Training Data Classifier: Step-by-Step Instructions

Part 1: Setting Up the Notebook

1. Go to the workshop notebook at colab.research.google.com/drive/1_-I2o_48VojD2tqsVC1Ob0-8qnm0fiHi.
2. Save a copy to your own Google Drive by going to **File → Save a copy in Drive**. A new tab will open with your personal copy of the notebook.
3. In your copy, change the runtime to GPU. Go to **Runtime → Change runtime type**. In the dropdown under "Hardware accelerator," select **T4 GPU**, then click **Save**.
4. Run the first cell by clicking the play button to its left. This cell installs the required software libraries and takes about one minute to complete.
5. Run the second cell. This cell loads the model code and starts the API server. When it finishes, it will print a public URL in the output area that looks something like:

```
Running on public URL: https://abc123def456.gradio.live
```

Copy this URL. You will paste it into the interface in the next section. The cell will show a green checkmark, but the server continues running in the background — do not close the Colab tab.

Part 2: Opening the Interface

6. Go to the Training Data Classifier interface at simulation-and-society.org/workshops/training_for_toxicity/training_data_classifier.html.
7. At the top of the page, paste the Gradio URL you copied from the notebook into the field labeled "Paste Gradio public URL" and click **Connect**. The status should read "Connected."

Part 3: Uploading Data and Training the Model

8. Download the file **toxic_labels.csv** from the workshop website, or use your own labeled CSV from Workshop 2.
9. In the interface, click the file upload area (or drag your CSV file onto it). The filename will appear once the file is selected.
10. Two dropdown menus will appear: one for the text column and one for the label column. The interface will attempt to auto-detect the correct columns from your CSV. Verify that the selections are correct before proceeding.
11. Click **Train Model**. The status will read "Reading CSV and training model..." while the model trains. This takes approximately 2–3 minutes. When training is complete, the page will display a dataset overview showing the total number of comments, the number of categories, and a label distribution with a toggle next to each category.

Part 4: Using the Interface

Testing Comments

12. Scroll down to the "Test the Model" section. Type any comment into the text field and click **Classify**.
13. The interface will return three forms of explanation:

Classification — The predicted label and confidence percentage, followed by a bar chart showing the probability the model assigned to every category.

Nearest Training Examples — The five labeled comments from the training data that are most similar to what you typed, along with their labels and similarity percentages.

Word Importance — A chart showing which words in your comment most affected the prediction. Each word was removed one at a time, and the change in confidence was measured. Positive values mean the word supported the prediction; negative values mean it worked against it.

Toggling Labels and Retraining

14. In the dataset overview, each category has a toggle switch. Green means the category is included in training; red means it is excluded.
15. Toggle any categories off and click **Retrain** to train a new model without those categories. The model retrains from scratch, which takes 2–3 minutes. After retraining, you can test the same comments again to see how the model's behavior has changed.
16. To re-include a previously excluded category, toggle it back on. The toggle will turn light green to indicate it has been re-enabled but the model has not yet been retrained with it. Click **Retrain** to include it again. After retraining, the toggle will return to full green.

Troubleshooting

"Connection failed" when pasting the Gradio URL. Make sure the URL includes `https://` and does not have a trailing slash. Make sure the second cell in the notebook has finished running and the Colab tab is still open.

The interface says "Training failed." Check the Colab notebook for error messages in the cell output. The most common cause is a mismatch between the selected column names and the actual columns in the CSV.

Training is taking a long time. Verify that you set the runtime to T4 GPU in step 4. Training on CPU is significantly slower.

The Gradio URL stopped working. Colab sessions time out after a period of inactivity (typically 30–90 minutes). If this happens, re-run the second cell in the notebook to generate a new URL, then paste the new URL into the interface and reconnect.

Students are seeing different results from each other. If one student retrains the model (by toggling labels and clicking Retrain), the model changes for everyone connected to the same notebook. Instructors may want to coordinate when retraining happens, or have each group run their own Colab notebook.