SESSION 6 REFERENCE GUIDE

Session 6: Same Space, Different Worlds — Instructor Guide

AI + Research Level 2 — Supplementary Material

Concept: OVERFITTING AND DOMAIN SHIFT

Space: Reuses Session 4 Sentiment Showdown (no new build this session)
Key Idea: Models trained on one type of text struggle with another type. The "world" the model learned from shapes what it can understand.

Time Breakdown (2 hours)

0:00–0:05 — Show-and-Tell

- Ask: "What settings recipes did you find? What worked best for what task?"
- Quick share of between-session experiments from the Text Playground.

0:05–0:15 — The Hook: "Same Models, Different Worlds"

- Open the Session 4 Sentiment Showdown Space.
- Say: "These three models haven't changed. Same weights, same training. But today we're going to give them text they've never seen before."
- Quick reminder of what each model was trained on:
- Movie reviews
- Tweets
- Product reviews
- Ask: "If the movie-review model has only ever read movie reviews, what happens when we give it a poem?"

0:15–0:50 — Domain Safari (35 minutes)

Paste pre-prepared texts from different domains (see full list below). For each one:

1. **Before pasting:** Ask students to predict — "Which model will handle this best? Will any of them get it right?"
2. **Paste and observe:** Run the text through all three models.
3. **After results:** "Were you right? What surprised you?"

Work through at least 5-6 domains. Let students pick which to try if time is tight.

Pacing: Spend ~5 minutes per domain. Don't rush — the discussion after each test is where learning happens.

0:50–1:10 — Pattern Recognition (20 minutes)

After testing multiple domains, step back and look for patterns:

- "Which model worked best across the most domains?"
- "Were there domains where ALL models struggled?"
- "What makes a domain 'hard' for these models?"

Pull up the model cards again (from Session 4):
- https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english
- https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest
- https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

Connect model card → training data → domain performance: "The Twitter model handles slang because it was trained on tweets. The product model handles star ratings because that's what it learned from."

1:10–1:30 — Student Challenge: Find a Domain Where ALL Models Fail (20 minutes)

Students suggest text types to test. The goal: find something that confuses every model.

Hints if they're stuck:
- Try mixing languages
- Try text with heavy irony that requires cultural context
- Try very short text (1-2 words)
- Try text with no sentiment at all (a recipe, a math problem)
- Try text from a domain none of the models ever saw (legal text, medical notes, code)

1:30–1:50 — Name the Concept: OVERFITTING AND DOMAIN SHIFT (20 minutes)

Overfitting:
- "When a model gets SO good at its training data that it can't handle anything else."
- Analogy: "Imagine studying only one teacher's test style. You ace THEIR tests. Then you take a test from a different teacher and bomb it — not because you don't know the material, but because you only learned one format."
- The movie review model is overfit to movie review language. It "thinks" everything is a movie review.

Domain Shift:
- "When the data a model encounters in the real world is different from what it was trained on."
- The model didn't get dumber — the world shifted under it.
- Analogy: "You learned to drive in a small town. Then someone puts you on a six-lane highway in another country where they drive on the other side. Same skill (driving), completely different domain."

Connect back to model evaluation (Session 4):
- "Remember when we asked 'which model is best?' Now we know the answer: best FOR WHAT?"
- Evaluation must include testing on the DOMAIN you care about, not just any test set.

1:50–1:55 — Notebook Time

Share the Colab link in the Zoom chat.

Walk through together:
1. Run the setup cell (loads all 3 sentiment models — takes a moment)
2. Run the first domain test (news article) together
3. "Now scroll down and try pasting your own text into the experiment cells"

Notebook skill being introduced: Recording observations in markdown cells — double-click a markdown cell to edit it, then run it to render.

Say: "The notebook has blank tables and observation cells for you to fill in. Double-click any green text cell to edit it. Write down what you noticed about each domain."

1:55–2:00 — Between-Session Suggestion

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

Pre-Prepared Domain Text Samples

Formal News Articles

Sample 1:

> *The Federal Reserve announced a quarter-point interest rate cut on Wednesday, signaling confidence that inflation is moving sustainably toward its 2 percent target. Markets responded with modest gains across major indexes.*

Sample 2:

> *The city council voted unanimously to approve the new zoning regulations, which will allow mixed-use development in previously residential-only areas. Opponents plan to appeal the decision.*

Discussion: "These are neutral/factual. The movie model has to pick POSITIVE or NEGATIVE — it has no neutral option. What does it do?"

Tweets / Social Media

Sample 1:

> *ngl this new update is mid at best □ they really thought they did something*

Sample 2:

> *bestie you did NOT just say that 😭😭😭 im screaming rn*

Discussion: "Which model handles slang and emoji best? (Twitter model should win here.) What about the product review model — has it ever seen '□'?"

Product Reviews (Amazon-style)

Sample 1:

> *Works exactly as advertised. Shipped on time. The build quality is decent for the price point. Would recommend for anyone on a budget who doesn't need premium features.*

Sample 2:

> *Bought this as a gift. Recipient seemed to like it. Packaging was nice. Took a star off because the color was slightly different from the photo.*

Discussion: "The product review model should be at home here. But does 'decent for the price' read as positive or negative?"

Song Lyrics

Sample 1 (melancholy pop):

> *I've been losing sleep over the things that I can't keep. The photographs are fading and the memories are deep. But I'll keep walking through the rain because the sun is just a dream away.*

Sample 2 (upbeat with dark lyrics):

> *Dancing on the ceiling, burning down the walls, laughing at the wreckage as the empire falls. We'll celebrate the ending with confetti made of ash.*

Discussion: "Song lyrics mix positive and negative imagery on purpose. Sample 2 sounds upbeat (dancing, laughing, celebrating) but is actually about destruction. Can the models tell?"

Student Essay Excerpts

Sample 1:

> *In conclusion, while both authors present compelling arguments, Smith's analysis is more thoroughly supported by evidence. However, Jones raises important counterpoints that cannot be ignored.*

Sample 2:

> *The experiment did not produce the expected results. The hypothesis was not supported by the data. However, the methodology was sound and the procedure could be replicated with adjusted variables.*

Discussion: "Academic writing is carefully balanced — not really positive or negative. It lives in a middle ground that the movie model doesn't understand."

Text Messages

Sample 1:

> *lol ok sure whatever u say *

Sample 2:

> *omg YES that's literally the best thing ever im so happy rn ahhh*

Discussion: "'lol ok sure whatever' — is that positive, negative, or sarcastic? Humans would need context. Models don't have that context."

Legal Text

Sample 1:

> *The party of the first part shall indemnify and hold harmless the party of the second part against any and all claims, damages, losses, costs, and expenses arising out of or relating to any breach of this agreement.*

Sample 2:

> *Nothing in this agreement shall be construed to limit the liability of either party for gross negligence, willful misconduct, or fraudulent misrepresentation.*

Discussion: "Legal text has no sentiment — it's purely functional. But the models HAVE to output something. What do they say? Why?"

Medical Notes

Sample 1:

> *Patient presents with acute onset of substernal chest pain radiating to the left arm. ECG shows ST-elevation in leads II, III, and aVF. Troponin levels are pending. Started on aspirin and heparin drip.*

Sample 2:

> *Follow-up visit. Patient reports significant improvement in symptoms since starting the new medication. Range of motion has increased. Recommend continuing current regimen.*

Discussion: "Sample 1 describes a heart attack in neutral clinical language. Sample 2 is genuinely positive news. Can the models tell the difference, or do clinical words confuse them?"

Poetry

Sample 1 (dark imagery, classic style):

> *I wandered lonely through the ash of things that used to gleam. The world had shed its golden mask and left a hollow dream.*

Sample 2 (joyful imagery):

> *The morning broke with strawberry light across the sleeping town. Each window caught a piece of sky and wore it like a crown.*

Discussion: "Poetry uses figurative language. 'Ash of things that used to gleam' is negative to a human, but does the model understand metaphor?"

Code Comments

Sample 1:

> *// HACK: This is a terrible workaround for the race condition. TODO: fix this properly before it breaks production again.*

Sample 2:

> *// Beautiful implementation of the merge sort algorithm. Clean, efficient, and well-tested. Props to @sarah for this one.*

Discussion: "Code comments are a completely alien domain for all three models. But they still contain sentiment. Can the models detect it through the technical noise?"

Meme Transcriptions

Sample 1:

> *Nobody: Absolutely nobody: My cat at 3am: knocks everything off the counter*

Sample 2:

> *Me: I should really go to bed early tonight. Also me at 2am: Let me just watch one more episode.*

Discussion: "Meme format is its own language. 'Nobody:' format, self-deprecating humor, relatable content. These aren't really positive or negative — they're funny. How do sentiment models handle humor?"

Discussion Questions by Domain

Domain	Key Question
News	"Can a model be useful if it has no 'neutral' option?"
Tweets	"Why does the Twitter model understand slang? What would happen if we trained a model only on Shakespeare?"
Product Reviews	"Is 'decent for the price' positive or negative? Is there a 'right' answer?"
Song Lyrics	"Should AI understand art? What would it need to learn that?"
Student Essays	"Academic writing is balanced on purpose. Is 'balanced' a sentiment?"
Text Messages	"How much of texting depends on who you're talking to?"
Legal Text	"If there's no sentiment, what should the model say? Is 'I don't know' an option?"
Medical Notes	"Where would domain shift be actually dangerous?"
Poetry	"Can AI understand metaphor? What would it need to?"
Code Comments	"These models never saw code. But humans wrote sentiment into comments. Weird, right?"
Memes	"Memes have their own grammar. Is that a 'domain'?"

Technical Notes

- No new Space to build or deploy this session.
- Students should use either the instructor's deployed Space or their own duplicate from Session 4.
- If students modified their duplicates (added models, etc.), that's fine — it adds to the discussion.
- Long text inputs should be kept under 512 characters (the model truncation limit).

Concept Connections

- **Session 4:** Introduced model comparison and evaluation — "which model is best?"
- **Session 6 (this session):** The answer is "best FOR WHAT DOMAIN?" — models are products of their training data.
- **Session 7 (upcoming):** Bias and fairness — training data doesn't just affect domain, it affects who the model works well for.