**Data Preprocessing and Model Training**

**1. Data Preprocessing**

**1.1 Loading the Dataset**

* The dataset is loaded from a CSV file (final\_labels.csv) using pandas.
* Missing values in the body column are dropped to ensure only valid comments are processed.

**1.2 Building Conversation Threads**

* A directed graph is created using NetworkX, where:
  + Each entry\_id represents a node (comment).
  + parent\_id establishes directed edges between parent and child comments.
* **Thread Reconstruction:**
  + Using depth-first search (DFS), full discussion threads are extracted.
  + Root comments (comments without a parent) are identified to reconstruct complete conversations.

**1.3 Summarization**

* A **T5 transformer model** (t5-base) is used for text summarization.
* To prevent input overload, text is limited to **512 words**.
* Summarization parameters:
  + max\_length = min(150, 0.75 \* word\_count)
  + min\_length = min(5, 0.3 \* word\_count)
  + do\_sample = False (Ensuring deterministic summarization)

**1.4 Context Mismatch Detection**

* A **Sentence Transformer model** (all-MiniLM-L6-v2) is used to compare comments with their parent.
* If **semantic similarity** between a comment and its parent is **< 0.5**, it is flagged as a mismatch.

**2. Model Training**

**2.1 Training Objective**

The model aims to:

* Reconstruct full discussion threads from fragmented comments.
* Generate meaningful summaries.
* Evaluate the quality of the generated summaries using NLP metrics.

**2.2 Performance Metrics**

1. **BLEU Score**: Measures word overlap between original and summarized text.
2. **ROUGE Score**: Evaluates the recall of key phrases in the summary.
3. **Perplexity**: Measures how well a language model predicts the summary’s fluency.
4. **Semantic Similarity**: Computes cosine similarity between the reconstructed thread and summary.

**2.3 Perplexity Calculation**

* A **GPT-2 language model** is used to compute perplexity.
* Perplexity is calculated as exp(loss), where loss is derived from GPT-2’s log-likelihood.
* **Handling NaN/Inf values:**
  + Short texts (<5 words) are ignored to prevent distorted calculations.
  + Infinite perplexities are replaced with NaN to avoid bias.

**2.4 Semantic Similarity Calculation**

* The same all-MiniLM-L6-v2 model is used to calculate cosine similarity between the original thread and generated summary.
* This evaluates how well the summary captures the original discussion’s context.

**2.5 Final Results**

The following averages are computed to assess model performance:

* **Average BLEU Score**
* **Average ROUGE-L Score**
* **Average Perplexity**
* **Average Semantic Similarity**

The processed data and results are saved to CSV files for further analysis.