CSAT Verbatims Analysis (FR)
CamemBERT Pipeline for Sentiment Classification

1. Context and Goal

Free provided anonymized French CSAT verbatims. The objective was to extract actionable insight from free text via sentiment classification. A hybrid setup (few-shot LLM prompting + local components) was initially considered; however, no API key was issued for `deepseek/deepseek-r1-distill-llama-8B`, and the hosted Hugging Face endpoint requires authentication. To ensure timely delivery, reproducibility, and on-prem feasibility, a **fully local** approach was adopted: **CamemBERT** fine-tuned for sentiment.

2. Data and Method Implemented

Data.

SQLite export containing verbatims (*commentaire*) and CSAT scores (1–5). CSAT values are used as supervision signals for sentiment.

Model choice.

CamemBERT was selected for its strong French performance, full offline execution (no external API), and suitability for sentence-level fine-tuning within the available GPU budget.

Labeling from CSAT.

CSAT 1–5 is mapped to three classes:

$$1-2 \rightarrow Negative, \quad 3 \rightarrow Neutral, \quad 4-5 \rightarrow Positive.$$

To handle score/text mismatches, a 4-class variant was also explored by splitting Positive into *Positive* and *Mixed-Positive* (high score with mixed/negative phrasing).

Preprocessing.

French accents and punctuation preserved; empty rows removed. Heavy normalization was avoided to keep CamemBERT tokenization effective.

Balancing & training.

Class imbalance was addressed via light over-/under-sampling to obtain a balanced dataset. Fine-tuning used AdamW and linear warmup/decay. All runs execute locally and are reproducible.

Why not zero-shot LLM in production.

Early zero-/few-shot trials (XLM-R/BART) and hosted DeepSeek were evaluated. Without company API tokens and with unstable outputs, these were not viable. A local CamemBERT approach was preferred for reproducibility and reliability.

3. Results (3-class vs. 4-class)

Validation compares the 3-class model (Negative/Neutral/Positive) with a 4-class variant adding *Mixed-Positive*.

Table 1: CamemBERT — 3 classes

Class	Precision	Recall	F1
Negative	0.88	0.79	0.84
Neutral	0.80	0.93	0.86
Positive	0.90	0.85	0.87
Macro Avg	0.86	0.86	0.86
Accuracy	85.6%		

Table 2: CamemBERT — 4 classes

Class	Precision	Recall	F1
Negative	0.64	0.67	0.65
Neutral	0.46	0.44	0.45
Positive	0.85	0.86	0.85
Mixed-Positive	0.79	0.76	0.77
Macro Avg	0.68	0.68	0.68
Accuracy	68.3%		

The 3-class model is reliable (macro-F1 0.86). Adding Mixed-Positive improves interpretability of score/text mismatches but increases task difficulty, reducing accuracy to 68.3%. Neutral remains the most challenging class, while Positive remains stable.

4. Evaluated Approaches

Zero-shot (XLM-R / BART MNLI).
Multilingual NLI zero-shot models were tested for topic tagging (Billing, Network, Support, ...). Issues observed: high "Other" rate, unstable thresholds, inconsistent handling of telecom-specific French.

Few-shot prompting.
Adding 2–6 French examples per class improved some edge cases but remained prompt-sensitive and noisy.

Hosted DeepSeek.
Endpoint required authentication. To ensure reproducibility and local execution, CamemBERT fine-tuning was prioritized.

5. Limitations and Mitigations

Score/text mismatch.
High CSAT with mixed or negative wording creates ambiguity; hence the introduction of the *Mixed-Positive* class.

Imbalance.
Positive comments dominate; mitigation uses dataset balancing (over-/under-sampling).

6. Next Steps and Conclusion

A fully local CamemBERT pipeline delivers reliable 3-class sentiment (with an optional 4th Mixed-Positive class). To harden it for production:

- **Score/Text mismatch:** keep the Mixed-Positive option; relabel a small set (100–300) to correct conflicts.
- **Imbalance:** add Neutral/Negative examples; optionally test focal loss.
- **Categories:** evolve from keywords to a small supervised multi-label classifier; later, with API access, augment using LLM few-shot prompting and structured JSON outputs.

With these steps, the system is deployment-ready for sentiment and can be cleanly extended to robust topic classification.