1 run

hate as no-offensive

alpha=2

training 1 epoch...

2.6523845195770264 seconds used......

1 training finished! average train loss: 2.4796295166015625

total clf loss: 16.81959366798401 total adv loss: 10.187425017356873

evaluating...

Confusion matrix, without normalization

[[ 3 9 1]

[ 8 20 1]

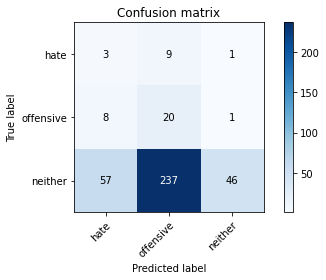
[ 57 237 46]]

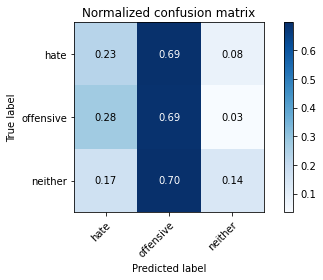
Normalized confusion matrix

[[0.23076923 0.69230769 0.07692308]

[0.27586207 0.68965517 0.03448276]

[0.16764706 0.69705882 0.13529412]]





precision recall f1-score support

hate 0.04 0.23 0.07 13

offensive 0.08 0.69 0.14 29

neither 0.96 0.14 0.24 340

accuracy 0.18 382

macro avg 0.36 0.35 0.15 382

weighted avg 0.86 0.18 0.22 382

average eval\_loss: 1.1362739145755767, accuracy: 0.1806282722513089

training 2 epoch...

2.643317461013794 seconds used......

2 training finished! average train loss: 2.303724726041158

total clf loss: 15.933063268661499 total adv loss: 9.311403512954712

evaluating...

Confusion matrix, without normalization

[[ 6 1 6]

[ 9 5 15]

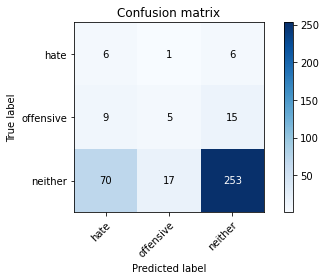
[ 70 17 253]]

Normalized confusion matrix

[[0.46153846 0.07692308 0.46153846]

[0.31034483 0.17241379 0.51724138]

[0.20588235 0.05 0.74411765]]





precision recall f1-score support

hate 0.07 0.46 0.12 13

offensive 0.22 0.17 0.19 29

neither 0.92 0.74 0.82 340

accuracy 0.69 382

macro avg 0.40 0.46 0.38 382

weighted avg 0.84 0.69 0.75 382

average eval\_loss: 0.9840408891439438, accuracy: 0.6910994764397905

training 3 epoch...

2.6432437896728516 seconds used......

3 training finished! average train loss: 2.1879575490951537

total clf loss: 15.170099020004272 total adv loss: 8.82463213801384

evaluating...

Confusion matrix, without normalization

[[ 7 4 2]

[ 10 10 9]

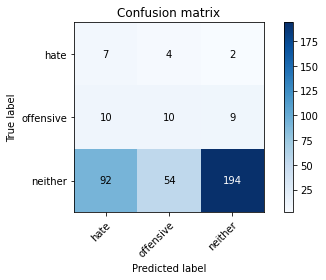
[ 92 54 194]]

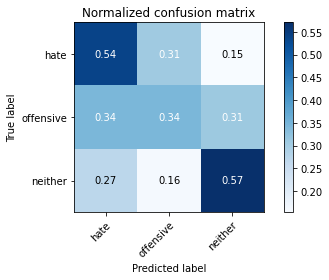
Normalized confusion matrix

[[0.53846154 0.30769231 0.15384615]

[0.34482759 0.34482759 0.31034483]

[0.27058824 0.15882353 0.57058824]]





precision recall f1-score support

hate 0.06 0.54 0.11 13

offensive 0.15 0.34 0.21 29

neither 0.95 0.57 0.71 340

accuracy 0.55 382

macro avg 0.39 0.48 0.34 382

weighted avg 0.86 0.55 0.65 382

average eval\_loss: 1.002968156337738, accuracy: 0.5523560209424084

training 4 epoch...

2.6782500743865967 seconds used......

4 training finished! average train loss: 1.8978057066599527

total clf loss: 13.692603528499603 total adv loss: 7.387241005897522

evaluating...

Confusion matrix, without normalization

[[ 8 4 1]

[ 11 11 7]

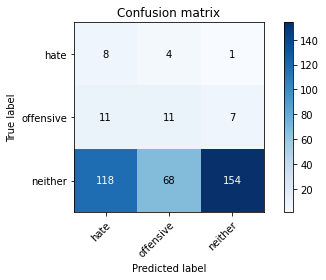
[118 68 154]]

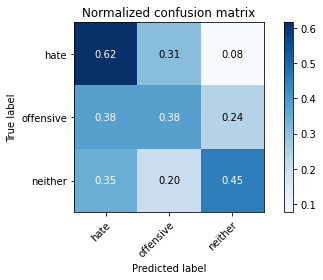
Normalized confusion matrix

[[0.61538462 0.30769231 0.07692308]

[0.37931034 0.37931034 0.24137931]

[0.34705882 0.2 0.45294118]]





precision recall f1-score support

hate 0.06 0.62 0.11 13

offensive 0.13 0.38 0.20 29

neither 0.95 0.45 0.61 340

accuracy 0.45 382

macro avg 0.38 0.48 0.31 382

weighted avg 0.86 0.45 0.56 382

average eval\_loss: 1.0334970653057098, accuracy: 0.45287958115183247

training 5 epoch...

2.659642457962036 seconds used......

5 training finished! average train loss: 1.4921827554702758

total clf loss: 11.341481447219849 total adv loss: 5.520629912614822

evaluating...

Confusion matrix, without normalization

[[ 8 4 1]

[ 8 11 10]

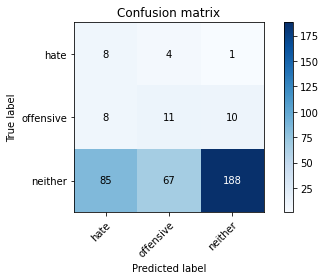
[ 85 67 188]]

Normalized confusion matrix

[[0.61538462 0.30769231 0.07692308]

[0.27586207 0.37931034 0.34482759]

[0.25 0.19705882 0.55294118]]





precision recall f1-score support

hate 0.08 0.62 0.14 13

offensive 0.13 0.38 0.20 29

neither 0.94 0.55 0.70 340

accuracy 0.54 382

macro avg 0.39 0.52 0.35 382

weighted avg 0.85 0.54 0.64 382

average eval\_loss: 0.9453805953264236, accuracy: 0.5418848167539267

model of 5 epoch

Confusion matrix, without normalization

[[ 30 9 2]

[ 29 29 37]

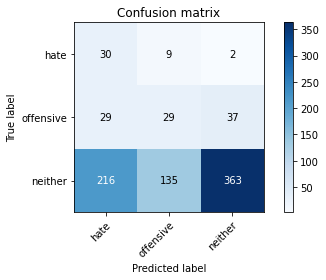
[216 135 363]]

Normalized confusion matrix

[[0.73170732 0.2195122 0.04878049]

[0.30526316 0.30526316 0.38947368]

[0.30252101 0.18907563 0.50840336]]





precision recall f1-score support

hate 0.11 0.73 0.19 41

offensive 0.17 0.31 0.22 95

neither 0.90 0.51 0.65 714

accuracy 0.50 850

macro avg 0.39 0.52 0.35 850

weighted avg 0.78 0.50 0.58 850

average eval\_loss: 0.990205351696458, accuracy: 0.4964705882352941

0.4964705882352941

2 run

training 1 epoch...

2.851351022720337 seconds used......

1 training finished! average train loss: 2.433317009607951

total clf loss: 16.875123739242554 total adv loss: 9.812315821647644

evaluating...

Confusion matrix, without normalization

[[ 0 13 0]

[ 3 26 0]

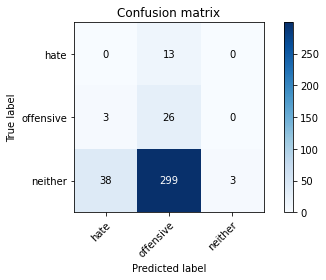
[ 38 299 3]]

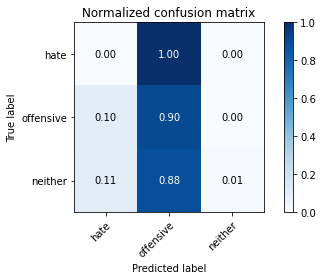
Normalized confusion matrix

[[0. 1. 0. ]

[0.10344828 0.89655172 0. ]

[0.11176471 0.87941176 0.00882353]]





precision recall f1-score support

hate 0.00 0.00 0.00 13

offensive 0.08 0.90 0.14 29

neither 1.00 0.01 0.02 340

accuracy 0.08 382

macro avg 0.36 0.30 0.05 382

weighted avg 0.90 0.08 0.03 382

average eval\_loss: 1.2257402777671813, accuracy: 0.07591623036649214

training 2 epoch...

2.824500560760498 seconds used......

2 training finished! average train loss: 2.263988717397054

total clf loss: 16.171553373336792 total adv loss: 8.894138395786285

evaluating...

Confusion matrix, without normalization

[[ 0 3 10]

[ 4 1 24]

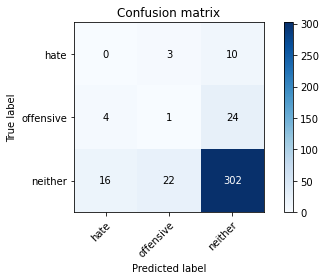
[ 16 22 302]]

Normalized confusion matrix

[[0. 0.23076923 0.76923077]

[0.13793103 0.03448276 0.82758621]

[0.04705882 0.06470588 0.88823529]]





precision recall f1-score support

hate 0.00 0.00 0.00 13

offensive 0.04 0.03 0.04 29

neither 0.90 0.89 0.89 340

accuracy 0.79 382

macro avg 0.31 0.31 0.31 382

weighted avg 0.80 0.79 0.80 382

average eval\_loss: 0.9661451488733291, accuracy: 0.7931937172774869

training 3 epoch...

2.8349881172180176 seconds used......

3 training finished! average train loss: 2.013710284233093

total clf loss: 15.299025475978851 total adv loss: 7.453314244747162

evaluating...

Confusion matrix, without normalization

[[ 5 7 1]

[ 9 17 3]

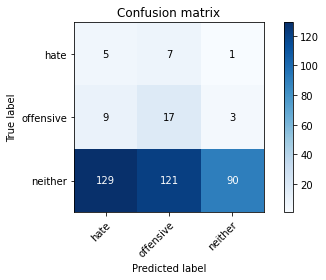
[129 121 90]]

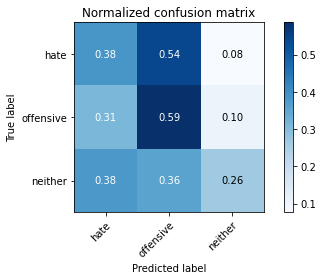
Normalized confusion matrix

[[0.38461538 0.53846154 0.07692308]

[0.31034483 0.5862069 0.10344828]

[0.37941176 0.35588235 0.26470588]]





precision recall f1-score support

hate 0.03 0.38 0.06 13

offensive 0.12 0.59 0.20 29

neither 0.96 0.26 0.41 340

accuracy 0.29 382

macro avg 0.37 0.41 0.22 382

weighted avg 0.86 0.29 0.39 382

average eval\_loss: 1.1304957151412964, accuracy: 0.2931937172774869

training 4 epoch...

2.836761236190796 seconds used......

4 training finished! average train loss: 1.699071725209554

total clf loss: 14.187517642974854 total adv loss: 5.6492791175842285

evaluating...

Confusion matrix, without normalization

[[ 7 4 2]

[ 10 12 7]

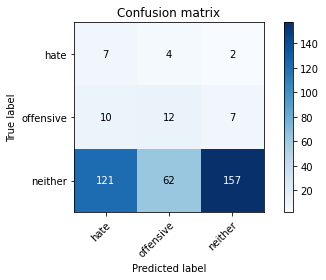
[121 62 157]]

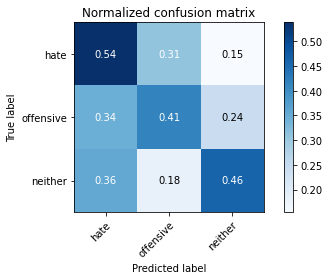
Normalized confusion matrix

[[0.53846154 0.30769231 0.15384615]

[0.34482759 0.4137931 0.24137931]

[0.35588235 0.18235294 0.46176471]]





precision recall f1-score support

hate 0.05 0.54 0.09 13

offensive 0.15 0.41 0.22 29

neither 0.95 0.46 0.62 340

accuracy 0.46 382

macro avg 0.38 0.47 0.31 382

weighted avg 0.86 0.46 0.57 382

average eval\_loss: 1.0307575970888139, accuracy: 0.4607329842931937

training 5 epoch...

2.864093780517578 seconds used......

5 training finished! average train loss: 1.1950484315554302

total clf loss: 12.20305609703064 total adv loss: 2.8613351806998253

evaluating...

Confusion matrix, without normalization

[[ 3 6 4]

[ 7 16 6]

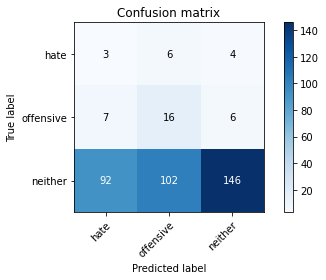
[ 92 102 146]]

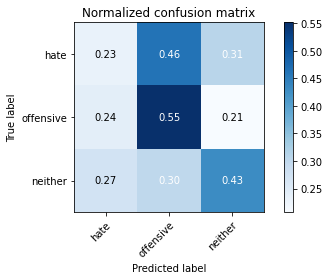
Normalized confusion matrix

[[0.23076923 0.46153846 0.30769231]

[0.24137931 0.55172414 0.20689655]

[0.27058824 0.3 0.42941176]]





precision recall f1-score support

hate 0.03 0.23 0.05 13

offensive 0.13 0.55 0.21 29

neither 0.94 0.43 0.59 340

accuracy 0.43 382

macro avg 0.36 0.40 0.28 382

weighted avg 0.84 0.43 0.54 382

average eval\_loss: 1.0736050069332124, accuracy: 0.4319371727748691

Confusion matrix, without normalization

[[ 26 11 4]

[ 31 46 18]

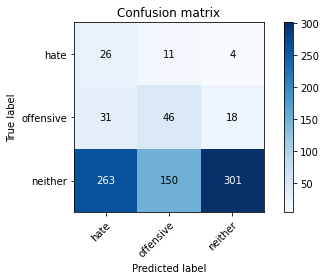
[263 150 301]]

Normalized confusion matrix

[[0.63414634 0.26829268 0.09756098]

[0.32631579 0.48421053 0.18947368]

[0.36834734 0.21008403 0.42156863]]





precision recall f1-score support

hate 0.08 0.63 0.14 41

offensive 0.22 0.48 0.30 95

neither 0.93 0.42 0.58 714

accuracy 0.44 850

macro avg 0.41 0.51 0.34 850

weighted avg 0.81 0.44 0.53 850

average eval\_loss: 1.0421017020247703, accuracy: 0.4388235294117647

[29]:

0.4388235294117647

3 run

training 1 epoch...

2.8807594776153564 seconds used......

1 training finished! average train loss: 2.408014520009359

total clf loss: 16.47092878818512 total adv loss: 9.824644446372986

evaluating...

Confusion matrix, without normalization

[[ 10 1 2]

[ 16 7 6]

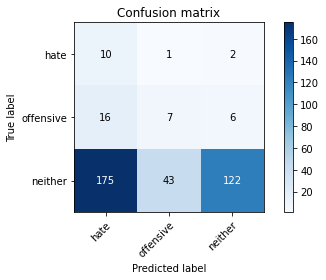
[175 43 122]]

Normalized confusion matrix

[[0.76923077 0.07692308 0.15384615]

[0.55172414 0.24137931 0.20689655]

[0.51470588 0.12647059 0.35882353]]





precision recall f1-score support

hate 0.05 0.77 0.09 13

offensive 0.14 0.24 0.18 29

neither 0.94 0.36 0.52 340

accuracy 0.36 382

macro avg 0.38 0.46 0.26 382

weighted avg 0.85 0.36 0.48 382

average eval\_loss: 1.0923223972320557, accuracy: 0.36387434554973824

training 2 epoch...

2.839048385620117 seconds used......

2 training finished! average train loss: 2.2360778013865152

total clf loss: 15.876765727996826 total adv loss: 8.832200676202774

evaluating...

Confusion matrix, without normalization

[[ 5 8 0]

[ 6 19 4]

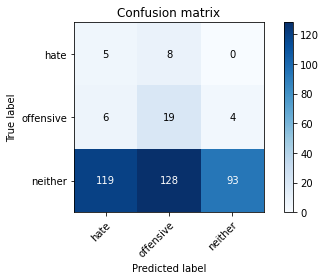
[119 128 93]]

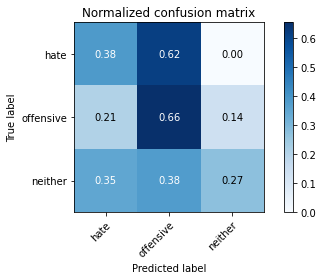
Normalized confusion matrix

[[0.38461538 0.61538462 0. ]

[0.20689655 0.65517241 0.13793103]

[0.35 0.37647059 0.27352941]]





precision recall f1-score support

hate 0.04 0.38 0.07 13

offensive 0.12 0.66 0.21 29

neither 0.96 0.27 0.43 340

accuracy 0.31 382

macro avg 0.37 0.44 0.23 382

weighted avg 0.86 0.31 0.40 382

average eval\_loss: 1.1413078248500823, accuracy: 0.306282722513089

training 3 epoch...

2.933938503265381 seconds used......

3 training finished! average train loss: 1.9915361404418945

total clf loss: 15.202381312847137 total adv loss: 7.33533039689064

evaluating...

Confusion matrix, without normalization

[[ 1 4 8]

[ 3 12 14]

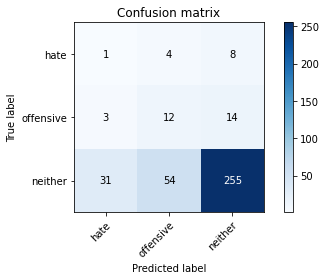
[ 31 54 255]]

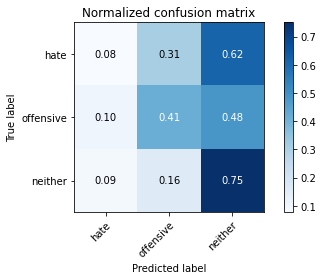
Normalized confusion matrix

[[0.07692308 0.30769231 0.61538462]

[0.10344828 0.4137931 0.48275862]

[0.09117647 0.15882353 0.75 ]]





precision recall f1-score support

hate 0.03 0.08 0.04 13

offensive 0.17 0.41 0.24 29

neither 0.92 0.75 0.83 340

accuracy 0.70 382

macro avg 0.37 0.41 0.37 382

weighted avg 0.83 0.70 0.76 382

average eval\_loss: 0.9506613403558731, accuracy: 0.7015706806282722

training 4 epoch...

2.8364734649658203 seconds used......

4 training finished! average train loss: 1.6902782201766968

total clf loss: 13.87440174818039 total adv loss: 5.739885941147804

evaluating...

Confusion matrix, without normalization

[[ 8 4 1]

[ 11 13 5]

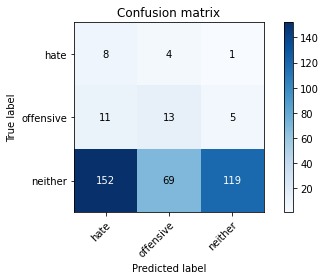
[152 69 119]]

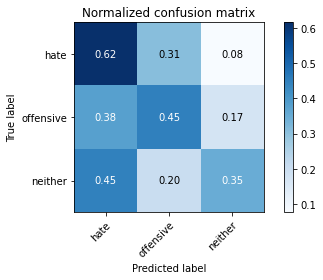
Normalized confusion matrix

[[0.61538462 0.30769231 0.07692308]

[0.37931034 0.44827586 0.17241379]

[0.44705882 0.20294118 0.35 ]]





precision recall f1-score support

hate 0.05 0.62 0.09 13

offensive 0.15 0.45 0.23 29

neither 0.95 0.35 0.51 340

accuracy 0.37 382

macro avg 0.38 0.47 0.27 382

weighted avg 0.86 0.37 0.48 382

average eval\_loss: 1.0820983290672301, accuracy: 0.36649214659685864

training 5 epoch...

2.840958595275879 seconds used......

5 training finished! average train loss: 1.2531107902526855

total clf loss: 12.216821670532227 total adv loss: 3.2899201698601246

evaluating...

Confusion matrix, without normalization

[[ 7 3 3]

[ 5 15 9]

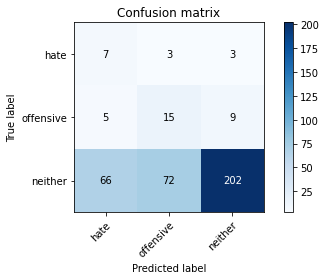
[ 66 72 202]]

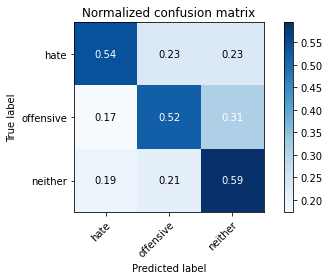
Normalized confusion matrix

[[0.53846154 0.23076923 0.23076923]

[0.17241379 0.51724138 0.31034483]

[0.19411765 0.21176471 0.59411765]]





precision recall f1-score support

hate 0.09 0.54 0.15 13

offensive 0.17 0.52 0.25 29

neither 0.94 0.59 0.73 340

accuracy 0.59 382

macro avg 0.40 0.55 0.38 382

weighted avg 0.86 0.59 0.67 382

average eval\_loss: 0.9512966752052308, accuracy: 0.5863874345549738

Confusion matrix, without normalization

[[ 24 10 7]

[ 67 7 21]

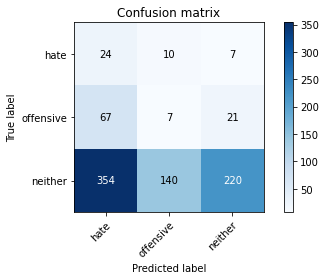
[354 140 220]]

Normalized confusion matrix

[[0.58536585 0.24390244 0.17073171]

[0.70526316 0.07368421 0.22105263]

[0.49579832 0.19607843 0.30812325]]





precision recall f1-score support

hate 0.05 0.59 0.10 41

offensive 0.04 0.07 0.06 95

neither 0.89 0.31 0.46 714

accuracy 0.30 850

macro avg 0.33 0.32 0.20 850

weighted avg 0.75 0.30 0.40 850

average eval\_loss: 1.1072106458419977, accuracy: 0.2952941176470588

[31]:

0.2952941176470588