BÁO CÁO HUẤN LUYỆN VÀ ĐÁNH GIÁ MÔ HÌNH PHÂN LOẠI TIN TỨC VỚI PhoBERT

1. Mục tiêu

Huấn luyện mô hình phân loại văn bản sử dụng PhoBERT-base với 3 cấu hình siêu tham số khác nhau.
Sau đó:

- Tính độ chính xác trung bình (mean accuracy) trên tập validation.
- Tính độ lệch chuẩn (standard deviation) để đánh giá độ ổn định.
- Ghi log biểu đồ Loss, Accuracy từng epoch để theo dõi quá trình huấn luyện.

2. Chuẩn bị dữ liệu

- Dataset: Dữ liệu tin tức từ báo Dân Trí (preprocessed_dan_tri.csv), đã được tiền xử lý.
- Tiền xử lý:
    - Chuyển nhãn (label) thành số nguyên bằng LabelEncoder.
    - Chia dữ liệu thành:
        - Train: 72%
        - Validation: 8%
        - Test: 20%

3. Mô hình

- Backbone: PhoBERT-base (vinai/phobert-base).
- Kiến trúc:
    - Thêm các lớp fully connected layers sau PhoBERT (num_layers thay đổi tùy cấu hình).
    - Dropout: 0.3 để tránh overfitting.
- Loss: CrossEntropyLoss.
- Optimizer: AdamW.
- Scheduler: ReduceLROnPlateau theo validation loss.

4. Cấu hình siêu tham số

Cấu hình	Batch size	Số epoch	Số lớp FC	Learning rate
1	32	3	1	3e-5
2	16	5	2	2e-5
3	8	4	3	5e-5

5. Kết quả đánh giá

Cấu hình	Validation Accuracy Mean	Validation Accuracy Std
1	0.9797	0.0028
2	0.9747	0.0057
3	0.9701	0.0073

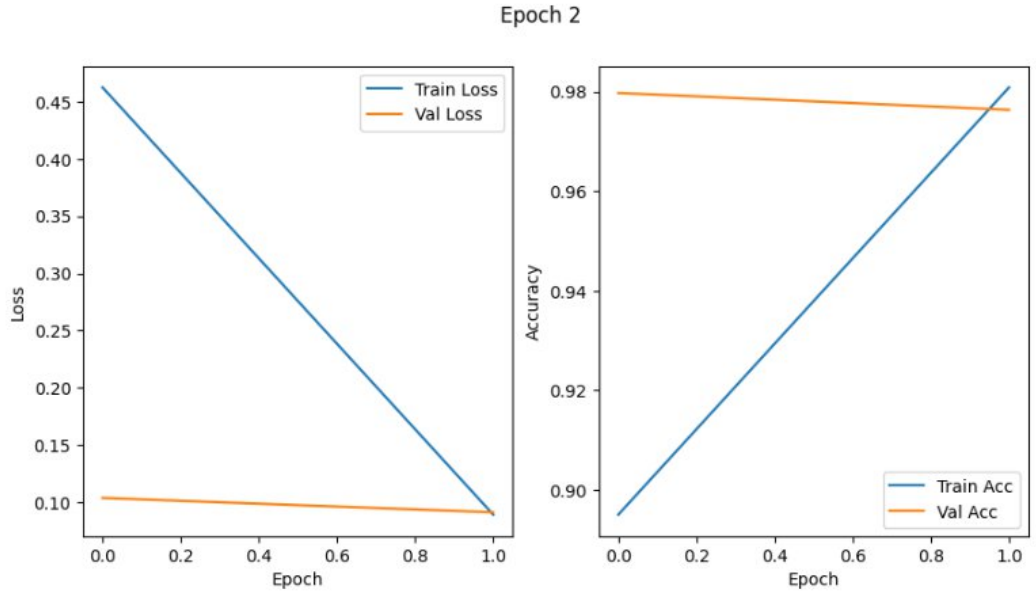
6. Phân tích kết quả

- Cấu hình 1 đạt độ chính xác cao nhất: 97.97% với độ lệch chuẩn thấp (0.28%), chứng tỏ mô hình rất ổn định qua nhiều lần huấn luyện.
- Cấu hình 2 mặc dù train nhiều epoch hơn và có thêm 1 lớp FC, nhưng accuracy giảm nhẹ, độ lệch chuẩn tăng ➔ Có dấu hiệu bắt đầu overfitting nhẹ.
- Cấu hình 3 với batch size nhỏ hơn (8) và learning rate lớn hơn (5e-5) dẫn đến accuracy thấp nhất (97.01%) và độ lệch chuẩn cao nhất (0.73%) ➔ Cấu hình này dễ bị overfitting, hiệu suất không cao khi training.

7. Biểu đồ huấn luyện

Epoch 2/3

✅ Train acc: 0.9808, loss: 0.0892 | Val acc: 0.9764, loss: 0.0913

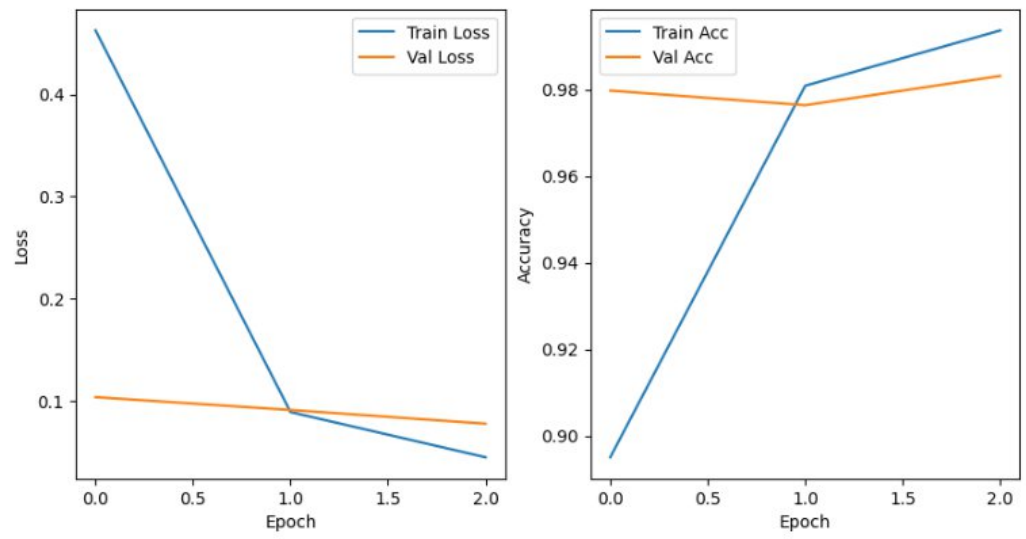


Epoch 3/3

✅ Train acc: 0.9936, loss: 0.0450 | Val acc: 0.9831, loss: 0.0779

Epoch 3/3

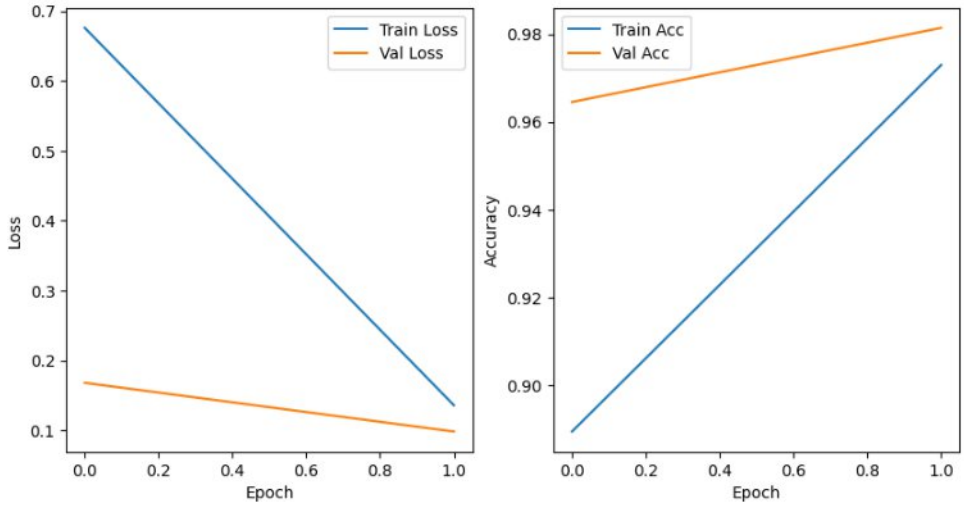
✅ Train acc: 0.9936, loss: 0.0450 | Val acc: 0.9831, loss: 0.0779



Epoch 2/5

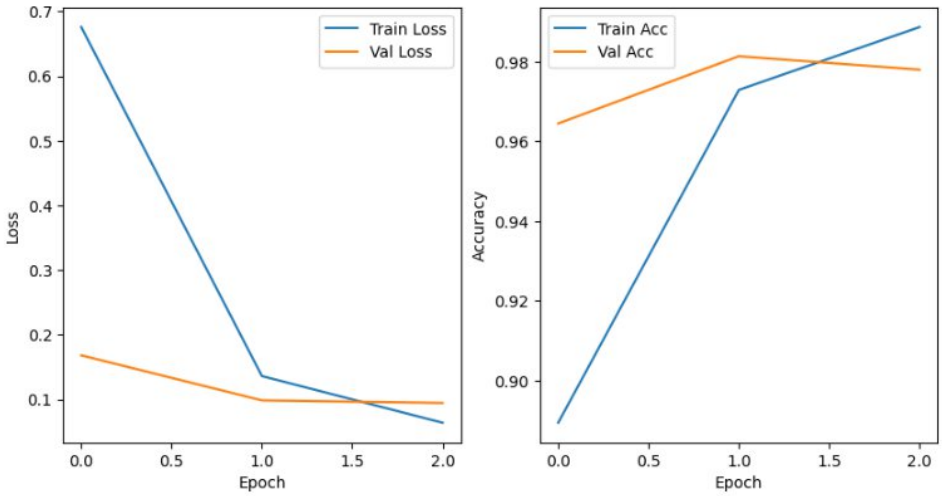
✅ Train acc: 0.9730, loss: 0.1361 | Val acc: 0.9814, loss: 0.0985

Epoch 2



Epoch 3/5

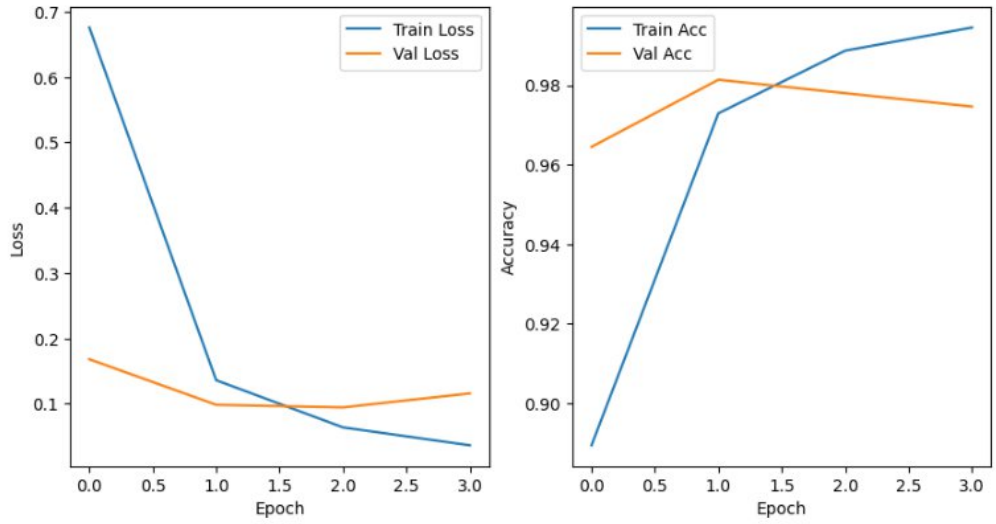
Epoch 3



Epoch 4/5

✅ Train acc: 0.9946, loss: 0.0365 | Val acc: 0.9747, loss: 0.1160

Epoch 4



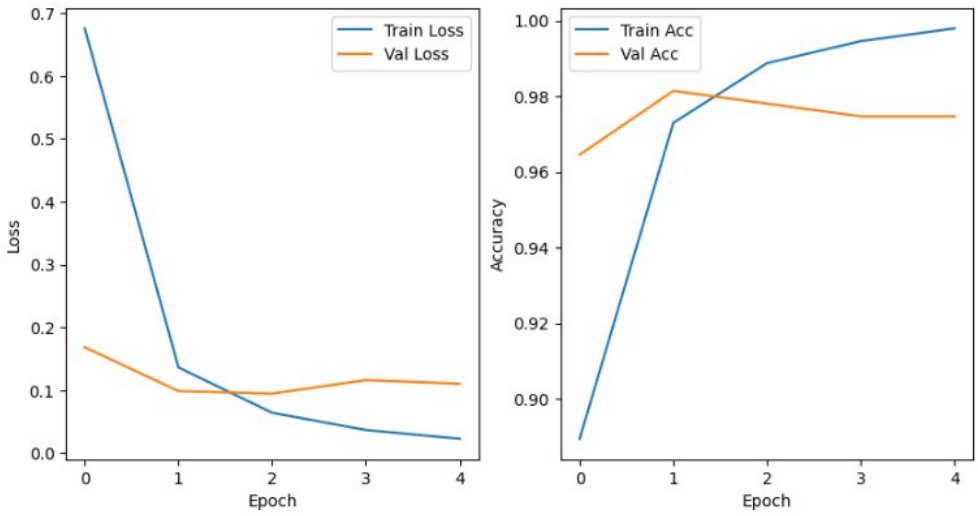
Epoch 5/5

✅ Train acc: 0.9979, loss: 0.0226 | Val acc: 0.9747, loss: 0.1102

Epoch 5/5

✅ Train acc: 0.9979, loss: 0.0226 | Val acc: 0.9747, loss: 0.1102

Epoch 5



- Độ dốc xuống rất nhanh chứng tỏ mô hình thật sự mạnh mẽ

8. Kết luận

- Cấu hình tốt nhất: Cấu hình 1 (batch_size=32, epochs=3, num_layers=1, lr=3e-5).
- Với mô hình PhoBERT, chỉ cần thêm một lớp FC đơn giản là đã đủ để đạt độ chính xác rất cao trên tập validation.
- Việc thêm nhiều lớp hoặc train lâu hơn không giúp cải thiện, ngược lại còn làm giảm độ ổn định.

- Cấu hình 3 cho thấy việc sử dụng quá ít sample trong batch và learning rate lớn có thể dẫn đến overfitting và giảm hiệu suất.

9. Hướng phát triển tiếp theo

- Fine-tuning thêm PhoBERT (unfreeze layers).
- Áp dụng kỹ thuật early stopping để tránh overfitting.
- Thử nghiệm các kỹ thuật tăng cường dữ liệu (data augmentation).
- Đánh giá chi tiết theo từng nhãn (classification report, confusion matrix).

🎯 Tổng kết

PhoBERT + đơn giản hóa đầu ra → đạt độ chính xác rất cao (gần 98%) cho bài toán phân loại tin tức tiếng Việt.