SESSION 2 REFERENCE GUIDE

Session 2: Swap the Engine

AI + Research Level 2 — Supplementary Material

Concept: TRAINING DATA AND REPRESENTATION
Space: Emotion Detector (evolved from Silly Phrase Finder)
Models used this session:
1. `valhalla/distilbart-mnli-12-3` (zero-shot — Session 1's model)
2. `distilbert-base-uncased-finetuned-sst-2-english` (sentiment — trained on movie reviews)
3. `j-hartmann/emotion-english-distilroberta-base` (emotion — trained on tweets)

Pre-built fallbacks: Have all three versions deployed under profplate/ before class.

Time Breakdown (2 hours)

0:00–0:10 — Show-and-Tell

Ask: "Did anyone modify their Space between sessions?"

If yes: share their screen (or have them paste the URL in chat). Celebrate it. Ask what labels they tried.

If no: that's fine. Show a modified version you prepared — e.g., one with labels like `["scariest", "funniest", "most boring"]`. Quickly demo it.

Transition: "Last week we built a Space. This week we're going to break it — by changing its brain."

0:10–0:25 — "What If We Swap the Engine?"

Open Session 1's Space (the Silly Phrase Finder). Remind students how it works.

Now open `app.py` in the Files tab and change the model line:

Before:

```
classifier = pipeline("zero-shot-classification", model="valhalla/distilbart-mnli-12-3")
```

After:

```
classifier = pipeline("sentiment-analysis", model="distilbert-base-uncased-finetuned-sst-2-english")
```

Also update the function to handle the different output format — sentiment returns `POSITIVE` / `NEGATIVE` instead of label scores.

Say: "Same interface, different engine. Let's see what happens."

Commit. Wait for rebuild. Try the same inputs from last week.

Talking points:
- "This model only knows POSITIVE and NEGATIVE. That's all it was trained to recognize."
- "It was trained on 50,000 movie reviews. So it thinks everything is either a good movie or a bad movie."
- Try: "The food was terrible but the service was excellent." Watch it struggle.

0:25–0:45 — Try the Emotion Model

Now swap again:

```
classifier = pipeline("text-classification", model="j-hartmann/emotion-english-distilroberta-base")
```

Same text, third model, third answer.

Test with the same inputs from the previous two models. The punchline: three models, three different answers to the same text.

Say: "Why do they disagree? They all read the same words. What's different?"

Answer to build toward: They were trained on different data, for different tasks. The zero-shot model was trained on general language understanding. The sentiment model was trained on movie reviews (positive/negative). The emotion model was trained on tweets (7 emotions).

0:45–1:05 — Model Card Reading Activity

Open the model cards for all three models:
- https://huggingface.co/valhalla/distilbart-mnli-12-3

- https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english
- https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

Give students structured questions to answer for each model:

Question	Model 1 (zero-shot)	Model 2 (sentiment)	Model 3 (emotion)
What was it trained on?			
How many examples?			
What language(s)?			
What task does it do?			
Who made it?			

Go through answers together. The key insight: **the training data determines what the model can see.**

1:05–1:25 — Labeling Challenge

The activity: Show 5–6 tweets on screen, one at a time. Students call out what emotion they think each one expresses. Write down their answers.

Pre-prepared tweets:

1. "Just got my test results back... I literally can't even right now"
   - Could be: joy (good results), fear (bad results), surprise (unexpected)
   - Students will disagree — that's the point

2. "My mom made her famous lasagna tonight and I'm not sharing with ANYONE"
   - Could be: joy, anger (possessive tone), neutral

3. "lol my flight got cancelled for the third time this week"
   - Could be: anger (frustrated), joy (sarcastic), sadness

4. "I stayed up until 4am reading and I have zero regrets"
   - Could be: joy, surprise (at themselves), neutral

5. "They really thought they could replace us with AI and we wouldn't notice"
   - Could be: anger, disgust, surprise

6. "Found my childhood diary today. I was a weird kid."
   - Could be: joy (nostalgia), surprise, neutral

After going through them:

Say: "You all disagreed on at least some of these. Here's the thing — the people who labeled the training data disagreed too. When the model gets it 'wrong,' sometimes it's because the humans who taught it couldn't agree either."

Then run the same tweets through the emotion model and compare to student answers.

1:25–1:45 — Rebuild with Emotion Model

Now build the final version of the Space together. This is the `app.py` in the session-02 folder.

Key changes from Session 1:
- Different pipeline type (`text-classification` instead of `zero-shot-classification`)
- Different output format (single label per sentence instead of label ranking)
- Labels each sentence individually

Walk through the code. Commit. Test with student-suggested inputs.

1:40–1:50 — Name the Concept: TRAINING DATA AND REPRESENTATION

Key points:
- "Training data is the textbook the model studied from. It can only know what it's been shown."
- "Representation means: what categories did the data use? Positive/negative is a representation. Seven emotions is a different representation. 'Silly vs. serious' is another."
- "Same text, different representation, different answer. The model doesn't 'understand' — it matches patterns from its training data."

Quick check: "If you wanted a model that detects sarcasm, what kind of training data would you need?"

1:50–2:00 — Notebook Time

Share the Colab link in the Zoom chat.

Walk through together:
1. "Click the link — remember, same as last week"

2. Run the setup cell together
3. Run the cell that loads all three models — point out that this takes a minute
4. Run the first comparison together and look at the output

Say: "The notebook has all three models you just saw. Try different inputs and see where they agree and disagree. Finish the experiments before next week."

Notebook skill being introduced: Running cells in order, reading output from multiple cells

GitHub skill being introduced: "Create a repo called `my-ai-portfolio` on GitHub, then upload this notebook to it."

2:00 — Wrap Up

Share the between-session challenge (see BETWEEN-SESSION.md).

What Could Go Wrong

Problem	Fix
Loading multiple models exceeds free CPU memory	Swap sequentially — only one model loaded at a time. Restart Space between swaps.
Sentiment model output format confuses students	Show the raw output: `[{"label": "POSITIVE", "score": 0.98}]` . It's simpler than zero-shot.
Model card is too technical	Focus on just the 5 questions in the table. Skip the technical details.
Students can't find the model card	Show them: go to huggingface.co, search the model name, click the model page.
Emotion model gives odd results on long text	It was trained on tweets (short text). That's a feature of this lesson, not a bug.

Key Vocabulary (introduce casually)

- **Training data** — the examples a model learned from
- **Representation** — the categories or labels the training data uses

- **Sentiment** — positive or negative feeling
- **Model card** — documentation that describes what a model does and how it was trained
- **Fine-tuned** — a model that was further trained on specific data for a specific task