Error Analysis Report: Common Mistakes Across Models

Overview

This report analyzes prediction errors across 18 models on the gold set (28 examples) to identify which examples are most difficult for sarcasm detection and what patterns emerge from model failures.

Models Included

Baselines (8 models):

- TF-IDF + Logistic Regression (4 conditions: context/text × original/normalized)
- TF-IDF + SVM (4 conditions: context/text × original/normalized)

Fine-tuned (4 models):

- XLM-RoBERTa fine-tuned (with context, without context)
- DeBERTa-v3-base fine-tuned (with context, without context)

Zero-shot LLMs (6 models):

- Gemma2:9b (with context, without context)
- Qwen2.5:14b (with context, without context)
- Phi3:medium (with context, without context)

Difficulty Distribution

Models Wrong	Number of Examples	Percentage
0	1	3.6%
1	1	3.6%
2	3	10.7%
3	4	14.3%
4	3	10.7%
5	2	7.1%
6	3	10.7%

Models Wrong	Number of Examples	Percentage
7	1	3.6%
8	1	3.6%
9	3	10.7%
10	1	3.6%
11	1	3.6%
12	2	7.1%
13	2	7.1%

Only 1 example (3.6%) was classified correctly by all 18 models. No example was misclassified by all models.

Hardest Examples

Examples with 13/18 models wrong

Example 1: (TRUE LABEL: LITERAL)

- Text: "RAJ: Leonard, may I present, live from New Delhi, Dr. and Mrs. V. M. Koothrappali."
- Error type: False Positive (13 models predicted sarcastic)
- Models correct: xlmr_finetuned_context, qwen2.5_14b_context, qwen2.5_14b_no_context, phi3_medium_context, phi3_medium_no_context

Example 2: (TRUE LABEL: LITERAL)

- Text: "LEONARD: Don't you think looking for a new city to live in is a bit of an overreaction?"
- Error type: False Positive (13 models predicted sarcastic)
- Models correct: xlmr_finetuned_no_context, deberta_finetuned_no_context, qwen2.5_14b_no_context, phi3_medium_context, phi3_medium_no_context

Examples with 12/18 models wrong

Example 3: (TRUE LABEL: SARCASTIC)

- Text: "PERSON: I don't know. Somebody bigger and... Yeah, sure. Why not you?"
- Error type: False Negative (12 models missed sarcasm)
- Models correct: gemma2_9b_context, gemma2_9b_no_context, qwen2.5_14b_context, qwen2.5_14b_no_context, phi3_medium_context, phi3_medium_no_context

Example 4: (TRUE LABEL: LITERAL)

- Text: "LEONARD: Technically, yes. But, if you'll notice... It's reversible!"
- Error type: False Positive (12 models predicted sarcastic)
- Models correct: xlmr_finetuned_context, deberta_finetuned_context, deberta_finetuned_no_context, qwen2.5_14b_no_context, phi3_medium_context, phi3_medium_no_context

Examples with 11/18 models wrong

Example 5: (TRUE LABEL: SARCASTIC)

- Text: "HOWARD: Yeah, terrific. The other astronauts would love to go hurtling through space with a guy named 'Crash.'"
- Error type: False Negative (11 models missed sarcasm)
- Models correct: deberta_finetuned_context, gemma2_9b_context, gemma2_9b_no_context, qwen2.5_14b_context, qwen2.5_14b_no_context, phi3_medium_context, phi3_medium_no_context

Examples with 10/18 models wrong

Example 6: (TRUE LABEL: SARCASTIC)

- Text: "CHANDLER: Oh Oh, I am convinced!"
- Error type: False Negative (10 models missed sarcasm)
- Models correct: tfidf_logreg_text_original, tfidf_logreg_text_normalized, tfidf_svm_text_original, xlmr_finetuned_no_context, deberta_finetuned_no_context, gemma2_9b_context, gemma2_9b_no_context, qwen2.5_14b_context

Easiest Example

Only 1 example was classified correctly by all 18 models:

- Text: "RACHEL: All right, let's do it."
- True label: LITERAL
- This straightforward, non-ambiguous statement was easy for all models.

Error Type Analysis

Among the hardest examples (≥9 models wrong):

True Label	Count	Primary Error Type
LITERAL	5	False Positive
SARCASTIC	5	False Negative

The hardest examples are evenly split between literal utterances misclassified as sarcastic (FP) and sarcastic utterances missed by models (FN).

Model Performance on Gold Set

Model	Accuracy
phi3_medium_context	0.893
phi3_medium_no_context	0.821
tfidf_logreg_text_original	0.750
deberta_finetuned_no_context	0.750
tfidf_svm_text_original	0.714
qwen2.5_14b_context	0.714
qwen2.5_14b_no_context	0.714
tfidf_logreg_context_original	0.643
xlmr_finetuned_no_context	0.643
gemma2_9b_context	0.643
gemma2_9b_no_context	0.643
tfidf_logreg_text_normalized	0.607
tfidf_svm_context_original	0.607
tfidf_svm_text_normalized	0.607
xlmr_finetuned_context	0.607
tfidf_logreg_context_normalized	0.500
tfidf_svm_context_normalized	0.500

Model	Accuracy
deberta_finetuned_context	0.500

Observations

1. Zero-shot LLMs outperformed other models on the hardest examples.

For the example where 12 models failed to detect sarcasm ("Yeah, sure. Why not you?"), only the six LLM models (Gemma2, Qwen2.5, Phi3) classified it correctly. All baseline and fine-tuned models failed.

2. Rhetorical questions are frequently misclassified as sarcastic.

The literal question "Don't you think looking for a new city to live in is a bit of an overreaction?" was predicted as sarcastic by 13 models. The questioning format appears to trigger false positive predictions.

3. Formal or theatrical speech patterns trigger false positives.

The literal utterance "Leonard, may I present, live from New Delhi, Dr. and Mrs. V. M. Koothrappali" was misclassified by 13 models. The formal, exaggerated presentation style was interpreted as sarcastic.

4. Dry sarcasm without lexical markers is frequently missed.

Sarcastic utterances like "Yeah, terrific" and "Oh Oh, I am convinced!" lack explicit sarcasm markers beyond the words themselves. These were missed by 10-11 models.

5. Context removal helped some models on hard examples.

For Example 2, models without context (xlmr_finetuned_no_context, deberta_finetuned_no_context, qwen2.5_14b_no_context) performed better than their context-inclusive counterparts.

6. Normalized name conditions performed worse.

Models using normalized names (SPEAKER: instead of character names) consistently appeared in the lower accuracy range, with tfidf_logreg_context_normalized and tfidf_svm_context_normalized at 50% accuracy.