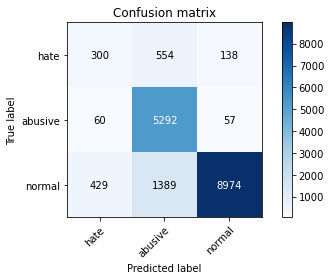
[ 429 1389 8974]]

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]

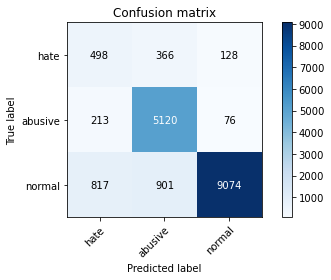
[ 817 901 9074]]

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]

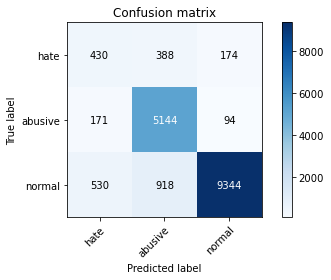
[ 530 918 9344]]

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]

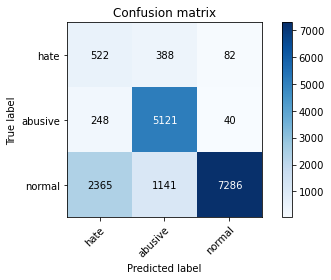
[2365 1141 7286]]

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]

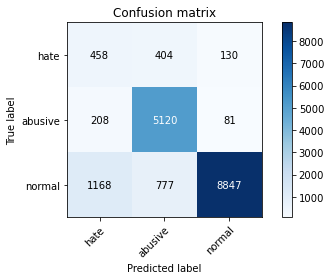
[1168 777 8847]]

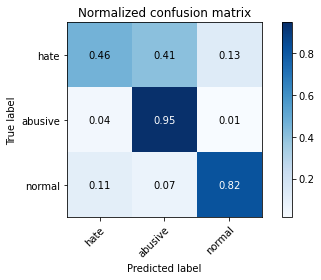
Normalized confusion matrix

[[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

2 run

training 1 epoch...

976.929187297821 seconds used......

1 training finished! average train loss: 0.9643658095394536

total clf loss: 431.50172393396497 total adv loss: 320.4365272540599

evaluating...

Confusion matrix, without normalization

[[ 337 469 186]

[ 82 5229 98]

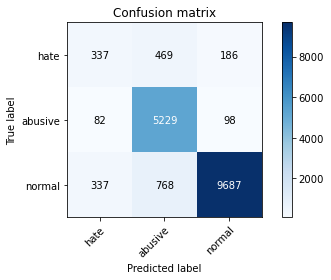
[ 337 768 9687]]

Normalized confusion matrix

[[0.33971774 0.47278226 0.1875 ]

[0.01515992 0.96672213 0.01811795]

[0.03122683 0.07116383 0.89760934]]





precision recall f1-score support

hate 0.45 0.34 0.39 992

abusive 0.81 0.97 0.88 5409

normal 0.97 0.90 0.93 10792

accuracy 0.89 17193

macro avg 0.74 0.73 0.73 17193

weighted avg 0.89 0.89 0.89 17193

average eval\_loss: 0.3785039882847043, accuracy: 0.8871633804455301

training 2 epoch...

976.5706403255463 seconds used......

2 training finished! average train loss: 0.7721761497702423

total clf loss: 344.89317671209574 total adv loss: 256.88335120119154

evaluating...

Confusion matrix, without normalization

[[ 440 442 110]

[ 151 5193 65]

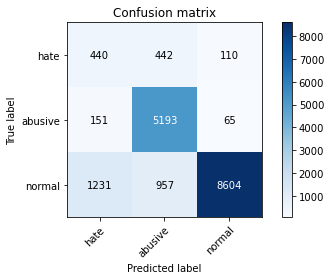
[1231 957 8604]]

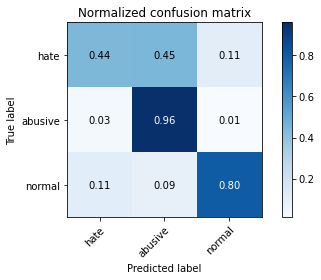
Normalized confusion matrix

[[0.44354839 0.44556452 0.1108871 ]

[0.02791644 0.96006656 0.01201701]

[0.11406597 0.0886768 0.79725723]]





precision recall f1-score support

hate 0.24 0.44 0.31 992

abusive 0.79 0.96 0.87 5409

normal 0.98 0.80 0.88 10792

accuracy 0.83 17193

macro avg 0.67 0.73 0.69 17193

weighted avg 0.88 0.83 0.84 17193

average eval\_loss: 0.5291764731670535, accuracy: 0.8280695631943232

training 3 epoch...

977.8858983516693 seconds used......

3 training finished! average train loss: 0.6227632114167908

total clf loss: 279.4650132637471 total adv loss: 206.52383867558092

evaluating...

Confusion matrix, without normalization

[[ 378 453 161]

[ 158 5157 94]

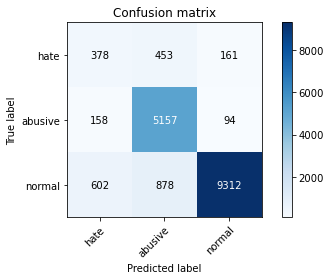
[ 602 878 9312]]

Normalized confusion matrix

[[0.38104839 0.45665323 0.16229839]

[0.02921057 0.95341098 0.01737844]

[0.05578206 0.08135656 0.86286138]]





precision recall f1-score support

hate 0.33 0.38 0.35 992

abusive 0.79 0.95 0.87 5409

normal 0.97 0.86 0.91 10792

accuracy 0.86 17193

macro avg 0.70 0.73 0.71 17193

weighted avg 0.88 0.86 0.87 17193

average eval\_loss: 0.4414458414712964, accuracy: 0.8635491188274298

training 4 epoch...

977.1389248371124 seconds used......

4 training finished! average train loss: 0.46039340567855475

total clf loss: 208.5427637565881 total adv loss: 151.7073518356774

evaluating...

Confusion matrix, without normalization

[[ 440 409 143]

[ 190 5126 93]

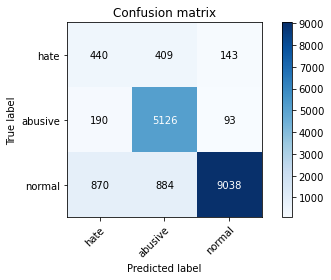
[ 870 884 9038]]

Normalized confusion matrix

[[0.44354839 0.41229839 0.14415323]

[0.03512664 0.94767979 0.01719357]

[0.08061527 0.08191253 0.8374722 ]]





precision recall f1-score support

hate 0.29 0.44 0.35 992

abusive 0.80 0.95 0.87 5409

normal 0.97 0.84 0.90 10792

accuracy 0.85 17193

macro avg 0.69 0.74 0.71 17193

weighted avg 0.88 0.85 0.86 17193

average eval\_loss: 0.5357564905463437, accuracy: 0.8494154597801431

training 5 epoch...

977.7617163658142 seconds used......

5 training finished! average train loss: 0.34418899121128105

total clf loss: 153.79134232038632 total adv loss: 114.47340777795762

evaluating...

Confusion matrix, without normalization

[[ 424 429 139]

[ 215 5115 79]

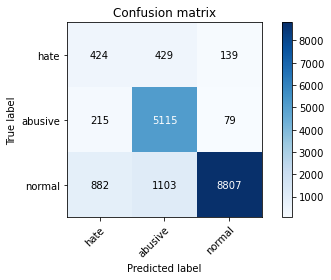
[ 882 1103 8807]]

Normalized confusion matrix

[[0.42741935 0.43245968 0.14012097]

[0.03974857 0.94564615 0.01460529]

[0.08172721 0.10220534 0.81606746]]





precision recall f1-score support

hate 0.28 0.43 0.34 992

abusive 0.77 0.95 0.85 5409

normal 0.98 0.82 0.89 10792

accuracy 0.83 17193

macro avg 0.67 0.73 0.69 17193

weighted avg 0.87 0.83 0.84 17193

average eval\_loss: 0.6211495865158043, accuracy: 0.8344093526435177