AI Lab I–II Research Proposal (v0.1)
Constraint-Aware Lightweight Multi-Label Emotion Detection

Roland Sándor Nagy

February 22, 2026

1 Motivation

Emotion detection from text is often treated as a standard multi-label classification problem, where models are evaluated primarily using global metrics (e.g., micro-F1). However, in realistic applications certain emotion categories may be rare but more critical than others.

This project investigates lightweight transformer-based emotion detection under explicit resource constraints, with a particular focus on improving performance on rare (tail) emotion categories. The goal is to move beyond unconstrained model comparison and instead study model and decision strategies within clearly defined deployment scenarios.

2 Conceptual Framing of Emotion

Emotion in text can be interpreted in multiple ways:

- **Writer-internal emotion:** the actual psychological state of the author at the time of writing.
- **Expressed emotion:** the emotion conveyed or expressed in the text.
- **Reader-evoked emotion:** the emotion elicited in the reader.

These interpretations are not necessarily equivalent. For example, a text may describe anger without the author being angry, or may evoke fear in the reader without explicitly expressing fear.

The GoEmotions dataset operationalizes emotion as the perceived expressed emotion in the text, based on human annotation.

Therefore, this project models *perceived expressed emotion*, rather than attempting to infer the author's internal psychological state or the reader's emotional reaction.

3 Dataset

Primary dataset:

- GoEmotions (28-label multi-label emotion classification dataset)

Optional extension (if needed for generalization analysis):

- Additional public emotion dataset (to be determined)

4 Scenario Definitions

4.1 Scenario B (Primary Proposal): Tail / Critical Emotion Focus

Application context: A wellbeing or monitoring system where rare but critical emotion categories must be detected reliably.

Objective:

- Maximize macro-F1.
- Prioritize recall on tail (low-frequency) emotion categories.

Resource constraints (initial proposal, subject to confirmation):

- Model size $\leq$ 120M parameters.
- GPU memory usage $\leq$ 12 GB.
- Inference latency target $<$ 100 ms per request (batch size = 1).

Optimization formulation:

Let $\mathcal{M}$ denote the set of candidate transformer models and $\mathcal{D}$ the set of decision strategies (thresholding and calibration methods). The goal is to solve:

$$\max_{m \in \mathcal{M},\, d \in \mathcal{D}} \text{Macro-F1}_{tail}(m, d)$$

subject to:

$$\text{Params}(m) \leq 120M,$$

$$\text{Memory}(m) \leq 12GB,$$

$$\text{Latency}(m) \leq 100ms.$$

Constraint-aware research question:

Given the above constraints, which transformer architectures and decision strategies satisfy the limits, and among those, which configuration maximizes performance on rare emotion categories?

This scenario serves as the primary experimental setting of the project.

4.2 Scenario A: Cost / Latency-Constrained Serving

Application context: A high-throughput text annotation API.

Objective:

- Maximize macro-F1 under strict latency and cost constraints.

Resource constraints (initial proposal):

- Latency (p95) $<$ 50 ms.
- Model size $\leq$ 60–100M parameters.
- GPU memory $\leq$ 12 GB.

Research question:

Which models provide the best performance under the defined latency and size constraints, and how does performance degrade as constraints become stricter?

4.3 Scenario C: Strict Model Size Cap (Edge-like Environment)

Application context: Deployment in a strongly resource-constrained environment (e.g., low-cost CPU server or edge device).

Objective:

- Maintain acceptable macro-F1 under strict model size limits.

Resource constraints (initial proposal):

- Model size $\leq$ 30M parameters.
- CPU inference time $<$ 200 ms per request.

Research question:
Under strict model size and inference constraints, which architectures remain viable, and how do decision strategies (thresholding, calibration) affect performance within these limits?

5 Core Methodological Components

Across scenarios, the following aspects will be studied:

- Multi-label training using transformer models (e.g., BERT-base, DistilBERT, MiniLM).
- Decision strategies:
  - Fixed threshold (0.5).
  - Global optimized threshold.
  - Per-class optimized threshold.
- Calibration techniques (e.g., temperature scaling).
- Comparison of multi-label vs single-label formulations (especially for tail categories).

6 Evaluation Protocol

- Micro-F1.
- Macro-F1.
- Macro-F1 computed specifically on tail labels.
- Per-class F1 and recall.
- Tail-label slice metrics.
- Latency measurement.
- Model parameter count and memory footprint.

Definition of tail labels:
Tail labels are defined as the bottom 30% of emotion categories ranked by frequency in the training set. Sensitivity analysis with alternative thresholds (e.g., bottom 20% and bottom 40%) will also be performed.

7 Hypotheses (Initial)

- H1:: Per-class threshold optimization yields statistically significant improvement in tail-label macro-F1 compared to a fixed 0.5 threshold.
- H2: Calibration improves stability and generalization of threshold-based decisions.
- H3: Single-label simplification reduces performance on rare emotion categories compared to multi-label modeling.

8 Expected Contributions

- A constraint-aware evaluation framework for lightweight multi-label emotion detection.
- Empirical analysis of thresholding and calibration under tail-priority settings.
- Comparative study of lightweight transformer models under explicit deployment constraints.

9 Preferred Direction

Among the three scenarios, Scenario B (Tail / Critical Emotion Focus) is proposed as the primary research direction due to its stronger methodological and application relevance. Scenario A elements (explicit resource constraints) are incorporated to ensure a well-defined and practically grounded problem setting.