TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

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

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

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

Tóm tắt nội dung

Đề tài xây dựng hệ thống phân loại cảm xúc đánh giá sản phẩm trên Amazon với khả năng tự động thu thập dữ liệu, huấn luyện và cập nhật mô hình theo thời gian. Hệ thống tích hợp công cụ n8n để thu thập dữ liệu định kỳ, sử dụng MLflow để theo dõi và quản lý mô hình, đồng thời hỗ trợ tái huấn luyện tự động khi hiệu suất giảm. Giải pháp giúp duy trì độ chính xác mô hình và giảm thiểu can thiệp thủ công trong môi trường thực tế.

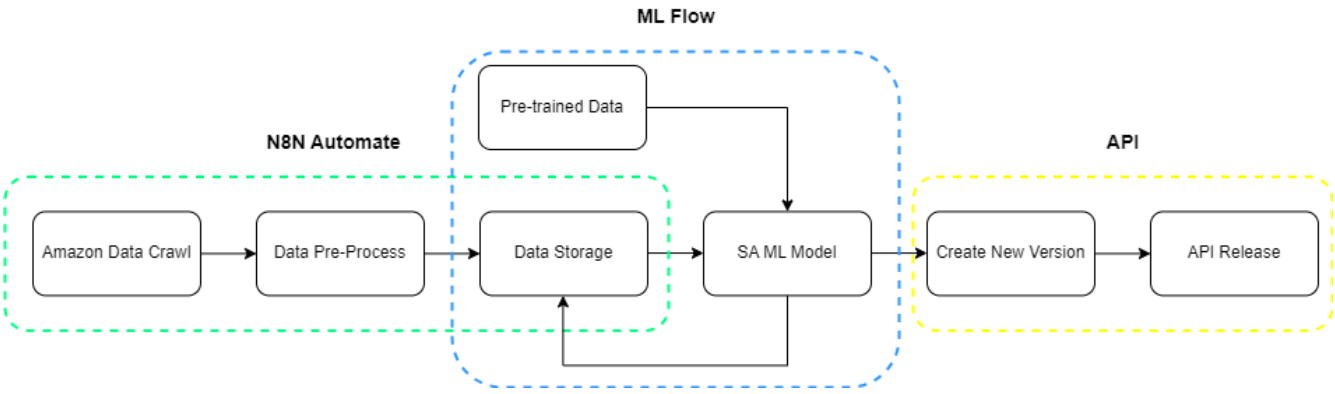
1 Giới thiệu

Trong thời đại thương mại điện tử phát triển mạnh mẽ, số lượng đánh giá từ người dùng trên các nền tảng như Amazon không ngừng gia tăng. Những đánh giá này không chỉ thể hiện quan điểm cá nhân mà còn là nguồn dữ liệu quan trọng giúp doanh nghiệp hiểu rõ hơn về trải nghiệm và cảm xúc của khách hàng đối với sản phẩm. Việc phân tích cảm xúc từ các đánh giá văn bản giúp các nhà bán lẻ và nhà sản xuất đưa ra quyết định cải thiện sản phẩm, dịch vụ, cũng như xây dựng chiến lược kinh doanh phù hợp hơn.

Tuy nhiên, các mô hình phân loại cảm xúc thường trở nên lỗi thời hoặc giảm hiệu quả theo thời gian do ngôn ngữ thay đổi, xu hướng người dùng mới, và sự đa dạng hóa sản phẩm. Vì vậy, nhu cầu tự động cập nhật mô hình phân loại cảm xúc là một vấn đề cấp thiết nhằm đảm bảo độ chính xác và khả năng thích ứng của hệ thống theo thời gian thực.

Đề tài hướng đến xây dựng một hệ thống tự động hoá quy trình cập nhật mô hình phân loại cảm xúc đánh giá sản phẩm trên nền tảng thương mại điện tử Amazon. Cụ thể:

- Phân loại cảm xúc (tích cực, tiêu cực, trung lập) từ các đánh giá sản phẩm dưới dạng văn bản.
- Tự động thu thập và xử lý dữ liệu mới từ Amazon.
- Tự động huấn luyện lại mô hình khi phát hiện độ chính xác giảm.
- Tự động lưu trữ, theo dõi và triển khai mô hình mới.
- Tối ưu toàn bộ pipeline bằng các công cụ hỗ trợ như MLflow và n8n để theo dõi và tự động hóa quy trình.



Hình 1: MLOps Flow

2 Phát biểu bài toán

Input

- Tập dữ liệu đánh giá sản phẩm trên Amazon (dưới dạng văn bản và nhãn cảm xúc nếu có).
- Ngưỡng hiệu suất mô hình (accuracy, F1-score) dùng để đánh giá khi nào cần cập nhật.

Output

- Mô hình phân loại cảm xúc được huấn luyện và cập nhật tự động.
- Kết quả phân loại cảm xúc cho mỗi đánh giá mới.
- Báo cáo theo dõi hiệu suất mô hình qua thời gian.

Yêu cầu hệ thống (Requirement)

- Hệ thống phải tự động thu thập hoặc nhận dữ liệu mới định kỳ.
- Theo dõi hiệu suất mô hình và kích hoạt lại quá trình huấn luyện nếu hiệu suất giảm dưới ngưỡng định sẵn.
- Cập nhật mô hình mới lên hệ thống một cách tự động và an toàn.
- Lưu trữ tất cả các phiên bản mô hình và thông tin liên quan để truy vết.
- Giao diện hoặc API đầu ra đơn giản để truy vấn kết quả phân loại.

3 Bộ dữ liệu

3.1 Dữ liệu Pre-trained

Bộ ngữ liệu gốc bao gồm gần 35 triệu những đánh giá sản phẩm thuộc sàn thương mai điện tử Amazon được thu thập trong vòng 18 năm (cập nhật đến tháng 3 năm 2013). Quá trình chi tiết được thực hiện bởi J. McAuley and J. Leskovec trong bài báo khoa học "Hidden factors and hidden topics: understanding rating dimensions with review text. RecSys, 2013" Sau đó, bộ ngữ liệu được tùy chỉnh phù hợp cho bài toán phân loại ngữ liệu bởi Xiang Zhang vào năm 2015 trong bài báo khoa học "Character-level Convolutional Networks for Text Classification. Advances in Neural Information Processing Systems 28 (NIPS 2015)". Bộ ngữ liệu được xây dựng bằng cách lấy điểm đánh giá 1 và 2 là tiêu cực, 4 và 5 là tích cực, các mẫu có điểm đánh giá là 3 được là trung lập.

3.2 Dữ liệu theo thời gian thực

Dữ liệu được thu thập trực tiếp từ trang thương mại điện tử Amazon của một cửa hàng kinh doanh sản phẩm Video Game. Dữ liệu được tự động thu thập với tần suất hàng ngày (vào lúc 7h sáng).

Thành phần ngữ liệu	**Chú giải**
Đánh giá chi tiết (Review Text)	Đoạn văn bản bằng tiếng Anh đánh giá được viết bởi người dùng, mô tả trải nghiệm về sản phẩm hoặc dịch vụ trên sàn thương mại điện tử Amazon.
Nhãn (Label)	Nhãn phân loại cảm xúc: 0 cho các đánh giá mang ý nghĩa tiêu cực 1 cho các đánh giá mang ý nghĩa trung lập 2 cho các đánh giá mang ý nghĩa tích cực

Bảng 1: Các thành phần ngữ liệu và chú giải

3.3 Thống kê dữ liệu

Cả dữ liệu Pre-Trained và dữ liệu theo thời gian thực được chia thành 3 tập train, validation và test theo tỉ lệ 7/2/1. Chi tiết dữ liệu Pre-Trained bao gồm: 2,786,980 mẫu (Train); 796,243 mẫu (Validation) và 398,124 mẫu (Test). Số lượng mẫu thuộc dữ liệu theo thời gian thực không cố định theo từng ngày.

4 Phương pháp

4.1 Mô tả chung

Phương pháp trên mô tả quy trình tự động cập nhật dữ liệu huấn luyện cho mô hình phân tích cảm xúc bằng cách tích hợp n8n, ML Flow và hệ thống API. Dữ liệu đánh giá từ Amazon được thu thập và xử lý tự động bằng n8n, sau đó được lưu trữ và đưa vào quy trình huấn luyện lại mô hình sentiment analysis trong ML Flow, kết hợp cùng dữ liệu huấn luyện sẵn có. Mô hình sau huấn luyện được đóng gói và triển khai dưới dạng API phiên bản mới, giúp hệ thống luôn phản ánh cảm xúc người dùng một cách kịp thời và chính xác mà không cần can thiệp thủ công.

4.2 Chi tiết các thành phần chính

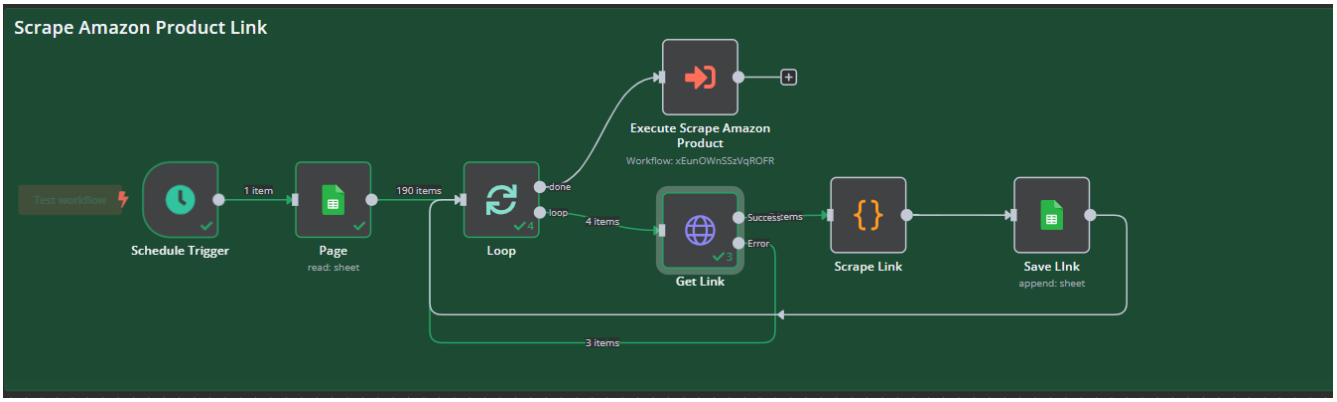
4.2.1 n8n - Workflow Automation

Trong hệ thống cập nhật dữ liệu huấn luyện cho mô hình phân tích cảm xúc, n8n đóng vai trò là công cụ tự động hóa toàn bộ quá trình thu thập và xử lý dữ liệu đầu vào. Với khả năng tích hợp đa dạng dịch vụ và lập lịch linh hoạt, n8n giúp đảm bảo dữ liệu đầu vào luôn được cập nhật định kỳ, đồng bộ và sẵn sàng phục vụ cho quá trình huấn luyện mô hình.

Thu thập dữ liệu từ Amazon

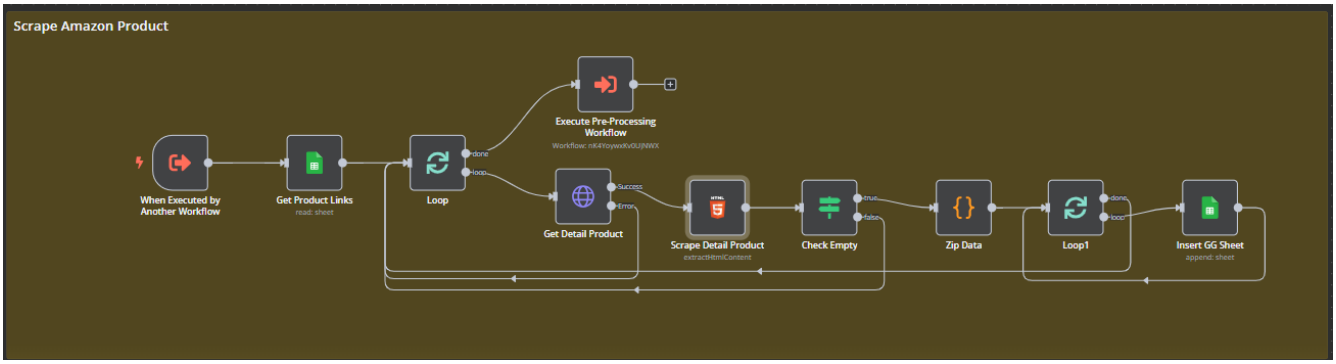
Quá trình thu thập dữ liệu từ Amazon được chia thành hai bước chính và được tự động hóa bằng các node trong n8n:

- Thu thập danh sách liên kết sản phẩm: Ở bước này, hệ thống duyệt qua các trang Amazon trong phạm vi cho phép (số lượng trang giới hạn) để lấy các URL sản phẩm. Việc này được thực hiện thông qua các node HTTP Request, kết hợp với bộ lọc hoặc vòng lặp để duyệt tuần tự từng trang. Kết quả của bước này là một danh sách các liên kết sản phẩm hợp lệ.



Hình 2: Flow thu thập danh sách liên kết sản phẩm

- Thu thập thông tin chi tiết sản phẩm: Sau khi có danh sách liên kết, hệ thống tiếp tục truy cập từng trang sản phẩm để thu thập thông tin chi tiết như: tên sản phẩm, mô tả, ngành hàng, giá, số sao đánh giá và nội dung bình luận của người dùng. Dữ liệu thô này sẽ được chuẩn hóa định dạng và chuyển tiếp sang bước tiền xử lý.

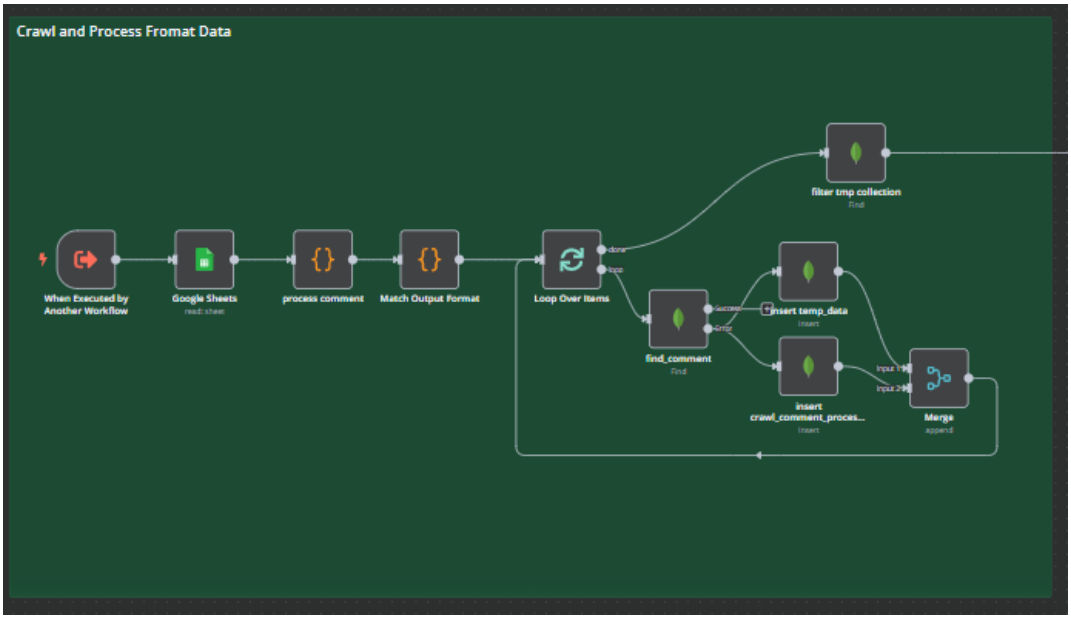


Hình 3: Flow thu thập chi tiết sản phẩm

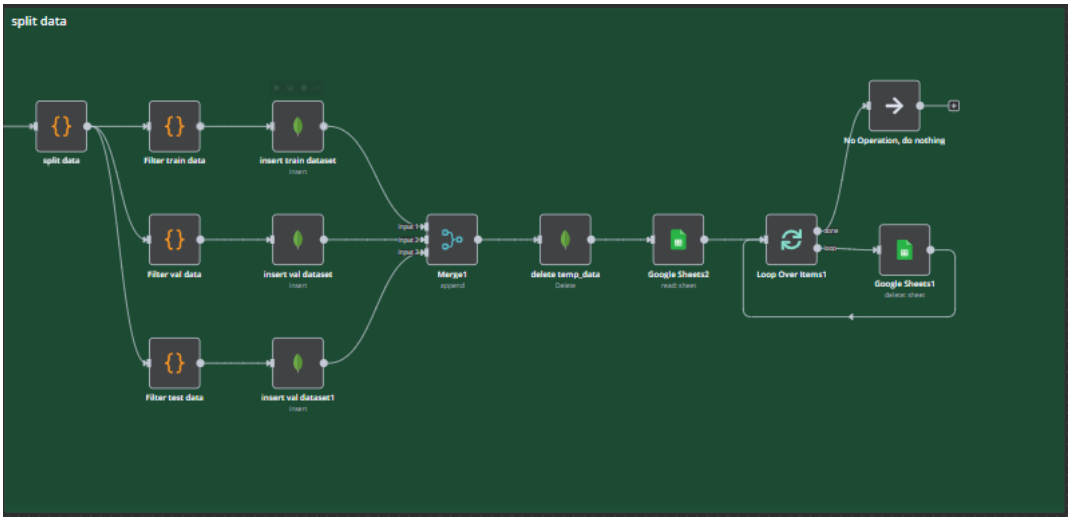
Hai bước trên được tự động hóa và lên lịch định kỳ vào 6 giờ sáng mỗi ngày, đảm bảo rằng dữ liệu luôn được cập nhật hằng ngày và phù hợp với nhu cầu huấn luyện lại mô hình sentiment analysis.

Tiền xử lý dữ liệu

Sau khi dữ liệu đánh giá sản phẩm được thu thập từ Amazon, hệ thống thực hiện và biến đổi dữ liệu về dạng phù hợp với việc huấn luyện mô hình học máy.



Hình 4: Flow xử lý định dạng dữ liệu



Hình 5: Flow chia dữ liệu

4.2.2 MLFlow

MLFlow được tích hợp vào hệ thống như một nền tảng theo dõi, quản lý và tái sử dụng các mô hình học máy trong quá trình phát triển hệ thống phân tích cảm xúc. Với khả năng ghi nhận toàn bộ thông tin liên quan đến quá trình huấn luyện, MLFlow giúp đảm bảo tính minh bạch, khả năng tái lập và tối ưu hóa liên tục cho mô hình.

Tiền xử lý dữ liệu

Trước khi dữ liệu đưa vào huấn luyện, các bước tiền xử lý dữ liệu bao gồm làm sạch và chuẩn hóa được triển khai với MLFlow:

- Lọc dữ liệu tiếng Anh: Các đánh giá không phải là tiếng Anh được loại bỏ nhằm đảm bảo đồng nhất ngôn ngữ trong việc huấn luyện và đánh giá mô hình phân loại.
- Xóa dữ liệu trùng lặp: Các đánh giá sản phẩm trùng lặp được xóa bỏ nhằm tăng chất lượng và độ tin cậy của bộ dữ liệu. Dữ liệu trùng lặp có thể dẫn đến các sai lệch trong việc phân loại mô hình.

- Làm sạch văn bản: Trong dữ liệu văn bản thô thường chứa các ký tự đặc biệt, biểu tượng cảm xúc,... Những ký tự này thường không mang nhiều ý nghĩa và có thể làm nhiễu dữ liệu. Các ký tự này được xóa đi và thay thế bằng khoảng trắng (một số ký tự đặc biệt mang ý nghĩa cảm xúc vẫn được giữ phục vụ cho việc rút trích đặc trưng.

- Tokenization: Đây là quá trình chia văn bản thành các từ riêng lẻ. Tokenization giúp biến đổi văn bản thô thành một dạng văn bản dễ xử lý hơn, đảm bảo dữ liệu được biểu diễn một cách nhất quán và chuẩn xác.

- Xóa stopwords: Stopwords là các từ thông dụng xuất hiện rất nhiều trong văn bản nhưng lại không mang nhiều ý nghĩa, nghĩa ngữ quan trọng. Loại bỏ các stopwords giúp giảm đáng kể kích thước của dữ liệu, tăng hiệu quả xử lý và lưu trữ, đồng thời giảm các chi phí tính toán.

- Part-of-Speech (POS) tagging: POS tagging là quá trình xác định và gán nhãn loại từ (như danh từ, động từ, tính từ,...) cho mỗi từ trong câu. POS tagging cung cấp thông tin ngữ pháp quan trọng, giúp mô hình đưa ra dự đoán chính xác hơn.

Theo dõi quá trình huấn luyện và tối ưu tham số

Trong hệ thống, MLFlow được sử dụng kết hợp với thư viện tối ưu siêu tham số `Optuna` để tự động hóa quá trình tìm kiếm các cấu hình mô hình tối ưu nhất. Toàn bộ quá trình huấn luyện, từ lựa chọn siêu tham số (như `n_estimators`, `learning_rate`, `max_depth`) cho đến đánh giá độ chính xác trên tập validation, đều được ghi lại chi tiết bằng các lệnh `mlflow.log_param()` và `mlflow.log_metric()`. Việc này giúp nhóm phát triển dễ dàng theo dõi các lần chạy thử nghiệm, so sánh hiệu quả giữa các cấu hình và tái sử dụng mô hình tốt nhất.

Lưu trữ và triển khai mô hình cuối cùng

Sau khi tìm được cấu hình tối ưu, mô hình `Logistic Regression` sẽ được huấn luyện lại trên toàn bộ tập dữ liệu (bao gồm tập train và validation), sau đó

đánh giá độ chính xác trên tập test. Mô hình huấn luyện cuối cùng được lưu trữ trực tiếp vào hệ thống MLFlow bằng `mlflow.sklearn.log_model()`, cho phép người dùng dễ dàng tải về, triển khai hoặc tái sử dụng trong tương lai.

Lợi ích của việc sử dụng MLFlow

- **Quản lý tập trung**: Tất cả các phiên bản mô hình, cấu hình siêu tham số, và các chỉ số đánh giá đều được quản lý tập trung tại giao diện MLFlow, giúp đơn giản hóa việc theo dõi và đánh giá hiệu quả mô hình qua các lần huấn luyện khác nhau.
- **Tái lập mô hình**: Với mỗi phiên bản mô hình, MLFlow ghi lại đầy đủ thông tin cần thiết để tái huấn luyện hoặc phục vụ cho việc kiểm thử lại trong tương lai.
- **Hỗ trợ triển khai**: MLFlow có thể tích hợp với nhiều nền tảng triển khai như Docker, REST API hoặc các hệ thống CI/CD, giúp đẩy nhanh quá trình đưa mô hình vào môi trường sản xuất.

4.2.3 Docker Hub

Docker Hub được sử đụng để tự động đóng gói mô hình mới nhất thành một API có thể triển khai.

Quy trình tự động đóng gói mô hình thành API

- **Đóng gói mô hình thành API**: Một script tự động sẽ sử dụng FastAPI để tạo một REST API phục vụ inference. Sau đó, toàn bộ thư mục chứa API và mô hình sẽ được cấu trúc thành một Dockerfile.
- **Đẩy lên Docker Hub**: Docker Image sau khi được build sẽ được đẩy lên Docker Hub bằng lệnh `docker push`. Sau đó, script sẽ tự động thực hiện các bước sau:

- Build Docker image từ thư mục API chứa model mới.
- Tag image với phiên bản mới
- Push image lên Docker Hub.

5 Thực nghiệm

Kết quả mô hình đạt được với Logistic Regression trên bộ dữ liệu Pre-Trained là 0.85 accuracy (đo trên tập Test) và 0.82 accuracy (đo trên tập Validation).

6 Kết luận

Trong đề tài này, nhóm đã xây dựng một hệ thống phân loại cảm xúc đánh giá sản phẩm với khả năng tự động hóa quy trình cập nhật mô hình theo thời gian. Bằng cách tích hợp các công cụ như n8n để thu thập dữ liệu định kỳ và MLflow để quản lý mô hình, hệ thống giúp giảm thiểu sự can thiệp thủ công, đảm bảo tính linh hoạt và thích ứng trước sự thay đổi của dữ liệu thực tế.

Việc sử dụng kết hợp dữ liệu huấn luyện ban đầu với dữ liệu thu thập mới giúp mô hình duy trì được hiệu suất ổn định khi áp dụng vào môi trường sản xuất. Ngoài ra, quy trình theo dõi hiệu suất và tự động tái huấn luyện cũng giúp hệ thống thích nghi tốt với các xu hướng và ngữ cảnh mới từ người dùng.

Trong tương lai, nhóm dự kiến sẽ mở rộng hệ thống để hỗ trợ đa ngôn ngữ, cải thiện khả năng xử lý ngữ nghĩa chuyên sâu, và tăng tính hiệu quả bằng cách áp dụng các kỹ thuật học sâu hiện đại hơn như transformer hoặc fine-tuning từ các mô hình ngôn ngữ lớn (LLMs).

Phân công công việc

Trương Huỳnh Thúy An

- Setup MongoDB Server
- Thiết kế các bảng dữ liệu
- Nhập dữ liệu Pre-Trained
- Thiết kế Flow n8n để tiền xử lý dữ liệu và lưu dữ liệu vào Databse

Trương Hồng Anh

- Thiết kế Flow n8n để tiền xử lý dữ liệu và lưu dữ liệu vào Database
- Thiết kế Flow n8n để thu thập dữ liệu từ sàn thương mại điện tử Amazon

Hoàng Đức Chung

- Thiết kế Flow n8n để thu thập dữ liệu từ sàn thương mại điện tử Amazon
- Phát triển Script đóng gói API tự động khi mô hình được cập nhật

Nguyễn Hải Đăng

- Phát triển và huấn luyện mô hình theo dữ liệu Pre-Trained
- Xây dựng mô hình tự động cập nhật với dữ liệu theo thời gian thực

Phan Thanh Hải

- Setup Server n8n, MongoDB Server
- Lên kế hoạch thực hiện đề tài
- Báo cáo và trình bày đề tài

Tài liệu

[1] n8n.io, *n8n Documentation*, 2023. Truy cập tại: `https://docs.n8n.io`

[2] Matei Zaharia, Andrew Chen, Aaron Davidson, Ali Ghodsi, Sue Ann Hong, Andrew Konwinski, Martin Murching, Tomas Nykodym, Paul Ogilvie, Shubham Parkhe, v.v. *MLflow: A Platform for the Machine Learning Lifecycle*, Proceedings of the 2nd International Workshop on Systems and Infrastructure for ML and AI, 2018. Truy cập tại: `https://mlflow.org`

[3] Docker Inc., *Docker Hub: Cloud-based Image Repository*, 2024. Truy cập tại: `https://hub.docker.com`

[4] Databricks, *MLflow Documentation*, 2024. Tài liệu chính thức tại: `https://mlflow.org/docs/latest/index.html`