ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN – TIN HỌC

-----o0o----

BÁO CÁO THỰC TẬP ĐỒ ÁN THỰC TẾ

ĐỀ TÀI: XỬ LÝ DỮ LIỆU LIFELOGGING CHO MÔ HÌNH NGÔN NGỮ LỚN (LLMS)

Họ và tên sinh viên: Nguyễn Minh Hùng
Mã số sinh viên: 21110301
Giảng viên phụ trách: PGS. TS. Nguyễn Thanh Bình

Thành phố Hồ Chí Minh, 2025

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN – TIN HỌC

-----o0o----

BÁO CÁO THỰC TẬP ĐỒ ÁN THỰC TẾ

ĐỀ TÀI: XỬ LÝ DỮ LIỆU LIFELOGGING CHO MÔ HÌNH NGÔN NGỮ LỚN (LLMS)

Họ và tên sinh viên: Nguyễn Minh Hùng
Mã số sinh viên: 21110301
Giảng viên phụ trách: PGS. TS. Nguyễn Thanh Bình
Nơi thực tập: Công ty Cổ phần giáo dục Trí tuệ nhân tạo và Khoa học dữ liệu
Thời gian thực tập: Từ 01/08/2024 đến 31/10/2024
Người hướng dẫn: ThS. Trần Quang Linh
Email: linh.tran3@mail.dcu.ie

LỜI NÓI ĐẦU

Em xin gửi đến quý thầy, cô và Ban chủ nhiệm khoa bản Báo cáo Thực tập đồ án thực tế với chủ đề "Xử lý dữ liệu Lifelogging cho mô hình ngôn ngữ lớn (LLMs)". Báo cáo này là thành quả của quá trình thực tập tại công ty Cổ phần giáo dục Trí tuệ nhân tạo và Khoa học dữ liệu.

Trong suốt quá trình học tập tại Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM em luôn mong muốn khám phá và áp dụng những kiến thức chuyên môn vào các bài toán thực tế, đặc biệt là trong lĩnh vực trí tuệ nhân tạo và khoa học dữ liệu. Được thực tập tại công ty không chỉ giúp em trải nghiệm môi trường làm việc chuyên nghiệp mà còn mang đến cơ hội tham gia làm việc trực tiếp với các dự án thực tế. Trong đó "Xử lý dữ liệu Lifelogging cho mô hình ngôn ngữ lớn (LLMs)" là một đề tài mang tính ứng dụng cao và đầy tiềm năng.

Lifelogging là lĩnh vực tập trung vào quá trình thu thập dữ liệu về cuộc sống cá nhân một cách liên tục thông qua các thiết bị đeo thông minh, điện thoại di động và các cảm biến khác. Công nghệ này mang lại tiềm năng to lớn trong việc giám sát sức khoẻ, theo dõi năng suất làm việc và hỗ trợ trí nhớ bằng cách ghi nhớ thông qua việc ghi nhận và phân tích dữ liệu từ cuộc sống hàng ngày.

Với ứng dụng trong các hệ thống Hỏi - Đáp (QA), dữ liệu lifelogging có thể được khai thác để tạo ra các phản hồi chính xác, hữu ích và mang tính cá nhân hoá, chẳng hạn như tư vấn sức khoẻ, tối ưu hoá thời gian làm việc hoặc hỗ trợ người dùng gợi nhắc các sự kiện quan trọng trong cuộc sống. Điều này yêu cầu một bộ dữ liệu được chú thích rõ ràng, chính xác và có cấu trúc tốt, đóng vai trò quan trọng làm nền tảng để huấn luyện, phát triển và đánh giá hiệu suất của các hệ thống AI tiên tiến, đồng thời thúc đẩy nghiên cứu đổi mới trong lĩnh vực này.

Trong dự án này, em đã tham gia xây dựng một bộ dữ liệu QA từ dữ liệu lifelog thông qua việc tạo ra mô tả chi tiết về các sự kiện, thiết kế các kịch bản câu hỏi và câu trả lời tương ứng từ dữ liệu lifelog, đồng thời giải quyết những thách thức đặc thù khi làm việc với dữ liệu dạng sự kiện cá nhân. Dự án không chỉ giúp em hiểu sâu hơn về cách tổ chức và xử lý dữ liệu mà còn hướng đến việc đóng góp vào quá trình nghiên cứu và phát triển các hệ thống thông minh nhằm nâng cao trải nghiệm người dùng.

Thời gian thực tập là cơ hội để em áp dụng kiến thức lý thuyết vào các vấn đề thực tế, phát triển kỹ năng xử lý vấn đề và tích luỹ các kỹ năng, kinh nghiệm hữu ích. Qua quá trình làm việc, em càng nhận thấy tầm quan trọng của việc trải nghiệm thực tế trong hành trình học tập và chuẩn bị trước khi bước vào môi trường làm việc trong tương lai.

Báo cáo này sẽ tập trung mô tả chi tiết các bước triển khai dự án và phạm vi công việc, từ xây dựng bộ dữ liệu và các thách thức kỹ thuật, những kết quả đạt được cũng như những kinh nghiệm có được trong quá trình thực hiện dự án. Em cũng đề xuất một số hướng phát triển tiềm năng cho lĩnh vực này trong tương lai.

Cuối cùng, em xin gửi lời cảm ơn đến quý thầy, cô và Ban chủ nhiệm khoa đã tạo điều kiện thuận lợi và hỗ trợ trong quá trình thực tập. Đồng thời, em xin chân thành cảm ơn đến các anh chị tại công ty Cổ phần giáo dục Trí tuệ nhân tạo và Khoa học dữ liệu đã tận tình chia sẻ kiến thức và cơ hội thực tập hữu ích này.

Em xin chân thành cảm ơn!

MỤC LỤC

I. TỔNG QUAN VỀ ĐƠN VỊ THỰC TẬP

1. Giới thiệu về công ty

Công ty Cổ phần giáo dục Trí tuệ nhân tạo và Khoa học dữ liệu là công ty cung cấp dịch vụ nghiên cứu, tư vấn và phát triển các dịch vụ liên quan đến trí tuệ nhân tạo, khoa học dữ liệu, khoa học máy tính, ứng dụng toán trong tính toán khoa học và khoa học công nghệ khác. Nghiên cứu, tư vấn và phát triển các dịch vụ liên quan đến quản lý dự án Công nghệ thông tin, Quản lý rủi ro và an toàn thông tin trong doanh nghiệp, Quản lý thông tin, Quản trị mạng và hệ thống. Nghiên cứu, tư vấn và phát triển các dịch vụ liên quan đến ngoại ngữ. Đối tượng của công ty không chỉ là các cá nhân mà còn là các doanh nghiệp, tổ chức.

2. Lĩnh vực hoạt động và sản phẩm/dịch vụ

Công ty đã hoạt động trong các lĩnh vực nghiên cứu khoa học và phát triển công nghệ, bao gồm:

- Khoa học tự nhiên;
- Khoa học kỹ thuật và công nghệ;
- Khoa học y tế, dược;
- Khoa học nông nghiệp;
- Nghiên cứu thị trường và thăm dò dư luận.

3. Cơ cấu tổ chức và phòng ban

Công ty hoạt động với cơ cấu tổ chức chuyên nghiệp và phong cách quản lý hiện đại. Công ty có các phòng ban và đội ngũ chuyên viên chất lượng cao, chịu trách nhiệm cho các hoạt động quản lý và vận hành dự án. Mỗi phòng ban chịu trách nhiệm cho một lĩnh vực cụ thể.

Bên cạnh đó, công ty cũng đặt mục tiêu phát triển bền vững và thúc đẩy sự phát triển chuyên nghiệp cho từng thành viên trong tổ chức. Điều này thể hiện qua việc công ty đề cao giá trị con người, xây dựng môi trường làm việc thân thiện và đa dạng cơ hội thăng tiến trong sự nghiệp.

II. NỘI DUNG THỰC TẬP

1. Thông tin tổng quan về quá trình thực tập

	Thứ hai	Thứ ba	Thứ tư	Thứ năm	Thứ sáu	Thứ bảy	Chủ nhật
Sáng	☑	☑	☑	☑	☑	☐	☐
Chiều	☑	☑	☑	☑	☑	☐	☐

Vị trí làm việc: Thực tập sinh.

Thời gian làm việc mỗi ngày: Tương ứng 4 giờ mỗi ngày.

Thời gian làm việc mỗi tuần: Từ thứ hai đến thứ sáu, tương ứng 20 giờ mỗi tuần.

Tổng thời gian làm việc: 3 tháng.

2. Quy trình làm việc

Quy trình bao gồm các công việc xoay quanh quá trình thu thập, xử lý, gắn nhãn và xây dựng bộ dữ liệu câu hỏi và câu trả lời (Q&A) từ dữ liệu lifelog. Các công việc được cụ thể như sau:

a. Thu thập dữ liệu lifelog

Thu thập dữ liệu từ các thiết bị đeo thông minh, điện thoại và các cảm biến khác. Dữ liệu bao gồm thông tin hoạt động hàng ngày, địa điểm, thời gian, hình ảnh chụp được vào thời gian đó, trạng thái sức khoẻ, lịch sử di chuyển. Có thể sử dụng các công cụ hỗ trợ như API từ thiết bị đeo, cơ sở dữ liệu cá nhân, nền tảng lưu trữ đám mây để có thể thu thập dữ liệu.

b. Xử lý dữ liệu

Sau khi có được dữ liệu, cần kiểm tra, làm sạch và chuẩn hoá dữ liệu để giải quyết vấn đề về nhiễu và lỗi. Các bước xử lý dữ liệu bào gồm việc loại bỏ dữ liệu bị trùng lặp hoặc không sử dụng được, chuẩn hoá định dạng dữ liệu (thời gian, địa điểm, …), ngoài ra có thể xử lý dữ liệu bị thiếu bằng cách sử dụng các công cụ được hỗ trợ bởi các thư viện Pandas và NumPy trong Python.

c. Xây dựng bộ dữ liệu

Trước tiên, tạo các mô tả chi tiết về các sự kiện theo ngày và giờ trong bộ dữ liệu lifelog: Dựa vào hình ảnh và thời gian (ngày, giờ, phút), địa điểm được cung cấp nhằm xác định có sự kiện gì diễn ra tại thời điểm đó và thêm mô tả các chi tiết cho sự kiện để xác định rõ bối cảnh của sự kiện.

Ví dụ:

- ID ảnh: 20190411_054316_000 *(05 giờ 43 phút 16 giây, ngày 11/04/2019)*;
  Sự kiện: Tôi đang sử dụng máy tính bàn tại Charm Hand & Foot Spa.
- ID ảnh: 20190711_085135_000 *(08 giờ 51 phút 35 giây, ngày 11/07/2019)*;
  Sự kiện: Tôi đang nói chuyện với một người đàn ông ở siêu thị.

Sau khi gắn nhãn các sự kiện diễn ra, ta tự tạo bộ câu hỏi và câu trả lời dựa trên các sự kiện có trong phần gắn nhãn. Các câu hỏi tập trung vào việc hỏi thời gian diễn ra, số lần diễn ra sự kiện đó, địa điểm, đúng / sai, … Từ đó tự tạo ra câu trả lời chính xác cho câu hỏi từ bộ dữ liệu.

Ví dụ:

- Câu hỏi: Tôi đã lái xe đến Dublin City University (DCU) trong bao lâu?
  Câu trả lời: 31 phút.
- Câu hỏi: Tôi đã nói chuyện với bao nhiêu người ở siêu thị?
  Câu trả lời: 2 người.
- Câu hỏi: Tôi đang làm gì vào lúc 18 giờ 39 phút?
  Câu trả lời: Tôi đang chạy bộ ở công viên.

Từ cấu trúc bộ dữ liệu Q&A tự tạo ở trên, mô hình ngôn ngữ lớn (LLMs) sẽ học và đưa ra chuỗi các câu hỏi từ bộ dữ liệu sự kiện. Từ đó kiểm tra câu hỏi mô hình đặt ra có phù hợp hay không, đồng thời xác định câu trả lời có chính xác hay không nhằm chỉnh sửa lại theo đúng với các mô tả sự kiện trong bộ dữ liệu để mô hình có thể đưa ra bộ dữ liệu ngày càng tốt hơn.

d. Đánh giá và cải thiện

Sau khi thực hiện quy trình trên, ta thực hiện đánh giá bộ dữ liệu về sự kiện được gắn nhãn và bộ dữ liệu Q&A, phân tích kết quả của mô hình ngôn ngữ học được dựa trên các bộ dữ liệu để xem xét cải thiện về độ chi tiết hay chính xác trong bộ dữ liệu hoặc về mô hình. Sau đó tổng hợp và đề xuất cải tiến mô hình để mang lại hiệu quả tốt nhất.

3. Sản phẩm

Dự án hướng đến việc xuất bản một bộ dữ liệu lifelog Q&A chất lượng (bao gồm các sự kiện được mô tả chi tiết và bộ dữ liệu Q&A), có khả năng trả lời các câu hỏi về hoạt động cụ thể trên bộ dữ liệu lifelog. Từ đó hy vọng khả năng ứng dụng vào thực tế như quản lý thời gian, giám sát sức khoẻ và hỗ trợ trí nhớ. Đồng thời, dự án cũng hướng mục tiêu có thể xuất bản các công trình nghiên cứu khoa học về bộ dữ liệu ở các hội nghị và bài báo khoa học uy tín ở các rank A – B.

III. NHẬN XÉT

1. Ưu điểm

Trong thời gian thực tập tại công ty, em không chỉ được tiếp cận với các dự án trong lĩnh vực khoa học dữ liệu, mà còn có cơ hội được tham gia trực tiếp vào các hoạt động nghiên cứu và phát triển. Việc tham gia vào các dự án thực tế, áp dụng các công nghệ và phương pháp hiện đại đã giúp em cải thiện năng lực chuyên môn, tư duy logic, kỹ năng giải quyết vấn đề thực tế.

Với môi trường làm việc chuyên nghiệp và sáng tạo tại công ty đã tạo điều kiện cho em phát triển tinh thần học hỏi không ngừng, kỹ năng quản lý thời gian và kỹ năng làm việc nhóm. Các trải nghiệm ở công ty trong vai trò là một thực tập sinh đã giúp cho em tích luỹ được những kinh nghiệm, đồng thời hiểu rõ hơn về quy trình làm việc phù hợp tại các dự án, công ty, doanh nghiệp.

Đồng thời, kết quả đạt được trong thời gian thực tập đã có đóng góp một phần nhỏ vào hoạt động nghiên cứu của dự án này, hỗ trợ nhóm nghiên cứu trong việc xử lý dữ liệu, qua đó giúp thúc đẩy tiến độ triển khai dự án được nhanh chóng và hiệu quả hơn.

2. Nhược điểm

Trong quá trình thực tập, một trong những khó khăn lớn nhất gặp phải là khối lượng công việc đôi lúc rất nhiều, đòi hỏi phải dành nhiều thời gian để có thể hoàn thành công việc được giao đúng tiến độ. Việc cân đối giữa lịch học tập trên trường và lịch làm việc đôi lúc không dễ dàng.

Ngoài ra, khi gặp phải các dữ liệu phức tạp, cụ thể chênh lệch thời gian thu thập được do thay đổi múi giờ, sự chênh lệch này nếu không chú ý sẽ dẫn đến việc sai sót trong bộ dữ liệu, từ đó mô hình sẽ không đưa ra câu hỏi và câu trả lời chính xác, nên bắt buộc cần phải được xử lý và tính toán lại để có thể đưa ra câu hỏi và câu trả lời về thời gian nhằm đảm bảo độ chính xác cao.

Tuy nhiên, chính những thách thức này đã giúp em rèn luyện được khả năng quản lý thời gian, sự kiên trì và tính kỷ luật trong công việc, từ đó tích luỹ được nhiều kinh nghiệm giúp dễ dàng thích nghi với môi trường công việc sau này.

KẾT LUẬN

Thời gian thực tập tại công ty đã mang đến cho em những trải nghiệm thực tế vô cùng ý nghĩa, giúp em tích luỹ được nhiều kiến thức và kỹ năng quý báu trong lĩnh vực khoa học dữ liệu. Những bài học và kinh nghiệm có được không chỉ đến từ công việc mà còn từ sự hỗ trợ tận tình tại công ty. Sau đây là những điểm nhấn và kết luận sau thời gian thực tập:

- Áp dụng lý thuyết đã được học vào thực tế: Thực tập là cơ hội để em áp dụng các kiến thức lý thuyết đã học tại trường thành các ứng dụng thực tiễn. Việc tham gia vào dự án giúp em hiểu rõ hơn về quy trình và các thách thức trong việc triển khai các giải pháp khoa học dữ liệu.
- Làm việc với dữ liệu thực tế: Em đã có cơ hội xử lý và phân tích các bộ dữ liệu phức tạp, đa dạng từ nhiều nguồn khác nhau. Việc thực hành trên dữ liệu thực tế không chỉ giúp nâng cao khả năng chuyên môn mà còn cải thiện tư duy phân tích và cách tiếp cận vấn đề.
- Sử dụng công nghệ và mô hình tiên tiến. Em đã áp dụng các công cụ học máy, bao gồm cả các mô hình ngôn ngữ lớn (LLMs) trong xử lý ngôn ngữ tự nhiên (NLP) để xây mô hình và tối ưu hoá kết quả. Điều này không chỉ giúp nâng cao kỹ năng mà còn giúp em nắm vững các phương pháp hiện đại trong khoa học dữ liệu.
- Đóng góp vào dự án: Qua các nhiệm vụ được giao, em đã có cơ hội đóng góp vào quá trình phát triển và hoàn thiện dự án hỗ trợ nhóm nghiên cứu trong việc tối ưu hoá quy trình và cải thiện chất lượng sản phẩm.
- Phát triển kỹ năng cá nhân: Ngoài việc nâng cao chuyên môn, thời gian thực tập cũng giúp em cải thiện các kỹ năng mềm như giao tiếp, làm việc nhóm và quản lý thời gian. Đây là những hành trang quan trọng để chuẩn bị cho sự nghiệp tương lai.

Tóm lại, quá trình thực tập tại công ty Cổ phần giáo dục Trí tuệ nhân tạo và Khoa học dữ liệu đã mang lại cho em một môi trường học tập thực tiễn, giúp củng cố kiến thức, phát triển kỹ năng và định hình tư duy nghề nghiệp. Em xin chân thành cảm ơn sự hỗ trợ và tạo điều kiện của công ty, đồng thời tin rằng những kinh nghiệm tích luỹ được sẽ là nền tảng quan trọng cho con đường sự nghiệp trong tương lai.