1 run

hate as no-offensive

alpha=2

training 1 epoch...

45.376911640167236 seconds used......

1 training finished! average train loss: 1.8793490748656423

total clf loss: 69.4890034198761 total adv loss: 36.670763432979584

evaluating...

Confusion matrix, without normalization

[[ 28 9 16]

[ 42 123 16]

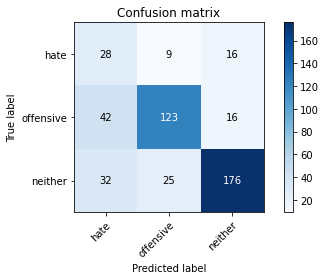
[ 32 25 176]]

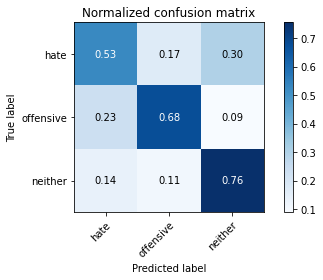
Normalized confusion matrix

[[0.52830189 0.16981132 0.30188679]

[0.2320442 0.67955801 0.08839779]

[0.13733906 0.10729614 0.75536481]]





precision recall f1-score support

hate 0.27 0.53 0.36 53

offensive 0.78 0.68 0.73 181

neither 0.85 0.76 0.80 233

accuracy 0.70 467

macro avg 0.63 0.65 0.63 467

weighted avg 0.76 0.70 0.72 467

average eval\_loss: 0.7075266391038895, accuracy: 0.7002141327623126

training 2 epoch...

43.006617307662964 seconds used......

2 training finished! average train loss: 1.272889508228553

total clf loss: 49.02213326096535 total adv loss: 23.858734726905823

evaluating...

Confusion matrix, without normalization

[[ 30 9 14]

[ 38 127 16]

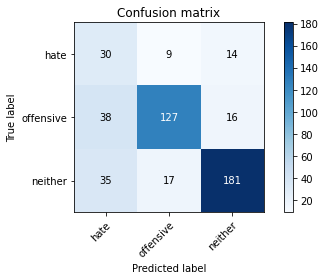
[ 35 17 181]]

Normalized confusion matrix

[[0.56603774 0.16981132 0.26415094]

[0.20994475 0.70165746 0.08839779]

[0.15021459 0.07296137 0.77682403]]





precision recall f1-score support

hate 0.29 0.57 0.38 53

offensive 0.83 0.70 0.76 181

neither 0.86 0.78 0.82 233

accuracy 0.72 467

macro avg 0.66 0.68 0.65 467

weighted avg 0.78 0.72 0.75 467

average eval\_loss: 0.6604242424170176, accuracy: 0.7237687366167024

training 3 epoch...

44.12295961380005 seconds used......

3 training finished! average train loss: 0.837271259803521

total clf loss: 35.316340267658234 total adv loss: 14.158137682825327

evaluating...

Confusion matrix, without normalization

[[ 33 10 10]

[ 32 136 13]

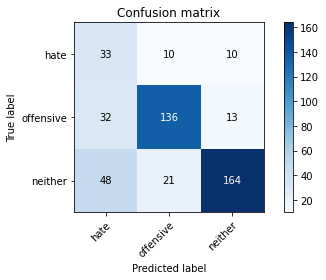
[ 48 21 164]]

Normalized confusion matrix

[[0.62264151 0.18867925 0.18867925]

[0.17679558 0.75138122 0.0718232 ]

[0.20600858 0.09012876 0.70386266]]





precision recall f1-score support

hate 0.29 0.62 0.40 53

offensive 0.81 0.75 0.78 181

neither 0.88 0.70 0.78 233

accuracy 0.71 467

macro avg 0.66 0.69 0.65 467

weighted avg 0.79 0.71 0.74 467

average eval\_loss: 0.725191400696834, accuracy: 0.7130620985010707

training 4 epoch...

43.534438610076904 seconds used......

4 training finished! average train loss: 0.48298447412487705

total clf loss: 23.579825200140476 total adv loss: 6.563497411669232

evaluating...

Confusion matrix, without normalization

[[ 28 10 15]

[ 37 125 19]

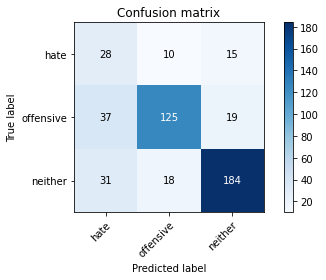
[ 31 18 184]]

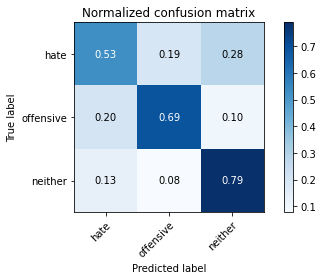
Normalized confusion matrix

[[0.52830189 0.18867925 0.28301887]

[0.20441989 0.69060773 0.10497238]

[0.13304721 0.07725322 0.78969957]]





precision recall f1-score support

hate 0.29 0.53 0.38 53

offensive 0.82 0.69 0.75 181

neither 0.84 0.79 0.82 233

accuracy 0.72 467

macro avg 0.65 0.67 0.65 467

weighted avg 0.77 0.72 0.74 467

average eval\_loss: 0.8264982979744673, accuracy: 0.721627408993576

training 5 epoch...

43.65590000152588 seconds used......

5 training finished! average train loss: 0.33613008397974464

total clf loss: 16.922501370310783 total adv loss: 4.311692480812781

evaluating...

Confusion matrix, without normalization

[[ 29 8 16]

[ 40 123 18]

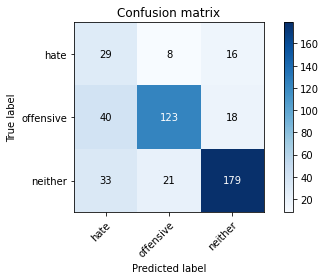
[ 33 21 179]]

Normalized confusion matrix

[[0.54716981 0.1509434 0.30188679]

[0.22099448 0.67955801 0.09944751]

[0.1416309 0.09012876 0.76824034]]





precision recall f1-score support

hate 0.28 0.55 0.37 53

offensive 0.81 0.68 0.74 181

neither 0.84 0.77 0.80 233

accuracy 0.71 467

macro avg 0.64 0.66 0.64 467

weighted avg 0.77 0.71 0.73 467

average eval\_loss: 0.9300861718753973, accuracy: 0.708779443254818

model of 3 epoch

Confusion matrix, without normalization

[[117 51 22]

[108 281 26]

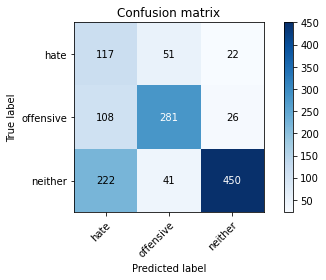
[222 41 450]]

Normalized confusion matrix

[[0.61578947 0.26842105 0.11578947]

[0.26024096 0.67710843 0.0626506 ]

[0.31136045 0.05750351 0.63113604]]





precision recall f1-score support

hate 0.26 0.62 0.37 190

offensive 0.75 0.68 0.71 415

neither 0.90 0.63 0.74 713

accuracy 0.64 1318

macro avg 0.64 0.64 0.61 1318

weighted avg 0.76 0.64 0.68 1318

average eval\_loss: 0.8598443513566797, accuracy: 0.6433990895295902

0.6433990895295902

2 run

training 1 epoch...

23.59068489074707 seconds used......

1 training finished! average train loss: 1.8951490254778611

total clf loss: 68.39754122495651 total adv loss: 37.81689238548279

evaluating...

Confusion matrix, without normalization

[[ 35 5 13]

[ 56 113 12]

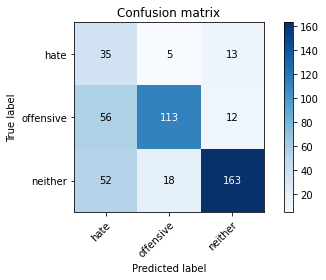
[ 52 18 163]]

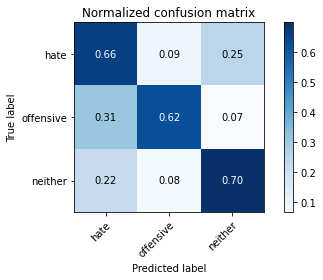
Normalized confusion matrix

[[0.66037736 0.09433962 0.24528302]

[0.30939227 0.62430939 0.06629834]

[0.22317597 0.07725322 0.69957082]]





precision recall f1-score support

hate 0.24 0.66 0.36 53

offensive 0.83 0.62 0.71 181

neither 0.87 0.70 0.77 233

accuracy 0.67 467

macro avg 0.65 0.66 0.61 467

weighted avg 0.78 0.67 0.70 467

average eval\_loss: 0.7582895718514919, accuracy: 0.6659528907922913

training 2 epoch...

23.6191143989563 seconds used......

2 training finished! average train loss: 1.3385444903060009

total clf loss: 50.02864933013916 total adv loss: 25.850365683436394

evaluating...

Confusion matrix, without normalization

[[ 25 10 18]

[ 27 134 20]

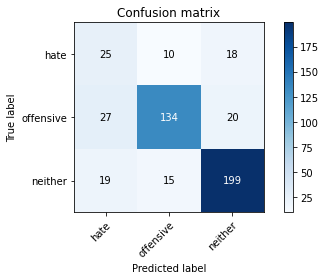
[ 19 15 199]]

Normalized confusion matrix

[[0.47169811 0.18867925 0.33962264]

[0.14917127 0.74033149 0.11049724]

[0.08154506 0.06437768 0.85407725]]





precision recall f1-score support

hate 0.35 0.47 0.40 53

offensive 0.84 0.74 0.79 181

neither 0.84 0.85 0.85 233

accuracy 0.77 467

macro avg 0.68 0.69 0.68 467

weighted avg 0.79 0.77 0.77 467

average eval\_loss: 0.6148097688953081, accuracy: 0.7665952890792291

training 3 epoch...

23.60957145690918 seconds used......

3 training finished! average train loss: 0.8664104558135334

total clf loss: 35.71084403991699 total adv loss: 15.068175304681063

evaluating...

Confusion matrix, without normalization

[[ 24 13 16]

[ 29 135 17]

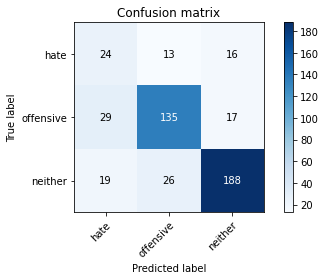
[ 19 26 188]]

Normalized confusion matrix

[[0.45283019 0.24528302 0.30188679]

[0.16022099 0.74585635 0.09392265]

[0.08154506 0.11158798 0.80686695]]





precision recall f1-score support

hate 0.33 0.45 0.38 53

offensive 0.78 0.75 0.76 181

neither 0.85 0.81 0.83 233

accuracy 0.74 467

macro avg 0.65 0.67 0.66 467

weighted avg 0.76 0.74 0.75 467

average eval\_loss: 0.6312394266327223, accuracy: 0.7430406852248393

training 4 epoch...

23.609922409057617 seconds used......

4 training finished! average train loss: 0.563982435158993

total clf loss: 26.02254205942154 total adv loss: 8.420061356911901

evaluating...

Confusion matrix, without normalization

[[ 35 9 9]

[ 50 118 13]

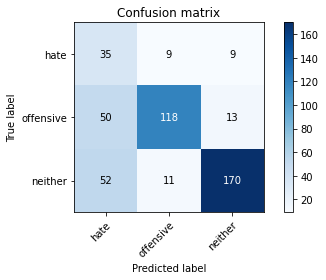
[ 52 11 170]]

Normalized confusion matrix

[[0.66037736 0.16981132 0.16981132]

[0.27624309 0.6519337 0.0718232 ]

[0.22317597 0.0472103 0.72961373]]





precision recall f1-score support

hate 0.26 0.66 0.37 53

offensive 0.86 0.65 0.74 181

neither 0.89 0.73 0.80 233

accuracy 0.69 467

macro avg 0.67 0.68 0.64 467

weighted avg 0.80 0.69 0.73 467

average eval\_loss: 0.8382350988686085, accuracy: 0.6916488222698073

training 5 epoch...

23.585565328598022 seconds used......

5 training finished! average train loss: 0.2933646182676679

total clf loss: 15.775320678949356 total adv loss: 3.2601951264805393

evaluating...

Confusion matrix, without normalization

[[ 38 7 8]

[ 50 123 8]

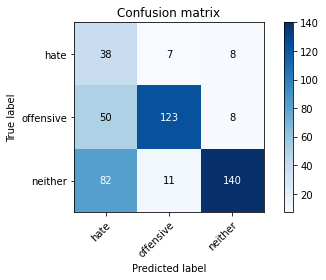
[ 82 11 140]]

Normalized confusion matrix

[[0.71698113 0.13207547 0.1509434 ]

[0.27624309 0.67955801 0.0441989 ]

[0.35193133 0.0472103 0.60085837]]





precision recall f1-score support

hate 0.22 0.72 0.34 53

offensive 0.87 0.68 0.76 181

neither 0.90 0.60 0.72 233

accuracy 0.64 467

macro avg 0.66 0.67 0.61 467

weighted avg 0.81 0.64 0.69 467

average eval\_loss: 1.0221013153592746, accuracy: 0.6445396145610278

Confusion matrix, without normalization

[[139 38 13]

[148 250 17]

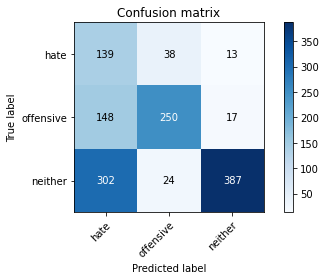
[302 24 387]]

Normalized confusion matrix

[[0.73157895 0.2 0.06842105]

[0.35662651 0.60240964 0.04096386]

[0.42356241 0.03366059 0.542777 ]]





precision recall f1-score support

hate 0.24 0.73 0.36 190

offensive 0.80 0.60 0.69 415

neither 0.93 0.54 0.68 713

accuracy 0.59 1318

macro avg 0.66 0.63 0.58 1318

weighted avg 0.79 0.59 0.64 1318

average eval\_loss: 1.2073357294906268, accuracy: 0.5887708649468892

[27]:

0.5887708649468892

3 run

training 1 epoch...

23.537813425064087 seconds used......

1 training finished! average train loss: 1.8694333653700979

total clf loss: 68.10053640604019 total adv loss: 36.98819960653782

evaluating...

Confusion matrix, without normalization

[[ 36 5 12]

[ 61 105 15]

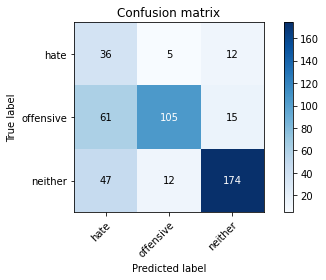
[ 47 12 174]]

Normalized confusion matrix

[[0.67924528 0.09433962 0.22641509]

[0.33701657 0.5801105 0.08287293]

[0.20171674 0.05150215 0.74678112]]





precision recall f1-score support

hate 0.25 0.68 0.37 53

offensive 0.86 0.58 0.69 181

neither 0.87 0.75 0.80 233

accuracy 0.67 467

macro avg 0.66 0.67 0.62 467

weighted avg 0.79 0.67 0.71 467

average eval\_loss: 0.7336169332265854, accuracy: 0.6745182012847966

training 2 epoch...

23.6349458694458 seconds used......

2 training finished! average train loss: 1.3020073418554508

total clf loss: 48.96502637863159 total adv loss: 24.993765771389008

evaluating...

Confusion matrix, without normalization

[[ 22 12 19]

[ 31 135 15]

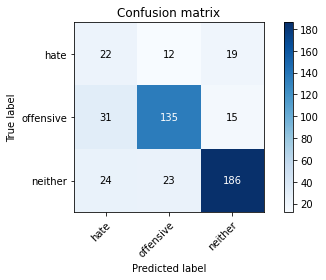
[ 24 23 186]]

Normalized confusion matrix

[[0.41509434 0.22641509 0.35849057]

[0.17127072 0.74585635 0.08287293]

[0.10300429 0.09871245 0.79828326]]





precision recall f1-score support

hate 0.29 0.42 0.34 53

offensive 0.79 0.75 0.77 181

neither 0.85 0.80 0.82 233

accuracy 0.73 467

macro avg 0.64 0.65 0.64 467

weighted avg 0.76 0.73 0.75 467

average eval\_loss: 0.6447433494031429, accuracy: 0.734475374732334

training 3 epoch...

23.592679023742676 seconds used......

3 training finished! average train loss: 0.8834271913296298

total clf loss: 35.424640417099 total adv loss: 15.85791288036853

evaluating...

Confusion matrix, without normalization

[[ 32 7 14]

[ 54 107 20]

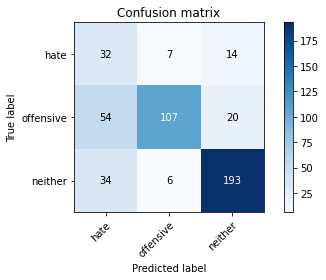
[ 34 6 193]]

Normalized confusion matrix

[[0.60377358 0.13207547 0.26415094]

[0.29834254 0.59116022 0.11049724]

[0.14592275 0.02575107 0.82832618]]





precision recall f1-score support

hate 0.27 0.60 0.37 53

offensive 0.89 0.59 0.71 181

neither 0.85 0.83 0.84 233

accuracy 0.71 467

macro avg 0.67 0.67 0.64 467

weighted avg 0.80 0.71 0.74 467

average eval\_loss: 0.8219041352470716, accuracy: 0.7109207708779444

training 4 epoch...

23.577460289001465 seconds used......

4 training finished! average train loss: 0.4951265403314641

total clf loss: 23.534392535686493 total adv loss: 7.047612245660275

evaluating...

Confusion matrix, without normalization

[[ 27 10 16]

[ 41 122 18]

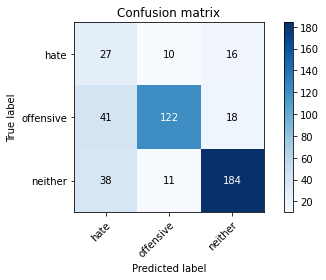
[ 38 11 184]]

Normalized confusion matrix

[[0.50943396 0.18867925 0.30188679]

[0.22651934 0.67403315 0.09944751]

[0.16309013 0.0472103 0.78969957]]





precision recall f1-score support

hate 0.25 0.51 0.34 53

offensive 0.85 0.67 0.75 181

neither 0.84 0.79 0.82 233

accuracy 0.71 467

macro avg 0.65 0.66 0.64 467

weighted avg 0.78 0.71 0.74 467

average eval\_loss: 0.8474269943932692, accuracy: 0.7130620985010707

training 5 epoch...

23.557306051254272 seconds used......

5 training finished! average train loss: 0.3005831566100058

total clf loss: 15.109378851950169 total adv loss: 3.8674705600715242

evaluating...

Confusion matrix, without normalization

[[ 30 11 12]

[ 34 131 16]

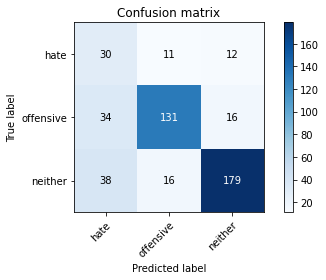
[ 38 16 179]]

Normalized confusion matrix

[[0.56603774 0.20754717 0.22641509]

[0.1878453 0.72375691 0.08839779]

[0.16309013 0.06866953 0.76824034]]





precision recall f1-score support

hate 0.29 0.57 0.39 53

offensive 0.83 0.72 0.77 181

neither 0.86 0.77 0.81 233

accuracy 0.73 467

macro avg 0.66 0.69 0.66 467

weighted avg 0.79 0.73 0.75 467

average eval\_loss: 0.9157316324611505, accuracy: 0.728051391862955

Confusion matrix, without normalization

[[137 19 34]

[182 194 39]

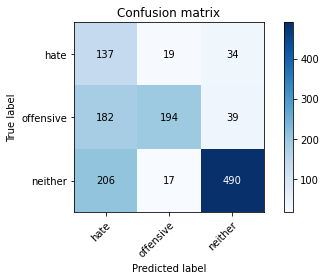
[206 17 490]]

Normalized confusion matrix

[[0.72105263 0.1 0.17894737]

[0.43855422 0.46746988 0.0939759 ]

[0.28892006 0.02384292 0.68723703]]





precision recall f1-score support

hate 0.26 0.72 0.38 190

offensive 0.84 0.47 0.60 415

neither 0.87 0.69 0.77 713

accuracy 0.62 1318

macro avg 0.66 0.63 0.58 1318

weighted avg 0.77 0.62 0.66 1318

average eval\_loss: 0.8306867579619089, accuracy: 0.6229135053110774

[29]:

0.6229135053110774