**Project Report**

1. **Introduction:**

With the rise of digital communication, email has become a primary channel for professional and personal correspondence. These emails often contain **Personally Identifiable Information (PII)** such as names, email addresses, phone numbers, and identification numbers.

At the same time, organizations need to **automatically classify emails** into operational categories like Incident, Request, Change, and Problem to streamline workflows.

This project addresses the need for a **robust multilingual pipeline** that can:

* Automatically **detect and mask PII**
* Classify emails into predefined categories
* Expose the functionality via a /classify **FastAPI endpoint**, deployable on **Hugging Face Spaces**

1. **Approach**

**PII Masking**

The solution combines Regex + Named Entity Recognition (NER) to detect and mask:

* Email addresses
* Phone numbers
* Dates of birth
* Aadhar numbers
* Credit card information
* CVV and expiry dates
* Full names (via NER)

Regex handles structured formats, while transformer**-**basedmultilingualNER (via Davlan/xlm-roberta-base-ner-hrl) detects person names (PER tags).

Masked text replaces sensitive segments with tags like [email], [phone\_number], [full\_name], etc., and stores original positions for unmasking or audit trails.

**🔸 Classification**

The masked emails are then passed into a fine-tuned transformer model that categorizes them into one of:

* **Incident**
* **Request**
* **Change**
* **Problem**

**🔹 3. Model Selection & Training Details**

**Base Model: microsoft/deberta-v3-base**

* Chosen for its contextual understanding and multilingual capability
* Supports high performance with relatively lower GPU demand

**Training Dataset**

* 24,003 labeled email samples
* Class imbalance handled viaclassweighting in loss function

| **Class** | **Count** |
| --- | --- |
| Incident | 9586 |
| Request | 6860 |
| Problem | 5037 |
| Change | 2517 |

**🧠 Hyperparameters**

* Epochs: 5–7
* Max length: 256
* Batch size: 8–16
* Optimizer: AdamW with linear learning rate scheduler
* Weighted cross-entropy loss to handle imbalance

**Results**

* **Validation Accuracy**: ~80.3%
* **F1 Score**: Peaked at **~80.1%**
* Balanced performance across all classes using weighted loss

### ****4. Challenges & Solutions****

| **Challenge** | **Solution** |
| --- | --- |
| PII overlap between regex and NER | Used offset tracking to prevent duplicate masking |
| Phone & CVV regex conflicts | Refined patterns and enforced masking order |
| Handling multilingual names | Used xlm-roberta-base-ner-hrl with aggregation\_strategy="simple" |
| Model file size exceeded 1GB | Switched from .safetensors to .bin format for deployment |
| Hugging Face Space file limits | Zipped and deployed model separately or used HF Hub |