ActiveFence | Take-Home Assignment: Feature Engineering for Text Classification

Goal

Use classical ML techniques and manual feature engineering to detect potentially toxic comments—without relying on deep learning or prebuilt classifiers. The mission is expected to take approximately 2–4 hours.

What to Do

1. Load the Dataset

Use the provided CSV file: `toxicity_toy_dataset.csv`
It contains ~500 short comments labeled as:

- `toxic = 1` → potentially rude, sarcastic, or inappropriate
- `toxic = 0` → neutral or respectful content

2. Engineer Your Own Features

Extract simple, interpretable features (e.g., comment length or swear word count — you may define your own clean list if needed).

3. Train a ML Model

Choose a classical ML model (e.g., logistic regression) or any suitable method to build a classifier that detects potentially toxic comments based on the extracted features.

What to Submit

- A **notebook or script** (clearly commented)
- A **Presentation** including:

- Results and key observations including the labels from your classifier
- Provide a description of your workflow—from research to results analysis. Also describe the methods you have considered and any additional approaches you have tried and your decision making process
- Strengths and weaknesses of your approach
- How would your approach change if you had more time?
- How you used AI tools (if at all)—include a few prompts you used during your research or development

---

Good Luck!