**Data Preprocessing and Model Training**

**1. Data Preprocessing**

**Dataset Loading**

* The dataset final\_labels.csv is loaded.
* Columns used: body, level\_1, level\_2, split, parent\_body.

**Contextual Combination**

* A new column context is created by concatenating parent\_body and body with a [SEP] separator to provide context.

**Label Encoding**

* level\_1 and level\_2 labels are mapped to integer values using dictionaries.

**Data Splitting**

* Data is split into **80% training** and **20% testing** using train\_test\_split().

**2. Dataset Creation**

A PyTorch Dataset class is implemented:

* Tokenizes text using roberta-large tokenizer.
* Stores input\_ids, attention\_mask, level\_1\_label, and level\_2\_label.

**3. Model Architecture**

**Base Model: roberta-large**

* **Two classifiers:**
  + classifier\_1 predicts level\_1 category.
  + classifier\_2 predicts level\_2 subcategory.
* **Loss Function:**
  + CrossEntropy loss for both levels.
  + Final loss = loss\_1 + loss\_2.

**4. Evaluation Metrics**

The compute\_metrics() function calculates:

* **Overall Accuracy:** Both level\_1 and level\_2 predictions are correct.
* **Precision:** Weighted average precision for both levels.
* **Recall:** Weighted average recall for both levels.
* **F1 Score:** Weighted average F1 score for both levels.

**5. Training Setup**

* **Epochs:** 4
* **Batch Size:** 8
* **Optimizer:** AdamW with weight decay (0.01)
* **Early Stopping:** Based on overall\_accuracy.

**6. Explainability & Fairness**

**SHAP Explainability**

* Uses SHAP (SHapley Additive exPlanations) to highlight important tokens in text predictions.
* **Metric:** Mean Absolute SHAP Value.

**Moderation Suggestions**

* Provides alternative, positive phrasing for potentially toxic comments.

**7. Example Outputs**

| **Feature** | **Result** |
| --- | --- |
| Explanation | SHAP highlights key tokens in text |
| Explainability Metric | Mean Absolute SHAP Score |
| Moderation Suggestion | Suggests rewording harsh comments positively |

**8. Next Steps**

1. **Fine-tune model** on more domain-specific data.
2. **Bias detection** for fairness improvement.
3. **Visualize SHAP explanations** interactively.
4. **Hyperparameter tuning** for better accuracy.