**Data Science Report: AI-powered Expense Tracker**

**1. Introduction**

The aim of this project was to develop an **AI Agent** that automatically extracts financial transactions from Gmail, parses key details (date, merchant, amount, transaction ID), and classifies them into **expense categories** (e.g., Food, Bills, Travel, Online Shopping). These transactions are then logged into a **Google Sheet**, providing users with a categorized expense record.

The core of this AI system is a **fine-tuned DistilBERT model**. We chose DistilBERT because:

* It is lighter and faster than BERT while retaining ~95% of its performance.
* It can be fine-tuned on domain-specific data (financial transactions) for higher accuracy than classical models like Naive Bayes.
* It meets the project requirement of using a fine-tuned model within an AI agent.

**2. Dataset**

**2.1 Sources**

* **Synthetic dataset (labeled\_transactions.csv / synthetic\_transactions.csv):**
  + Categories: Bills, Credit, Entertainment, Food, Miscellaneous, Online Shopping, Stationery, Transfer, Travel.
  + Created ~950 labeled examples, each row containing:
    - Date, Merchant, Amount, RawText, Category.
* **Real Gmail transactions:**
  + Raw text examples such as:
    - "Dear Customer, Rs.180.00 has been debited from account 2715 to VPA q962361699@ybl MANASI BARMAN on 07-09-25..."
    - "Dear Customer, Rs.20.00 has been debited from account 2715 to VPA paytmqr281005050101hzvclx6i8l5f@paytm GAUTAM BHUYAN..."
  + These examples highlight real-world noise: unstructured merchant names, VPA IDs, abbreviations.

**2.2 Preprocessing**

* Converted text to lowercase.
* Removed extra spaces, special characters, HTML entities.
* Preserved merchant names and currency formats (important for classification).
* Parsed merchant names separately for dictionary matching.

**3. Methodology**

**3.1 Pipeline Overview**

1. **Gmail API** → Fetch last 100 emails containing keywords like "transaction".
2. **Parser** → Extract Date, Merchant, Amount, Transaction\_ID, and RawText.
3. **Categorizer** →
   * First check **Merchant Dictionary** (e.g., “GAUTAM BHUYAN → Stationery”).
   * Otherwise, use **DistilBERT fine-tuned classifier**.
   * If confidence < threshold, mark as "Uncategorized" for later user feedback.
4. **Google Sheets Logging** → Append clean row [Date, Merchant, Amount, RawText, Category, Transaction\_ID] if not duplicate.

**4. Model Training & Fine-tuning**

**4.1 Models Tried**

* **Naive Bayes:**
  + Fast, works well with structured synthetic data.
  + Accuracy ~94% on synthetic dataset.
  + However, struggled with real noisy Gmail text → misclassified merchants like "OpenAI LLC" as "Food".
* **Logistic Regression (baseline):**
  + Similar to NB in performance, slightly slower.
* **Fine-tuned DistilBERT:**
  + Pretrained on large text corpora.
  + Fine-tuned on our labeled dataset (~950 rows).
  + Used Hugging Face Transformers with Trainer API.
  + Achieved **~90% validation accuracy** after fine-tuning.

**4.2 Training Setup**

* Hardware: Google Colab GPU.
* Epochs: 3–5.
* Batch size: 16.
* Optimizer: AdamW.
* Loss: CrossEntropy.
* Libraries: transformers, torch, datasets.

**5. Results**

**5.1 Quantitative**

* **Naive Bayes:**
  + Accuracy: ~0.94 on synthetic dataset.
  + Misclassified noisy real-world transactions.
* **DistilBERT (fine-tuned):**
  + Validation Accuracy: ~0.90.
  + More robust to noisy Gmail raw text.
  + Predictions included multiple classes (Food, Credit, Stationery, Entertainment).

Example predictions (DistilBERT):

* "Paid Rs.450 to Swiggy" → **Food** (Confidence 0.26).
* "Uber ride payment of Rs.320" → **Travel** (Confidence 0.18).
* "Electricity bill payment Rs.1200" → **Bills** (Confidence 0.23).
* "Bought Rs.2500 worth of clothes on Amazon" → **Online Shopping** (Confidence 0.25).

**5.2 Qualitative**

* The model learned general categories well (Food, Travel, Online Shopping).
* Struggled with **local merchants** (e.g., "GAUTAM BHUYAN" stationery shop).
* Added a **Merchant Dictionary** to handle local recurring names.
* Introduced a fallback "Uncategorized" category for low-confidence predictions.

**6. Evaluation Methodology**

**Quantitative Metrics**

* **Accuracy, Precision, Recall, F1-score** (via sklearn classification\_report).
* DistilBERT: balanced performance across all 9 categories.
* Naive Bayes: high precision but biased towards synthetic patterns.

**Qualitative Checks**

* Manually inspected Gmail transactions parsed and logged.
* Verified that merchants like "Swiggy" consistently map to **Food**.
* Validated Google Sheets log consistency (headers + no duplicate Transaction IDs).

**7. Limitations**

* Synthetic dataset does not fully represent noisy real-world financial text.
* DistilBERT predictions skewed towards high-frequency categories (Food, Credit).
* Gmail raw text often lacks explicit merchant/category cues.
* Large model files (safetensors ~255 MB) difficult to manage in GitHub without Git LFS.

**8. Future Work**

* Expand labelled dataset with **real Gmail transactions**.
* Use **parameter-efficient fine-tuning (LoRA/PEFT)** to reduce model size.
* Implement **user feedback loop**:
  + Ask user to manually categorize "Uncategorized" rows.
  + Update dictionary + incremental fine-tuning.
* Add **visual analytics (PowerBI / Data Studio dashboards)** for expense insights.
* Deploy as a lightweight **Flask/FastAPI service** with Google OAuth.