Assignment A2 : Evaluation and Error Analysis

Individual assignment • 20% individual-report component Released: 26 Nov 2026 Due: 3 Dec 2026, 23:59

In class we saw that a single accuracy number can hide what a model actually does — especially who it works for and who it fails. In this assignment you will **evaluate three toxicity classifiers properly** and **find where and for whom they break**. The question to answer: *which model would you deploy, and which group does it let down?*

Data. We use the comment text from the Kaggle competition *Jigsaw Unintended Bias in Toxicity Classification* (`https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification`). Load it with `load_textclf()` and `load_predictions()` from the `starter/` bundle (join the competition first and set up Kaggle credentials). *Content note: the comments can be offensive.* You get the `validation` split (with gold toxicity labels and the identity columns) and three models' predictions over it. This is pure analysis — no training, runs in seconds.

1. Tasks

1. **Score the models.** For each of the three models, compute precision, recall and F1, both **macro** and **micro** averaged.
2. **Look for who it fails.** Break F1 down **per identity subgroup** and report the gap between the best- and worst-served group for each model.
3. **Read the errors.** Pick one model and fill `a2_errors.csv` with **20 of its mistakes**, each tagged with a short `category` (use at least three categories).

What to submit & how it's graded. Submit `a2_metrics.json` (per model: macro/micro precision, recall, F1, and F1 per subgroup) and `a2_errors.csv` (20 rows). The metrics are checked against a hidden reference within `1e-3`; the error file is checked for 20 valid rows and $\geq$3 categories. *Metrics: precision, recall, F1 (macro & micro) and the per-subgroup F1 gap.*

2. Reflection (required, max 200 words)

Submit `reflection.md` (or `.pdf`). It must answer these three fixed questions and include the AI-use declaration. The reflection is graded for honesty and insight, not length.

1. **What did you know before?** — your starting point on this topic.
2. **What was hardest, and how did you get unstuck?**
3. **What does your result say, and what would you try next?**

AI-use declaration. State which parts (if any) were generated with AI tools, and *how you verified them.* Using AI is allowed; undisclosed or unverified use is not.

3. Submission

Upload a single `zip` to Brightspace named `A2_<studentnumber>.zip`, containing the required output file(s) listed above, your `code/` folder, and `reflection.md`. Run the format checker in `starter/` on your zip first — it validates file names, schemas, and value ranges locally before you submit. Auto-grading scores your output file against the hidden reference; the format checker passing is a precondition for grading.