KẾT LUẬN

Đồ án đã xây dựng thành công một hệ thống nhận dạng và phân loại giấy tờ tiếng Việt trên cơ sở kết hợp công nghệ OCR với mô hình ngôn ngữ hiện đại. Cụ thể, hệ thống sử dụng PaddleOCR để trích xuất văn bản từ ảnh tài liệu, PhoBERT để chuyển văn bản thành vector ngữ nghĩa, và Logistic Regression để phân loại loại giấy tờ. Ngoài ra, ứng dụng còn được triển khai trên nền tảng Streamlit, giúp người dùng dễ dàng tải ảnh, xem kết quả OCR, và nhận phân loại trực tiếp qua giao diện web.

So với các giải pháp đơn thuần chỉ dựa trên OCR hoặc các mô hình phân loại truyền thống, hệ thống này mang lại ưu thế vượt trội nhờ khả năng hiểu ngữ nghĩa tiếng Việt. Các thử nghiệm cho thấy kết quả phân loại chính xác hơn và ổn định hơn so với khi chỉ dùng từ khóa hay đặc trưng thủ công. Bên cạnh đó, việc kết hợp PaddleOCR và Tesseract cũng giúp tăng độ tin cậy của văn bản trích xuất.

Trong quá trình thực hiện, sinh viên đã đạt được nhiều kết quả tích cực: xây dựng pipeline xử lý trọn vẹn từ ảnh đến nhãn tài liệu, áp dụng thành công PhoBERT vào bài toán phân loại tài liệu, và phát triển ứng dụng web thân thiện. Tuy nhiên, vẫn còn một số hạn chế như: tập dữ liệu huấn luyện còn hạn chế về số lượng và tính đa dạng, hệ thống chưa xử lý tốt chữ viết tay hoặc ảnh chất lượng thấp, và tốc độ xử lý phụ thuộc nhiều vào cấu hình phần cứng.

Nhìn chung, đồ án đã hoàn thành các mục tiêu đề ra, đưa ra một giải pháp khả thi và thực tiễn cho bài toán nhận dạng và phân loại giấy tờ trong bối cảnh chuyển đổi số hiện nay.

0.1 Hướng phát triển

Trong tương lai, hệ thống có thể được hoàn thiện và mở rộng theo các hướng sau:

- **Hoàn thiện và mở rộng dữ liệu huấn luyện:** Thu thập thêm nhiều loại giấy tờ đa dạng về bố cục, chất lượng ảnh để cải thiện độ chính xác phân loại.
- **Xử lý chữ viết tay:** Nghiên cứu tích hợp các mô hình OCR tiên tiến hơn (ví dụ TrOCR hoặc mô hình Transformer chuyên cho handwriting) để nhận dạng được cả chữ viết tay.
- **Cải thiện khả năng chịu lỗi:** Tăng khả năng xử lý đối với ảnh bị mờ, nghiêng, ánh sáng kém thông qua các kỹ thuật tiền xử lý nâng cao hoặc mô hình OCR mạnh hơn.
- **Phân loại nâng cao:** Thay Logistic Regression bằng các mô hình học sâu tiên

tiến (ví dụ Fine-tune PhoBERT hoặc BERT phân loại trực tiếp) để tăng độ chính xác.

- **Triển khai thực tế:** Đưa hệ thống lên môi trường server hoặc cloud, tích hợp API để kết nối với các hệ thống quản lý tài liệu, hành chính điện tử hoặc phần mềm doanh nghiệp.
- **Giao diện người dùng:** Phát triển thêm các chức năng nâng cao trên ứng dụng web như quản lý dữ liệu huấn luyện, báo cáo thống kê, và tuỳ chọn cấu hình mô hình.

Với những hướng phát triển trên, hệ thống không chỉ dừng lại ở mức thử nghiệm mà còn có khả năng trở thành một giải pháp thực tiễn, hỗ trợ đắc lực cho quá trình số hóa và quản lý tài liệu trong nhiều lĩnh vực.