Sentiment Analysis with BERT

Report

Introduction to Machine Learning and NLP

Pipeline Explanation

The code implements a sentiment analysis pipeline using a pre-trained BERT model fine-tuned on the IMDb movie review dataset. The dataset is loaded using the Hugging Face `datasets` library and preprocessing is performed with the `bert-base-uncased` tokenizer. Tokenization involves padding and truncating each input to a maximum length of 512 tokens, ensuring compatibility with the BERT architecture.

The model is fine-tuned using the `Trainer` API provided by Hugging Face's `transformers` library. A batch size of 8 and 2 epochs are used for training to ensure that the code can run on limited memory environments such as Google Colab or mid-range GPUs (as in case of my Laptop). During training, evaluation metrics such as accuracy and F1-score are computed to assess model performance. These metrics are also evaluated explicitly on the test set after training.

The pipeline is modular, cleanly structured and includes a demonstration of inference on a sample input. The model and tokenizer are saved locally and can be reused without retraining.

Challenges and Difficulties Faced

Below are some challenges I encountered in building this pipeline:

1. **Dataset Loading Errors**:
   Some environments (especially Colab without secrets) may throw warnings or errors while downloading datasets using `load_dataset("imdb")`.
2. **Hugging Face `datasets` issues**:
   When using newer versions, `set_format(type="torch")` is necessary for `Trainer` to work correctly with tokenized datasets. Skipping this results in runtime errors during training.
3. **Circular Import Errors (e.g., in pandas)**:
   This was observed locally when environment or conda packages conflicted. Reinstalling `pandas` or running in a clean environment helped resolve it.

4. **Environment Setup**:
   Installing compatible versions of `transformers`, `datasets`, `scikit-learn` and `torch` is critical. Local environments without GPU or limited pip versions had trouble running the full IMDb dataset.
5. **Colab vs Local Conflicts**:
   While Colab has most dependencies pre-installed, local environments often require manual setup with pip. GPU access also affects performance drastically.
6. **HF_TOKEN and Authentication Warnings**:
   Though not blocking for public datasets/models, these can confuse users during the first-time run. They are harmless but frequent.
7. **Training Time**:
   Full IMDb dataset is large (25k+ samples). Running for multiple epochs without a GPU can be very slow. A subset may be used for testing.
8. **Evaluation Bug**:
   If metrics like `accuracy` and `f1_score` are not explicitly returned in the `compute_metrics` function or if labels are not properly formatted, evaluation will silently fail.

Resolution

All the above issues were handled through:

- Careful environment setup and dependency control.
- Breaking down the pipeline into modular functions.
- Testing on both Colab and local environments.
- Using subsets of the dataset for debugging.
- Explicit error handling and adding `set_format("torch")`.