1 run

hate as no-offensive

alpha=2

training 1 epoch...

197.38122296333313 seconds used......

1 training finished! average train loss: 2.0882363711532794

total clf loss: 152.1793835759163 total adv loss: 82.61627238988876

evaluating...

Confusion matrix, without normalization

[[ 40 43 38]

[ 12 82 13]

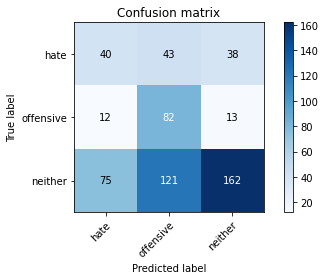
[ 75 121 162]]

Normalized confusion matrix

[[0.33057851 0.3553719 0.31404959]

[0.11214953 0.76635514 0.12149533]

[0.20949721 0.33798883 0.45251397]]





precision recall f1-score support

hate 0.31 0.33 0.32 121

offensive 0.33 0.77 0.46 107

neither 0.76 0.45 0.57 358

accuracy 0.48 586

macro avg 0.47 0.52 0.45 586

weighted avg 0.59 0.48 0.50 586

average eval\_loss: 1.000156287352244, accuracy: 0.48464163822525597

training 2 epoch...

195.66008853912354 seconds used......

2 training finished! average train loss: 1.744502986732282

total clf loss: 131.36636477708817 total adv loss: 66.89904476702213

evaluating...

Confusion matrix, without normalization

[[ 67 33 21]

[ 20 81 6]

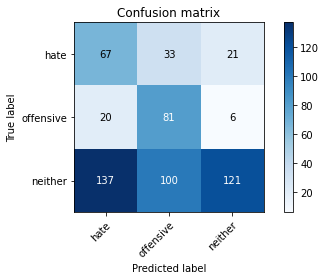
[137 100 121]]

Normalized confusion matrix

[[0.55371901 0.27272727 0.17355372]

[0.18691589 0.75700935 0.05607477]

[0.38268156 0.27932961 0.33798883]]





precision recall f1-score support

hate 0.30 0.55 0.39 121

offensive 0.38 0.76 0.50 107

neither 0.82 0.34 0.48 358

accuracy 0.46 586

macro avg 0.50 0.55 0.46 586

weighted avg 0.63 0.46 0.46 586

average eval\_loss: 0.9964302579561869, accuracy: 0.4590443686006826

training 3 epoch...

195.45994520187378 seconds used......

3 training finished! average train loss: 1.3492386184240643

total clf loss: 109.76300635933876 total adv loss: 47.660631977021694

evaluating...

Confusion matrix, without normalization

[[ 60 17 44]

[ 27 51 29]

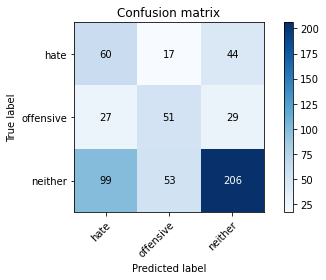
[ 99 53 206]]

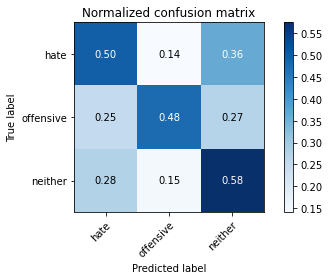
Normalized confusion matrix

[[0.49586777 0.14049587 0.36363636]

[0.25233645 0.47663551 0.27102804]

[0.27653631 0.14804469 0.57541899]]





precision recall f1-score support

hate 0.32 0.50 0.39 121

offensive 0.42 0.48 0.45 107

neither 0.74 0.58 0.65 358

accuracy 0.54 586

macro avg 0.49 0.52 0.50 586

weighted avg 0.59 0.54 0.56 586

average eval\_loss: 0.9330986728270848, accuracy: 0.5409556313993175

training 4 epoch...

195.7139744758606 seconds used......

4 training finished! average train loss: 0.9025989610113596

total clf loss: 81.72667497396469 total adv loss: 27.734183401276823

evaluating...

Confusion matrix, without normalization

[[ 63 28 30]

[ 20 76 11]

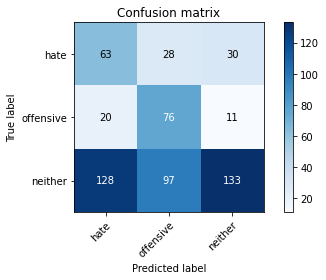
[128 97 133]]

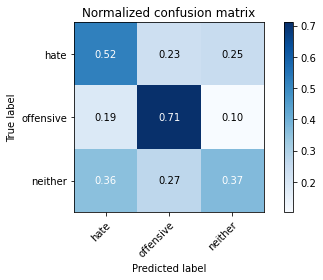
Normalized confusion matrix

[[0.52066116 0.23140496 0.24793388]

[0.18691589 0.71028037 0.10280374]

[0.3575419 0.27094972 0.37150838]]





precision recall f1-score support

hate 0.30 0.52 0.38 121

offensive 0.38 0.71 0.49 107

neither 0.76 0.37 0.50 358

accuracy 0.46 586

macro avg 0.48 0.53 0.46 586

weighted avg 0.60 0.46 0.47 586

average eval\_loss: 1.3106244683265686, accuracy: 0.4641638225255973

training 5 epoch...

196.13809323310852 seconds used......

5 training finished! average train loss: 0.5349977854639292

total clf loss: 58.68038946390152 total adv loss: 11.319637115506339

evaluating...

Confusion matrix, without normalization

[[ 74 17 30]

[ 32 59 16]

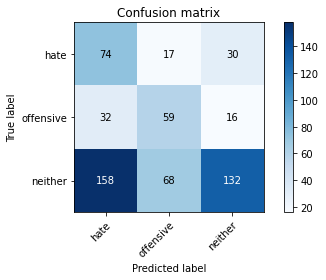
[158 68 132]]

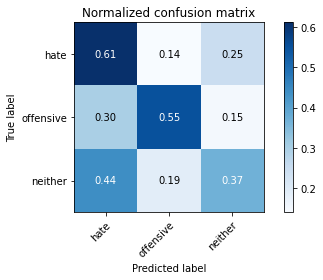
Normalized confusion matrix

[[0.61157025 0.14049587 0.24793388]

[0.29906542 0.55140187 0.14953271]

[0.44134078 0.18994413 0.36871508]]





precision recall f1-score support

hate 0.28 0.61 0.38 121

offensive 0.41 0.55 0.47 107

neither 0.74 0.37 0.49 358

accuracy 0.45 586

macro avg 0.48 0.51 0.45 586

weighted avg 0.59 0.45 0.47 586

average eval\_loss: 1.6675564964612326, accuracy: 0.4522184300341297

model of 5 epoch

Confusion matrix, without normalization

[[ 96 18 10]

[ 64 90 10]

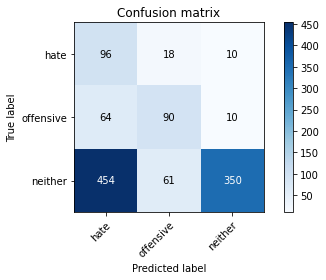
[454 61 350]]

Normalized confusion matrix

[[0.77419355 0.14516129 0.08064516]

[0.3902439 0.54878049 0.06097561]

[0.52485549 0.07052023 0.40462428]]





precision recall f1-score support

hate 0.16 0.77 0.26 124

offensive 0.53 0.55 0.54 164

neither 0.95 0.40 0.57 865

accuracy 0.46 1153

macro avg 0.54 0.58 0.46 1153

weighted avg 0.80 0.46 0.53 1153

average eval\_loss: 1.4547050400026913, accuracy: 0.46487424111014747

2 run

training 1 epoch...

105.6728150844574 seconds used......

1 training finished! average train loss: 2.0833970982777443

total clf loss: 151.62452179193497 total adv loss: 82.52591872215271

evaluating...

Confusion matrix, without normalization

[[ 55 49 17]

[ 12 90 5]

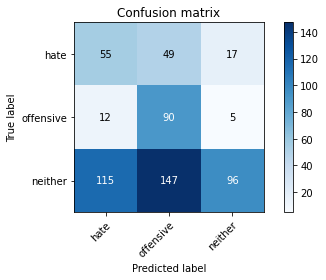
[115 147 96]]

Normalized confusion matrix

[[0.45454545 0.40495868 0.14049587]

[0.11214953 0.8411215 0.04672897]

[0.32122905 0.41061453 0.26815642]]





precision recall f1-score support

hate 0.30 0.45 0.36 121

offensive 0.31 0.84 0.46 107

neither 0.81 0.27 0.40 358

accuracy 0.41 586

macro avg 0.48 0.52 0.41 586

weighted avg 0.62 0.41 0.41 586

average eval\_loss: 1.058538939555486, accuracy: 0.4112627986348123

training 2 epoch...

105.56983613967896 seconds used......

2 training finished! average train loss: 1.7348898766856444

total clf loss: 130.58655893802643 total adv loss: 66.55835101008415

evaluating...

Confusion matrix, without normalization

[[ 62 36 23]

[ 12 80 15]

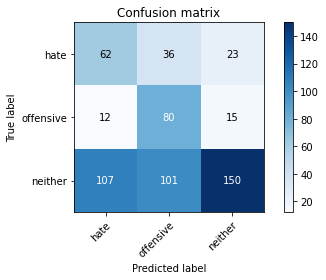
[107 101 150]]

Normalized confusion matrix

[[0.51239669 0.29752066 0.19008264]

[0.11214953 0.74766355 0.14018692]

[0.29888268 0.28212291 0.41899441]]





precision recall f1-score support

hate 0.34 0.51 0.41 121

offensive 0.37 0.75 0.49 107

neither 0.80 0.42 0.55 358

accuracy 0.50 586

macro avg 0.50 0.56 0.48 586

weighted avg 0.63 0.50 0.51 586

average eval\_loss: 0.9507249852021535, accuracy: 0.49829351535836175

training 3 epoch...

105.60153532028198 seconds used......

3 training finished! average train loss: 1.4099491458190114

total clf loss: 112.28856194019318 total adv loss: 51.01185445487499

evaluating...

Confusion matrix, without normalization

[[ 66 21 34]

[ 20 66 21]

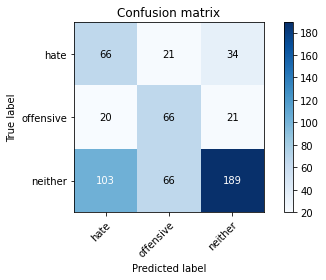
[103 66 189]]

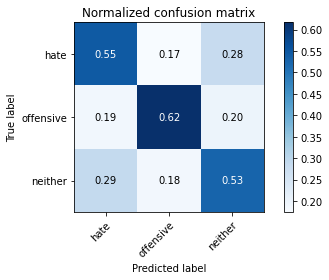
Normalized confusion matrix

[[0.54545455 0.17355372 0.28099174]

[0.18691589 0.61682243 0.19626168]

[0.2877095 0.18435754 0.52793296]]





precision recall f1-score support

hate 0.35 0.55 0.43 121

offensive 0.43 0.62 0.51 107

neither 0.77 0.53 0.63 358

accuracy 0.55 586

macro avg 0.52 0.56 0.52 586

weighted avg 0.62 0.55 0.56 586

average eval\_loss: 0.933578469355901, accuracy: 0.5477815699658704

training 4 epoch...

105.58109736442566 seconds used......

4 training finished! average train loss: 0.9384764478002724

total clf loss: 84.7720365524292 total adv loss: 28.93819185346365

evaluating...

Confusion matrix, without normalization

[[ 78 21 22]

[ 28 65 14]

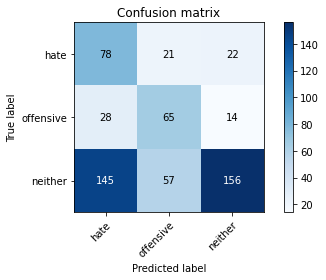
[145 57 156]]

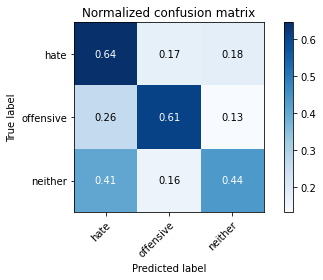
Normalized confusion matrix

[[0.6446281 0.17355372 0.18181818]

[0.26168224 0.60747664 0.13084112]

[0.40502793 0.15921788 0.43575419]]





precision recall f1-score support

hate 0.31 0.64 0.42 121

offensive 0.45 0.61 0.52 107

neither 0.81 0.44 0.57 358

accuracy 0.51 586

macro avg 0.53 0.56 0.50 586

weighted avg 0.64 0.51 0.53 586

average eval\_loss: 1.142509259780248, accuracy: 0.5102389078498294

training 5 epoch...

105.6093418598175 seconds used......

5 training finished! average train loss: 0.6146036221792823

total clf loss: 62.0295822173357 total adv loss: 15.695083950093249

evaluating...

Confusion matrix, without normalization

[[ 66 16 39]

[ 31 54 22]

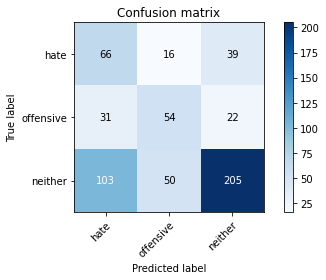
[103 50 205]]

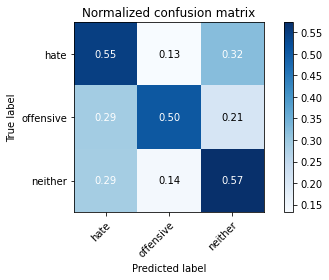
Normalized confusion matrix

[[0.54545455 0.1322314 0.32231405]

[0.28971963 0.5046729 0.20560748]

[0.2877095 0.1396648 0.5726257 ]]





precision recall f1-score support

hate 0.33 0.55 0.41 121

offensive 0.45 0.50 0.48 107

neither 0.77 0.57 0.66 358

accuracy 0.55 586

macro avg 0.52 0.54 0.51 586

weighted avg 0.62 0.55 0.57 586

average eval\_loss: 1.2731338063875834, accuracy: 0.5546075085324232

Confusion matrix, without normalization

[[ 91 24 9]

[ 46 111 7]

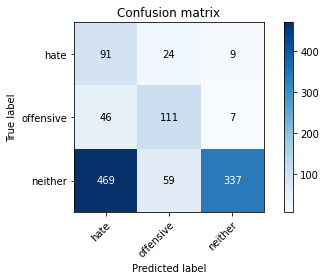
[469 59 337]]

Normalized confusion matrix

[[0.73387097 0.19354839 0.07258065]

[0.2804878 0.67682927 0.04268293]

[0.54219653 0.06820809 0.38959538]]





precision recall f1-score support

hate 0.15 0.73 0.25 124

offensive 0.57 0.68 0.62 164

neither 0.95 0.39 0.55 865

accuracy 0.47 1153

macro avg 0.56 0.60 0.47 1153

weighted avg 0.81 0.47 0.53 1153

average eval\_loss: 1.0807580279892888, accuracy: 0.46747614917606245

[23]:

0.46747614917606245

3 run

training 1 epoch...

105.61793303489685 seconds used......

1 training finished! average train loss: 2.086354234500935

total clf loss: 152.34310364723206 total adv loss: 82.39137071371078

evaluating...

Confusion matrix, without normalization

[[ 42 22 57]

[ 22 65 20]

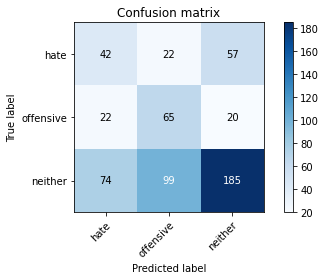
[ 74 99 185]]

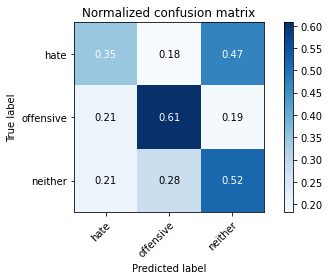
Normalized confusion matrix

[[0.34710744 0.18181818 0.47107438]

[0.20560748 0.60747664 0.18691589]

[0.20670391 0.27653631 0.51675978]]





precision recall f1-score support

hate 0.30 0.35 0.32 121

offensive 0.35 0.61 0.44 107

neither 0.71 0.52 0.60 358

accuracy 0.50 586

macro avg 0.45 0.49 0.45 586

weighted avg 0.56 0.50 0.51 586

average eval\_loss: 0.9238885581493378, accuracy: 0.49829351535836175

training 2 epoch...

105.67695379257202 seconds used......

2 training finished! average train loss: 1.7146994193133556

total clf loss: 129.6832014322281 total adv loss: 65.47555553913116

evaluating...

Confusion matrix, without normalization

[[ 68 20 33]

[ 37 59 11]

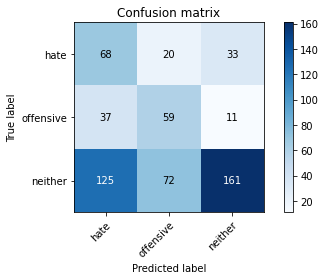
[125 72 161]]

Normalized confusion matrix

[[0.56198347 0.16528926 0.27272727]

[0.34579439 0.55140187 0.10280374]

[0.34916201 0.20111732 0.44972067]]





precision recall f1-score support

hate 0.30 0.56 0.39 121

offensive 0.39 0.55 0.46 107

neither 0.79 0.45 0.57 358

accuracy 0.49 586

macro avg 0.49 0.52 0.47 586

weighted avg 0.61 0.49 0.51 586

average eval\_loss: 0.9242610851923625, accuracy: 0.49146757679180886

training 3 epoch...

105.56494975090027 seconds used......

3 training finished! average train loss: 1.2922403514385223

total clf loss: 105.95527812838554 total adv loss: 45.2326273098588

evaluating...

Confusion matrix, without normalization

[[ 67 22 32]

[ 27 68 12]

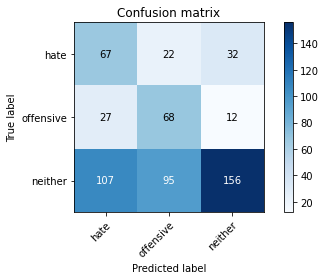
[107 95 156]]

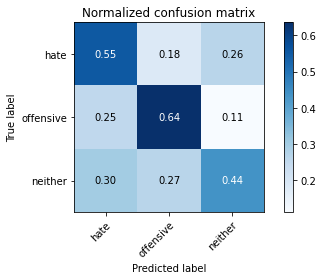
Normalized confusion matrix

[[0.55371901 0.18181818 0.26446281]

[0.25233645 0.63551402 0.11214953]

[0.29888268 0.26536313 0.43575419]]





precision recall f1-score support

hate 0.33 0.55 0.42 121

offensive 0.37 0.64 0.47 107

neither 0.78 0.44 0.56 358

accuracy 0.50 586

macro avg 0.49 0.54 0.48 586

weighted avg 0.61 0.50 0.51 586

average eval\_loss: 1.0389032145341237, accuracy: 0.49658703071672355

training 4 epoch...

105.59541392326355 seconds used......

4 training finished! average train loss: 0.80481604663165

total clf loss: 77.5672737956047 total adv loss: 22.382382664596662

evaluating...

Confusion matrix, without normalization

[[ 71 19 31]

[ 29 60 18]

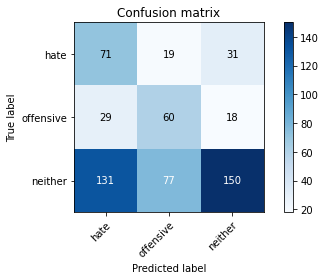
[131 77 150]]

Normalized confusion matrix

[[0.58677686 0.15702479 0.25619835]

[0.27102804 0.56074766 0.1682243 ]

[0.36592179 0.2150838 0.41899441]]





precision recall f1-score support

hate 0.31 0.59 0.40 121

offensive 0.38 0.56 0.46 107

neither 0.75 0.42 0.54 358

accuracy 0.48 586

macro avg 0.48 0.52 0.47 586

weighted avg 0.59 0.48 0.50 586

average eval\_loss: 1.147988780339559, accuracy: 0.47952218430034127

training 5 epoch...

105.63885831832886 seconds used......

5 training finished! average train loss: 0.5651685641588349

total clf loss: 59.654458209872246 total adv loss: 13.125581867061555

evaluating...

Confusion matrix, without normalization

[[ 52 23 46]

[ 18 70 19]

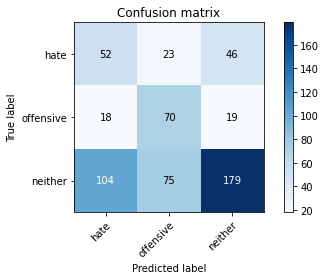
[104 75 179]]

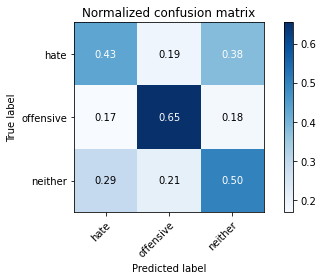
Normalized confusion matrix

[[0.42975207 0.19008264 0.38016529]

[0.1682243 0.65420561 0.17757009]

[0.29050279 0.20949721 0.5 ]]





precision recall f1-score support

hate 0.30 0.43 0.35 121

offensive 0.42 0.65 0.51 107

neither 0.73 0.50 0.59 358

accuracy 0.51 586

macro avg 0.48 0.53 0.49 586

weighted avg 0.59 0.51 0.53 586

average eval\_loss: 1.329411248366038, accuracy: 0.5136518771331058

Confusion matrix, without normalization

[[ 83 26 15]

[ 42 110 12]

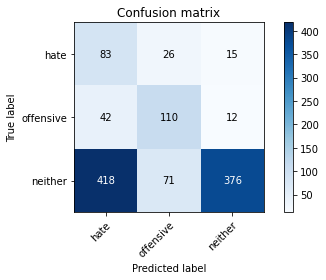
[418 71 376]]

Normalized confusion matrix

[[0.66935484 0.20967742 0.12096774]

[0.25609756 0.67073171 0.07317073]

[0.48323699 0.08208092 0.43468208]]





precision recall f1-score support

hate 0.15 0.67 0.25 124

offensive 0.53 0.67 0.59 164

neither 0.93 0.43 0.59 865

accuracy 0.49 1153

macro avg 0.54 0.59 0.48 1153

weighted avg 0.79 0.49 0.56 1153

average eval\_loss: 1.0113835591694404, accuracy: 0.4934952298352125

[25]:

0.4934952298352125