1 run

hate as no-offensive

alpha=2

training 1 epoch...

52.24550533294678 seconds used......

1 training finished! average train loss: 2.210364180141025

total clf loss: 93.78907692432404 total adv loss: 52.57185012102127

evaluating...

Confusion matrix, without normalization

[[ 55 18 17]

[ 22 37 8]

[ 69 72 203]]

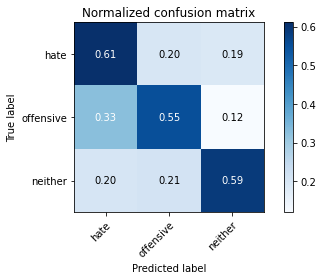
Normalized confusion matrix

[[0.61111111 0.2 0.18888889]

[0.32835821 0.55223881 0.11940299]

[0.2005814 0.20930233 0.59011628]]





precision recall f1-score support

hate 0.38 0.61 0.47 90

offensive 0.29 0.55 0.38 67

neither 0.89 0.59 0.71 344

accuracy 0.59 501

macro avg 0.52 0.58 0.52 501

weighted avg 0.72 0.59 0.62 501

average eval\_loss: 0.9307772540129148, accuracy: 0.5888223552894212

training 2 epoch...

50.65625739097595 seconds used......

2 training finished! average train loss: 1.6303175780508252

total clf loss: 71.07245767116547 total adv loss: 37.828061774373055

evaluating...

Confusion matrix, without normalization

[[ 50 17 23]

[ 17 41 9]

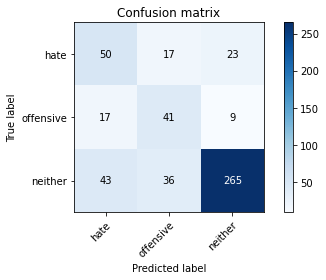
[ 43 36 265]]

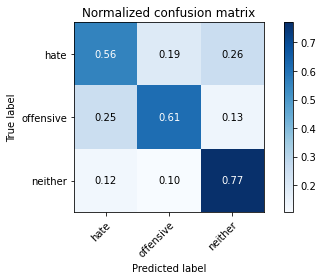
Normalized confusion matrix

[[0.55555556 0.18888889 0.25555556]

[0.25373134 0.6119403 0.13432836]

[0.125 0.10465116 0.77034884]]





precision recall f1-score support

hate 0.45 0.56 0.50 90

offensive 0.44 0.61 0.51 67

neither 0.89 0.77 0.83 344

accuracy 0.71 501

macro avg 0.59 0.65 0.61 501

weighted avg 0.75 0.71 0.73 501

average eval\_loss: 0.695737150999216, accuracy: 0.7105788423153693

training 3 epoch...

51.06855225563049 seconds used......

3 training finished! average train loss: 1.0388903061548869

total clf loss: 49.35186918079853 total adv loss: 22.07412936910987

evaluating...

Confusion matrix, without normalization

[[ 54 18 18]

[ 16 42 9]

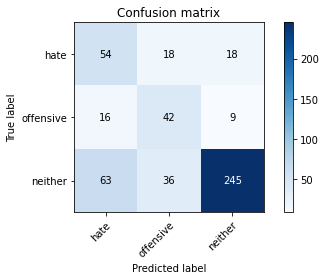
[ 63 36 245]]

Normalized confusion matrix

[[0.6 0.2 0.2 ]

[0.23880597 0.62686567 0.13432836]

[0.18313953 0.10465116 0.7122093 ]]





precision recall f1-score support

hate 0.41 0.60 0.48 90

offensive 0.44 0.63 0.52 67

neither 0.90 0.71 0.80 344

accuracy 0.68 501

macro avg 0.58 0.65 0.60 501

weighted avg 0.75 0.68 0.70 501

average eval\_loss: 0.771972154195492, accuracy: 0.6806387225548902

training 4 epoch...

51.129534006118774 seconds used......

4 training finished! average train loss: 0.5874094247817994

total clf loss: 32.77831430733204 total adv loss: 10.044266858836636

evaluating...

Confusion matrix, without normalization

[[ 56 17 17]

[ 20 38 9]

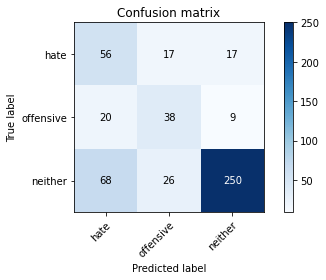
[ 68 26 250]]

Normalized confusion matrix

[[0.62222222 0.18888889 0.18888889]

[0.29850746 0.56716418 0.13432836]

[0.19767442 0.0755814 0.72674419]]





precision recall f1-score support

hate 0.39 0.62 0.48 90

offensive 0.47 0.57 0.51 67

neither 0.91 0.73 0.81 344

accuracy 0.69 501

macro avg 0.59 0.64 0.60 501

weighted avg 0.75 0.69 0.71 501

average eval\_loss: 0.7684619415264863, accuracy: 0.6866267465069861

training 5 epoch...

51.07679200172424 seconds used......

5 training finished! average train loss: 0.275903914620479

total clf loss: 21.09832403808832 total adv loss: 1.8665141132951248

evaluating...

Confusion matrix, without normalization

[[ 47 24 19]

[ 12 47 8]

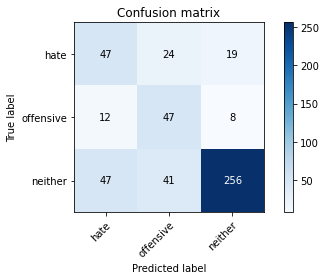
[ 47 41 256]]

Normalized confusion matrix

[[0.52222222 0.26666667 0.21111111]

[0.17910448 0.70149254 0.11940299]

[0.13662791 0.11918605 0.74418605]]





precision recall f1-score support

hate 0.44 0.52 0.48 90

offensive 0.42 0.70 0.53 67

neither 0.90 0.74 0.82 344

accuracy 0.70 501

macro avg 0.59 0.66 0.61 501

weighted avg 0.76 0.70 0.72 501

average eval\_loss: 0.8315352568259606, accuracy: 0.6986027944111777

model of 4 epoch

Confusion matrix, without normalization

[[ 428 119 226]

[ 111 199 120]

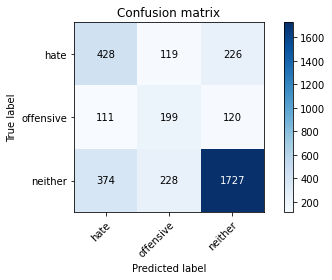
[ 374 228 1727]]

Normalized confusion matrix

[[0.55368693 0.15394567 0.2923674 ]

[0.25813953 0.4627907 0.27906977]

[0.16058394 0.09789609 0.74151997]]





precision recall f1-score support

hate 0.47 0.55 0.51 773

offensive 0.36 0.46 0.41 430

neither 0.83 0.74 0.78 2329

accuracy 0.67 3532

macro avg 0.56 0.59 0.57 3532

weighted avg 0.70 0.67 0.68 3532

average eval\_loss: 0.813398297560417, accuracy: 0.6664779161947905

2 run

training 1 epoch...

25.100884199142456 seconds used......

1 training finished! average train loss: 2.225915727350447

total clf loss: 95.16163659095764 total adv loss: 52.58539018034935

evaluating...

Confusion matrix, without normalization

[[ 44 17 29]

[ 20 25 22]

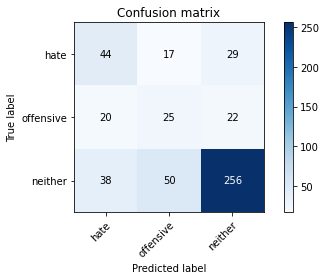
[ 38 50 256]]

Normalized confusion matrix

[[0.48888889 0.18888889 0.32222222]

[0.29850746 0.37313433 0.32835821]

[0.11046512 0.14534884 0.74418605]]





precision recall f1-score support

hate 0.43 0.49 0.46 90

offensive 0.27 0.37 0.31 67

neither 0.83 0.74 0.79 344

accuracy 0.65 501

macro avg 0.51 0.54 0.52 501

weighted avg 0.69 0.65 0.66 501

average eval\_loss: 0.9102282432409433, accuracy: 0.6487025948103793

training 2 epoch...

25.160584926605225 seconds used......

2 training finished! average train loss: 1.7065812322828504

total clf loss: 73.17076855897903 total adv loss: 40.210771560668945

evaluating...

Confusion matrix, without normalization

[[ 64 16 10]

[ 27 35 5]

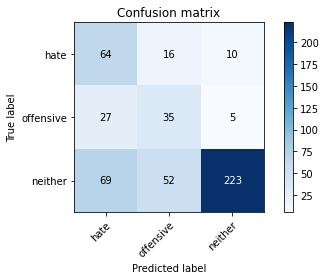
[ 69 52 223]]

Normalized confusion matrix

[[0.71111111 0.17777778 0.11111111]

[0.40298507 0.52238806 0.07462687]

[0.2005814 0.15116279 0.64825581]]





precision recall f1-score support

hate 0.40 0.71 0.51 90

offensive 0.34 0.52 0.41 67

neither 0.94 0.65 0.77 344

accuracy 0.64 501

macro avg 0.56 0.63 0.56 501

weighted avg 0.76 0.64 0.67 501

average eval\_loss: 0.8492832458936251, accuracy: 0.6427145708582834

training 3 epoch...

25.129168033599854 seconds used......

3 training finished! average train loss: 1.0689986228942872

total clf loss: 50.18456055223942 total adv loss: 23.01265785098076

evaluating...

Confusion matrix, without normalization

[[ 56 13 21]

[ 17 38 12]

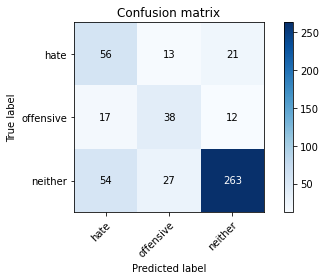
[ 54 27 263]]

Normalized confusion matrix

[[0.62222222 0.14444444 0.23333333]

[0.25373134 0.56716418 0.17910448]

[0.15697674 0.07848837 0.76453488]]





precision recall f1-score support

hate 0.44 0.62 0.52 90

offensive 0.49 0.57 0.52 67

neither 0.89 0.76 0.82 344

accuracy 0.71 501

macro avg 0.61 0.65 0.62 501

weighted avg 0.75 0.71 0.73 501

average eval\_loss: 0.6579433011894043, accuracy: 0.7125748502994012

training 4 epoch...

25.113286018371582 seconds used......

4 training finished! average train loss: 0.5249730398257574

total clf loss: 31.82028552889824 total adv loss: 7.713643947383389

evaluating...

Confusion matrix, without normalization

[[ 47 24 19]

[ 11 46 10]

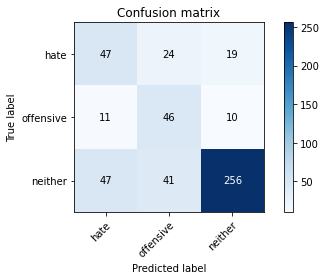
[ 47 41 256]]

Normalized confusion matrix

[[0.52222222 0.26666667 0.21111111]

[0.1641791 0.68656716 0.14925373]

[0.13662791 0.11918605 0.74418605]]





precision recall f1-score support

hate 0.45 0.52 0.48 90

offensive 0.41 0.69 0.52 67

neither 0.90 0.74 0.81 344

accuracy 0.70 501

macro avg 0.59 0.65 0.60 501

weighted avg 0.75 0.70 0.71 501

average eval\_loss: 0.7427153724890488, accuracy: 0.6966067864271457

training 5 epoch...

25.26009964942932 seconds used......

5 training finished! average train loss: 0.30008782520890237

total clf loss: 21.560032904148102 total adv loss: 2.723935614543734

evaluating...

Confusion matrix, without normalization

[[ 63 17 10]

[ 24 38 5]

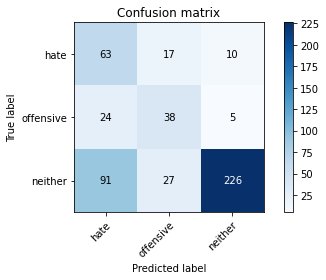
[ 91 27 226]]

Normalized confusion matrix

[[0.7 0.18888889 0.11111111]

[0.35820896 0.56716418 0.07462687]

[0.26453488 0.07848837 0.65697674]]





precision recall f1-score support

hate 0.35 0.70 0.47 90

offensive 0.46 0.57 0.51 67

neither 0.94 0.66 0.77 344

accuracy 0.65 501

macro avg 0.59 0.64 0.58 501

weighted avg 0.77 0.65 0.68 501

average eval\_loss: 0.8953264195185441, accuracy: 0.6526946107784432

Confusion matrix, without normalization

[[ 486 129 158]

[ 141 205 84]

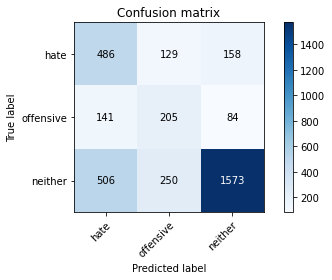
[ 506 250 1573]]

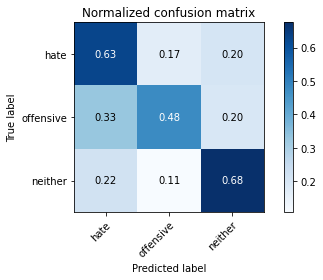
Normalized confusion matrix

[[0.62871928 0.16688228 0.20439845]

[0.32790698 0.47674419 0.19534884]

[0.21726063 0.10734221 0.67539717]]





precision recall f1-score support

hate 0.43 0.63 0.51 773

offensive 0.35 0.48 0.40 430

neither 0.87 0.68 0.76 2329

accuracy 0.64 3532

macro avg 0.55 0.59 0.56 3532

weighted avg 0.71 0.64 0.66 3532

average eval\_loss: 0.9555819628939117, accuracy: 0.6409966024915063

[29]:

3 run

training 1 epoch...

25.17029309272766 seconds used......

1 training finished! average train loss: 2.2275289522276984

total clf loss: 93.29429984092712 total adv loss: 53.591652899980545

evaluating...

Confusion matrix, without normalization

[[ 48 17 25]

[ 19 33 15]

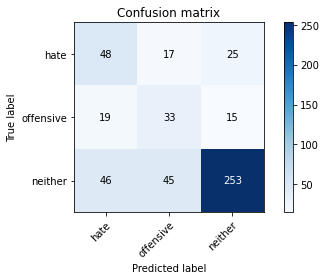
[ 46 45 253]]

Normalized confusion matrix

[[0.53333333 0.18888889 0.27777778]

[0.28358209 0.49253731 0.2238806 ]

[0.13372093 0.13081395 0.73546512]]





precision recall f1-score support

hate 0.42 0.53 0.47 90

offensive 0.35 0.49 0.41 67

neither 0.86 0.74 0.79 344

accuracy 0.67 501

macro avg 0.55 0.59 0.56 501

weighted avg 0.72 0.67 0.68 501

average eval\_loss: 0.8622129559516907, accuracy: 0.6666666666666666

training 2 epoch...

25.222296476364136 seconds used......

2 training finished! average train loss: 1.6048082027170394

total clf loss: 69.28256765007973 total adv loss: 37.575085535645485

evaluating...

Confusion matrix, without normalization

[[ 48 18 24]

[ 13 39 15]

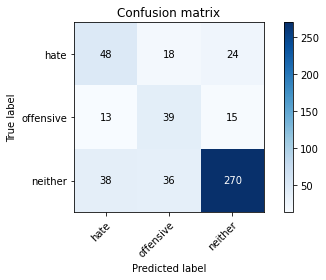
[ 38 36 270]]

Normalized confusion matrix

[[0.53333333 0.2 0.26666667]

[0.19402985 0.58208955 0.2238806 ]

[0.11046512 0.10465116 0.78488372]]





precision recall f1-score support

hate 0.48 0.53 0.51 90

offensive 0.42 0.58 0.49 67

neither 0.87 0.78 0.83 344

accuracy 0.71 501

macro avg 0.59 0.63 0.61 501

weighted avg 0.74 0.71 0.72 501

average eval\_loss: 0.674267685184112, accuracy: 0.7125748502994012

training 3 epoch...

25.08798623085022 seconds used......

3 training finished! average train loss: 0.961686780055364

total clf loss: 47.38359108567238 total adv loss: 19.58410974126309

evaluating...

Confusion matrix, without normalization

[[ 49 27 14]

[ 13 47 7]

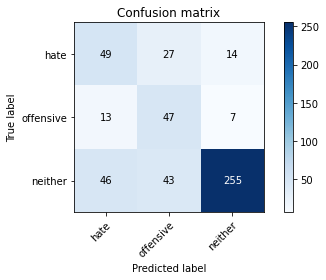
[ 46 43 255]]

Normalized confusion matrix

[[0.54444444 0.3 0.15555556]

[0.19402985 0.70149254 0.10447761]

[0.13372093 0.125 0.74127907]]





precision recall f1-score support

hate 0.45 0.54 0.49 90

offensive 0.40 0.70 0.51 67

neither 0.92 0.74 0.82 344

accuracy 0.70 501

macro avg 0.59 0.66 0.61 501

weighted avg 0.77 0.70 0.72 501

average eval\_loss: 0.7533414077300292, accuracy: 0.7005988023952096

training 4 epoch...

25.13395071029663 seconds used......

4 training finished! average train loss: 0.4880644153389666

total clf loss: 30.090030923485756 total adv loss: 6.917883139103651

evaluating...

Confusion matrix, without normalization

[[ 62 18 10]

[ 19 42 6]

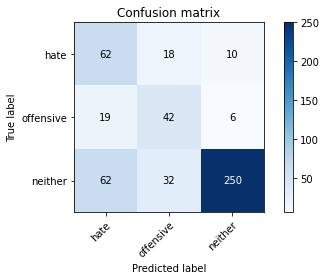
[ 62 32 250]]

Normalized confusion matrix

[[0.68888889 0.2 0.11111111]

[0.28358209 0.62686567 0.08955224]

[0.18023256 0.09302326 0.72674419]]





precision recall f1-score support

hate 0.43 0.69 0.53 90

offensive 0.46 0.63 0.53 67

neither 0.94 0.73 0.82 344

accuracy 0.71 501

macro avg 0.61 0.68 0.63 501

weighted avg 0.78 0.71 0.73 501

average eval\_loss: 0.7494014682104955, accuracy: 0.7065868263473054

training 5 epoch...

25.123844146728516 seconds used......

5 training finished! average train loss: 0.3266397119810184

total clf loss: 21.342507753521204 total adv loss: 4.02753306995146

evaluating...

Confusion matrix, without normalization

[[ 59 14 17]

[ 21 37 9]

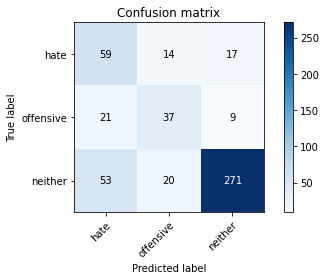
[ 53 20 271]]

Normalized confusion matrix

[[0.65555556 0.15555556 0.18888889]

[0.31343284 0.55223881 0.13432836]

[0.15406977 0.05813953 0.7877907 ]]





precision recall f1-score support

hate 0.44 0.66 0.53 90

offensive 0.52 0.55 0.54 67

neither 0.91 0.79 0.85 344

accuracy 0.73 501

macro avg 0.63 0.67 0.64 501

weighted avg 0.78 0.73 0.75 501

average eval\_loss: 0.8219390098865216, accuracy: 0.7325349301397206

Confusion matrix, without normalization

[[ 387 148 238]

[ 93 216 121]

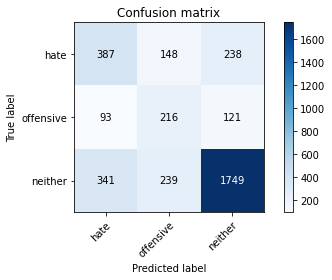
[ 341 239 1749]]

Normalized confusion matrix

[[0.50064683 0.19146184 0.30789133]

[0.21627907 0.50232558 0.28139535]

[0.14641477 0.10261915 0.75096608]]





precision recall f1-score support

hate 0.47 0.50 0.49 773

offensive 0.36 0.50 0.42 430

neither 0.83 0.75 0.79 2329

accuracy 0.67 3532

macro avg 0.55 0.58 0.56 3532

weighted avg 0.69 0.67 0.68 3532

average eval\_loss: 0.8569261945236875, accuracy: 0.665911664779162

[32]:

0.665911664779162