Assignment 8: Sentiment Analysis with BERT

Overview

This assignment implements sentiment analysis on the Stocktwits dataset using BERT for binary classification of investor messages as bullish or bearish.

Score: 95/100 ⭐⭐

Dataset

Stocktwits Dataset

- `stocktwits_train_100k.csv` - 100,000 training records
- `stocktwits_test_20k.csv` - 20,000 test records

Columns:

- **bull**: `1` = Bullish, `-1` = Bearish (target variable)
- **len**: Number of words (use to determine `max_seq_len`)
- **msg**: Message text (input)
  - `$` prefix = company ticker symbol
  - Contains HTML entities (`&gt`, `&#39;`)

Preprocessing: Convert HTML entities to improve performance

- Reference: https://www.w3schools.com/HTML/html_entities.asp

Task Requirements

- **Input**: `msg` column
- **Target**: `bull` column
- **Epochs**: 20
- Design your own classifier on top of BERT
- Plot training and validation accuracy/loss curves

Code Requirements

⚠️ **CRITICAL**: Must use and modify reference code or get **0 points**

Reference Code:

- https://colab.research.google.com/drive/1pTuQhug6Dhl9XalKB0zUGf4FIdYFlpcX#scrollTo=BJR6t_gCQe_x
- http://mccormickml.com/2019/07/22/BERT-fine-tuning/

Use new **Hugging Face API** to simplify:

3.3 Tokenize Dataset

```python
max_length = 64
inputs = tokenizer(sentences,
                   padding=True,
                   max_length=max_length,
                   truncation=True,
                   return_tensors="pt")
print(inputs)
```

3.4 Training & Validation Split

```python
labels = torch.tensor(labels)
dataset = TensorDataset(inputs['input_ids'],
                        inputs['attention_mask'],
                        labels)
```

Penalties

- **-10 points**: No results in notebook
- **-10 points**: Not using template or not filling fields
- **0 points**: Not using reference code

Deliverables

- Jupyter notebook with:
  - Data preprocessing
  - BERT tokenization
  - Custom classifier
  - 20 epochs training
  - Accuracy/loss plots
  - Results displayed

Important:

- ✅ Use reference code
- ✅ Use template
- ✅ Show results
- ❌ Don't compress file

Requirements

```bash
pip install transformers torch pandas numpy matplotlib scikit-learn
```

Note: Training is very time consuming. Use GPU if available.

Assignment completed as part of Deep Learning coursework