# CSAT Verbatims Analysis (FR)
# CamemBERT Pipeline for Sentiment Classification

## 1. Context and Goal

Free provided anonymized French CSAT verbatims. The objective is to extract actionable insight from free text via sentiment classification. A hybrid setup (few-shot LLM prompting plus local components) was initially considered; however, no API key was issued for `deepseek/deepseek-r1-distill-llama-8b:fp8`, 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 with verbatims (*commentaire*) and CSAT scores (1–5). CSAT is used to derive supervision signals for sentiment.

**Model choice.**
**CamemBERT** was selected for its strong FR performance, local/offline execution (no external API), and suitability for sentence-level fine-tuning within the available GPU budget.

**Labeling from CSAT (supervision).**
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.**
Preserve French accents and punctuation; drop empty/placeholder rows, duplicates, and ultra-short texts ($\leq$3 tokens). Heavy normalization is avoided to keep CamemBERT tokenization effective.

**Balancing & training.**
Class imbalance is addressed via (i) class-weighted loss and (ii) mild oversampling of minority classes. Fine-tuning uses AdamW, linear warmup/decay, and early stopping on validation macro-F1. All runs execute locally and are reproducible.

**Why not LLM zero-shot in production.**
Zero-/few-shot trials (XLM-R/BART) and DeepSeek on HF were assessed. Without company API/tokens, hosted LLMs were not viable; early trials also produced excessive "Other" or noisy assignments. CamemBERT fine-tuning was therefore preferred for reliability and delivery.

## 3. Results (3 classes vs. 4 classes)

Validation compares the 3-class model (Negative/Neutral/Positive) with a 4-class variant that adds *Mixed-Positive* to capture high-score comments with mixed/negative phrasing.

Table 1: CamemBERT — 3 classes

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| 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 (adds Mixed-Positive)

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| 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* is stable. With additional *Mixed-Positive* examples and light augmentation, the 4-class gap is expected to narrow.

# 4. Evaluated Approaches

**Zero-shot topic classification (XLM-R / BART MNLI).**
Multilingual NLI zero-shot models (`joeddav/xlm-roberta-large-xnli`, `facebook/bart-large-mnli`) were tested to tag business categories (Billing, Network, Support, ...). Observed issues: (i) a high share of "Other" (∼75%), (ii) unstable label thresholds across batches, and (iii) inconsistent handling of telecom-specific French terms (e.g., *RIO, portabilité, eSIM*) and very short comments, leading to noisy, inconsistent labels.

**Few-/N-shot prompting.**
Adding 2–6 French examples per class helped a few edge cases but outputs remained sensitive to prompt wording and noisy. At scale it was inefficient without a small labeled set to calibrate per-class thresholds, so quality stayed below target.

**Hosted DeepSeek (HF).**
`deepseek/deepseek-r1-distill-llama-8b:fp8` on Hugging Face requires authentication; no company endpoint/API was available in the project window. To keep the workflow reproducible and fully local, the pipeline prioritized CamemBERT fine-tuning for sentiment with a lightweight, interpretable keyword layer for categories.

# 5. Limitations and Mitigations

**Score/Text mismatch.**
High CSAT with mixed or negative wording creates ambiguity. A *4th class (Mixed–Positive)* was introduced to separate "resolved-but-complaint" cases from clean positives, improving downstream interpretability.

**Imbalance.**
Positive comments dominate, biasing the classifier toward the majority class. Mitigation used (i) class-weighted loss and (ii) mild oversampling of minority classes, which improved recall for Neutral/Negative without materially hurting Positive.

# 6. Next Steps and Conclusion

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

- **Score/Text mismatch**: keep the *Mixed-Positive* option; relabel a small set (100–300) to correct conflicts; add simple "issue resolved" patterns; display CSAT and model outputs side-by-side in dashboards.
- **Imbalance**: retain class-weighted loss and targeted sampling; add Neutral/Negative examples; optionally test focal loss; track per-class F1 over time.
- **Categories**: evolve from keywords to a small supervised multi-label classifier (6–8 topics); later, with API access, augment via few-shot LLM and structured JSON outputs.

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