hate as offensive

alpha=0.05

1 run...

training 1 epoch...

289.5330562591553 seconds used......

1 training finished! average train loss: 0.871721005640771

total clf loss: 1234.2966566681862 total adv loss: 905.3841196596622

evaluating...

Confusion matrix, without normalization

[[ 0 160 862]

[ 0 1132 4275]

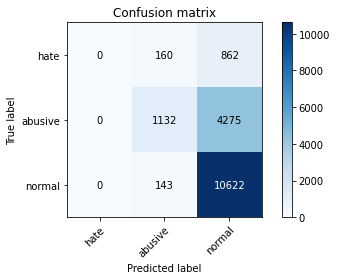
[ 0 143 10622]]

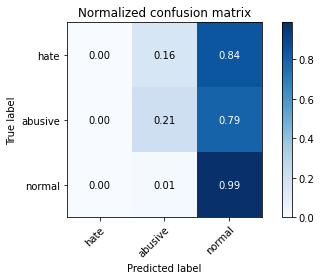
Normalized confusion matrix

[[0. 0.15655577 0.84344423]

[0. 0.20935824 0.79064176]

[0. 0.01328379 0.98671621]]





/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

precision recall f1-score support

hate 0.00 0.00 0.00 1022

abusive 0.79 0.21 0.33 5407

normal 0.67 0.99 0.80 10765

accuracy 0.68 17194

macro avg 0.49 0.40 0.38 17194

weighted avg 0.67 0.68 0.61 17194

average eval\_loss: 0.7654597707959109, accuracy: 0.683610561823892

training 2 epoch...

289.5554177761078 seconds used......

2 training finished! average train loss: 0.7154024191899908

total clf loss: 1006.712755933404 total adv loss: 618.0771241337061

evaluating...

Confusion matrix, without normalization

[[ 0 445 577]

[ 0 3731 1676]

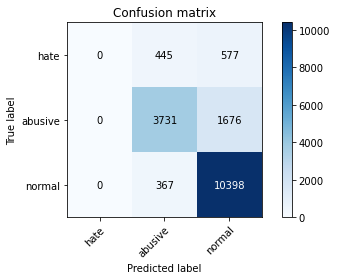
[ 0 367 10398]]

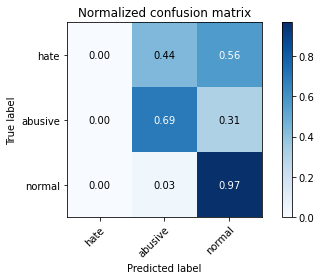
Normalized confusion matrix

[[0. 0.43542074 0.56457926]

[0. 0.69003144 0.30996856]

[0. 0.03409196 0.96590804]]





/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

precision recall f1-score support

hate 0.00 0.00 0.00 1022

abusive 0.82 0.69 0.75 5407

normal 0.82 0.97 0.89 10765

accuracy 0.82 17194

macro avg 0.55 0.55 0.55 17194

weighted avg 0.77 0.82 0.79 17194

average eval\_loss: 0.5640160883062107, accuracy: 0.8217401419099686

training 3 epoch...

289.65828824043274 seconds used......

3 training finished! average train loss: 0.6090066109693819

total clf loss: 856.9667772203684 total adv loss: 525.6351921260357

evaluating...

Confusion matrix, without normalization

[[ 0 524 498]

[ 0 4438 969]

[ 0 612 10153]]

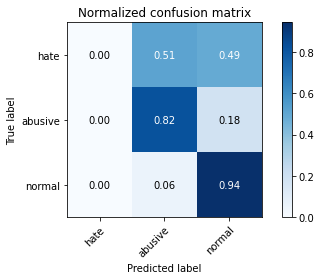
Normalized confusion matrix

[[0. 0.51272016 0.48727984]

[0. 0.82078787 0.17921213]

[0. 0.05685091 0.94314909]]





/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

precision recall f1-score support

hate 0.00 0.00 0.00 1022

abusive 0.80 0.82 0.81 5407

normal 0.87 0.94 0.91 10765

accuracy 0.85 17194

macro avg 0.56 0.59 0.57 17194

weighted avg 0.80 0.85 0.82 17194

average eval\_loss: 0.5358290240168572, accuracy: 0.8486099802256601

training 4 epoch...

289.70270705223083 seconds used......

4 training finished! average train loss: 0.5541439097582245

total clf loss: 779.4668437764049 total adv loss: 472.2910287603736

evaluating...

Confusion matrix, without normalization

[[ 0 484 538]

[ 0 4330 1077]

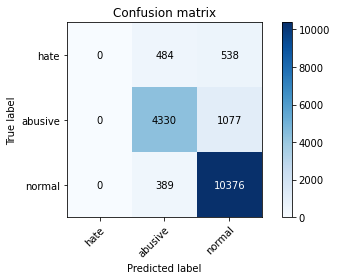
[ 0 389 10376]]

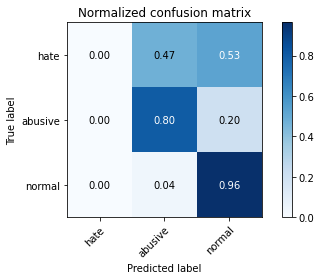
Normalized confusion matrix

[[0. 0.47358121 0.52641879]

[0. 0.80081376 0.19918624]

[0. 0.03613562 0.96386438]]





/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

precision recall f1-score support

hate 0.00 0.00 0.00 1022

abusive 0.83 0.80 0.82 5407

normal 0.87 0.96 0.91 10765

accuracy 0.86 17194

macro avg 0.57 0.59 0.58 17194

weighted avg 0.80 0.86 0.83 17194

average eval\_loss: 0.4762173557350802, accuracy: 0.855298359892986

training 5 epoch...

289.5990149974823 seconds used......

5 training finished! average train loss: 0.5159597052507037

total clf loss: 725.5669226646423 total adv loss: 435.95768519863486

evaluating...

Confusion matrix, without normalization

[[ 0 583 439]

[ 0 4777 630]

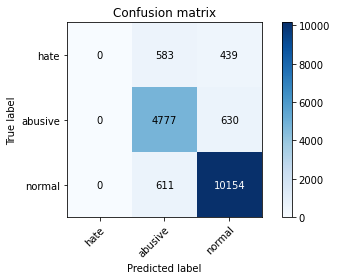
[ 0 611 10154]]

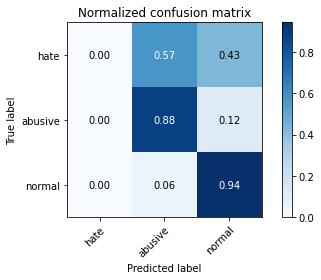
Normalized confusion matrix

[[0. 0.5704501 0.4295499 ]

[0. 0.88348437 0.11651563]

[0. 0.05675801 0.94324199]]





/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/opt/conda/lib/python3.7/site-packages/sklearn/metrics/\_classification.py:1245: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

precision recall f1-score support

hate 0.00 0.00 0.00 1022

abusive 0.80 0.88 0.84 5407

normal 0.90 0.94 0.92 10765

accuracy 0.87 17194

macro avg 0.57 0.61 0.59 17194

weighted avg 0.82 0.87 0.84 17194

average eval\_loss: 0.4823178365653337, accuracy: 0.8683843201116669

2 run...

training 1 epoch...