with adv

normal examples = abusive examples

training 1 epoch...

1920.0919613838196 seconds used......

1 training finished! average train loss: 1.0012927455483918

total clf loss: 1235.0734509378672 total adv loss: 481.38206324446946

evaluating...

Confusion matrix, without normalization

[[ 270 429 13 277]

[ 80 5180 41 148]

[ 5 172 1677 948]

[ 104 552 1184 8919]]

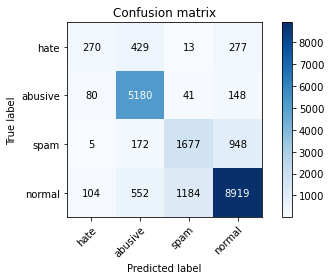
Normalized confusion matrix

[[0.27300303 0.43377149 0.01314459 0.28008089]

[0.01468159 0.95063314 0.00752432 0.02716095]

[0.00178444 0.06138473 0.59850107 0.33832976]

[0.00966633 0.05130588 0.1100474 0.82898039]]





precision recall f1-score support

hate 0.59 0.27 0.37 989

abusive 0.82 0.95 0.88 5449

spam 0.58 0.60 0.59 2802

normal 0.87 0.83 0.85 10759

accuracy 0.80 19999

macro avg 0.71 0.66 0.67 19999

weighted avg 0.80 0.80 0.80 19999

average eval\_loss: 0.5130999229997396, accuracy: 0.8023401170058503

training 2 epoch...

1919.9121131896973 seconds used......

2 training finished! average train loss: 0.8276756730601021

total clf loss: 1038.6545942500234 total adv loss: 389.04675455531105

evaluating...

Confusion matrix, without normalization

[[ 381 356 19 233]

[ 162 5115 40 132]

[ 9 185 1840 768]

[ 263 502 1476 8518]]

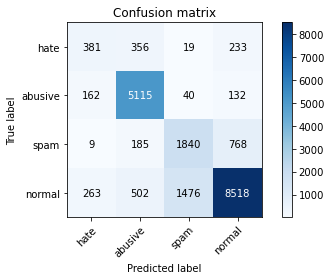
Normalized confusion matrix

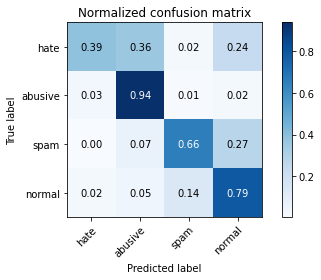
[[0.38523761 0.35995956 0.01921132 0.23559151]

[0.02973023 0.93870435 0.0073408 0.02422463]

[0.00321199 0.06602427 0.6566738 0.27408994]

[0.02444465 0.04665861 0.13718747 0.79170927]]





precision recall f1-score support

hate 0.47 0.39 0.42 989

abusive 0.83 0.94 0.88 5449

spam 0.55 0.66 0.60 2802

normal 0.88 0.79 0.83 10759

accuracy 0.79 19999

macro avg 0.68 0.69 0.68 19999

weighted avg 0.80 0.79 0.79 19999

average eval\_loss: 0.5252449019029737, accuracy: 0.7927396369818491

training 3 epoch...

1919.9666624069214 seconds used......

3 training finished! average train loss: 0.6865066270282437

total clf loss: 876.7084203995764 total adv loss: 315.08681326406077

evaluating...

Confusion matrix, without normalization

[[ 443 318 14 214]

[ 227 5016 27 179]

[ 11 175 1627 989]

[ 379 451 1159 8770]]

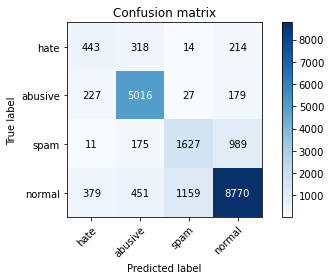
Normalized confusion matrix

[[0.4479272 0.32153691 0.01415571 0.21638018]

[0.04165902 0.92053588 0.00495504 0.03285006]

[0.00392577 0.06245539 0.58065667 0.35296217]

[0.03522632 0.04191839 0.10772377 0.81513152]]





precision recall f1-score support

hate 0.42 0.45 0.43 989

abusive 0.84 0.92 0.88 5449

spam 0.58 0.58 0.58 2802

normal 0.86 0.82 0.84 10759

accuracy 0.79 19999

macro avg 0.67 0.69 0.68 19999

weighted avg 0.80 0.79 0.79 19999

average eval\_loss: 0.535277524895966, accuracy: 0.7928396419820991

training 4 epoch...

1921.996776342392 seconds used......

4 training finished! average train loss: 0.5322935469872859

total clf loss: 685.21324409917 total adv loss: 241.58554521691985

evaluating...

Confusion matrix, without normalization

[[ 394 314 19 262]

[ 223 4883 118 225]

[ 11 146 1696 949]

[ 252 347 1385 8775]]

Normalized confusion matrix

[[0.3983822 0.31749242 0.01921132 0.26491405]

[0.04092494 0.89612773 0.02165535 0.04129198]

[0.00392577 0.05210564 0.60528194 0.33868665]

[0.02342225 0.03225207 0.12872944 0.81559625]]





precision recall f1-score support

hate 0.45 0.40 0.42 989

abusive 0.86 0.90 0.88 5449

spam 0.53 0.61 0.56 2802

normal 0.86 0.82 0.84 10759

accuracy 0.79 19999

macro avg 0.67 0.68 0.67 19999

weighted avg 0.79 0.79 0.79 19999

average eval\_loss: 0.632131822951138, accuracy: 0.7874393719685985

training 5 epoch...

1924.5800461769104 seconds used......

5 training finished! average train loss: 0.39136815790536733

total clf loss: 503.4136204365641 total adv loss: 177.81974326260388

evaluating...

Confusion matrix, without normalization

[[ 394 345 35 215]

[ 212 4999 98 140]

[ 17 163 1987 635]

[ 386 444 2337 7592]]

Normalized confusion matrix

[[0.3983822 0.34883721 0.03538928 0.2173913 ]

[0.03890622 0.91741604 0.01798495 0.02569279]

[0.00606709 0.05817273 0.70913633 0.22662384]

[0.03587694 0.04126778 0.2172135 0.70564179]]





precision recall f1-score support

hate 0.39 0.40 0.39 989

abusive 0.84 0.92 0.88 5449

spam 0.45 0.71 0.55 2802

normal 0.88 0.71 0.79 10759

accuracy 0.75 19999

macro avg 0.64 0.68 0.65 19999

weighted avg 0.79 0.75 0.76 19999

average eval\_loss: 0.7948656122013926, accuracy: 0.7486374318715936

normal examples = spam examples

training 1 epoch...

1605.42764544487 seconds used......

1 training finished! average train loss: 1.058505166483243

total clf loss: 1021.6324029415846 total adv loss: 442.89695324748755

evaluating...

Confusion matrix, without normalization

[[ 195 503 20 271]

[ 28 5225 44 152]

[ 2 177 1903 720]

[ 86 544 1683 8446]]

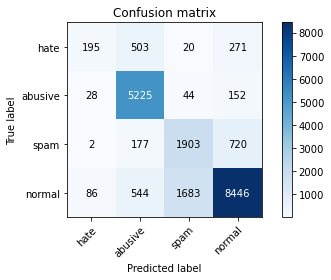
Normalized confusion matrix

[[1.97168857e-01 5.08594540e-01 2.02224469e-02 2.74014156e-01]

[5.13855753e-03 9.58891540e-01 8.07487612e-03 2.78950266e-02]

[7.13775874e-04 6.31691649e-02 6.79157744e-01 2.56959315e-01]

[7.99330793e-03 5.05623199e-02 1.56427177e-01 7.85017195e-01]]





precision recall f1-score support

hate 0.63 0.20 0.30 989

abusive 0.81 0.96 0.88 5449

spam 0.52 0.68 0.59 2802

normal 0.88 0.79 0.83 10759

accuracy 0.79 19999

macro avg 0.71 0.66 0.65 19999

weighted avg 0.80 0.79 0.78 19999

average eval\_loss: 0.5395916346311569, accuracy: 0.7884894244712236

training 2 epoch...

1604.7587487697601 seconds used......

2 training finished! average train loss: 0.8625752203224369

total clf loss: 841.5923002660275 total adv loss: 356.38412268832326

evaluating...

Confusion matrix, without normalization

[[ 408 414 26 141]

[ 150 5165 54 80]

[ 25 186 2180 411]

[ 684 640 2389 7046]]

Normalized confusion matrix

[[0.41253792 0.41860465 0.02628918 0.14256825]

[0.02752799 0.94788035 0.00991008 0.01468159]

[0.0089222 0.06638116 0.7780157 0.14668094]

[0.06357468 0.05948508 0.22204666 0.65489358]]





precision recall f1-score support

hate 0.32 0.41 0.36 989

abusive 0.81 0.95 0.87 5449

spam 0.47 0.78 0.59 2802

normal 0.92 0.65 0.76 10759

accuracy 0.74 19999

macro avg 0.63 0.70 0.65 19999

weighted avg 0.80 0.74 0.75 19999

average eval\_loss: 0.6540880582630634, accuracy: 0.7399869993499675

training 3 epoch...

1605.3938689231873 seconds used......

3 training finished! average train loss: 0.7169193034861382

total clf loss: 707.60245988518 total adv loss: 292.143063002266

evaluating...

Confusion matrix, without normalization

[[ 442 345 25 177]

[ 224 5052 67 106]

[ 21 156 2027 598]

[ 607 491 1970 7691]]

Normalized confusion matrix

[[0.44691608 0.34883721 0.02527806 0.17896866]

[0.04110846 0.92714259 0.01229583 0.01945311]

[0.00749465 0.05567452 0.72341185 0.21341899]

[0.05641788 0.04563621 0.18310252 0.71484339]]





precision recall f1-score support

hate 0.34 0.45 0.39 989

abusive 0.84 0.93 0.88 5449

spam 0.50 0.72 0.59 2802

normal 0.90 0.71 0.80 10759

accuracy 0.76 19999

macro avg 0.64 0.70 0.66 19999

weighted avg 0.80 0.76 0.77 19999

average eval\_loss: 0.6226886205971242, accuracy: 0.7606380319015951

training 4 epoch...

1606.8876192569733 seconds used......

4 training finished! average train loss: 0.5509000708112343

total clf loss: 546.8323058523238 total adv loss: 222.94481087173335

evaluating...

Confusion matrix, without normalization

[[ 405 356 38 190]

[ 201 5057 73 118]

[ 15 168 2188 431]

[ 372 540 2650 7197]]

Normalized confusion matrix

[[0.40950455 0.35995956 0.03842265 0.19211325]

[0.0368875 0.92806019 0.01339695 0.02165535]

[0.00535332 0.05995717 0.78087081 0.1538187 ]

[0.0345757 0.05019054 0.24630542 0.66892834]]





precision recall f1-score support

hate 0.41 0.41 0.41 989

abusive 0.83 0.93 0.87 5449

spam 0.44 0.78 0.56 2802

normal 0.91 0.67 0.77 10759

accuracy 0.74 19999

macro avg 0.65 0.70 0.65 19999

weighted avg 0.80 0.74 0.75 19999

average eval\_loss: 0.7649540381133556, accuracy: 0.7423871193559678

training 5 epoch...

1606.5232000350952 seconds used......

5 training finished! average train loss: 0.410420323260021

total clf loss: 402.89664256572723 total adv loss: 168.34039017837495

evaluating...

Confusion matrix, without normalization

[[ 480 311 27 171]

[ 337 4891 110 111]

[ 30 161 1957 654]

[ 763 492 2100 7404]]

Normalized confusion matrix

[[0.48533873 0.31445905 0.0273003 0.17290192]

[0.06184621 0.89759589 0.02018719 0.02037071]

[0.01070664 0.05745896 0.69842969 0.23340471]

[0.07091737 0.04572916 0.19518543 0.68816805]]





precision recall f1-score support

hate 0.30 0.49 0.37 989

abusive 0.84 0.90 0.87 5449

spam 0.47 0.70 0.56 2802

normal 0.89 0.69 0.78 10759

accuracy 0.74 19999

macro avg 0.62 0.69 0.64 19999

weighted avg 0.79 0.74 0.75 19999

average eval\_loss: 0.7823891351670027, accuracy: 0.7366368318415921