**Title**:

*SBERT-GPT Hybrid Model for Semantic Multi-Label Text Analysis*

**Description**:

This repository contains the code and resources for a hybrid machine learning model integrating SBERT and GPT to perform context-aware, multi-label annotation of short texts. This implementation supports semantic label extraction and metadata enrichment for translated Quranic verses. It combines Sentence-BERT (SBERT) for semantic similarity and OpenAI’s GPT model for contextual metadata generation.

**Project Overview**

Goal:

Automatically generate meaningful metadata labels for Quranic verses (Surah Al-Baqarah) using AI.

Approach:

1. Use SBERT to embed verse texts and compute semantic similarity.

2. Use GPT to generate enriched metadata for each verse.

3. Compare generated labels with content and construct a label recurrence matrix.

**Dataset Information**

Source: [Kaggle Quran Dataset](https://www.kaggle.com/zusmani/the-holy-quran/version/3?select=Dataset-Verse-by-Verse.xlsx)

Used Subset: Only Surah Al-Baqarah was used.

File Used: QuranDS.csv

**Code Structure**

**SBERT\_Quran\_verses.ipynb**

* Use SBERT to compute embeddings of verses.
* Perform cosine similarity to identify semantically close verses.
* Build recurrence matrix for label propagation.

**GPT\_Quran\_Labelling.ipynb**

* Preprocess verses and call OpenAI GPT to generate context-based labels.
* Save enriched metadata into CSV.
* GPT API Requirement: you must have an OpenAI API key.

1. Register at [https://platform.openai.com](https://platform.openai.com/).
2. Generate your API key.
3. Set the API key in your environment: export OPENAI\_API\_KEY='your-api-key-here'

**Usage Instructions**

Step 1: Prepare your environment and dependencies

Step 2: Load the dataset and run SBERT + GPT code

Open notebooks and run all cells in order

**Requirements**

- Python 3.7+

- Dependencies:

- `openai`

- `sentence-transformers`

- `pandas`

- `numpy`

- `scikit-learn`

**Methodology**

1. **Text Cleaning**: Tokenization, normalization.
2. **Label Embedding**: Use SBERT to encode labels and verses.
3. **GPT Label Generation**: Fine-tuned prompts sent to OpenAI API.

**Citation**

If you use this code or dataset, please cite:

*Muhammad Tasaddaq Latif et al. (2025), “A Hybrid SBERT-GPT Model for Semantic Multi-Label Tagging of Short Texts”*