AG News Topic Classification Using BERT

1. Problem Statement

The rapid growth of online news has made it challenging for users to efficiently sort and access relevant articles. Manual classification of news headlines into categories such as World, Sports, Business, and Science/Technology is time-consuming and prone to errors. This project addresses the need for an automated system that accurately classifies news headlines, improving information retrieval and user experience.

2. Objective

The primary objectives of this project are:

1. Develop an automated news classification system using a pre-trained BERT transformer model**.**
2. Fine-tune the model on the AG News dataset for accurate prediction of four categories: World, Sports, Business, and Science/Technology.
3. Evaluate the model using accuracy and F1-score to ensure reliability.
4. Deploy the model using Gradio**,** enabling real-time user interaction.
5. Demonstrate the effectiveness of transfer learning in NLP-based text classification tasks.

3. Dataset Loading & Preprocessing

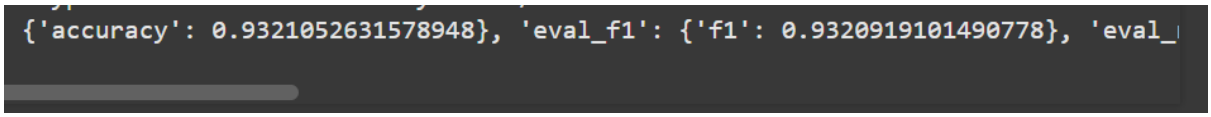
- **Dataset:** AG News dataset from Hugging Face Datasets.
- **Contents:** 4 classes – World, Sports, Business, Sci/Tech; over 120,000 training samples and 7,600 test samples.
- **Preprocessing Steps**
    - Load dataset using datasets.load_dataset("ag_news").
    - Tokenize headlines using BERT tokenizer (bert-base-uncased) with truncation and padding.
    - Prepare **DataCollatorWithPadding** for dynamic batching during training.

4. Model Development & Training

- **Model:** Pre-trained BERT (bert-base-uncased) with a classification head for 4 labels.
- **Fine-Tuning:**
    - Training using **Hugging Face Trainer** API.
    - Hyperparameters:
        - batch_size: 16
        - num_epochs: 2–3
        - learning_rate: 2e-5
        - weight_decay: 0.01
    - Training subset used for faster iterations (~20k samples for training, 5k for evaluation).

5. Evaluation with Relevant Metrics

- Metrics used:
  - **Accuracy**: Measures proportion of correct predictions.
  - **F1-score (weighted)**: Balances precision and recall across imbalanced classes.



6. Gradio Deployment

- An interactive interface built using **Gradio**.
- Users can input a news headline and get predicted category along with confidence scores.
- **Sample interface features:**
  - **Input:** Textbox for news headline
  - **Output:** Label showing probabilities for World, Sports, Business, Sci/Tech
  - Shareable public link for demo.