Tự động hóa cập nhật mô hình phân loại cảm xúc đánh giá sản phẩm trên Amazon

CS317.P22 – Phát triển và vận hành hệ thống máy học

T.H.T. An	T.H. Anh	H.Đ. Chung	N.H. Đăng	P.T. Hải
22520033	22520084	22520161	22520189	22520390

Khoa Khoa học Máy tính
Trường Đại học Công nghệ Thông tin, ĐHQG-HCM

June 4, 2025

Overview

Overview

Bối cảnh và động lực

- Thương mại điện tử ngày càng phát triển mạnh mẽ $\rightarrow$ Số lượng đánh giá người dùng ngày càng nhiều.
- Các bài đánh giá sản phẩm chứa nhiều thông tin quan trọng về trải nghiệm và cảm xúc của khách hàng.
- Phân tích cảm xúc giúp doanh nghiệp:
  - Cải thiện sản phẩm và dịch vụ
  - Định hướng chiến lược kinh doanh

Vấn đề đặt ra

- Mô hình phân loại cảm xúc có thể giảm hiệu quả theo thời gian do:
  - Ngôn ngữ và hành vi người dùng thay đổi
  - Sự đa dạng hóa sản phẩm
- Cần thiết xây dựng hệ thống có khả năng tự động cập nhật để:
  - Giữ độ chính xác cao
  - Thích ứng với dữ liệu mới

Mục tiêu đề tài

Xây dựng hệ thống **tự động hóa toàn bộ pipeline phân loại cảm xúc** cho các đánh giá sản phẩm trên Amazon, bao gồm:

- Phân loại cảm xúc: tích cực, tiêu cực, trung lập
- Tự động thu thập và xử lý dữ liệu mới
- Tự động huấn luyện lại khi phát hiện giảm độ chính xác
- Lưu trữ, theo dõi và triển khai mô hình mới
- Tối ưu hóa bằng công cụ: **MLflow**, **n8n**

Tổng quan hệ thống MLOps

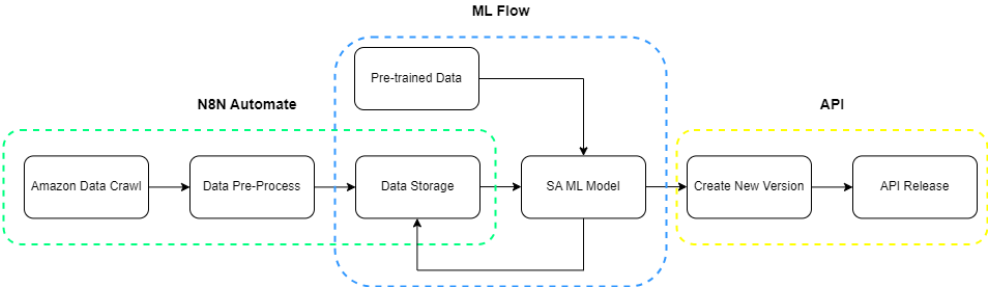


Figure: Quy trình MLOps trong đề tài

Phát biểu bài toán

Input:

- Đánh giá sản phẩm từ Amazon (văn bản, nhãn cảm xúc nếu có)
- Ngưỡng hiệu suất mô hình (Accuracy, F1-score)

Output:

- Mô hình phân loại cảm xúc được cập nhật tự động
- Kết quả phân loại cho đánh giá mới
- Báo cáo hiệu suất theo thời gian

Yêu cầu hệ thống (Requirement)

- Tự động thu thập hoặc nhận dữ liệu mới định kỳ
- Theo dõi hiệu suất mô hình: Kích hoạt lại huấn luyện khi hiệu suất giảm dưới ngưỡng
- Tự động cập nhật mô hình mới lên hệ thống một cách an toàn
- Lưu trữ toàn bộ các phiên bản mô hình để truy vết
- Cung cấp giao diện/API đơn giản để truy vấn kết quả

Overview

Dữ liệu Pre-trained

- Dữ liệu gốc gần **35 triệu đánh giá** từ Amazon (thu thập trong 18 năm, đến tháng 3/2013). Được công bố bởi:
  - J. McAuley & J. Leskovec – RecSys 2013
  - Xiang Zhang – NIPS 2015
- Dữ liệu được trích xuất từ **category "Video Games"** trong tập đánh giá của Amazon, phù hợp cho bài toán phân tích cảm xúc theo lĩnh vực giải trí.
- Tiền xử lý nhãn theo điểm đánh giá:
  - 1–2 sao: **Tiêu cực**
  - 3 sao: **Trung lập**
  - 4–5 sao: **Tích cực**
- Sau xử lý, dữ liệu được chia theo tỷ lệ **7:2:1**:

Tập	Train	Validation	Test	Tổng cộng
Số mẫu	2,786,980	796,243	398,124	**3,981,347**

Dữ liệu Theo Thời Gian Thực

- Nguồn: trang Amazon của một cửa hàng kinh doanh **Video Game**.
- Thu thập tự động mỗi ngày lúc **7h sáng**.
- Kết hợp cùng dữ liệu Pre-trained để huấn luyện và cập nhật mô hình.

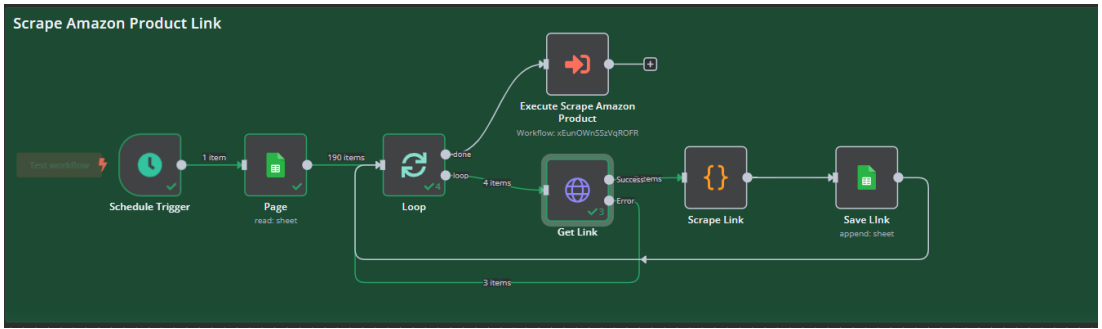
Thành phần	**Chú giải**
Review Text	Văn bản tiếng Anh do người dùng viết, mô tả trải nghiệm với sản phẩm.
Label	Nhãn cảm xúc: 0 – Tiêu cực 1 – Trung lập 2 – Tích cực

Overview

Tổng quan hệ thống

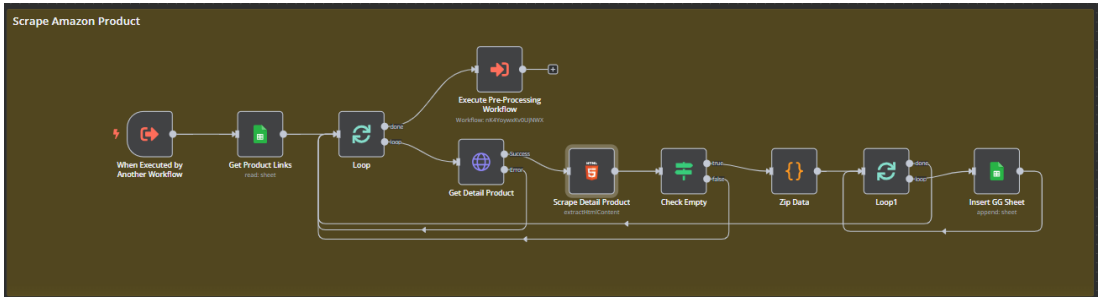
- Tự động cập nhật dữ liệu huấn luyện mô hình phân tích cảm xúc.
- Kết hợp: **n8n**, **MLFlow**, **FastAPI**.
- Dữ liệu đánh giá từ Amazon được thu thập $\rightarrow$ tiền xử lý $\rightarrow$ huấn luyện $\rightarrow$ triển khai API mới.

Quy trình tổng thể - n8n: Thu thập dữ liệu từ Amazon



Flow thu thập danh sách liên kết sản phẩm

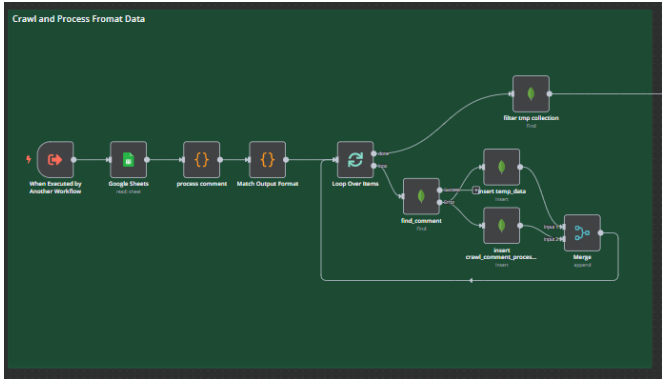
Quy trình tổng thể - n8n: Thu thập dữ liệu từ Amazon



Flow thu thập chi tiết sản phẩm

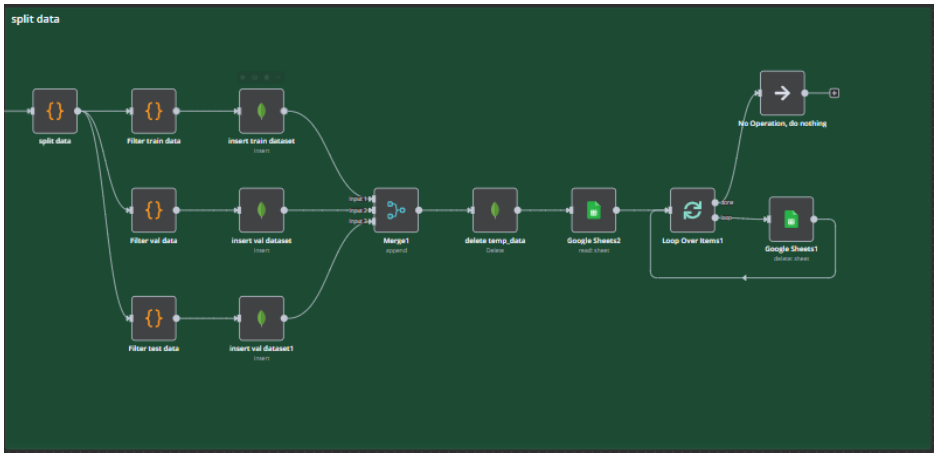
Quy trình tổng thể - n8n: Flow xử lý định dạng dữ liệu

Dữ liệu được xử lý nhằm làm sạch và biến đổi về dạng phù hợp với việc huấn luyện mô hình học máy.



Flow thu thập chi tiết sản phẩm

Quy trình tổng thể - n8n: Flow chia dữ liệu



Flow thu thập chi tiết sản phẩm

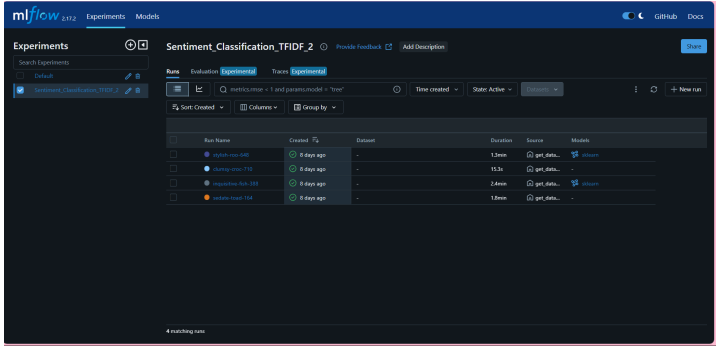
MLFlow trong hệ thống phân tích cảm xúc

Quản lý và theo dõi mô hình học máy

- Quy trình xử lý dữ liệu trước khi huấn luyện gồm:
  - Chuẩn hóa văn bản, lọc dữ liệu tiếng Anh.
  - Làm sạch văn bản (loại bỏ ký tự đặc biệt, HTML tag, v.v.).
  - Tách từ (Tokenization), loại bỏ từ dừng (stopwords).
  - Gán nhãn từ loại (Part-of-Speech tagging).
- MLFlow hỗ trợ theo dõi quá trình huấn luyện, tối ưu và lưu trữ mô hình.
- Kết hợp với `Optuna` để tìm kiếm siêu tham số tối ưu:
  - Ghi nhận tham số bằng `mlflow.log_param()`.
  - Ghi nhận độ chính xác bằng `mlflow.log_metric()`.
- Mô hình Logistic Regression sau khi tối ưu sẽ được:
  - Huấn luyện lại trên toàn bộ dữ liệu.
  - Lưu bằng `mlflow.sklearn.log_model()`.

Lợi ích của MLFlow

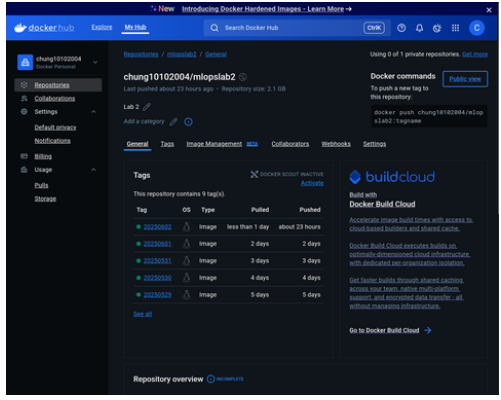
- **Quản lý tập trung**: Toàn bộ mô hình, siêu tham số và chỉ số được lưu trữ rõ ràng.
- **Tái lập mô hình**: Dễ dàng tái huấn luyện hoặc kiểm thử lại các mô hình cũ.
- **Triển khai linh hoạt**: Hỗ trợ Docker, REST API, CI/CD để triển khai nhanh.



Triển khai mô hình với Docker Hub

Tự động đóng gói và triển khai mô hình dưới dạng API

- Sử dụng **FastAPI** để tạo REST API phục vụ inference.
- Tự động đóng gói mô hình và API vào một Docker Image:
  - Cấu trúc thư mục và viết Dockerfile.
  - Build image từ thư mục chứa model mới.
- Đẩy Docker Image lên Docker Hub:
  - Tag image với phiên bản mới.
  - Sử dụng lệnh `docker push`.



Kết quả thực nghiệm

- Mô hình sử dụng: **Logistic Regression**
- Bộ dữ liệu: **Pre-Trained Data**
- Kết quả đạt được:
  - Accuracy trên tập **Test**: **0.85**
  - Accuracy trên tập **Validation**: **0.82**

Overview

Kết luận

- Hệ thống phân loại cảm xúc có khả năng **tự động cập nhật mô hình theo thời gian**.
- Tích hợp:
  - **n8n**: Thu thập dữ liệu định kỳ.
  - **MLflow**: Quản lý và theo dõi mô hình.
- Kết hợp dữ liệu huấn luyện cũ và mới giúp duy trì hiệu suất mô hình.
- Hệ thống linh hoạt và thích ứng với sự thay đổi dữ liệu thực tế.

Hướng phát triển tương lai

- Hỗ trợ **đa ngôn ngữ**.
- Cải thiện xử lý ngữ nghĩa chuyên sâu.
- Áp dụng các mô hình học sâu tiên tiến như:
  - **Transformer**
  - **Fine-tuning** từ các mô hình ngôn ngữ lớn (**LLMs**)