Báo cáo tóm tắt Lab02

1. Đề bài

Xây dựng một mô hình học sâu để phân loại cảm xúc (sentiment classification) từ dữ liệu đánh giá phim IMDB. Mục tiêu là phân loại các review thành hai nhãn: **positive** và **negative**.

2. Tiền xử lý dữ liệu

- **Tải dữ liệu**: Dữ liệu được lưu trữ trên Google Drive và được đọc dưới dạng file CSV (IMDB Dataset.csv).
- **Khám phá dữ liệu**:
    - Dữ liệu gốc có 50,000 dòng với 2 cột: review và sentiment.
    - Phân phối nhãn: 25,000 positive và 25,000 negative.
- **Cân bằng dữ liệu**:
    - Lấy mẫu ngẫu nhiên 5,000 dòng từ mỗi lớp để tạo thành tập dữ liệu cân bằng.
- **Tiền xử lý văn bản**:
    - Loại bỏ các thẻ HTML và ký tự không phải chữ cái.
    - Chuyển toàn bộ văn bản thành chữ thường.
    - Sử dụng bộ lọc stopwords (từ không cần thiết) và lemmatizer để chuẩn hóa từ.
- **Tách tập dữ liệu**:
    - Chia thành tập train (5,000 dòng) và test (5,000 dòng) theo tỷ lệ 50:50.
- **Tokenization và Padding**:
    - Token hóa văn bản (biểu diễn dưới dạng số nguyên) với tối đa 10,000 từ phổ biến nhất.
    - Đệm các câu về độ dài cố định 200 ký tự.

3. Mô hình

- **Kiến trúc mô hình**:

- Lớp Embedding để biểu diễn từ dưới dạng vector.
- Các lớp LSTM (Long Short-Term Memory) được bọc trong cấu trúc Bidirectional để nắm bắt thông tin theo cả hai chiều.
- Lớp Dropout để giảm thiểu overfitting.
- Lớp Dense (đầu ra) với hàm kích hoạt sigmoid để phân loại nhị phân.

- **Hàm mất mát**: Binary Crossentropy.
- **Đánh giá**: Độ chính xác (accuracy).

4. Siêu tham số

Các cấu hình siêu tham số đã thử nghiệm bao gồm:

1. **Baseline**:
   - Batch size: 64
   - Learning rate: 0.001
   - Số lớp ẩn: 1
   - Số neuron/lớp: 64
   - Dropout: 0.2
   - Optimizer: Adam
   - Epochs: 5
2. **Deeper Network**:
   - Số lớp ẩn: 2 (các thông số khác tương tự Baseline).
3. **RMSprop Optimizer**:
   - Optimizer: RMSprop (các thông số khác tương tự Baseline).
4. **Higher Learning Rate**:
   - Learning rate: 0.01 (các thông số khác tương tự Baseline).
5. **More Neurons**:
   - Số neuron/lớp: 128 (các thông số khác tương tự Baseline).

5. Kết quả

Cấu hình	Độ chính xác trung bình	Độ lệch chuẩn
Baseline	83.77%	0.51%
Deeper Network	82.25%	0.60%
RMSprop Optimizer	83.13%	0.65%
Higher Learning Rate	80.71%	1.05%
More Neurons	82.10%	0.86%

- **Cấu hình tốt nhất**: **Baseline** với độ chính xác trung bình là **83.77%** và độ lệch chuẩn thấp nhất (0.51%).

Confusion Matrix (Baseline):

- Negative: Precision = 79%, Recall = 88%.
- Positive: Precision = 86%, Recall = 76%.
- Accuracy = 82% trên tập test.

6. Nhận xét

- **Hiệu suất**: Mô hình Baseline với cấu hình đơn giản (1 lớp ẩn, Adam optimizer) đạt hiệu suất cao nhất trong các thử nghiệm. Điều này cho thấy việc tăng số lớp hoặc thay đổi optimizer không cải thiện đáng kể.
- **Dữ liệu**: Việc tiền xử lý văn bản giúp tăng hiệu quả của mô hình, đặc biệt là loại bỏ stopwords và lemmatization.
- **Hạn chế**: Precision và Recall chưa cân bằng giữa hai lớp, đặc biệt là với lớp Positive. Có thể thử các kỹ thuật như tăng cường dữ liệu (data augmentation) hoặc điều chỉnh trọng số lớp (class weights) để cải thiện.
- **Hướng phát triển**:

- Thử nghiệm thêm các kiến trúc mô hình khác như GRU hoặc Transformer.
- Tăng thời gian huấn luyện (nhiều epoch hơn) hoặc áp dụng kỹ thuật fine-tuning trên các mô hình ngôn ngữ lớn như BERT.