Project Report: RAG-Driven SMS Fraud Detection

Program: Intel Unnati GenAI Program

Developer: [Your Name]

Technical Stack: Python, Hugging Face, Zephyr-7B-Beta, FAISS, LangChain

---

1. Introduction

With the rise of sophisticated "smishing" (SMS phishing) attacks, traditional keyword-based filters are no longer sufficient. This project demonstrates a **Retrieval-Augmented Generation (RAG)** approach to identify fraudulent messages by providing an LLM with real-time context from a verified dataset of fraud patterns.

2. Problem Statement

Large Language Models (LLMs) can sometimes hallucinate or fail to recognize very recent or specific fraud patterns not present in their training data. By using a RAG-based system, we bridge this gap by retrieving the most relevant fraud examples from a local CSV dataset before classification.

3. Methodology & Architecture

The project follows a modular RAG pipeline:

- **Data Preprocessing:** A CSV dataset of labeled SMS (Fraud/Ham) is cleaned and tokenized using Python.
- **Vector Embeddings:** We use the sentence-transformers library from **Hugging Face** to convert text into 384-dimensional dense vectors.
- **Vector Store: FAISS** is utilized for efficient similarity searches.
- **Retrieval:** For every incoming user query, the system retrieves the top $K=3$ most similar examples from the vector store.
- **Augmentation & Generation:** The retrieved context is injected into a system prompt. The **Zephyr-7B-Beta** model (quantized to 4-bit for notebook efficiency) then generates a verdict with a detailed reasoning for its decision.

4. Model Configuration

- **Model ID:** HuggingFaceH4/zephyr-7b-beta

- **Quantization:** 4-bit NormalFloat (NF4) via bitsandbytes.
- **Parameters:** Temperature set to 0.1 to ensure deterministic and factual outputs.

5. Results & Conclusion

The system successfully identifies complex fraud attempts that lack common "spam" keywords by recognizing the semantic intent of the message. This project showcases the power of combining open-source LLMs with retrieval systems to solve real-world cybersecurity challenges.