Báo cáo Tóm tắt Kết quả Phân loại Tin tức Tiếng Việt

1. Đề bài:

- Xây dựng một mô hình học máy để phân loại tin tức tiếng Việt vào các thể loại khác nhau (thời sự, thế giới, bất động sản, giáo dục).
- Sử dụng tập dữ liệu được thu thập từ trang web VnExpress.
- Đánh giá mô hình dựa trên các độ đo Accuracy, F1-score, và Confusion Matrix.

2. Tiền xử lý dữ liệu:

- **Làm sạch văn bản:** Loại bỏ HTML, các ký tự đặc biệt và dấu câu thừa.
- **Chuẩn hóa Unicode:** Sử dụng NFC để chuẩn hóa Unicode.
- **Tách từ:** Sử dụng thư viện underthesea để tách từ tiếng Việt.
- **Xây dựng từ điển:** Tạo từ điển từ các từ xuất hiện trong dữ liệu huấn luyện.
- **Chuyển đổi văn bản thành số:** Biểu diễn văn bản bằng các chỉ số tương ứng trong từ điển.
- **Chia dữ liệu:** Chia dữ liệu thành tập huấn luyện, tập validation, và tập kiểm tra.

3. Mô hình:

- Mô hình được sử dụng là **LSTM kết hợp với cơ chế Attention**.
- LSTM được sử dụng để trích xuất các đặc trưng từ chuỗi văn bản.
- Cơ chế Attention giúp mô hình tập trung vào các phần quan trọng của văn bản để đưa ra dự đoán chính xác hơn.

4. Siêu tham số:

- Đã thử nghiệm với 3 cấu hình siêu tham số khác nhau (config_1, config_2, config_3).
- Các siêu tham số bao gồm:
    - hidden_dim: Số đơn vị ẩn trong LSTM.
    - attention_dim: Số chiều của lớp Attention.
    - embedding_dim: Số chiều của vector nhúng từ.
    - num_layers: Số lớp LSTM.
    - batch_size: Kích thước batch.

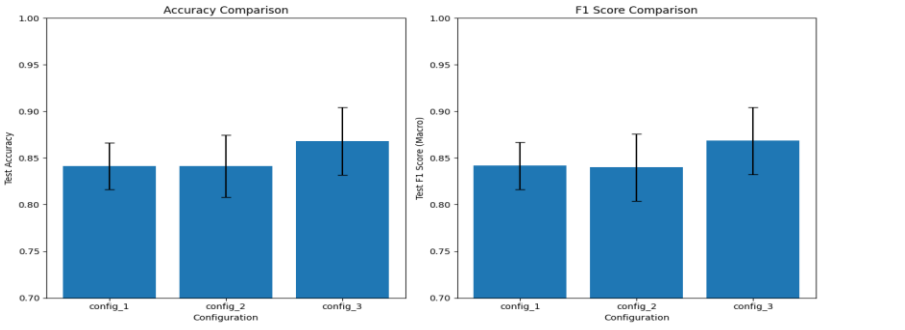
- dropout: Tỷ lệ dropout.
- learning_rate: Tốc độ học.
- optimizer: Bộ tối ưu hóa (Adam, AdamW, RMSprop).

5. Kết quả:

Config	Avg Test Accuracy	Std Test Accuracy	Avg Test F1	Std Test F1
config_1	0.8414	0.0249	0.8418	0.0254
config_2	0.8414	0.0331	0.8399	0.0359
config_3	0.8683	0.0363	0.8685	0.0359

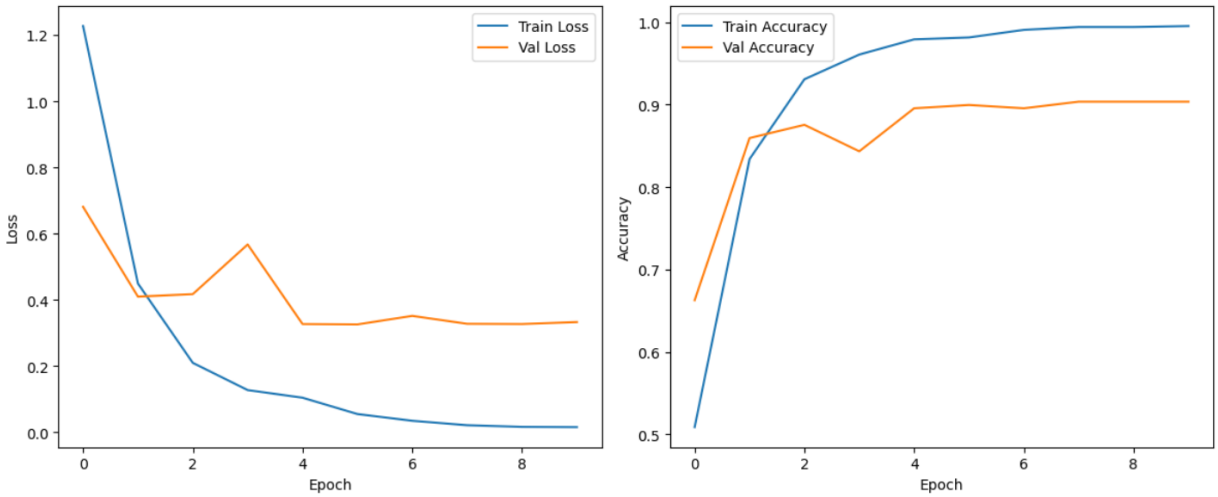
6. Nhận xét:

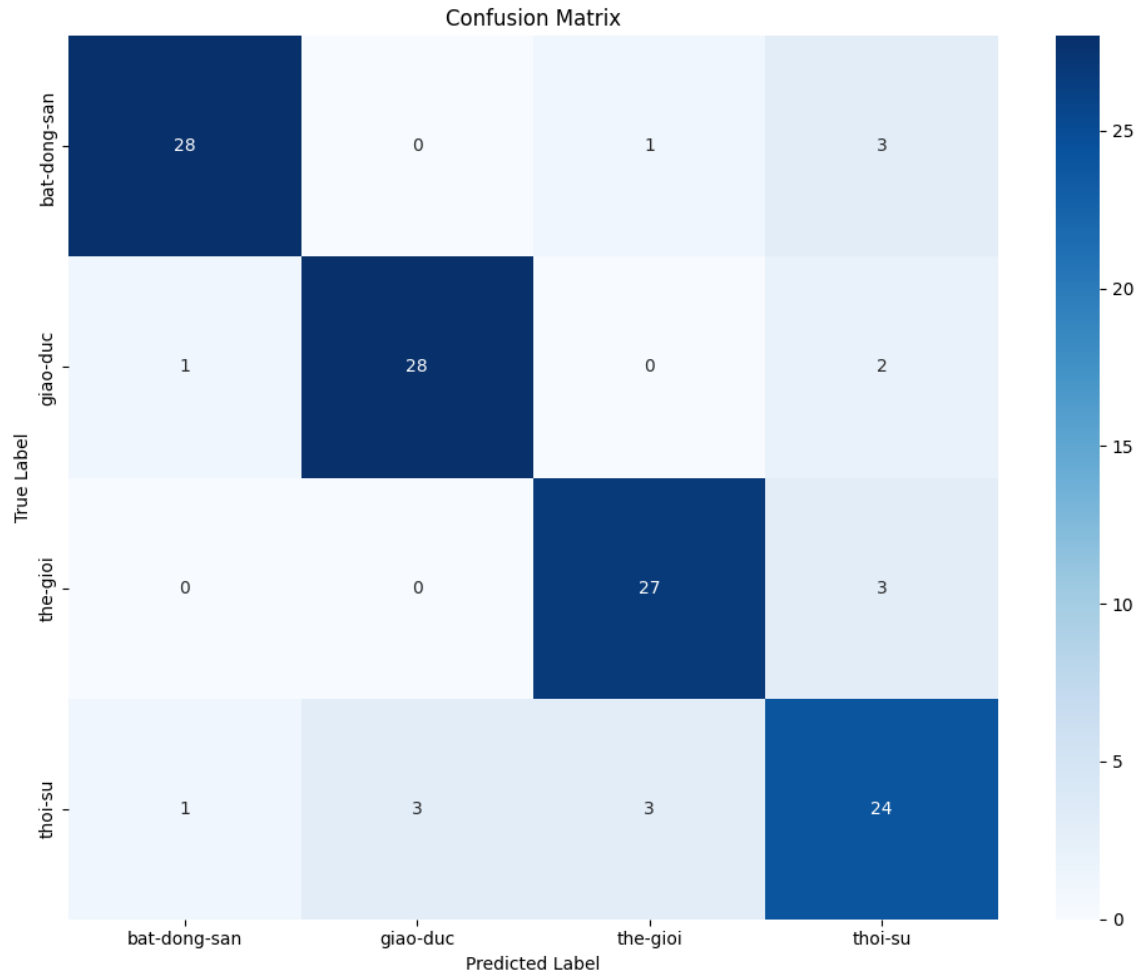
- Mô hình **config_3** đạt được kết quả tốt nhất với **Accuracy trung bình là 0.8683** và **F1-score trung bình là 0.8685**.
- Việc sử dụng tốc độ học khác nhau và bộ tối ưu hóa RMSprop đã cải thiện đáng kể hiệu suất của mô hình.
- Mô hình **config_2** (Larger model) không mang lại sự cải thiện đáng kể so với mô hình baseline.
- Có thể tiếp tục cải thiện mô hình bằng cách thử nghiệm với các cấu hình siêu tham số khác, sử dụng kỹ thuật tăng cường dữ liệu, hoặc tinh chỉnh kiến trúc mô hình.



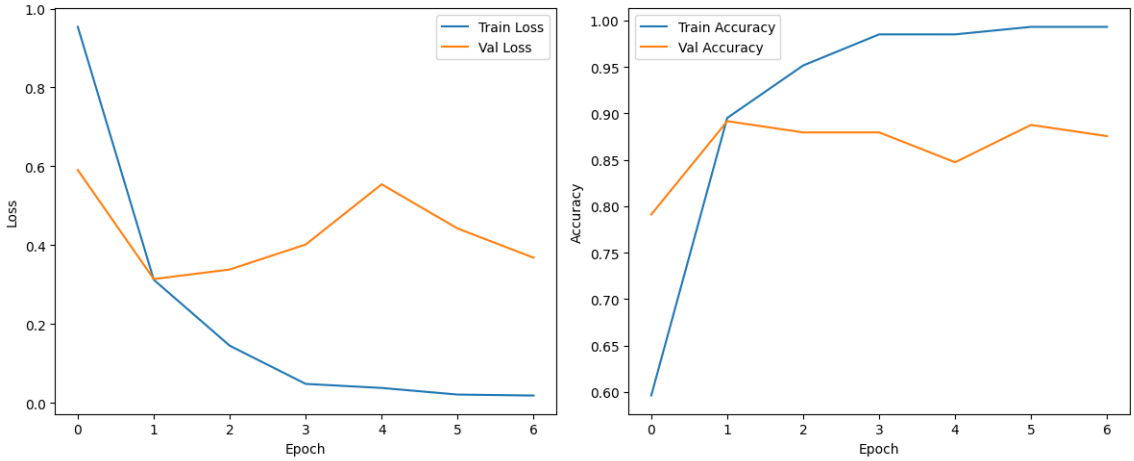
So sánh accuracy và f1-score

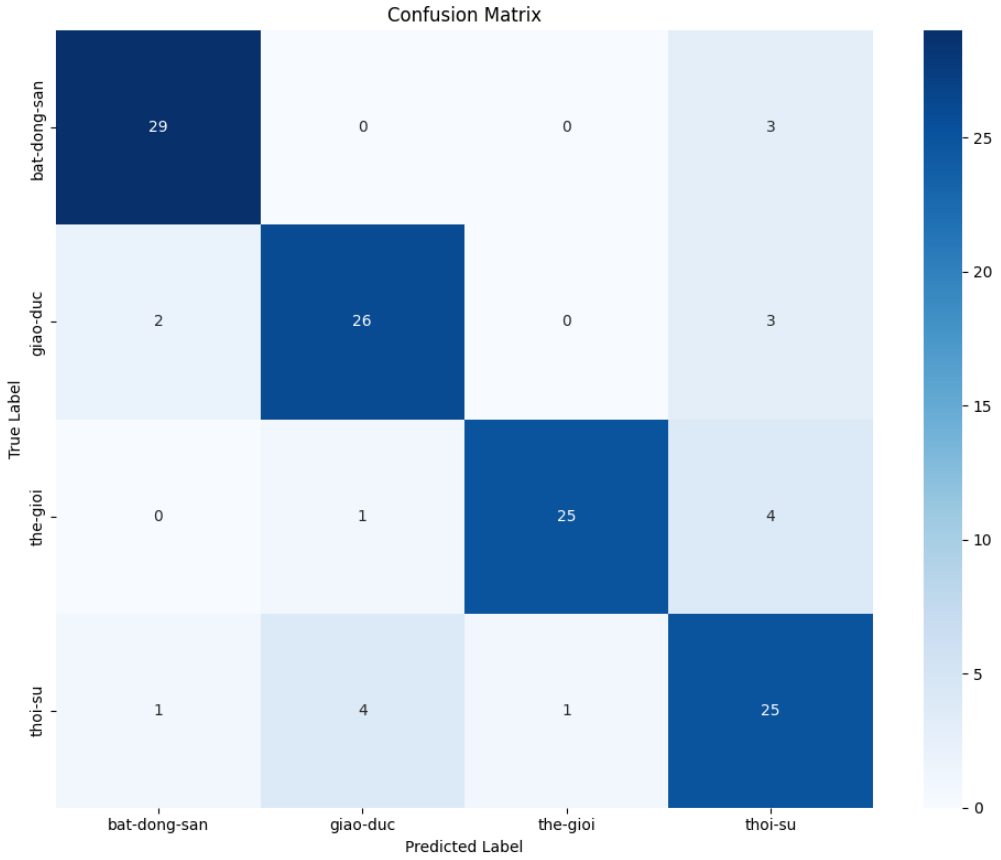
Một số biểu đồ, confusion maxtrix đánh giá các config



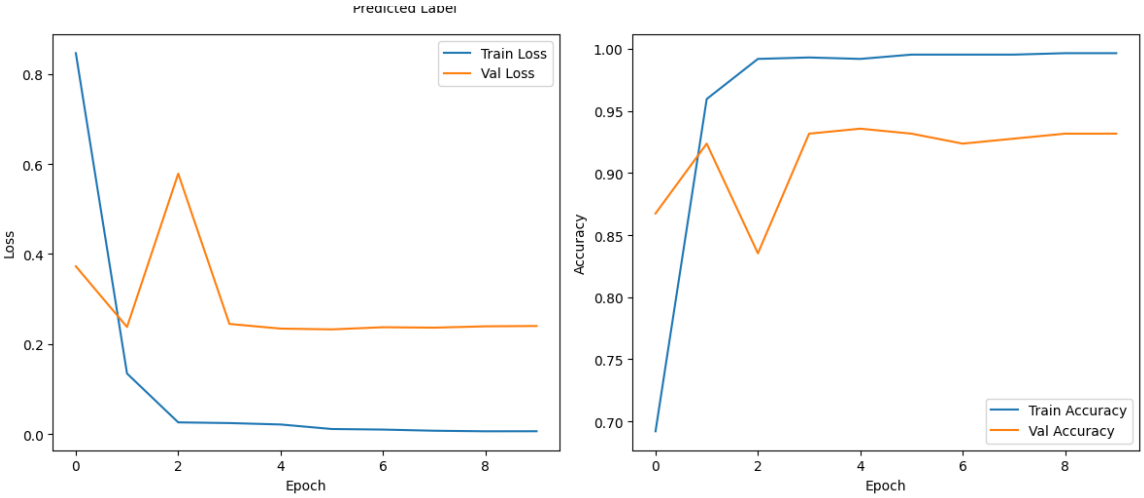


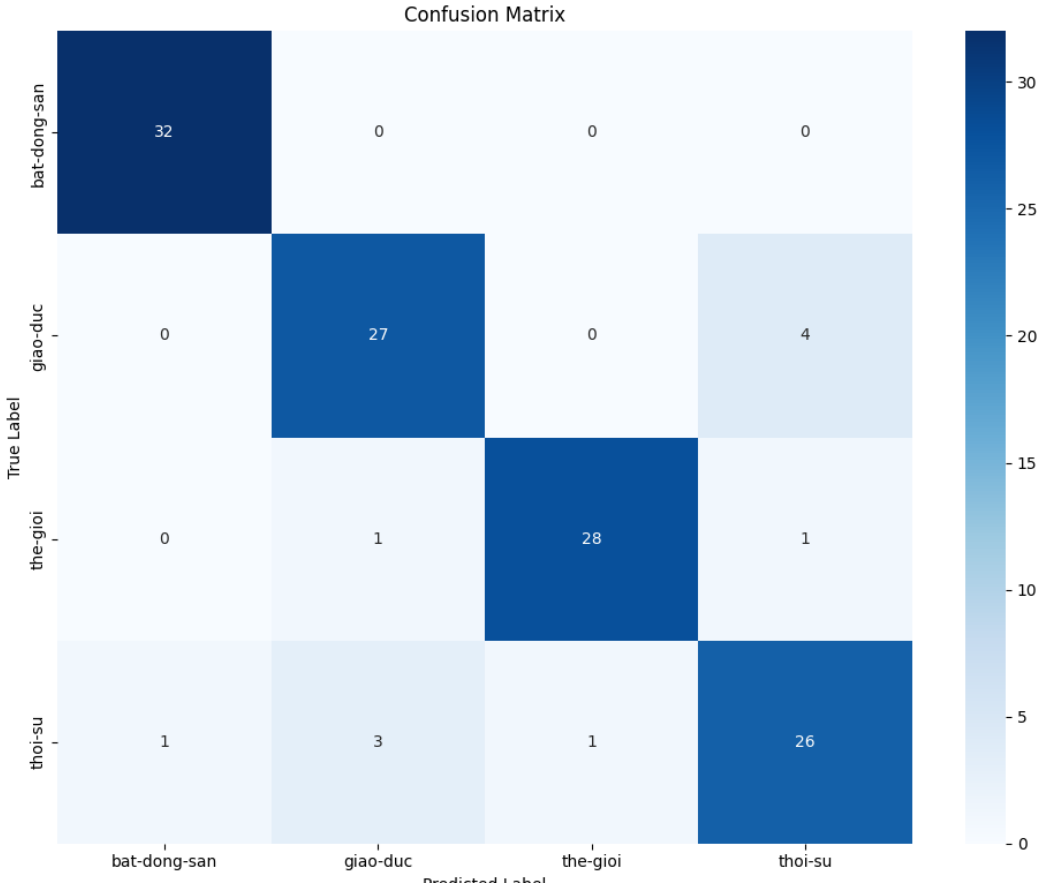
Config 1 – run 1





Config 2 – run 1





Config 3 – run 1