**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

For the classification task, you need to predict a binary label (0 or 1) for each of the six possible types of comment toxicity (toxic, severe\_toxic, obscene, threat, insult, identity\_hate) for every id in the test set.

**File descriptions**

* **train.csv** - the training set, contains comments with their binary labels
* **test.csv** - the test set, you must predict the toxicity for these comments.
* **sample\_submission.csv** - a sample submission file in the correct format

**Submission File**  
The submission file must include a header and follow the format below, with the columns in the specified order:

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

00001cee341fdb12,1,0,0,1,0,0

0000247867823ef7,1,0,1,0,1,1

00013b17ad220c46,0,0,1,0,0,0

00017563c3f7919a,1,1,1,0,0,0

00017695ad8997eb,1,0,0,1,0,1

…

**Evaluation**

The evaluation metric will be the average f1\_score over all toxicity categories

with average='macro'. Your model should therefore be good in detecting all toxicitys. Here is the evaluation code example:

f1\_scores = {}

for label in label\_columns:

f1\_scores[label] = f1\_score(

merged\_df[f"{label}\_pred"],

merged\_df[f"{label}\_true"],

average='macro')

# Compute the average F1 score across all labels

avg\_f1\_score = sum(f1\_scores.values()) / len(f1\_scores)

st.write(f"Average F1 Score across all labels: \*\*{avg\_f1\_score \* 100:.2f}%\*\*")