TÓM TẮT NỘI DUNG ĐỒ ÁN

Trong bối cảnh chuyển đổi số hiện nay, việc xử lý và quản lý giấy tờ thủ công gặp nhiều hạn chế như tốn thời gian, dễ sai sót và khó lưu trữ, tìm kiếm. Các hướng tiếp cận truyền thống dựa vào nhập liệu thủ công hay những hệ thống OCR cơ bản chỉ mới dừng ở mức nhận dạng ký tự, chưa đảm bảo độ chính xác cao khi phân loại đa dạng loại giấy tờ như căn cước công dân, hóa đơn, hợp đồng… Bên cạnh đó, nhiều giải pháp hiện có chưa khai thác được sức mạnh của các mô hình ngôn ngữ hiện đại, dẫn đến hạn chế trong việc hiểu ngữ cảnh và nội dung của văn bản trích xuất.

Để khắc phục những hạn chế này, đề tài lựa chọn hướng tiếp cận kết hợp công nghệ OCR hiện đại và mô hình ngôn ngữ tiền huấn luyện. Cụ thể, hệ thống sử dụng PaddleOCR để trích xuất văn bản từ ảnh giấy tờ, sau đó ứng dụng PhoBERT – một mô hình ngôn ngữ mạnh cho tiếng Việt – để chuyển văn bản thành vector đặc trưng. Tiếp theo, các vector này được đưa vào mô hình Logistic Regression nhằm huấn luyện và phân loại các loại giấy tờ. Việc lựa chọn hướng tiếp cận này vừa tận dụng được ưu điểm của OCR trong việc nhận diện ký tự, vừa phát huy khả năng hiểu ngữ nghĩa của PhoBERT, từ đó cải thiện đáng kể hiệu quả phân loại.

Giải pháp của sinh viên được triển khai thành hai phần: mô-đun huấn luyện cho phép xây dựng mô hình phân loại từ bộ dữ liệu giấy tờ, và ứng dụng giao diện web phát triển bằng Streamlit để người dùng dễ dàng tải ảnh, thực hiện OCR, xem văn bản trích xuất và kết quả phân loại. Hệ thống hỗ trợ cả PaddleOCR và Tesseract để tăng tính linh hoạt.

Đóng góp chính của ĐATN là xây dựng một quy trình tự động, hiệu quả và thân thiện với người dùng cho bài toán nhận dạng và phân loại giấy tờ tiếng Việt. Kết quả thử nghiệm cho thấy mô hình hoạt động ổn định, nhận diện và phân loại chính xác nhiều loại giấy tờ phổ biến, góp phần giảm thiểu công sức nhập liệu thủ công và hỗ trợ quá trình số hóa tài liệu trong thực tiễn.

Sinh viên thực hiện
(Ký và ghi rõ họ tên)
Trương Thị Trang Linh