CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

0.1 Đặt vấn đề

Trong bối cảnh các cơ quan, doanh nghiệp đang đẩy mạnh chuyển đổi số nhằm tối ưu hóa quy trình vận hành, nhu cầu số hóa và tự động hóa xử lý giấy tờ trở nên ngày càng cấp thiết. Công việc nhận diện và phân loại các loại văn bản—như CMND/CCCD, giấy phép lái xe, hóa đơn, hợp đồng, phiếu thu/chi hay công văn—nếu thực hiện thủ công thường mất nhiều thời gian, dễ xảy ra sai sót và khó khăn trong việc lưu trữ, tra cứu. Hiện nay, nhiều đơn vị vẫn còn dựa vào hình thức nhập liệu tay hoặc tìm kiếm truyền thống, dẫn đến quy trình xử lý chậm, thiếu minh bạch về dữ liệu và khó kiểm soát lịch sử thay đổi tài liệu.

Từ thực tế đó, đề tài "Xây dựng hệ thống nhận diện và phân loại giấy tờ" được triển khai với mục tiêu phát triển một giải pháp số hóa toàn diện cho vòng đời tài liệu: từ khâu tiếp nhận ảnh/chứng từ, tiền xử lý (chỉnh nghiêng, khử nhiễu, tăng cường chất lượng), nhận dạng vùng văn bản, OCR trích xuất nội dung, cho đến phân loại tự động theo từng loại giấy tờ và trích xuất các trường thông tin quan trọng (họ tên, số định danh, ngày cấp, số tiền,…). Hệ thống hướng đến việc rút ngắn thời gian xử lý, nâng cao độ chính xác, chuẩn hóa cấu trúc dữ liệu và hỗ trợ việc lưu trữ – tra cứu tập trung, phục vụ cả công tác quản lý nội bộ lẫn khai thác dữ liệu lâu dài.

0.2 Mục tiêu và phạm vi đề tài

Xuất phát từ bài toán đã nêu ở phần 0.1, mục tiêu tổng quát của đề tài là phát triển một hệ thống nhận dạng và phân loại giấy tờ tiếng Việt tự động, ứng dụng kết hợp công nghệ OCR và mô hình ngôn ngữ hiện đại. Cụ thể, hệ thống cần đạt được các mục tiêu sau:

- Trích xuất được văn bản từ ảnh tài liệu với độ chính xác cao, ngay cả khi chất lượng ảnh ở mức trung bình (nhiễu, bóng mờ nhẹ).
- Chuyển đổi văn bản trích xuất thành vector đặc trưng ngữ nghĩa bằng mô hình PhoBERT, nhằm nắm bắt nội dung thay vì chỉ dựa trên từ khóa rời rạc.
- Xây dựng bộ phân loại dựa trên hồi quy Logistic (Logistic Regression) để gán nhãn loại giấy tờ phù hợp với văn bản đặc trưng đầu vào.
- Triển khai ứng dụng giao diện web trực quan, thân thiện với người dùng, cho phép tải ảnh, xem văn bản OCR và nhận kết quả phân loại tức thì.
- Bảo đảm khả năng mở rộng: khi có loại giấy tờ mới, hệ thống chỉ cần huấn luyện lại với dữ liệu bổ sung mà không phải thay đổi toàn bộ kiến trúc.

Phạm vi của đề tài tập trung vào các loại giấy tờ tiếng Việt có định dạng in ấn phổ biến, ví dụ căn cước công dân, hóa đơn, hợp đồng. Đề tài chưa xử lý các trường hợp khó như chữ viết tay, ảnh mờ nặng, bị gập mép, mất góc lớn hoặc chứa nhiều ngôn ngữ khác nhau. Ngoài ra, do giới hạn thời gian và dữ liệu, hệ thống được thử nghiệm chủ yếu trên tập dữ liệu do sinh viên thu thập và tiền xử lý, không triển khai trên quy mô công nghiệp lớn.

0.3 Yêu cầu phi chức năng

Yêu cầu phi chức năng (từ quy trình trên) - Tốc độ phân loại tài liệu nhanh, đảm bảo đáp ứng nhu cầu xử lý hàng loạt.

- OCR nhận diện ký tự chính xác cao

- Có thể mở rộng để hỗ trợ nhiều loại tài liệu khác nhau trong tương lai.

- Giao diện đơn giản, thông báo lỗi rõ ràng

0.4 Định hướng giải pháp

Từ những mục tiêu đã đề ra, định hướng giải pháp của đề tài được xác định như sau. Hệ thống gồm ba thành phần chính: OCR, trích xuất đặc trưng và phân loại.

Thành phần thứ nhất sử dụng **PaddleOCR** để trích xuất văn bản từ ảnh. PaddleOCR hỗ trợ tốt tiếng Việt, có khả năng nhận dạng các dòng chữ bị xoay, nghiêng và đã được tối ưu cho nhiều tình huống thực tế. Thành phần này giúp chuẩn hóa đầu vào cho bước tiếp theo.

Thành phần thứ hai sử dụng **PhoBERT**, một mô hình ngôn ngữ tiền huấn luyện dựa trên kiến trúc BERT, được huấn luyện trên kho dữ liệu lớn tiếng Việt. PhoBERT có khả năng biểu diễn văn bản thành vector ẩn, phản ánh cả ngữ nghĩa và ngữ cảnh. Đây là lợi thế quan trọng để phân loại tài liệu chính xác, ngay cả khi các loại giấy tờ có từ ngữ trùng lặp.

Thành phần thứ ba là **mô hình Logistic Regression**. Lý do chọn Logistic Regression thay vì mạng nơ-ron phức tạp là vì mô hình này đơn giản, dễ huấn luyện, tốc độ nhanh, dễ giải thích và vẫn đảm bảo hiệu quả cao trong bài toán phân loại nhị phân hoặc đa lớp với dữ liệu vector từ PhoBERT.

Đóng góp chính của đề tài là xây dựng một **quy trình xử lý trọn vẹn** từ ảnh tài liệu đến kết quả phân loại nhãn, đồng thời phát triển ứng dụng **Streamlit** để cung cấp giao diện thực nghiệm. Người dùng có thể tải ảnh, xem kết quả OCR, văn bản chuẩn hóa, và nhãn phân loại tương ứng. Đây là giải pháp khép kín, dễ triển khai và có tính thực tiễn cao.

0.5 Bố cục đồ án

Phần còn lại của báo cáo được tổ chức như sau.

Chương 2 trình bày cơ sở lý thuyết và các công trình nghiên cứu liên quan. Phần này tập trung giới thiệu nền tảng lý thuyết của OCR, các hướng tiếp cận xử lý ngôn ngữ tự nhiên, đặc biệt là các mô hình BERT và PhoBERT cho tiếng Việt, cùng với các thuật toán phân loại văn bản phổ biến.

Chương 3 phân tích bài toán và yêu cầu hệ thống. Nội dung bao gồm khảo sát nhu cầu thực tiễn, phân tích các kịch bản sử dụng, đặc tả yêu cầu chức năng và phi chức năng, cùng với việc xác định dữ liệu đầu vào – đầu ra.

Chương 4 đi sâu vào thiết kế hệ thống. Phần này mô tả kiến trúc tổng thể, sơ đồ luồng xử lý, thiết kế các mô-đun chức năng (OCR, xử lý ngôn ngữ, phân loại, giao diện người dùng), đồng thời phân tích cách các thành phần phối hợp với nhau để đảm bảo mục tiêu đặt ra.

Chương 5 trình bày quá trình hiện thực hóa. Cụ thể, sinh viên mô tả việc cài đặt các thành phần chính trong code (train.py và app.py), các bước huấn luyện mô hình, tiền xử lý dữ liệu, triển khai ứng dụng web bằng Streamlit, và tích hợp các mô-đun lại thành hệ thống hoàn chỉnh.

Chương 6 tổng kết đồ án, rút ra những đóng góp chính và hạn chế. Ngoài ra, chương này đề xuất hướng phát triển trong tương lai, như cải tiến khả năng xử lý chữ viết tay, tăng khả năng chịu lỗi với ảnh chất lượng kém, mở rộng hệ thống phân loại cho nhiều loại giấy tờ hơn và tích hợp với hệ thống quản lý dữ liệu quy mô lớn.