TRƯỜNG ĐẠI HỌC QUỐC TẾ HỒNG BÀNG

BỘ MÔN CÔNG NGHỆ THÔNG TIN

□🕮□

BÁO CÁO THƯỜNG KỲ

MÁY HỌC

ĐỀ TÀI: Xây dựng mô-đun đánh giá chất lượng câu trả lời của AI trợ giảng bằng Machine Learning

Giảng viên hướng dẫn: Trần Sơn Hải

Họ Và Tên Sinh Viên	**Mã Số Sinh Viên**
Trần Đình Bảo Huy	2311110136
Nguyễn Đỗ Anh Khoa	2311110118

TP. Hồ Chí Minh, 202

MỤC LỤC

MỞ ĐẦU

1. **Vấn đề nghiên cứu:**

Các hệ thống AI trợ giảng hiện nay vẫn thường sinh ra câu trả lời không chính xác, thiếu liên quan hoặc gây hiểu nhầm cho người học.

2. **Mục tiêu:**

Đề tài hướng tới việc xây dựng một mô hình Machine Learning nhằm đánh giá và chấm điểm chất lượng câu trả lời của AI trước khi hiển thị cho người học.

3. **Phương pháp:**

Nghiên cứu tiến hành thu thập dữ liệu câu hỏi – câu trả lời, trích xuất các đặc trưng ngôn ngữ phù hợp và huấn luyện mô hình phân loại hoặc chấm điểm chất lượng câu trả lời.

Nghiên cứu sử dụng kết hợp các đặc trưng thống kê (TF-IDF), đặc trưng ngữ nghĩa (semantic similarity giữa câu hỏi và câu trả lời) và các đặc trưng thủ công như độ dài câu trả lời, mức độ bao phủ từ vựng và các tín hiệu không chắc chắn nhằm nâng cao hiệu quả phân loại.

4. **Kết quả:**

Mô hình được đánh giá bằng chỉ số F1-macro thông qua phương pháp cross-validation, cho thấy khả năng phân biệt hiệu quả giữa câu trả lời chất lượng tốt và kém.

5. **Ý nghĩa:**

Kết quả nghiên cứu giúp nâng cao chất lượng phản hồi của AI trợ giảng, tăng khả năng ứng dụng thực tế trong môi trường giáo dục số.

CHƯƠNG 1: GIỚI THIỆU

1.1. Bối cảnh và động cơ nghiên cứu

Trong những năm gần đây, trí tuệ nhân tạo (Artificial Intelligence – AI), đặc biệt là các mô hình xử lý ngôn ngữ tự nhiên, đã được ứng dụng rộng rãi trong lĩnh vực giáo dục. Một trong những ứng dụng tiêu biểu là hệ thống AI trợ giảng, có khả năng hỗ trợ sinh viên trả lời câu hỏi, giải thích khái niệm và hướng dẫn học tập mọi lúc, mọi nơi. AI trợ giảng góp phần giảm tải cho giảng viên, cá nhân hóa trải nghiệm học tập và nâng cao khả năng tiếp cận tri thức của người học.

Tuy nhiên, bên cạnh những lợi ích, các hệ thống AI trợ giảng hiện nay vẫn tồn tại nhiều hạn chế. Đặc biệt, các mô hình ngôn ngữ lớn (Large Language Models – LLMs) có thể sinh ra những câu trả lời không chính xác, thiếu liên quan đến câu hỏi, hoặc mang tính suy đoán và gây hiểu nhầm cho sinh viên. Hiện tượng này thường được gọi là "hallucination" – khi AI tạo ra thông tin nghe có vẻ hợp lý nhưng không đúng về mặt nội dung.

Trong bối cảnh giáo dục, việc cung cấp thông tin sai lệch có thể gây ảnh hưởng nghiêm trọng đến quá trình học tập và nhận thức của người học. Do đó, nhu cầu đặt ra là cần có một cơ chế kiểm soát chất lượng câu trả lời của AI trước khi hiển thị cho sinh viên. Thay vì chỉ dựa vào bản thân mô hình sinh câu trả lời, một mô-đun đánh giá độc lập dựa trên Machine Learning có thể đóng vai trò như một "bộ lọc", giúp đánh giá mức độ phù hợp và chất lượng của câu trả lời.

Xuất phát từ nhu cầu thực tiễn đó, đề tài này tập trung vào việc xây dựng mô-đun đánh giá chất lượng câu trả lời của AI trợ giảng bằng các kỹ thuật Machine Learning truyền thống kết hợp với các đặc trưng ngữ nghĩa hiện đại.

1.2. Mô tả hệ thống AI trợ giảng Ai-trogiang

Hệ thống AI trợ giảng Ai-trogiang được thiết kế như một nền tảng hỗ trợ học tập cho sinh viên, cho phép người học đặt câu hỏi liên quan đến nội dung môn học và nhận được câu trả lời tự động từ AI. Về mặt kiến trúc, hệ thống bao gồm ba thành phần chính: giao diện người dùng (frontend), hệ thống xử lý phía máy chủ (backend) và mô-đun AI.

Trong đó, mô-đun AI đảm nhiệm việc tiếp nhận câu hỏi, truy xuất tri thức (nếu có) và sinh ra câu trả lời. Mô-đun đánh giá chất lượng câu trả lời được đặt ở giai

đoạn sau khi AI sinh câu trả lời nhưng trước khi kết quả được gửi đến người dùng. Mô-đun này có nhiệm vụ phân tích cặp câu hỏi – câu trả lời và đưa ra đánh giá về chất lượng, từ đó quyết định xem câu trả lời có nên được hiển thị hay cần yêu cầu AI sinh lại câu trả lời khác.

1.3. Mục tiêu và phạm vi đề tài

Mục tiêu chính của đề tài là xây dựng một mô hình Machine Learning có khả năng đánh giá chất lượng câu trả lời của AI trợ giảng dựa trên cặp dữ liệu câu hỏi – câu trả lời. Mô hình được thiết kế để phân loại câu trả lời thành hai nhóm chính: câu trả lời tốt (đúng, liên quan, rõ ràng) và câu trả lời kém chất lượng (sai, không liên quan hoặc mơ hồ).

Đề tài không tập trung vào việc huấn luyện hay cải thiện mô hình sinh ngôn ngữ (LLM), mà chỉ tập trung vào việc đánh giá đầu ra của hệ thống AI. Phạm vi dữ liệu được sử dụng trong đề tài ở mức nhỏ, phục vụ mục đích nghiên cứu và minh họa quy trình xây dựng mô-đun đánh giá.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. Bài toán đánh giá chất lượng câu trả lời

Trong các hệ thống hỏi – đáp tự động (Question Answering Systems), đặc biệt là các hệ thống AI trợ giảng, chất lượng của câu trả lời đóng vai trò then chốt đối với trải nghiệm và hiệu quả học tập của người dùng. Bài toán đánh giá chất lượng câu trả lời có thể được hiểu là việc xác định mức độ phù hợp, chính xác và hữu ích của một câu trả lời đối với một câu hỏi cụ thể.

Về mặt Machine Learning, đây là một bài toán học có giám sát (supervised learning), trong đó mỗi cặp dữ liệu gồm (câu hỏi, câu trả lời) được gán một nhãn chất lượng. Nhãn này có thể được biểu diễn dưới nhiều dạng khác nhau, chẳng hạn như phân loại nhị phân (tốt / kém), phân loại đa lớp hoặc chấm điểm liên tục theo thang đo (ví dụ từ 1 đến 5). Trong phạm vi đề tài này, bài toán được đơn giản hóa thành bài toán phân loại nhị phân nhằm phù hợp với quy mô dữ liệu và mục tiêu nghiên cứu ban đầu.

Đầu vào của mô hình là cặp câu hỏi – câu trả lời đã được tiền xử lý và biểu diễn dưới dạng vector đặc trưng. Đầu ra của mô hình là dự đoán về chất lượng câu trả lời, cho biết liệu câu trả lời đó có đáp ứng được yêu cầu về nội dung, mức độ liên quan và tính rõ ràng hay không. Việc giải quyết bài toán này giúp hệ thống AI trợ giảng có thể tự động phát hiện và hạn chế các câu trả lời sai lệch, lan man hoặc mang tính suy đoán.

2.2. Biểu diễn văn bản trong Machine Learning

Biểu diễn văn bản là một bước quan trọng trong các bài toán xử lý ngôn ngữ tự nhiên, bởi các mô hình Machine Learning truyền thống chỉ có thể làm việc với dữ liệu dạng số. Do đó, văn bản cần được chuyển đổi thành các vector số sao cho vẫn giữ được thông tin ngữ nghĩa cần thiết.

Phương pháp đầu tiên được sử dụng trong đề tài là TF-IDF (Term Frequency – Inverse Document Frequency). TF-IDF đo lường mức độ quan trọng của một từ trong một văn bản dựa trên tần suất xuất hiện của từ đó trong văn bản và trong toàn bộ tập dữ liệu. Phương pháp này có ưu điểm là đơn giản, dễ triển khai và hiệu quả đối với các tập dữ liệu nhỏ. Trong đề tài, TF-IDF được áp dụng trên

chuỗi kết hợp của câu hỏi và câu trả lời nhằm nắm bắt được các từ khóa quan trọng liên quan đến nội dung trả lời.

Tuy nhiên, TF-IDF chỉ phản ánh thông tin thống kê về từ vựng mà chưa thể hiện được mối quan hệ ngữ nghĩa sâu giữa các câu. Do đó, đề tài kết hợp thêm phương pháp biểu diễn ngữ nghĩa ở mức câu thông qua Sentence Embedding. Cụ thể, mô hình Sentence-BERT (SBERT) được sử dụng để ánh xạ câu hỏi và câu trả lời vào không gian vector ngữ nghĩa có chiều cố định. Từ các vector này, độ tương đồng ngữ nghĩa giữa câu hỏi và câu trả lời được tính bằng cosine similarity.

Việc kết hợp giữa TF-IDF và Sentence Embedding cho phép mô hình vừa khai thác được thông tin từ vựng, vừa nắm bắt được mức độ liên quan về mặt ngữ nghĩa giữa câu hỏi và câu trả lời.

2.3. Các mô hình Machine Learning sử dụng

Trong phạm vi nghiên cứu, Logistic Regression được lựa chọn làm mô hình phân loại chính cho bài toán đánh giá chất lượng câu trả lời. Logistic Regression là một mô hình phân loại tuyến tính, trong đó xác suất một mẫu dữ liệu thuộc về một lớp được mô hình hóa thông qua hàm sigmoid.

Ưu điểm của Logistic Regression là đơn giản, dễ huấn luyện, ít yêu cầu tài nguyên tính toán và hoạt động hiệu quả với tập dữ liệu có kích thước nhỏ. Ngoài ra, mô hình còn cho phép phân tích mức độ đóng góp của từng đặc trưng vào kết quả dự đoán, từ đó giúp người nghiên cứu hiểu rõ hơn về hành vi của mô hình.

Mặc dù các mô hình phức tạp hơn như Support Vector Machine hay các mô hình học sâu có thể mang lại hiệu quả cao hơn trong những tập dữ liệu lớn, Logistic Regression vẫn là lựa chọn phù hợp cho giai đoạn nghiên cứu ban đầu và mục tiêu minh họa của đề tài.

2.4. Các chỉ số đánh giá mô hình

Để đánh giá hiệu quả của mô hình phân loại, đề tài sử dụng các chỉ số đánh giá phổ biến trong Machine Learning. Precision đo lường tỷ lệ các câu trả lời được dự đoán là tốt và thực sự tốt. Recall phản ánh khả năng mô hình phát hiện được các câu trả lời tốt trong toàn bộ tập dữ liệu. F1-score là trung bình điều hòa giữa Precision và Recall, cho phép đánh giá cân bằng giữa hai chỉ số này.

Bên cạnh đó, confusion matrix được sử dụng để minh họa trực quan số lượng dự đoán đúng và sai của mô hình đối với từng lớp. Thông qua confusion matrix, có thể phân tích chi tiết các trường hợp mô hình dự đoán sai để rút ra những hạn chế và hướng cải thiện.

CHƯƠNG 3: XÂY DỰNG DỮ LIỆU

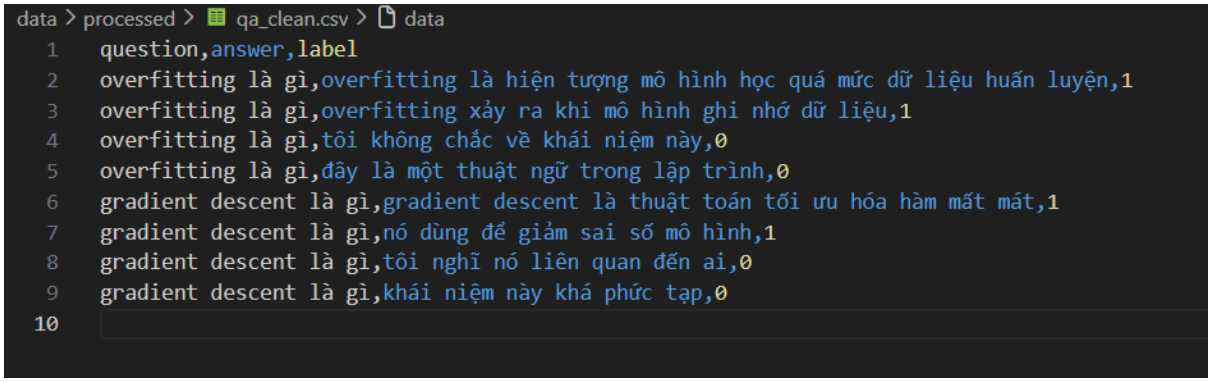
3.1. Thu thập dữ liệu

Dữ liệu đóng vai trò nền tảng trong mọi bài toán Machine Learning, đặc biệt là các bài toán xử lý ngôn ngữ tự nhiên. Đối với bài toán đánh giá chất lượng câu trả lời của AI trợ giảng, dữ liệu không chỉ cần phản ánh đúng nội dung chuyên môn mà còn phải thể hiện được sự khác biệt rõ ràng giữa các câu trả lời chất lượng cao và các câu trả lời kém chất lượng.

Trong phạm vi đề tài này, nhóm tiến hành xây dựng tập dữ liệu thủ công dựa trên các câu hỏi và câu trả lời liên quan đến những khái niệm cơ bản trong lĩnh vực Machine Learning. Các câu hỏi được lựa chọn là những câu hỏi thường gặp của sinh viên trong quá trình học tập, ví dụ như các câu hỏi yêu cầu định nghĩa, giải thích hoặc mô tả khái niệm. Điều này giúp đảm bảo rằng dữ liệu có tính thực tiễn và phù hợp với bối cảnh ứng dụng của hệ thống AI trợ giảng.

Các câu trả lời trong tập dữ liệu được xây dựng theo hai hướng. Thứ nhất là các câu trả lời đúng, ngắn gọn, bám sát nội dung câu hỏi và sử dụng thuật ngữ chuyên ngành chính xác. Thứ hai là các câu trả lời mang tính mơ hồ, không đầy đủ, thiếu liên quan hoặc thể hiện sự không chắc chắn. Việc chủ động xây dựng cả hai loại câu trả lời này giúp mô hình học được sự khác biệt giữa câu trả lời tốt và câu trả lời kém.

Mặc dù quy mô dữ liệu trong đề tài còn hạn chế, tập dữ liệu vẫn đáp ứng được mục tiêu nghiên cứu ban đầu là minh họa quy trình xây dựng mô-đun đánh giá chất lượng câu trả lời và kiểm chứng tính khả thi của phương pháp đề xuất. Trong các nghiên cứu tiếp theo, tập dữ liệu này có thể được mở rộng thông qua việc thu thập log câu hỏi – câu trả lời thực tế từ hệ thống AI trợ giảng.



Hình 3.1. Dữ liệu câu hỏi – câu trả lời sau khi tiền xử lý.

3.2. Gán nhãn dữ liệu

Sau khi thu thập và xây dựng các cặp câu hỏi – câu trả lời, bước tiếp theo là gán nhãn dữ liệu. Đây là bước quan trọng trong bài toán học có giám sát, bởi chất lượng của nhãn sẽ ảnh hưởng trực tiếp đến hiệu quả huấn luyện của mô hình.

Trong đề tài, quá trình gán nhãn được thực hiện hoàn toàn thủ công dựa trên các tiêu chí đánh giá đã được xác định trước. Một câu trả lời được gán nhãn chất lượng tốt khi thỏa mãn các điều kiện sau: nội dung trả lời đúng với kiến thức chuyên môn, có mức độ liên quan cao đến câu hỏi, diễn đạt rõ ràng và không gây hiểu nhầm cho người học. Ngược lại, các câu trả lời mang tính chung chung, trả lời sai trọng tâm, hoặc thể hiện sự không chắc chắn sẽ được gán nhãn kém chất lượng.

Việc gán nhãn thủ công cho phép kiểm soát tốt chất lượng dữ liệu, đặc biệt trong bối cảnh dữ liệu có quy mô nhỏ. Tuy nhiên, phương pháp này cũng tồn tại hạn chế nhất định do yếu tố chủ quan của người gán nhãn. Trong các hệ thống lớn hơn, quá trình gán nhãn có thể được thực hiện bởi nhiều người và áp dụng cơ chế đồng thuận để giảm sai lệch.

3.3. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước không thể thiếu trong các bài toán xử lý ngôn ngữ tự nhiên. Mục tiêu của bước này là loại bỏ nhiễu, chuẩn hóa dữ liệu và giúp mô hình học được các đặc trưng quan trọng một cách hiệu quả hơn.

Trong đề tài, dữ liệu văn bản được tiền xử lý thông qua các bước cơ bản. Trước hết, toàn bộ văn bản được chuyển về chữ thường nhằm tránh việc cùng một từ nhưng được xem là khác nhau do khác biệt về chữ hoa – chữ thường. Tiếp theo, các dấu câu và ký tự đặc biệt được loại bỏ để giảm số lượng từ vựng không cần thiết.

Sau quá trình tiền xử lý, dữ liệu được lưu lại và sử dụng cho các bước trích xuất đặc trưng và huấn luyện mô hình. Mặc dù đề tài chưa áp dụng các kỹ thuật tiền xử lý nâng cao như loại bỏ stopwords tiếng Việt hay chuẩn hóa dấu tiếng Việt, các bước tiền xử lý cơ bản này vẫn đủ để đảm bảo tính nhất quán của dữ liệu trong phạm vi nghiên cứu.

CHƯƠNG 4: THIẾT KẾ VÀ HUẤN LUYỆN MÔ HÌNH

4.1. Kiến trúc tổng thể của module đánh giá

Mô-đun đánh giá chất lượng câu trả lời được thiết kế như một thành phần độc lập trong hệ thống AI trợ giảng, có thể dễ dàng tích hợp vào nhiều kiến trúc chatbot hoặc hệ thống hỏi – đáp khác nhau. Về mặt tổng thể, mô-đun này hoạt động theo một pipeline tuần tự, trong đó mỗi bước đảm nhận một nhiệm vụ cụ thể và đầu ra của bước trước sẽ là đầu vào của bước sau.

Quy trình bắt đầu khi hệ thống AI trợ giảng sinh ra một câu trả lời cho câu hỏi của người học. Cặp dữ liệu gồm câu hỏi và câu trả lời này sẽ được chuyển đến mô-đun đánh giá. Tại đây, dữ liệu văn bản trước hết được tiền xử lý và chuẩn hóa để đảm bảo tính nhất quán. Sau đó, mô-đun tiến hành trích xuất đặc trưng nhằm chuyển đổi dữ liệu văn bản thành các vector số phù hợp với mô hình Machine Learning.

Các vector đặc trưng sau khi được xây dựng sẽ được đưa vào mô hình phân loại đã được huấn luyện trước đó. Mô hình sẽ dự đoán nhãn chất lượng của câu trả lời và trả về kết quả đánh giá kèm theo mức độ tin cậy. Dựa trên kết quả này, hệ thống AI trợ giảng có thể quyết định hiển thị câu trả lời cho người học hoặc yêu cầu AI sinh lại câu trả lời khác.

Thiết kế theo dạng pipeline giúp mô-đun đánh giá có tính linh hoạt cao, dễ dàng mở rộng và thay thế từng thành phần khi cần thiết, chẳng hạn như thay đổi phương pháp trích xuất đặc trưng hoặc mô hình phân loại.

4.2. Trích xuất đặc trưng

Trích xuất đặc trưng là bước quan trọng nhất trong quá trình xây dựng mô hình đánh giá chất lượng câu trả lời, bởi hiệu quả của mô hình phụ thuộc lớn vào khả năng biểu diễn thông tin của các đặc trưng đầu vào. Trong đề tài này, nhóm sử dụng kết hợp nhiều loại đặc trưng khác nhau nhằm phản ánh chất lượng câu trả lời ở cả mức độ từ vựng, ngữ nghĩa và hình thức diễn đạt.

Trước hết, đặc trưng TF-IDF được sử dụng để biểu diễn thông tin thống kê về từ vựng. Cụ thể, câu hỏi và câu trả lời được nối lại thành một chuỗi văn bản duy nhất, sau đó áp dụng phương pháp TF-IDF để tạo ra vector đặc trưng. Cách tiếp cận này giúp mô hình

học được các từ khóa quan trọng xuất hiện trong câu trả lời và mối liên hệ của chúng với nội dung câu hỏi.

Bên cạnh TF-IDF, đề tài sử dụng độ tương đồng ngữ nghĩa giữa câu hỏi và câu trả lời làm một đặc trưng quan trọng. Độ tương đồng này được tính bằng cosine similarity giữa các vector embedding của câu hỏi và câu trả lời, trong đó các embedding được sinh ra bởi mô hình Sentence-BERT. Đặc trưng này cho phép mô hình đánh giá mức độ liên quan về mặt ngữ nghĩa, ngay cả khi câu trả lời không sử dụng đúng các từ khóa xuất hiện trong câu hỏi.

Ngoài các đặc trưng tự động, đề tài còn thiết kế một số đặc trưng thủ công nhằm phản ánh chất lượng câu trả lời ở khía cạnh hình thức. Các đặc trưng này bao gồm độ dài câu trả lời, tỷ lệ bao phủ từ vựng giữa câu hỏi và câu trả lời, số lượng từ thể hiện sự không chắc chắn và mức độ lặp từ trong câu trả lời. Việc kết hợp các đặc trưng thủ công với đặc trưng ngữ nghĩa giúp mô hình có cái nhìn toàn diện hơn về chất lượng câu trả lời.

4.3. Huấn luyện mô hình

Sau khi trích xuất đặc trưng, dữ liệu được đưa vào giai đoạn huấn luyện mô hình. Trong đề tài, mô hình Logistic Regression được lựa chọn do tính đơn giản, dễ triển khai và phù hợp với quy mô dữ liệu hiện có. Logistic Regression mô hình hóa xác suất một câu trả lời thuộc về lớp chất lượng tốt thông qua hàm sigmoid, từ đó cho phép đưa ra quyết định phân loại.

Do tập dữ liệu có kích thước nhỏ, việc chia tập train và test cố định có thể dẫn đến kết quả đánh giá không ổn định. Vì vậy, đề tài sử dụng phương pháp cross-validation để đánh giá mô hình. Cụ thể, dữ liệu được chia thành nhiều phần, trong đó mỗi phần lần lượt được sử dụng làm tập kiểm tra trong khi các phần còn lại dùng để huấn luyện. Kết quả đánh giá cuối cùng được tính bằng trung bình các lần đánh giá, giúp phản ánh chính xác hơn khả năng tổng quát hóa của mô hình.

Sau khi đánh giá, mô hình được huấn luyện lại trên toàn bộ tập dữ liệu nhằm tận dụng tối đa thông tin sẵn có. Các tham số của mô hình được lựa chọn sao cho đảm bảo mô hình hội tụ và tránh hiện tượng quá khớp.

4.4. Tích hợp vào hệ thống Ai-trogiang

Sau khi hoàn tất quá trình huấn luyện và đánh giá, mô hình Machine Learning cần được lưu trữ để phục vụ cho giai đoạn triển khai và sử dụng trong môi trường

thực tế. Việc lưu trữ mô hình không chỉ giúp tiết kiệm thời gian và tài nguyên tính toán mà còn đảm bảo tính nhất quán giữa quá trình huấn luyện và quá trình suy luận khi hệ thống được đưa vào vận hành.

Trong đề tài này, mô hình Logistic Regression sau khi huấn luyện trên toàn bộ tập dữ liệu được lưu lại dưới dạng tệp nhị phân thông qua thư viện joblib. Song song với đó, bộ trích xuất đặc trưng TF-IDF cũng được lưu trữ riêng biệt. Việc tách riêng mô hình và bộ vectorizer giúp đảm bảo rằng dữ liệu đầu vào trong giai đoạn suy luận sẽ được biểu diễn theo đúng không gian đặc trưng mà mô hình đã được huấn luyện trước đó.

Quá trình suy luận được thiết kế sao cho mô-đun đánh giá có thể hoạt động độc lập với quá trình huấn luyện. Khi hệ thống AI trợ giảng sinh ra một câu trả lời mới, mô-đun đánh giá sẽ tiếp nhận cặp câu hỏi – câu trả lời, thực hiện các bước tiền xử lý và trích xuất đặc trưng giống hệt như trong giai đoạn huấn luyện. Sau đó, vector đặc trưng được đưa vào mô hình đã lưu để dự đoán nhãn chất lượng của câu trả lời.

Ngoài nhãn phân loại (tốt hoặc kém), mô hình còn trả về xác suất dự đoán tương ứng với từng lớp. Giá trị xác suất này được sử dụng như một chỉ số thể hiện mức độ tin cậy của mô hình đối với dự đoán đưa ra. Thông tin về mức độ tin cậy giúp hệ thống AI trợ giảng có thêm cơ sở để đưa ra quyết định, chẳng hạn như chấp nhận câu trả lời, yêu cầu AI sinh lại câu trả lời khác hoặc cảnh báo người dùng về độ tin cậy của nội dung.

Về mặt tích hợp hệ thống, mô-đun đánh giá chất lượng câu trả lời được triển khai như một thành phần trung gian trong luồng xử lý của chatbot AI trợ giảng. Sau khi mô hình sinh ngôn ngữ tạo ra câu trả lời, mô-đun đánh giá sẽ được gọi để kiểm tra chất lượng trước khi phản hồi được gửi đến người học. Cách tiếp cận này giúp giảm thiểu rủi ro cung cấp thông tin sai lệch và nâng cao độ tin cậy tổng thể của hệ thống.

Thiết kế mô-đun theo hướng độc lập và có thể tái sử dụng cũng tạo điều kiện thuận lợi cho việc mở rộng trong tương lai. Chẳng hạn, mô-đun đánh giá có thể được áp dụng cho nhiều môn học khác nhau hoặc thay thế mô hình Logistic Regression bằng các mô hình phức tạp hơn mà không làm thay đổi kiến trúc tổng thể của hệ thống. Đây là một ưu điểm quan trọng giúp hệ thống AI trợ giảng có

khả năng phát triển lâu dài và thích ứng với các yêu cầu mới trong môi trường giáo dục số.

CHƯƠNG 5: ĐÁNH GIÁ VÀ THỰC NGHIỆM

5.1. Thiết lập thực nghiệm

Mục tiêu của quá trình thực nghiệm là đánh giá hiệu quả của mô-đun đánh giá chất lượng câu trả lời được đề xuất, đồng thời kiểm chứng mức độ phù hợp của các đặc trưng và mô hình Machine Learning đã lựa chọn. Việc thiết lập thực nghiệm được xây dựng sao cho phản ánh đúng kịch bản sử dụng thực tế của hệ thống AI trợ giảng, trong đó mô-đun đánh giá đóng vai trò kiểm soát chất lượng đầu ra của AI.

Thực nghiệm được tiến hành trong môi trường lập trình Python với các thư viện phổ biến trong lĩnh vực Machine Learning và xử lý ngôn ngữ tự nhiên, bao gồm pandas, numpy, scikit-learn và sentence-transformers. Việc sử dụng các thư viện này giúp đảm bảo tính ổn định, khả năng tái lập kết quả và phù hợp với các chuẩn mực nghiên cứu hiện nay.

Tập dữ liệu sử dụng trong thực nghiệm chính là tập dữ liệu đã được xây dựng, gán nhãn và tiền xử lý ở Chương 3. Do quy mô dữ liệu còn hạn chế, đề tài không chia tập dữ liệu theo cách train/test truyền thống, mà sử dụng phương pháp cross-validation để đánh giá mô hình. Cách tiếp cận này cho phép tận dụng tối đa dữ liệu hiện có và giảm thiểu sự phụ thuộc vào cách chia dữ liệu ngẫu nhiên.

Các chỉ số đánh giá được lựa chọn bao gồm F1-score và confusion matrix. F1-score được ưu tiên sử dụng vì đây là chỉ số cân bằng giữa precision và recall, phù hợp với bài toán phân loại nhị phân trong bối cảnh dữ liệu có quy mô nhỏ. Confusion matrix được sử dụng để phân tích chi tiết các trường hợp dự đoán đúng và sai của mô hình.

Hình 5.1. Kết quả đánh giá mô hình bằng chỉ số F1-macro thông qua cross-validation.

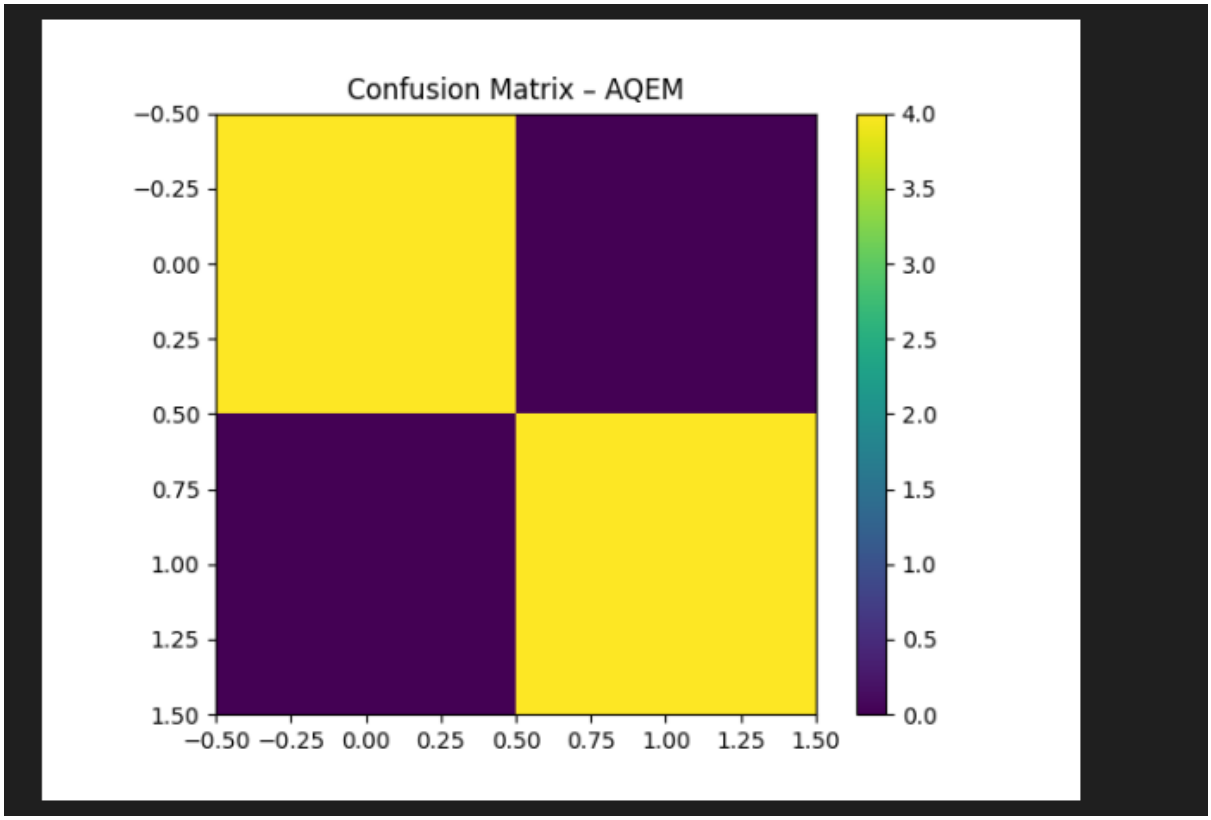
5.2. Kết quả thực nghiệm

Kết quả thực nghiệm cho thấy mô hình Logistic Regression kết hợp với các đặc trưng TF-IDF, độ tương đồng ngữ nghĩa và đặc trưng thủ công có khả năng phân

biệt tương đối tốt giữa các câu trả lời chất lượng tốt và các câu trả lời kém chất lượng. Thông qua các vòng cross-validation, mô hình đạt được giá trị F1-score ổn định, cho thấy khả năng học được các đặc trưng quan trọng của bài toán.

Confusion matrix được xây dựng để minh họa chi tiết kết quả dự đoán của mô hình. Ma trận này cho phép quan sát số lượng câu trả lời tốt được dự đoán đúng, số lượng câu trả lời kém bị phát hiện chính xác, cũng như các trường hợp mô hình dự đoán sai. Kết quả cho thấy mô hình hoạt động hiệu quả trong việc phát hiện các câu trả lời không liên quan hoặc mang tính suy đoán.

Ngoài ra, việc sử dụng xác suất dự đoán của mô hình giúp cung cấp thêm thông tin về mức độ tin cậy của mỗi dự đoán. Điều này đặc biệt hữu ích trong bối cảnh triển khai thực tế, khi hệ thống có thể dựa vào mức độ tin cậy để đưa ra các quyết định linh hoạt hơn.



Hình 5.2. Ma trận nhầm lẫn của mô hình phân loại chất lượng câu trả lời.

5.3. Phân tích kết quả

Từ kết quả thực nghiệm, có thể nhận thấy rằng việc kết hợp nhiều loại đặc trưng khác nhau mang lại hiệu quả rõ rệt cho bài toán đánh giá chất lượng câu trả lời.

Các đặc trưng TF-IDF giúp mô hình nắm bắt được các từ khóa quan trọng, trong khi độ tương đồng ngữ nghĩa cho phép phát hiện mức độ liên quan giữa câu hỏi và câu trả lời ngay cả khi cách diễn đạt khác nhau.

Các đặc trưng thủ công, mặc dù đơn giản, nhưng đóng vai trò bổ trợ quan trọng trong việc phát hiện các câu trả lời mang tính mơ hồ hoặc thiếu chắc chắn. Điều này cho thấy rằng việc kết hợp giữa kiến thức miền và các kỹ thuật học máy có thể mang lại hiệu quả cao hơn so với việc chỉ sử dụng một loại đặc trưng duy nhất.

Tuy nhiên, kết quả thực nghiệm cũng cho thấy một số hạn chế. Do quy mô dữ liệu nhỏ, khả năng tổng quát hóa của mô hình vẫn còn hạn chế và chưa thể đảm bảo hiệu quả trong các kịch bản phức tạp hơn. Một số trường hợp mô hình dự đoán sai xuất phát từ việc câu trả lời có nội dung đúng một phần nhưng cách diễn đạt chưa rõ ràng.

Những phân tích này cho thấy mô-đun đánh giá chất lượng câu trả lời có tiềm năng ứng dụng thực tế, đồng thời cũng chỉ ra các hướng cải thiện trong tương lai, chẳng hạn như mở rộng tập dữ liệu, sử dụng các mô hình học sâu hoặc bổ sung thêm các đặc trưng ngữ cảnh.

CHƯƠNG 6: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1. Kết luận

Trong bối cảnh các hệ thống AI trợ giảng ngày càng được ứng dụng rộng rãi trong giáo dục số, vấn đề đảm bảo chất lượng và độ tin cậy của câu trả lời do AI sinh ra trở nên đặc biệt quan trọng. Đề tài này đã tập trung nghiên cứu và xây dựng một mô-đun đánh giá chất lượng câu trả lời của AI trợ giảng dựa trên các kỹ thuật Machine Learning, với mục tiêu đóng vai trò như một lớp kiểm soát chất lượng trước khi thông tin được cung cấp cho người học.

Thông qua quá trình nghiên cứu, đề tài đã làm rõ bài toán đánh giá chất lượng câu trả lời dưới góc độ học máy có giám sát, trong đó đầu vào là cặp câu hỏi – câu trả lời và đầu ra là nhãn thể hiện mức độ chất lượng. Trên cơ sở đó, nhóm đã xây dựng một tập dữ liệu minh họa, thực hiện gán nhãn thủ công và tiến hành tiền xử lý nhằm đảm bảo tính nhất quán và phù hợp cho việc huấn luyện mô hình.

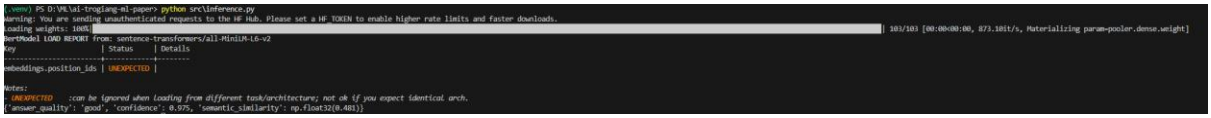
Về mặt phương pháp, đề tài đã đề xuất cách tiếp cận kết hợp giữa các đặc trưng thống kê truyền thống và các đặc trưng ngữ nghĩa hiện đại. Cụ thể, việc sử dụng TF-IDF giúp mô hình khai thác được thông tin từ vựng và các từ khóa quan trọng, trong khi Sentence-BERT và độ tương đồng cosine cho phép đánh giá mức độ liên quan về mặt ngữ nghĩa giữa câu hỏi và câu trả lời. Bên cạnh đó, các đặc trưng thủ công được thiết kế dựa trên kiến thức miền cũng góp phần phản ánh chất lượng câu trả lời ở khía cạnh hình thức và cách diễn đạt.

Mô hình Logistic Regression được lựa chọn và huấn luyện trong đề tài cho thấy khả năng phân loại tương đối tốt trong phạm vi dữ liệu thử nghiệm. Kết quả thực nghiệm cho thấy mô hình có thể phát hiện hiệu quả các câu trả lời kém chất lượng, không liên quan hoặc mang tính suy đoán, qua đó góp phần giảm thiểu nguy cơ cung cấp thông tin sai lệch cho người học. Việc sử dụng cross-validation giúp đảm bảo kết quả đánh giá có độ tin cậy cao hơn so với các phương pháp đánh giá đơn giản.

Quan trọng hơn, đề tài không chỉ dừng lại ở việc xây dựng và đánh giá mô hình, mà còn xem xét khả năng tích hợp mô-đun đánh giá vào hệ thống AI trợ giảng thực tế. Thiết kế mô-đun theo hướng độc lập, dễ mở rộng và tái sử dụng cho thấy tính khả thi của giải pháp trong các hệ thống giáo dục số. Mô-đun đánh giá có thể

được xem như một thành phần hỗ trợ quan trọng, giúp nâng cao độ tin cậy, tính nhất quán và mức độ an toàn của các hệ thống AI trợ giảng trong thực tế.

Tổng kết lại, đề tài đã đạt được các mục tiêu đề ra ban đầu, bao gồm việc xây dựng mô-đun đánh giá chất lượng câu trả lời dựa trên Machine Learning, kiểm chứng tính hiệu quả của phương pháp thông qua thực nghiệm và đề xuất hướng tích hợp vào hệ thống AI trợ giảng. Những kết quả đạt được cho thấy đề tài có ý nghĩa cả về mặt học thuật lẫn ứng dụng, đồng thời đặt nền tảng cho các nghiên cứu và phát triển sâu hơn trong tương lai.



Hình 6.1. Kết quả suy luận đánh giá chất lượng câu trả lời của mô-đun AQEM.

6.2. Hướng phát triển

Mặc dù đề tài đã đạt được các mục tiêu nghiên cứu ban đầu và chứng minh được tính khả thi của mô-đun đánh giá chất lượng câu trả lời của AI trợ giảng, vẫn còn nhiều hướng phát triển tiềm năng có thể được tiếp tục nghiên cứu và mở rộng trong tương lai nhằm nâng cao hiệu quả và phạm vi ứng dụng của hệ thống.

Thứ nhất, một hướng phát triển quan trọng là mở rộng và đa dạng hóa tập dữ liệu huấn luyện. Trong phạm vi đề tài, dữ liệu được xây dựng thủ công với quy mô còn hạn chế, chủ yếu phục vụ mục đích minh họa và kiểm chứng phương pháp. Trong tương lai, hệ thống có thể thu thập dữ liệu thực tế từ log câu hỏi – câu trả lời của người dùng khi sử dụng AI trợ giảng. Việc mở rộng dữ liệu không chỉ giúp mô hình học được nhiều tình huống phong phú hơn mà còn cải thiện khả năng tổng quát hóa khi triển khai trong môi trường thực tế.

Thứ hai, mô-đun đánh giá có thể được mở rộng từ bài toán phân loại nhị phân sang bài toán chấm điểm đa mức. Thay vì chỉ phân loại câu trả lời thành tốt hoặc kém, hệ thống có thể đánh giá chất lượng theo nhiều mức độ khác nhau, chẳng hạn như rất tốt, chấp nhận được hoặc kém. Cách tiếp cận này cho phép hệ thống AI trợ giảng đưa ra các phản hồi linh hoạt hơn, ví dụ như chỉ cảnh báo người học khi độ tin cậy thấp hoặc tự động yêu cầu AI cải thiện câu trả lời khi điểm chất lượng chưa đạt ngưỡng mong muốn.

Thứ ba, về mặt mô hình, các phương pháp Machine Learning nâng cao và các mô hình học sâu có thể được nghiên cứu và áp dụng. Các mô hình như Support Vector Machine, Random Forest hoặc các mô hình dựa trên kiến trúc Transformer fine-tuning trực tiếp cho bài toán đánh giá chất lượng câu trả lời có tiềm năng mang lại hiệu quả cao hơn, đặc biệt khi tập dữ liệu được mở rộng. Việc so sánh và đánh giá các mô hình này sẽ giúp lựa chọn được phương pháp phù hợp nhất cho từng bối cảnh triển khai cụ thể.

Thứ tư, mô-đun đánh giá chất lượng câu trả lời có thể được tích hợp sâu hơn vào vòng lặp tương tác của hệ thống AI trợ giảng. Thay vì chỉ đóng vai trò đánh giá sau khi AI sinh câu trả lời, mô-đun này có thể được sử dụng để phản hồi ngược lại cho hệ thống sinh ngôn ngữ, từ đó hỗ trợ việc cải thiện prompt hoặc điều chỉnh chiến lược sinh câu trả lời. Cách tiếp cận này giúp xây dựng một hệ thống AI trợ giảng có khả năng tự cải thiện theo thời gian.

Cuối cùng, về mặt ứng dụng, mô-đun đánh giá chất lượng câu trả lời không chỉ giới hạn trong phạm vi AI trợ giảng mà còn có thể được áp dụng cho nhiều hệ thống hỏi – đáp khác nhau, chẳng hạn như chatbot tư vấn, hệ thống hỗ trợ khách hàng hoặc các nền tảng giáo dục trực tuyến. Điều này cho thấy tiềm năng ứng dụng rộng rãi của hướng nghiên cứu và mở ra nhiều cơ hội phát triển trong tương lai.

TÀI LIỆU THAM KHẢO

[1] Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

[2] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[3] Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*, O'Reilly Media, 2019.

[4] Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[5] Reimers, N. and Gurevych, I., "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[6] Hugging Face, *Sentence Transformers Documentation*. Truy cập tại: https://www.sbert.net

[7] Manning, C. D., Raghavan, P., and Schütze, H., *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[8] Jurafsky, D. and Martin, J. H., *Speech and Language Processing*, Pearson, 2023.

PHỤ LỤC

Phụ lục này trình bày tổng quan về cấu trúc thư mục và các thành phần chính của hệ thống mô-đun đánh giá chất lượng câu trả lời của AI trợ giảng được xây dựng trong đề tài.

Thư mục `data` chứa các tập dữ liệu phục vụ cho quá trình huấn luyện và đánh giá mô hình, bao gồm dữ liệu thô ban đầu và dữ liệu đã được tiền xử lý. Các dữ liệu này bao gồm các cặp câu hỏi – câu trả lời được gán nhãn chất lượng.

Thư mục `src` chứa các tệp mã nguồn chính của hệ thống, bao gồm:

- `preprocess.py`: thực hiện các bước tiền xử lý văn bản như chuẩn hóa chữ, loại bỏ ký tự không cần thiết và chuẩn bị dữ liệu đầu vào.

- `features.py`: trích xuất các đặc trưng TF-IDF, đặc trưng ngữ nghĩa dựa trên Sentence-BERT và các đặc trưng thủ công.

- `train.py`: huấn luyện mô hình Machine Learning và lưu trữ mô hình sau khi huấn luyện.

- `evaluate.py`: đánh giá mô hình thông qua các chỉ số như precision, recall, F1-score và confusion matrix.

- `inference.py`: triển khai chức năng suy luận và đánh giá chất lượng câu trả lời trong môi trường sử dụng thực tế.

Thư mục `results` lưu trữ các mô hình đã huấn luyện, vectorizer và kết quả đánh giá, phục vụ cho việc phân tích và triển khai hệ thống.