**Project: Toxic comment classification**

**Task**

You are provided with a large number of comments which have been labeled by human raters for toxic behavior. The types of toxicity are:

* toxic
* severe\_toxic
* obscene
* threat
* insult
* identity\_hate

You must create a model which predicts a probability of each type of toxicity for each comment.

**File descriptions**

* **train.csv** - the training set, contains comments with their binary labels
* **test.csv** - the test set, you must predict the toxicity probabilities for these comments. To deter hand labeling, the test set contains some comments which are not included in scoring.
* **sample\_submission.csv** - a sample submission file in the correct format

**Submission File**

For each id in the test set, you must predict a probability for each of the six possible types of comment toxicity (toxic, severe\_toxic, obscene, threat, insult, identity\_hate). The columns must be in the same order as shown below. The file should contain a header and have the following format:

id,toxic,severe\_toxic,obscene,threat,insult,identity\_hate

00001cee341fdb12,0.5,0.5,0.5,0.5,0.5,0.5

0000247867823ef7,0.5,0.5,0.5,0.5,0.5,0.5

etc.

**Evaluation**

Submissions are evaluated on the mean column-wise ROC AUC. In other words, the score is the average of the individual AUCs of each predicted column.