Báo cáo

1. Đề bài.

Bài toán đặt ra là xây dựng một hệ thống phân loại bài báo tiếng Việt từ các trang báo điện tử lớn như VnExpress và Dân Trí.

Hệ thống gồm các bước:

- Thu thập dữ liệu tin tức từ nhiều chuyên mục.
- Tiền xử lý dữ liệu bằng tokenizer tiếng Việt.
- Vector hóa văn bản bằng TF-IDF.
- Huấn luyện mô hình phân loại dựa trên PyTorch.
- Đánh giá và so sánh các cấu hình siêu tham số khác nhau.

2. Thu thập dữ liệu.

2.1. Crawl dữ liệu từ VnExpress.

Sử dụng thư viện requests, BeautifulSoup, newspaper3k.

Crawl từ 30 chuyên mục như Thời sự, Kinh doanh, Thể thao, Công nghệ...

Lưu trữ mỗi bài báo gồm: Tiêu đề, Nội dung, Nhãn (chuyên mục).

Crawl tối đa 10 trang mỗi chuyên mục (~ khoảng vài trăm bài mỗi chuyên mục).

2.2. Crawl dữ liệu từ Dân Trí.

Tương tự, crawl từ 6 chuyên mục lớn như Xã hội, Kinh doanh, Thể thao...

Crawl tối đa 20 trang mỗi chuyên mục.

3. Tiền xử lý dữ liệu.

3.1. Làm sạch văn bản.

Xóa bỏ HTML tags, ký tự đặc biệt, chuẩn hóa chữ thường.

Tách từ tiếng Việt bằng underthesea.word_tokenize.

3.2. Encode label.

Ánh xạ tên chuyên mục (string) thành số nguyên (LabelEncoder).

Lưu lại label2id và id2label vào file JSON để dễ sử dụng.

3.3. Vector hóa văn bản.

Áp dụng TF-IDF (TfidfVectorizer) với tối đa 5000 features.

Dữ liệu đưa vào TF-IDF là: Tiêu đề + Nội dung.

3.4. Chia tập train/test.

Chia cố định 80% train, 20% test (train_test_split).

Giữ tỉ lệ nhãn đồng đều giữa train/test bằng stratify=y.

4. Huấn luyện mô hình.

4.1. Mô hình sử dụng.

Mô hình phân loại đơn giản (NewsClassifier) gồm:

3 lớp Linear (Fully Connected).

Hàm kích hoạt ReLU.

Sử dụng CrossEntropyLoss và Adam Optimizer.

4.2. Cố định các yếu tố ngẫu nhiên.

Set random.seed, numpy.random.seed, torch.manual_seed để đảm bảo tính reproducibility.

4.3. Theo dõi bằng Weights & Biases (wandb).

Log loss, accuracy mỗi epoch.

Quản lý các lần chạy với các cấu hình khác nhau.

4.4. Các cấu hình thử nghiệm.

Config	Learning Rate	Batch Size	Hidden Size
1	0.01	32	128
2	0.005	64	256
3	0.001	32	64

Mỗi cấu hình chạy 3 lần để lấy trung bình và độ lệch chuẩn accuracy.

5. Kết quả.

Config	Accuracy Trung Bình	Độ Lệch Chuẩn
1	90.25%	0.26%
2	90.45%	0.34%
3	90.53%	0.20%

Nhận xét:

- Config 3 đạt độ chính xác cao nhất (90.53%) và ổn định nhất (độ lệch chuẩn thấp nhất).
- Learning rate nhỏ (0.001) giúp mô hình hội tụ tốt hơn, tránh bị overfitting.
- Hidden size nhỏ (64) vẫn đủ sức học các đặc trưng từ TF-IDF mà không bị quá phức tạp.

6. Kết luận và hướng phát triển.

Kết luận: Bài toán phân loại tin tức tiếng Việt có thể được giải quyết hiệu quả bằng TF-IDF + MLP đơn giản.

Hướng phát triển:

- Thử embedding từ Word2Vec, FastText, hoặc BERT.
- Sử dụng mô hình phức tạp hơn như RNN, LSTM, Transformer.
- Ứng dụng kỹ thuật Early stopping hoặc Dropout để cải thiện độ ổn định hơn nữa.
- Crawl thêm dữ liệu từ nhiều nguồn báo khác.