1 run

hate as no-offensive

alpha=2

training 1 epoch...

21.925204515457153 seconds used......

1 training finished! average train loss: 2.267572147505624

total clf loss: 58.60013872385025 total adv loss: 34.19195154309273

evaluating...

Confusion matrix, without normalization

[[ 26 6 13]

[ 60 111 15]

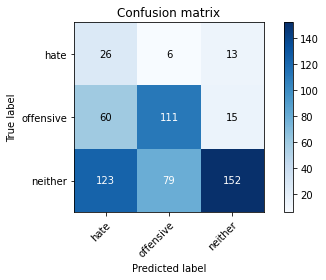
[123 79 152]]

Normalized confusion matrix

[[0.57777778 0.13333333 0.28888889]

[0.32258065 0.59677419 0.08064516]

[0.34745763 0.22316384 0.42937853]]





precision recall f1-score support

hate 0.12 0.58 0.20 45

offensive 0.57 0.60 0.58 186

neither 0.84 0.43 0.57 354

accuracy 0.49 585

macro avg 0.51 0.53 0.45 585

weighted avg 0.70 0.49 0.55 585

average eval\_loss: 1.0110040525595347, accuracy: 0.49401709401709404

training 2 epoch...

22.353015661239624 seconds used......

2 training finished! average train loss: 2.0054234990051816

total clf loss: 50.76464593410492 total adv loss: 30.76953473687172

evaluating...

Confusion matrix, without normalization

[[ 39 1 5]

[ 78 86 22]

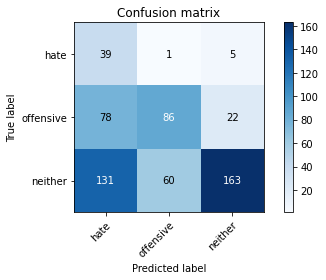
[131 60 163]]

Normalized confusion matrix

[[0.86666667 0.02222222 0.11111111]

[0.41935484 0.46236559 0.11827957]

[0.3700565 0.16949153 0.46045198]]





precision recall f1-score support

hate 0.16 0.87 0.27 45

offensive 0.59 0.46 0.52 186

neither 0.86 0.46 0.60 354

accuracy 0.49 585

macro avg 0.53 0.60 0.46 585

weighted avg 0.72 0.49 0.55 585

average eval\_loss: 0.9757328867912293, accuracy: 0.49230769230769234

training 3 epoch...

22.601354360580444 seconds used......

3 training finished! average train loss: 1.4695519006678037

total clf loss: 37.85035026073456 total adv loss: 22.222278088331223

evaluating...

Confusion matrix, without normalization

[[ 18 13 14]

[ 8 147 31]

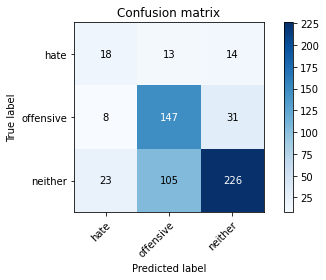
[ 23 105 226]]

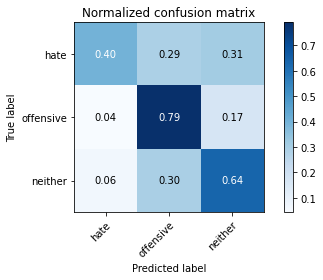
Normalized confusion matrix

[[0.4 0.28888889 0.31111111]

[0.04301075 0.79032258 0.16666667]

[0.06497175 0.29661017 0.63841808]]





precision recall f1-score support

hate 0.37 0.40 0.38 45

offensive 0.55 0.79 0.65 186

neither 0.83 0.64 0.72 354

accuracy 0.67 585

macro avg 0.59 0.61 0.59 585

weighted avg 0.71 0.67 0.67 585

average eval\_loss: 0.7292805582284927, accuracy: 0.6683760683760683

training 4 epoch...

22.720101833343506 seconds used......

4 training finished! average train loss: 0.9134044407733849

total clf loss: 24.432701244950294 total adv loss: 13.358973808586597

evaluating...

Confusion matrix, without normalization

[[ 18 13 14]

[ 15 136 35]

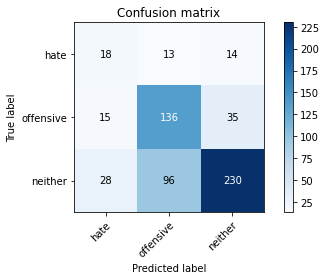
[ 28 96 230]]

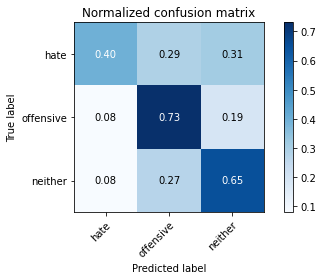
Normalized confusion matrix

[[0.4 0.28888889 0.31111111]

[0.08064516 0.7311828 0.18817204]

[0.07909605 0.27118644 0.64971751]]





precision recall f1-score support

hate 0.30 0.40 0.34 45

offensive 0.56 0.73 0.63 186

neither 0.82 0.65 0.73 354

accuracy 0.66 585

macro avg 0.56 0.59 0.57 585

weighted avg 0.70 0.66 0.67 585

average eval\_loss: 0.8009018460909526, accuracy: 0.6564102564102564

training 5 epoch...

22.837148189544678 seconds used......

5 training finished! average train loss: 0.6549666706206543

total clf loss: 16.617399476468563 total adv loss: 10.030367065221071

evaluating...

Confusion matrix, without normalization

[[ 32 4 9]

[ 35 123 28]

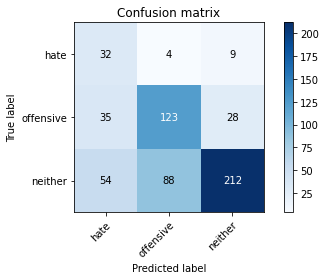
[ 54 88 212]]

Normalized confusion matrix

[[0.71111111 0.08888889 0.2 ]

[0.18817204 0.66129032 0.15053763]

[0.15254237 0.24858757 0.59887006]]





precision recall f1-score support

hate 0.26 0.71 0.39 45

offensive 0.57 0.66 0.61 186

neither 0.85 0.60 0.70 354

accuracy 0.63 585

macro avg 0.56 0.66 0.57 585

weighted avg 0.72 0.63 0.65 585

average eval\_loss: 1.0214560121297835, accuracy: 0.6273504273504273

model of 2 model

Confusion matrix, without normalization

[[ 45 3 3]

[ 79 68 22]

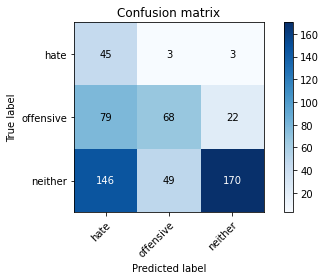
[146 49 170]]

Normalized confusion matrix

[[0.88235294 0.05882353 0.05882353]

[0.46745562 0.40236686 0.13017751]

[0.4 0.13424658 0.46575342]]





precision recall f1-score support

hate 0.17 0.88 0.28 51

offensive 0.57 0.40 0.47 169

neither 0.87 0.47 0.61 365

accuracy 0.48 585

macro avg 0.54 0.58 0.45 585

weighted avg 0.72 0.48 0.54 585

average eval\_loss: 1.0039280931154886, accuracy: 0.48376068376068376

0.48376068376068376

2 run

training 1 epoch...

14.012221097946167 seconds used......

1 training finished! average train loss: 2.2847513450043544

total clf loss: 59.047960460186005 total adv loss: 34.44905775785446

evaluating...

Confusion matrix, without normalization

[[ 29 14 2]

[ 41 143 2]

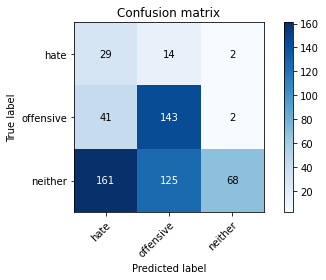
[161 125 68]]

Normalized confusion matrix

[[0.64444444 0.31111111 0.04444444]

[0.22043011 0.7688172 0.01075269]

[0.45480226 0.35310734 0.1920904 ]]





precision recall f1-score support

hate 0.13 0.64 0.21 45

offensive 0.51 0.77 0.61 186

neither 0.94 0.19 0.32 354

accuracy 0.41 585

macro avg 0.53 0.54 0.38 585

weighted avg 0.74 0.41 0.40 585

average eval\_loss: 1.0901832342147828, accuracy: 0.41025641025641024

training 2 epoch...

14.020027875900269 seconds used......

2 training finished! average train loss: 1.939704567193985

total clf loss: 49.325459241867065 total adv loss: 29.64899817109108

evaluating...

Confusion matrix, without normalization

[[ 29 5 11]

[ 32 117 37]

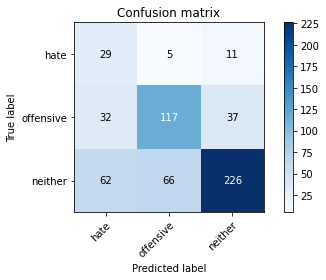
[ 62 66 226]]

Normalized confusion matrix

[[0.64444444 0.11111111 0.24444444]

[0.17204301 0.62903226 0.19892473]

[0.17514124 0.18644068 0.63841808]]





precision recall f1-score support

hate 0.24 0.64 0.35 45

offensive 0.62 0.63 0.63 186

neither 0.82 0.64 0.72 354

accuracy 0.64 585

macro avg 0.56 0.64 0.56 585

weighted avg 0.72 0.64 0.66 585

average eval\_loss: 0.8340918779373169, accuracy: 0.6358974358974359

training 3 epoch...

14.115302324295044 seconds used......

3 training finished! average train loss: 1.3858679756522179

total clf loss: 35.46074190735817 total adv loss: 21.073932334780693

evaluating...

Confusion matrix, without normalization

[[ 15 7 23]

[ 8 119 59]

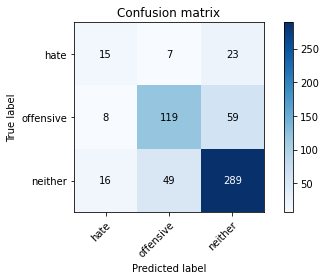
[ 16 49 289]]

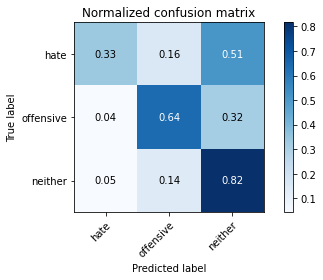
Normalized confusion matrix

[[0.33333333 0.15555556 0.51111111]

[0.04301075 0.63978495 0.3172043 ]

[0.04519774 0.13841808 0.81638418]]





precision recall f1-score support

hate 0.38 0.33 0.36 45

offensive 0.68 0.64 0.66 186

neither 0.78 0.82 0.80 354

accuracy 0.72 585

macro avg 0.61 0.60 0.60 585

weighted avg 0.72 0.72 0.72 585

average eval\_loss: 0.671064481139183, accuracy: 0.7230769230769231

training 4 epoch...

14.017977952957153 seconds used......

4 training finished! average train loss: 0.9177812023886612

total clf loss: 24.2332776337862 total adv loss: 13.581234730780125

evaluating...

Confusion matrix, without normalization

[[ 17 12 16]

[ 10 133 43]

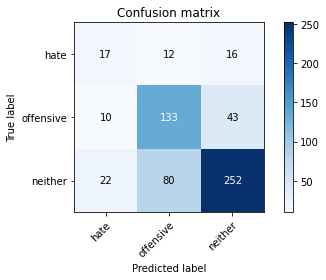
[ 22 80 252]]

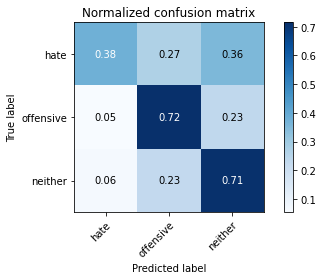
Normalized confusion matrix

[[0.37777778 0.26666667 0.35555556]

[0.05376344 0.71505376 0.2311828 ]

[0.06214689 0.2259887 0.71186441]]





precision recall f1-score support

hate 0.35 0.38 0.36 45

offensive 0.59 0.72 0.65 186

neither 0.81 0.71 0.76 354

accuracy 0.69 585

macro avg 0.58 0.60 0.59 585

weighted avg 0.70 0.69 0.69 585

average eval\_loss: 0.7909988174835841, accuracy: 0.6871794871794872

training 5 epoch...

14.017144203186035 seconds used......

5 training finished! average train loss: 0.591921810859016

total clf loss: 17.793370611965656 total adv loss: 7.677125383168459

evaluating...

Confusion matrix, without normalization

[[ 23 10 12]

[ 19 134 33]

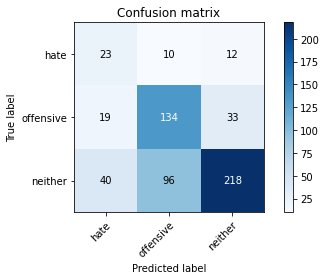
[ 40 96 218]]

Normalized confusion matrix

[[0.51111111 0.22222222 0.26666667]

[0.10215054 0.72043011 0.17741935]

[0.11299435 0.27118644 0.61581921]]





precision recall f1-score support

hate 0.28 0.51 0.36 45

offensive 0.56 0.72 0.63 186

neither 0.83 0.62 0.71 354

accuracy 0.64 585

macro avg 0.56 0.62 0.57 585

weighted avg 0.70 0.64 0.66 585

average eval\_loss: 1.08884193499883, accuracy: 0.6410256410256411

Confusion matrix, without normalization

[[ 37 5 9]

[ 42 91 36]

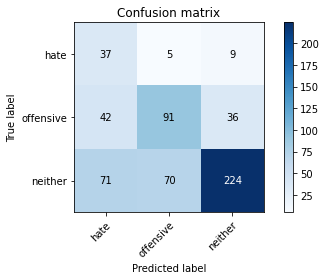
[ 71 70 224]]

Normalized confusion matrix

[[0.7254902 0.09803922 0.17647059]

[0.24852071 0.53846154 0.21301775]

[0.19452055 0.19178082 0.61369863]]





precision recall f1-score support

hate 0.25 0.73 0.37 51

offensive 0.55 0.54 0.54 169

neither 0.83 0.61 0.71 365

accuracy 0.60 585

macro avg 0.54 0.63 0.54 585

weighted avg 0.70 0.60 0.63 585

average eval\_loss: 0.8662843028704326, accuracy: 0.6017094017094017

[26]:

0.6017094017094017

3 run

training 1 epoch...

13.985543966293335 seconds used......

1 training finished! average train loss: 2.3147126819406236

total clf loss: 59.8131148815155 total adv loss: 34.90539747476578

evaluating...

Confusion matrix, without normalization

[[ 6 17 22]

[ 10 146 30]

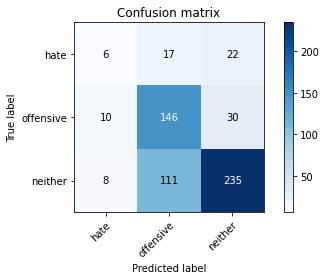
[ 8 111 235]]

Normalized confusion matrix

[[0.13333333 0.37777778 0.48888889]

[0.05376344 0.78494624 0.16129032]

[0.02259887 0.31355932 0.66384181]]





precision recall f1-score support

hate 0.25 0.13 0.17 45

offensive 0.53 0.78 0.63 186

neither 0.82 0.66 0.73 354

accuracy 0.66 585

macro avg 0.53 0.53 0.51 585

weighted avg 0.68 0.66 0.66 585

average eval\_loss: 0.8859731296698252, accuracy: 0.6615384615384615

training 2 epoch...

14.014824628829956 seconds used......

2 training finished! average train loss: 1.952700702207429

total clf loss: 50.36096692085266 total adv loss: 29.495135694742203

evaluating...

Confusion matrix, without normalization

[[ 31 3 11]

[ 61 83 42]

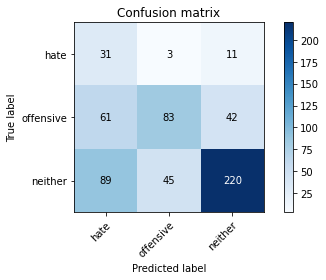
[ 89 45 220]]

Normalized confusion matrix

[[0.68888889 0.06666667 0.24444444]

[0.32795699 0.44623656 0.22580645]

[0.25141243 0.12711864 0.62146893]]





precision recall f1-score support

hate 0.17 0.69 0.27 45

offensive 0.63 0.45 0.52 186

neither 0.81 0.62 0.70 354

accuracy 0.57 585

macro avg 0.54 0.59 0.50 585

weighted avg 0.70 0.57 0.61 585

average eval\_loss: 0.8740121225516001, accuracy: 0.5709401709401709

training 3 epoch...

14.01753568649292 seconds used......

3 training finished! average train loss: 1.4004887991717883

total clf loss: 36.23469293117523 total adv loss: 21.096339985728264

evaluating...

Confusion matrix, without normalization

[[ 21 10 14]

[ 16 130 40]

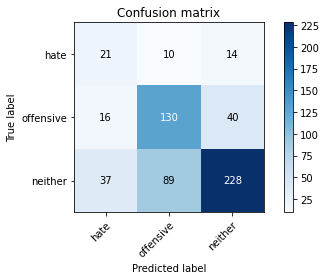
[ 37 89 228]]

Normalized confusion matrix

[[0.46666667 0.22222222 0.31111111]

[0.08602151 0.69892473 0.21505376]

[0.10451977 0.25141243 0.6440678 ]]





precision recall f1-score support

hate 0.28 0.47 0.35 45

offensive 0.57 0.70 0.63 186

neither 0.81 0.64 0.72 354

accuracy 0.65 585

macro avg 0.55 0.60 0.57 585

weighted avg 0.69 0.65 0.66 585

average eval\_loss: 0.7822575747966767, accuracy: 0.6478632478632479

training 4 epoch...

13.99333381652832 seconds used......

4 training finished! average train loss: 0.8739782614367348

total clf loss: 23.401167064905167 total adv loss: 12.77080781199038

evaluating...

Confusion matrix, without normalization

[[ 21 16 8]

[ 17 154 15]

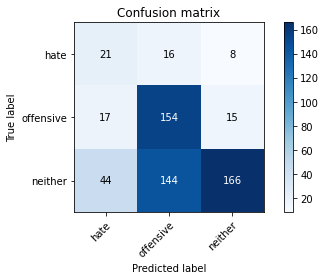
[ 44 144 166]]

Normalized confusion matrix

[[0.46666667 0.35555556 0.17777778]

[0.09139785 0.82795699 0.08064516]

[0.12429379 0.40677966 0.46892655]]





precision recall f1-score support

hate 0.26 0.47 0.33 45

offensive 0.49 0.83 0.62 186

neither 0.88 0.47 0.61 354

accuracy 0.58 585

macro avg 0.54 0.59 0.52 585

weighted avg 0.71 0.58 0.59 585

average eval\_loss: 1.0572785913944245, accuracy: 0.582905982905983

training 5 epoch...

14.040800094604492 seconds used......

5 training finished! average train loss: 0.5078846716455051

total clf loss: 14.910453770309687 total adv loss: 6.765543946065009

evaluating...

Confusion matrix, without normalization

[[ 27 8 10]

[ 32 118 36]

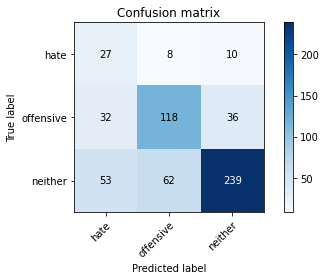
[ 53 62 239]]

Normalized confusion matrix

[[0.6 0.17777778 0.22222222]

[0.17204301 0.6344086 0.19354839]

[0.14971751 0.17514124 0.67514124]]





precision recall f1-score support

hate 0.24 0.60 0.34 45

offensive 0.63 0.63 0.63 186

neither 0.84 0.68 0.75 354

accuracy 0.66 585

macro avg 0.57 0.64 0.57 585

weighted avg 0.73 0.66 0.68 585

average eval\_loss: 1.0403944849967957, accuracy: 0.6564102564102564

Confusion matrix, without normalization

[[ 37 5 9]

[ 62 70 37]

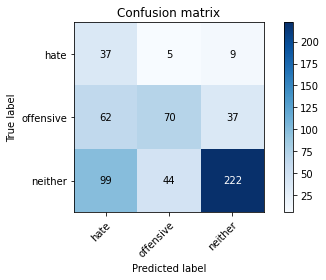
[ 99 44 222]]

Normalized confusion matrix

[[0.7254902 0.09803922 0.17647059]

[0.36686391 0.41420118 0.21893491]

[0.27123288 0.12054795 0.60821918]]





precision recall f1-score support

hate 0.19 0.73 0.30 51

offensive 0.59 0.41 0.49 169

neither 0.83 0.61 0.70 365

accuracy 0.56 585

macro avg 0.53 0.58 0.49 585

weighted avg 0.70 0.56 0.60 585

average eval\_loss: 0.9031471093495687, accuracy: 0.5623931623931624

[28]:

0.5623931623931624