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

BÁO CÁO SEMINAR CHO KHOA HỌC DỮ LIỆU

Sentiment and Emotion Analysis
Natural Language Processing

Lê Nguyễn Quỳnh Anh	22280002
Mai Thị Kim Ngân	22280058
Hồ Trần Anh Thư	22280088

Người hướng dẫn: ThS Huỳnh Thanh Sơn

Mục lục

1 Giới thiệu bài toán

1.1 Bối cảnh

Trong những năm gần đây, sự phát triển nhanh chóng của Trí tuệ nhân tạo (AI) và Xử lý ngôn ngữ tự nhiên (NLP) đã thúc đẩy nhiều ứng dụng thực tiễn đặc biệt là trong y tế, từ phân tích hồ sơ bệnh án, hỗ trợ tư vấn, đến tự động hóa chăm sóc khách hàng và theo dõi sức khỏe tinh thần. Trong đó, **phân tích cảm xúc** (*sentiment analysis*) đóng vai trò như một "cảm biến xã hội", giúp hệ thống nhận diện trạng thái cảm xúc của người dùng trong các tương tác giữa bệnh nhân và nhân viên y tế (hoặc tổng đài/tư vấn viên). Việc phân loại cảm xúc theo các mức *tích cực/trung tính/tiêu cực* không chỉ hỗ trợ đo lường mức độ hài lòng dịch vụ, mà còn phản ánh mức độ lo âu, căng thẳng, mệt mỏi, cũng như các dấu hiệu tâm lý tiêu cực tiềm ẩn. Ở góc độ vận hành, kết quả phân tích cảm xúc có thể hỗ trợ phân luồng cuộc gọi, ưu tiên các trường hợp rủi ro cao, đánh giá chất lượng tư vấn theo thời gian thực và cải thiện trải nghiệm chăm sóc sức khỏe một cách có hệ thống.

Tuy nhiên, y tế là một miền dữ liệu nhạy cảm, nơi quyết định sai hoặc diễn giải sai có thể dẫn đến hệ quả nghiêm trọng. Khác với các miền như thương mại điện tử hay mạng xã hội, dữ liệu y tế thường giàu thuật ngữ chuyên ngành, đòi hỏi ngữ cảnh sâu (lịch sử bệnh, diễn tiến triệu chứng) và thể hiện cảm xúc đa chiều (bệnh nhân có thể vừa lo lắng vừa cố gắng bình tĩnh; vừa biết ơn vừa bất an). Thực tế cho thấy ngay cả con người cũng có thể bất đồng khi gán nhãn cảm xúc cho hội thoại y tế, bởi cảm xúc trong chăm sóc sức khỏe không "rõ ràng" như các phản hồi dịch vụ thông thường. Do đó, bài toán đặt ra yêu cầu nghiêm ngặt về thiết kế nhãn, quy trình gán nhãn và cách đánh giá mô hình trong bối cảnh y tế.

Ngoài thách thức kỹ thuật, y tế còn đặt ra yêu cầu cao về **độ tin cậy** và **trách nhiệm giải trình**. Các bên liên quan (bác sĩ, điều dưỡng, nhân viên tổng đài, quản trị bệnh viện, thậm chí bệnh nhân) thường không chấp nhận một mô hình hoạt động như "hộp đen", chỉ đưa ra nhãn hoặc xác suất mà không có cơ sở rõ ràng. Khi hệ thống dự đoán "tiêu cực", người dùng cần biết: tiêu cực vì điều gì, dựa trên phát ngôn nào, và liệu mô hình có nhầm lẫn với các câu mang tính "thông tin – hướng dẫn" vốn thường trung tính hay không. Thiếu minh bạch khiến mô hình khó được tin tưởng, khó tích hợp vào quy trình nghiệp vụ và khó đáp ứng yêu cầu quản trị rủi ro. Do đó, cùng với việc cải thiện độ chính xác, cộng đồng nghiên cứu ngày càng quan tâm đến các hướng tiếp cận giúp mô hình *giải thích được quyết định* theo cách có ý nghĩa với người dùng.

1.2 Lý do, động lực và bài toán nghiên cứu

Từ bối cảnh trên, có thể nhận thấy một "khoảng cách giữa nghiên cứu và triển khai", giữa nhu cầu thực tiễn của hệ thống y tế và cách tiếp cận phân tích cảm xúc truyền thống. Phần lớn mô hình hiện nay được tối ưu theo hướng chỉ trả về một nhãn cảm xúc (tích cực, trung tính hoặc tiêu cực), đôi khi kèm theo xác suất. Cách tiếp cận này có thể phù hợp với các bài toán rủi ro thấp, nhưng trong y tế – nơi quyết định có thể tác động trực tiếp đến chất lượng chăm sóc và an toàn bệnh nhân – một nhãn dự đoán đơn lẻ là chưa đủ để hỗ trợ ra quyết định. Khi thiếu cơ chế giải thích, mô hình dễ trở thành một "hộp đen" khó kiểm chứng, khó truy vết lỗi và khó đáp ứng yêu cầu trách nhiệm giải trình, từ đó hạn chế khả năng tích hợp vào quy trình nghiệp vụ y tế.

Vì vậy, động lực trọng tâm của nghiên cứu là đưa *Explainable AI* (XAI) vào bài toán phân tích cảm xúc trong y tế theo hướng có khả năng giải thích và có cơ sở. Thay vì chỉ xuất nhãn, mô hình sẽ cung cấp thêm một đoạn giải thích ngắn gọn dựa trên nội dung hội thoại, cho biết vì sao mô hình đưa ra nhãn đó. Trong bối cảnh y tế, khả năng

giải thích không chỉ giúp tăng độ tin cậy, mà còn là cơ chế giảm rủi ro: khi mô hình dự đoán sai, phần giải thích giúp người dùng phát hiện sai lệch sớm và hạn chế việc sử dụng kết quả một cách máy móc.

Xuất phát từ các mục tiêu trên, báo cáo này tập trung vào bài toán *Sentiment Reasoning* – một thiết lập mở rộng của *sentiment analysis*, trong đó mô hình thực hiện đồng thời hai nhiệm vụ:

1. **Sentiment Classification**: dự đoán nhãn cảm xúc thuộc ba lớp *Positive, Neutral, Negative*.
2. **Rationale Generation**: sinh ra đoạn *rationale* giải thích lý do cho nhãn dự đoán dựa trên đầu vào.

Trên nền tảng đó, mục tiêu của nhóm là xây dựng một quy trình thực nghiệm có hệ thống nhằm:

(i) tái lập các mô hình nền (*baseline*) theo các nhóm kiến trúc *Encoder*, *Encoder–Decoder* và *Decoder/LLM*;

(ii) đề xuất cải tiến hướng tới việc đồng thời nâng cao hiệu suất phân loại và tăng tính minh bạch của mô hình.

Các khái niệm nền tảng, kiến trúc mô hình và thước đo đánh giá sẽ được trình bày ở Chương 2; tổng quan công trình liên quan và bộ dữ liệu ở Chương 3; phương pháp và thiết lập thí nghiệm của nhóm ở Chương 4; kết quả và phân tích ở Chương 5; và cuối cùng là kết luận cùng định hướng phát triển ở Chương 6.

2 Cơ sở phát triển và định nghĩa

2.1 Sentiment Analysis và Explainable AI trong y tế

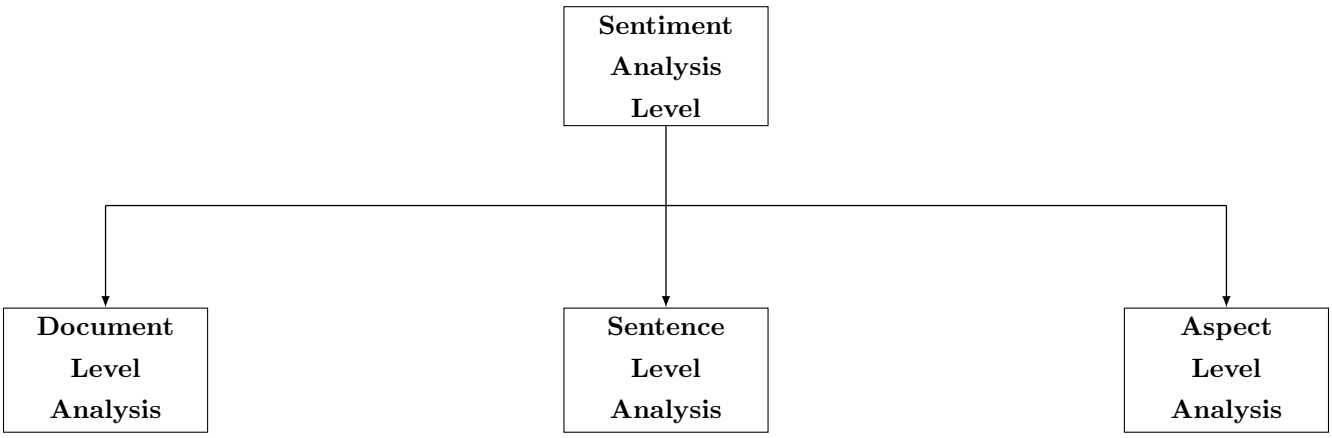
Mục tiêu của phần này là xây dựng một nền tảng khái niệm đầy đủ và để củng cố các chương sau của báo cáo. Cụ thể, nhóm làm rõ hai điểm lý thuyết quan trọng xuyên suốt bài báo cáo là: **phân tích cảm xúc** (*Sentiment analysis*) và **Explainable AI** (XAI)

2.1.1 Sentiment Analysis (SA)

Trong các tài liệu và nghiên cứu SA thường được đặt song hành với thuật ngữ *opinion mining*. Tác giả mô tả đây là một hướng nghiên cứu nhằm khai thác và mô hình hóa các "ý kiến/cảm xúc" trong văn bản để phục vụ các tác vụ như truy vấn, tóm tắt và hỗ trợ đưa ra quyết định. Ở mức khái quát hơn thì định nghĩa của sentiment analysis/opinion mining là nghiên cứu tính toán về các ý kiến, cảm xúc, thái độ, đánh giá, và các trạng thái chủ quan khác mà con người biểu đạt trong ngôn ngữ tự nhiên. Từ góc nhìn học máy, hai định nghĩa này thống nhất ở một điểm cốt lõi: SA sẽ biến tín hiệu "chủ quan" của con người (cảm xúc/thái độ) thành một dạng biểu diễn có thể học được và ra quyết định được (thường là nhãn phân loại, điểm số hoặc cấu trúc đánh giá).

Tuy nhiên, khi áp dụng ở lĩnh vực y tế, ý nghĩa của SA cần được hiểu rộng hơn so với các ví dụ đơn giản như đánh giá sản phẩm. Một tổng quan hệ thống cho thấy dữ liệu sức khỏe thường đến từ nhiều nguồn không đồng nhất (cộng đồng bệnh nhân, phản hồi dịch vụ, mạng xã hội, hội thoại hỗ trợ), và bản chất "cảm xúc" ở đây gắn chặt với tình trạng sức khỏe, trải nghiệm điều trị và bối cảnh tương tác.Nhấn mạnh thêm rằng *medical sentiment* không chỉ đơn thuần là tích cực/tiêu cực theo kiểu "thích/không thích", mà là sự biểu đạt cảm xúc trong những tình huống đặc thù: đau đớn, lo lắng, sợ hãi, hy vọng, nhẹ nhõm, thất vọng, hoặc thậm chí sự mơ hồ khi bệnh nhân chưa hiểu tình trạng của mình. Điều này dẫn tới một hệ quả quan trọng về mặt định nghĩa: SA trong y tế không chỉ quan tâm đến "định hướng cảm xúc" (polarity) mà còn quan tâm đến **ngữ cảnh và hàm ý lâm sàng** của phát ngôn.

Một cách hệ thống, SA có thể được thiết lập ở nhiều mức. Ở mức tài liệu (*document-level*), mục tiêu là gán nhãn cho toàn bộ một văn bản. Ở mức câu hoặc lượt phát ngôn (*sentence/utterance-level*), mô hình dự đoán cảm xúc cho từng đơn vị phát biểu, phù hợp với dữ liệu hội thoại. Ở mức khía cạnh (*aspect-level*), mô hình tách đối tượng/khía cạnh (ví dụ "thuốc", "bác sĩ", "chi phí") và gán nhãn cảm xúc tương ứng cho từng khía cạnh; thiết lập này hữu ích khi một văn bản chứa nhiều quan điểm khác nhau. Dù thiết lập nào, cốt lõi vẫn là ánh xạ từ dữ liệu ngôn ngữ sang một đầu ra biểu diễn cảm xúc nhằm phục vụ giám sát, tổng hợp, hoặc ra quyết định.



Hình 2.1: Sentiment Analysis Level

Trong báo cáo này, chúng tôi sử dụng SA đc theo thiết lập **phân loại 3 lớp** *Positive/Neutral/Negative* trên **transcript hội thoại y tế**. "Transcript" ở đây có thể là bản ghi do con người tạo (human transcript) hoặc văn bản sinh ra từ hệ thống nhận dạng tiếng nói tự động (ASR transcript).

2.1.2 Explainable AI (XAI): định nghĩa, mục tiêu và cách hiểu trong y tế

Trong khi SA trả lời câu hỏi "kết quả là gì?", XAI tập trung trả lời câu hỏi "tại sao lại ra kết quả đó?". XAI trở thành một chủ đề trung tâm bởi sự phát triển của các mô hình học sâu và mô hình ngôn ngữ lớn: chúng đạt hiệu năng cao nhưng thường khó diễn giải trực tiếp, tạo ra rào cản khi triển khai trong các miền rủi ro cao. Doshi-Velez và Kim đề xuất một hướng tiếp cận mang tính "khoa học" cho interpretability, nhấn mạnh rằng một lời giải thích chỉ có ý nghĩa khi nó được đánh giá trong bối cảnh nhiệm vụ và đối tượng người dùng cụ thể, thay vì những phát biểu chung chung.

Trong y tế, nhu cầu XAI trở nên cấp thiết vì quyết định dựa trên AI có thể ảnh hưởng trực tiếp tới chẩn đoán, điều trị và an toàn bệnh nhân. Tổng quan hệ thống của Antoniadi và cộng sự về XAI trong các hệ thống hỗ trợ quyết định lâm sàng (CDSS) nhấn mạnh rằng XAI đóng vai trò như một cầu nối giữa mô hình và người dùng chuyên môn: giải thích tốt giúp tăng niềm tin, hỗ trợ phát hiện sai sót, và cải thiện khả năng tích hợp vào quy trình lâm sàng; đồng thời, nghiên cứu cũng chỉ ra rằng đánh giá XAI cần quan tâm đến mức độ hữu ích thực tế và sự phù hợp với workflow. Tổng quan của Loh và cộng sự tiếp tục củng cố quan điểm rằng XAI trong y tế không chỉ nhằm "làm mô hình dễ hiểu", mà còn gắn với quản trị rủi ro, giảm thiên lệch, và cải thiện khả năng chấp nhận khi triển khai AI trong môi trường có ràng buộc chặt chẽ.

Trong bối cảnh ngôn ngữ, XAI có nhiều hình thức. Đối với y tế, lời giải thích dạng ngôn ngữ tự nhiên có lợi thế vì nó gần với cách con người lập luận và dễ tích hợp vào quy trình báo cáo/giám sát. Tuy nhiên, nó cũng đi kèm rủi ro "hợp lý hóa": lời giải thích nghe thuyết phục nhưng không phản ánh trung thực cơ chế quyết định của mô hình. Do đó, các khảo sát trong y tế nhấn mạnh rằng đánh giá XAI cần xem xét đồng thời **độ trung thực** (fidelity) và **tính hữu ích** (usefulness/actionability), cũng như ưu tiên đánh giá với người dùng mục tiêu trong workflow thực.

2.2 Mô hình và kiến trúc

2.2.1 Mô hình ASR (Automatic Speech Recognition)

Trong hệ thống Sentiment Reasoning, thành phần ASR Model đóng vai trò là khối xử lý tiền đề, có nhiệm vụ chuyển đổi dữ liệu âm thanh đầu vào (Audio signal) thành văn bản (ASR Transcript). Văn bản này sau đó sẽ trở thành dữ liệu đầu vào cho mô hình phân loại cảm xúc.

Việc tích hợp mô hình ASR giúp hệ thống có khả năng xử lý đa phương thức (multimodal), không chỉ giới hạn ở dữ liệu văn bản thuần túy (Human Transcript) mà còn mở rộng ra các kịch bản tương tác trực tiếp bằng giọng nói trong môi trường y tế.

PhoWhisper

PhoWhisper là bộ mô hình nhận dạng giọng nói tự động (ASR) chuyên biệt cho tiếng Việt, được tinh chỉnh (fine-tuned) từ kiến trúc Whisper trên tập dữ liệu tiếng Việt đa quy mô và đa vùng miền.

Đặc điểm: Mô hình được huấn luyện trên 844 giờ dữ liệu âm thanh thực tế, bao phủ đa dạng giọng nói các vùng miền Việt Nam. PhoWhisper được phát hành với năm phiên bản cấu hình khác nhau (Tiny, Base, Small, Medium, Large) nhằm tối ưu hóa cho các điều kiện tài nguyên tính toán từ thiết bị cá nhân đến hệ thống máy chủ. Đặc biệt, để tăng cường tính bền vững (robustness), 50% dữ liệu huấn luyện đã được bổ sung tiếng ồn môi trường.

Vai trò trong hệ thống: PhoWhisper đóng vai trò là lõi xử lý ngôn ngữ âm thanh. Với khả năng hiểu sâu đặc trưng âm học tiếng Việt, mô hình đảm nhận nhiệm vụ chuyển đổi chính xác các đoạn hội thoại y tế phức tạp thành văn bản, cung cấp đầu vào chất lượng.

Sherpa

Sherpa (sherpa-onnx) là một framework suy luận ASR mã nguồn mở thuộc hệ sinh thái k2-fsa, tập trung vào triển khai/inference các mô hình end-to-end.

Đặc điểm: Thư viện sử dụng ONNX Runtime và có thể chạy hoàn toàn cục bộ (không cần Internet), mô hình cho phép chuyển đổi âm thanh theo thời gian thực (streaming) hoặc theo tệp (non-streaming) với độ chính xác cao và độ trễ thấp.

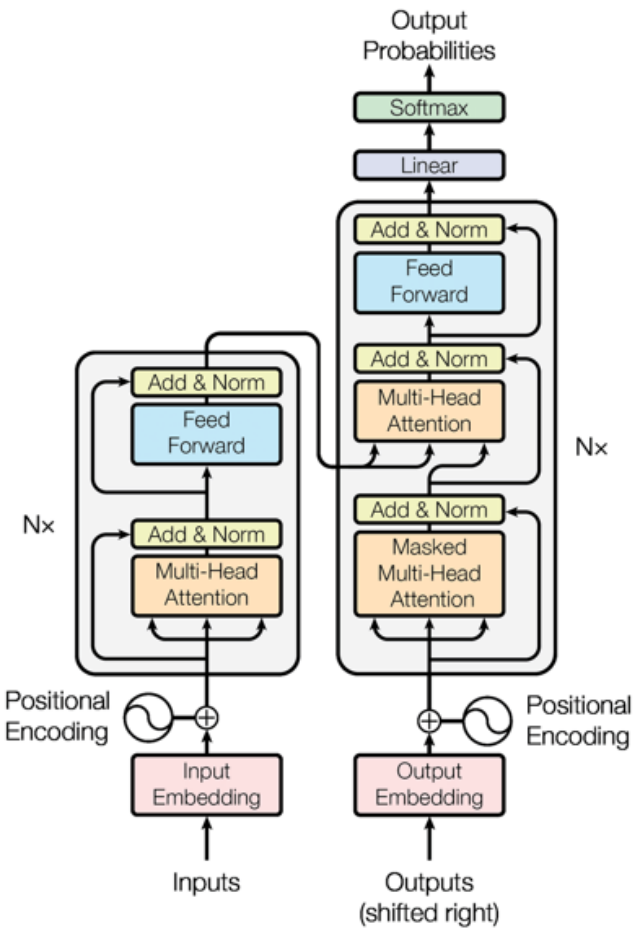
Vai trò trong hệ thống: Sherpa đóng vai trò là hạ tầng hỗ trợ (infrastructure). Mô hình này cung cấp bản dịch thuật (transcript) thô từ các audio y tế, làm cơ sở để mô hình Sentiment Reasoning phân tích các nhãn cảm xúc và đưa ra các giải thích (rationale) tương ứng.

2.2.2 Kiến trúc Transformer

Kiến trúc Transformers, được giới thiệu lần đầu trong nghiên cứu "Attention is All You Need", là nền tảng cốt lõi cho hầu hết các mô hình ngôn ngữ hiện đại. Thay vì sử dụng các cơ chế tuần tự (Recurrence), kiến trúc này dựa hoàn toàn trên cơ chế Attention để xử lý dữ liệu song song và nắm bắt ngữ cảnh toàn cục.

Dựa trên sơ đồ kiến trúc trên, mô hình bao gồm các thành phần trọng tâm sau:

- **Positional Encoding:** Do Transformers không xử lý văn bản theo thứ tự tuần tự như RNN, thành phần này giúp mô hình tiếp nhận thông tin về vị trí của các từ trong câu.



- **Multi-Head Attention:** Cho phép mô hình đồng thời tập trung vào các phần khác nhau của chuỗi đầu vào, giúp hiểu được các mối quan hệ ngữ nghĩa phức tạp trong thuật ngữ y tế.
- **Add & Norm (Residual Connections):** Các kết nối tắt kết hợp với chuẩn hóa lớp giúp việc huấn luyện các mạng cực sâu trở nên ổn định hơn, tránh hiện tượng triệt tiêu đạo hàm.
- **Feed Forward:** Mỗi lớp Encoder và Decoder đều chứa một mạng thần kinh truyền thẳng giúp biến đổi phi tuyến tính các đặc trưng đã học được.

Các mô hình trong nghiên cứu được chia thành ba nhóm chính:

- **Nhánh Encoder (Trái):** Chỉ sử dụng các khối Encoder để tạo ra vector biểu diễn ngữ cảnh. Các mô hình như PhoBERT và ViHealthBERT thuộc nhóm này, tối ưu cho việc phân loại label.
- **Cấu trúc đầy đủ Encoder-Decoder:** Kết hợp cả hai nhánh để thực hiện các tác vụ sinh văn bản có điều kiện (Sequence-to-Sequence). Các mô hình như ViT5 và BARTpho sử dụng cấu trúc này để vừa phân loại vừa sinh lời giải thích (Rationale).
- **Nhánh Decoder (Phải):** Chỉ sử dụng các khối Decoder với cơ chế Masked Multi-Head Attention. Đây là kiến trúc của các mô hình ngôn ngữ lớn (LLMs) như Vistral-7B, Llama3-8B và Qwen3-8B, cho phép thực hiện đồng thời các tác vụ suy luận phức tạp.

2.2.3 Mô hình Encoder

Mô hình PhoBERT

PhoBERT là mô hình ngôn ngữ tiền huấn luyện (pre-trained language model) đầu tiên và được thiết kế dành riêng cho tiếng Việt, được phát triển dựa trên kiến trúc RoBERTa (một biến thể tối ưu của BERT).

Một số đặc điểm chính nổi bật của mô hình PhoBERT:

- **Tối ưu cho tiếng Việt:** PhoBERT được huấn luyện trên tập dữ liệu tiếng Việt khổng lồ (khoảng 20GB văn bản từ Wikipedia và các trang tin tức) ⇒ giúp mô hình nắm bắt tốt các đặc thù về ngữ pháp, từ vựng và sắc thái biểu cảm trong tiếng Việt.
- **Xử lý cấp độ từ (Word-level):** PhoBERT sử dụng phương pháp tách từ (tokenization) dựa trên đơn vị từ thay vì âm tiết đơn lẻ ⇒ giúp mô hình hiểu được các từ ghép phức tạp trong tiếng Việt - một yếu tố cực kỳ quan trọng trong việc phân tích cảm xúc.
- **Kiến trúc Transformer mạnh mẽ:** Với cơ chế Self-Attention, PhoBERT có khả năng nắm bắt mối quan hệ giữa các từ trong một câu dài (ví dụ: các đoạn hội thoại y tế), từ đó hỗ trợ tốt cho việc tạo ra các giải thích (Rationale Generation) có tính logic cao.

Vai trò trong hệ thống: Trong quy trình thực nghiệm, PhoBERT đóng vai trò là bộ mã hóa (Encoder). Mô hình nhận đầu vào là các ASR Transcript hoặc Human Transcript, sau đó chuyển đổi chúng thành các vector biểu diễn đặc trưng (embeddings). Các vector này sẽ được đưa vào các lớp phân loại phía sau để đưa ra kết quả cuối cùng:

- **Sentiment Classification:** Xác định nhãn cảm xúc Positive/Neutral/Negative.
- **Rationale Generation:** Trích xuất hoặc tạo ra các cụm từ quan trọng giải thích tại sao mô hình lại đưa ra quyết định đó.

Mô hình ViHealthBERT

ViHealthBERT là mô hình ngôn ngữ tiền huấn luyện được thiết kế chuyên biệt cho lĩnh vực y tế tiếng Việt. Mô hình được xây dựng dựa trên kiến trúc BERT và được huấn luyện trên kho dữ liệu văn bản y khoa quy mô lớn (bao gồm hồ sơ bệnh án, báo cáo lâm sàng và các văn bản y tế tổng hợp), giúp mô hình nắm bắt tốt các đặc trưng chuyên ngành.

Một số đặc điểm chính nổi bật của mô hình ViHealthBERT:

- **Tối ưu cho ngôn ngữ y khoa:** *ViHealthBERT* được huấn luyện trên tập dữ liệu giàu thuật ngữ y học, giúp mô hình hiểu sâu hơn các khái niệm lâm sàng, triệu chứng, quy trình điều trị, cũng như các mối quan hệ giữa các thực thể y tế trong câu.
- **Xử lý cấp độ từ và cụm từ chuyên ngành:** Nhờ phương pháp tách từ phù hợp với ngữ cảnh và đặc thù thuật ngữ y học, mô hình có khả năng nhận diện và diễn giải chính xác các cụm thuật ngữ phức tạp (ví dụ: *tăng huyết áp*, *viêm phổi cộng đồng*, *rối loạn đông máu*).
- **Khai thác sức mạnh Transformer:** Với kiến trúc Attention của BERT, *ViHealthBERT* nắm bắt hiệu quả ngữ nghĩa dài hạn trong các đoạn mô tả y khoa, từ đó hỗ trợ tốt cho các tác vụ suy luận như phân loại, trích xuất thông tin hoặc tạo giải thích.

Vai trò trong hệ thống: Trong quy trình thực nghiệm, ViHealthBERT đóng vai trò là bộ mã hóa (Encoder). Các transcript từ ASR hoặc văn bản nhập từ bác sĩ/nhân viên y tế được đưa vào ViHealthBERT để chuyển thành

vector biểu diễn ngữ nghĩa chuyên ngành y tế. Các vector này tiếp tục được sử dụng trong các tác vụ sau:

- **Sentiment Classification**: Xác định cảm xúc của bác sĩ/bệnh nhân trong hội thoại y tế (Positive/Neutral/Negative).
- **Rationale Generation**: Trích xuất hoặc tạo ra các cụm từ, câu quan trọng giải thích vì sao mô hình đưa ra quyết định.

2.2.4 Mô hình Encoder-Decoder

Mô hình ViT5

ViT5 (Vietnamese T5) là một mô hình ngôn ngữ dựa trên kiến trúc Transformer với cấu trúc Encoder-Decoder, là phiên bản T5 cho tiếng Việt, được thiết kế chuyên biệt cho các tác vụ xử sinh (ví dụ: tóm tắt, NER dạng text-to-text, QA, sinh câu trả lời/giải thích).

Đặc điểm nổi bật của mô hình ViT5 cho bài toán Sentiment Reasoning:

- **Kiến trúc Encoder-Decoder:** ViT5 sở hữu cả Encoder và Decoder, cho phép mô hình không chỉ hiểu ngữ cảnh mà còn có khả năng sinh văn bản một cách trôi chảy. Đây là yếu tố quyết định để thực hiện tác vụ Rationale Generation - tạo ra các câu giải thích cho nhãn cảm xúc đã chọn.
- **Tiền huấn luyện quy mô lớn:** ViT5 được huấn luyện trên tập dữ liệu tiếng Việt khổng lồ thông qua các tác vụ như phục hồi từ bị mất (span corruption) ⇒ giúp mô hình nắm vững cấu trúc câu và sự liên kết logic trong tiếng Việt.
- **Linh hoạt trong tác vụ (Text-to-Text framework):** ViT5 xem mọi bài toán là chuyển đổi từ chuỗi văn bản này sang chuỗi văn bản khác. Trong hệ thống này, đầu vào là Transcript kèm theo prompt hướng dẫn và đầu ra là một chuỗi kết hợp bao gồm cả Label và Rationale (Ví dụ: "Label: Negative | Rationale: Vì bệnh nhân tỏ ra lo lắng về kết quả xét nghiệm").

Vai trò trong hệ thống: ViT5 được sử dụng như một mô hình sinh (Generative Model), ViT5 giúp hệ thống trở nên "thông minh" hơn bằng cách diễn giải lý do đằng sau mỗi dự đoán.

Mô hình BARTPho

BARTPho là mô hình pre-trained sequence-to-sequence dành riêng cho tiếng Việt, xây dựng dựa trên kiến trúc BART (Bidirectional and Auto-Regressive Transformers) với cơ chế denoising auto-encoder, được huấn luyện trên dữ liệu tiếng Việt lớn để học cách tái tạo và sinh văn bản tự nhiên.

BARTPho có hai phiên bản chính: *BARTpho-syllable* – xử lý theo đơn vị âm tiết tiếng Việt, và *BARTpho-word* – xử lý theo đơn vị từ. Cả hai đều dùng kiến trúc "large" của BART với encoder và decoder đầy đủ, giúp mô hình học biểu diễn sâu và mạnh cho các tác vụ sequence-to-sequence.

Các đặc điểm nổi bật của BARTPho trong bài toán này:

- **Cơ chế khử nhiễu (Denoising):** BARTPho được huấn luyện bằng cách phục hồi các đoạn văn bản gốc từ các biến thể bị làm nhiễu (như xáo trộn thứ tự câu hoặc che khuất từ) ⇒ giúp mô hình học được cấu trúc ngữ pháp và sự liên kết logic chặt chẽ trong tiếng Việt.
- **Sự kết hợp giữa Bidirectional và Auto-regressive:** Bộ mã hóa của BARTPho sử dụng cơ chế hai chiều (tương tự BERT) để hiểu ngữ cảnh, trong khi bộ giải mã hoạt động theo kiểu tự hồi quy (tương tự GPT) để sinh văn bản ⇒ giúp mô hình vừa phân loại cảm xúc chính xác, vừa diễn đạt lý do một cách mạch lạc.

- **Hiệu quả với tiếng Việt:** BARTPho kế thừa bộ từ điển (vocabulary) và cách xử lý ngôn ngữ đặc thù của tiếng Việt, giúp khắc phục nhược điểm của các mô hình đa ngôn ngữ khi xử lý các sắc thái biểu cảm phức tạp trong transcript y tế.

Vai trò trong hệ thống: Trong kiến trúc Sentiment Reasoning, BARTPho đóng vai trò là một mô hình sinh (Generative Model) tiềm năng. Nó nhận đầu vào là các bản hội thoại từ khối ASR và thực hiện đồng thời hoặc tuần tự hai nhiệm vụ:

- Sentiment Classification: Trích xuất đặc trưng từ Encoder để dự đoán nhãn cảm xúc.
- Rationale Generation: Sử dụng Decoder để chuyển đổi các đặc trưng đó thành văn bản giải thích tự do.

2.2.5 Mô hình Decoder

Mô hình Qwen3-8B

Qwen3-8B là một mô hình ngôn ngữ lớn (LLM) thuộc thế hệ thứ 3 của dòng Qwen (Alibaba Cloud), với kiến trúc Decoder-only và quy mô 8.2 tỷ tham số. Mô hình này được huấn luyện trên dữ liệu đa ngôn ngữ và thiết kế để xử lý các tác vụ NLP phức tạp như hỏi-đáp, tóm tắt, suy luận, đối thoại đa vòng, và chain-of-thought.

Các đặc điểm kỹ thuật nổi bật của Qwen3-8B:

- **Chế độ suy luận kép (Thinking Mode):** Qwen3 có khả năng chuyển đổi linh hoạt giữa "Thinking mode" (suy luận từng bước - Chain-of-Thought) và "Non-thinking mode", giúp cân bằng giữa tốc độ và chiều sâu của output.
- **Kiến trúc tối ưu:** Mô hình sử dụng cơ chế Grouped Query Attention (GQA) giúp tăng tốc độ xử lý và tiết kiệm bộ nhớ, cùng với SwiGLU activation và RoPE (Rotary Positional Embeddings) để xử lý ngữ cảnh dài lên đến 128,000 tokens (thông qua kỹ thuật YaRN).
- **Khả năng đa ngôn ngữ và chuyên biệt hóa:** Qwen3-8B được huấn luyện trên tập dữ liệu khổng lồ (36 nghìn tỷ tokens), cho phép xử lý văn bản trong hơn 100 ngôn ngữ và hỗ trợ cực tốt tiếng Việt.

Vai trò trong hệ thống:

- **Zero-shot/Few-shot Reasoning:** Qwen3-8B có thể thực hiện suy luận cảm xúc ngay lập tức mà không cần huấn luyện lại quá nhiều, nhờ vào lượng kiến thức khổng lồ có sẵn.
- **Sinh và tinh chỉnh Rationale phức tạp:** Tạo ra các lời giải thích chi tiết, mang tính diễn dịch sâu từ transcript và nhãn cảm xúc; đồng thời đóng vai trò hậu xử lý để kiểm chứng và làm phong phú kết quả từ các mô hình nhỏ (ViT5, BARTPho).
- **Xử lý hội thoại dài:** Với cửa sổ ngữ cảnh lớn, Qwen3-8B hỗ trợ tóm tắt và phân tích cảm xúc từ những đoạn hội thoại y tế dài hơi mà không lo bị mất thông tin do giới hạn ký tự.

Mô hình Vistral-7B

Vistral-7B là mô hình LLM thuộc dòng Vistral, được thiết kế tối ưu cho tiếng Việt và các ứng dụng NLP đa nhiệm. Mô hình có 7 tỷ tham số, sử dụng kiến trúc Decoder-only và được huấn luyện từ đầu (from-scratch) trên tập dữ liệu tiếng Việt quy mô rất lớn, bao gồm báo chí, mạng xã hội, tài liệu học thuật và hội thoại thực tế.

Các đặc điểm kỹ thuật nổi bật của Vistral-7B:

- **Tối ưu cho tiếng Việt:** *Vistral-7B* được huấn luyện chủ yếu trên dữ liệu tiếng Việt, giúp mô hình hiểu tốt cấu trúc ngữ pháp, từ vựng và các đặc trưng ngữ nghĩa của ngôn ngữ.

- **Hiệu quả tính toán:** Kiến trúc gọn nhẹ với 7B tham số cho phép mô hình triển khai trên các GPU phổ thông, phù hợp với những ứng dụng đòi hỏi tốc độ xử lý nhanh.
- **Khả năng suy luận và tóm tắt:** Nhờ chiến lược huấn luyện đa nhiệm, mô hình xử lý hiệu quả các tác vụ y tế như tóm tắt hội thoại, phân loại cảm xúc, trích xuất thực thể và suy luận ngữ cảnh.

Vai trò trong hệ thống. Vistral-7B đóng vai trò hỗ trợ các tác vụ NLP phía sau, đặc biệt là tóm tắt transcript y tế, kiểm chứng và chuẩn hóa nội dung đầu ra từ các mô hình ASR hoặc encoder (như ViT5, BARTPho). Mô hình giúp tạo ra văn bản mạch lạc, đầy đủ thông tin và giữ được ngữ nghĩa gốc từ hội thoại y tế.

Mô hình Llama 3.1 8B

Llama 3.1 8B là mô hình LLM thế hệ mới của Meta, sở hữu 8 tỷ tham số với kiến trúc *Decoder-only*. Mô hình được thiết kế chú trọng vào khả năng suy luận, tính ổn định và hiệu suất đa ngôn ngữ, trong đó có tiếng Việt. Llama 3.1 8B được huấn luyện trên tập dữ liệu hàng nghìn tỷ token, bao gồm văn bản mở, hội thoại và nội dung kỹ thuật.

Các đặc điểm kỹ thuật nổi bật của Llama 3.1 8B:

- **Khả năng suy luận mạnh:** Llama 3.1 cải thiện đáng kể mạch suy luận so với thế hệ 3.0, cho phép mô hình phân tích, giải thích và trả lời các câu hỏi y tế dài với tính logic và độ chính xác cao.
- **Hiệu suất đa ngôn ngữ và thích ứng tốt:** Dù không được huấn luyện riêng cho tiếng Việt, mô hình vẫn cho hiệu quả cao nhờ tiền huấn luyện trên corpora đa ngữ quy mô lớn, từ đó hoạt động tốt trên các tác vụ tiếng Việt.
- **Tối ưu cho hội thoại:** Llama 3.1 8B được tinh chỉnh đặc biệt cho hội thoại, hỗ trợ xử lý ngữ cảnh dài, duy trì mạch hội thoại và sinh phản hồi tự nhiên.

Vai trò trong hệ thống. Llama 3.1 8B đảm nhiệm vai trò *reasoning model*, được sử dụng để phân tích cảm xúc phức tạp trong hội thoại y tế và tạo lập lời giải thích (*rationale*) thuyết phục. Mô hình cũng có thể kiểm chứng và làm giàu kết quả từ các mô hình nhỏ hơn như ViT5, BARTPho và ViHealthBERT.

2.3 Các chỉ số đánh giá

Bài toán trong báo cáo này gồm 2 thành phần đầu vào là audio và text, hai thành phần đầu ra là một nhãn phân loại cảm xúc rời rạc thuộc ba lớp *Positive/Neutral/Negative*, và đoạn văn bản giải thích. Vì vậy, việc đánh giá cần sử dụng hai nhóm thước đo tương ứng: thước đo phân loại cho nhãn cảm xúc và thước đo sinh văn bản/độ tương đồng ngữ nghĩa cho rationale.

2.3.1 Thước đo cho ASR

Để đánh giá chất lượng transcript từ ASR, nhóm sử dụng các chỉ số chuẩn trong nhận dạng tiếng nói, dựa trên căn chỉnh giữa transcript tham chiếu (thường là *human transcript*) và transcript sinh ra bởi ASR. Các lỗi được phân rã thành ba loại thao tác chỉnh sửa cơ bản:

- **Substitution (S)**: thay thế một từ/ký tự bằng một từ/ký tự khác,
- **Deletion (D)**: bỏ sót (xóa) một từ/ký tự,

- **Insertion (I)**: chèn thêm (thừa) một từ/ký tự.

Ngoài ra, C biểu thị số đơn vị được nhận dạng đúng (*Correct*), và N là tổng số đơn vị trong transcript tham chiếu.

WER (Word Error Rate). WER là thước đo phổ biến nhất ở mức từ:

$$\text{WER} = \frac{S + D + I}{N}. \tag{2.1}$$

WER càng thấp thì transcript càng gần với tham chiếu. Trong pipeline Sentiment Reasoning, WER phản ánh trực tiếp mức độ nhiễu mà mô hình sentiment phải đối mặt khi nhận transcript từ ASR.

CER (Character Error Rate). CER đo tỷ lệ lỗi ở mức ký tự (hữu ích khi đánh giá sai khác nhỏ về chính tả/dấu hoặc các biến thể từ vựng):

$$\text{CER} = \frac{S_c + D_c + I_c}{N_c}, \tag{2.2}$$

trong đó S_c, D_c, I_c là số phép thay thế/xóa/chèn ở mức ký tự, và N_c là tổng số ký tự trong transcript tham chiếu.

MER (Match Error Rate). MER đo tỷ lệ lỗi dựa trên tổng số thao tác căn chỉnh so với tổng số đơn vị đã căn chỉnh:

$$\text{MER} = \frac{S + D + I}{S + D + I + C}. \tag{2.3}$$

MER thường gần với WER nhưng nhấn mạnh góc nhìn "tỷ lệ lỗi trên tổng số phần tử căn chỉnh"

2.3.2 Thước đo cho Sentiment Classification

Giả sử tập đánh giá có N mẫu, nhãn thật là y_i và nhãn dự đoán là $\hat{y}_i$. Khi đó, **Accuracy** đo tỷ lệ dự đoán đúng trên toàn bộ tập:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\hat{y}_i = y_i\right), \tag{2.4}$$

trong đó $\mathbb{I}(\cdot)$ là hàm chỉ thị.

Tuy nhiên, trong các bài toán phân loại nhiều lớp, đặc biệt khi phân phối lớp có thể không cân bằng hoặc khi một lớp (ví dụ *Neutral*) có ranh giới mơ hồ, Accuracy có thể che khuất việc mô hình hoạt động kém ở một lớp cụ thể. Do đó, báo cáo này sử dụng thêm **Precision**, **Recall** và **F1-score** theo từng lớp để đánh giá chi tiết.

Với mỗi lớp $c \in \{NEG, NEU, POS\}$, ta định nghĩa:

- TP_c: số mẫu thuộc lớp c được dự đoán đúng là c,
- FP_c: số mẫu không thuộc lớp c nhưng bị dự đoán nhầm thành c,
- FN_c: số mẫu thuộc lớp c nhưng bị dự đoán nhầm sang lớp khác.

Khi đó:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \tag{2.5}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \tag{2.6}$$

$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}. \tag{2.7}$$

Trong bảng kết quả, chúng tôi báo cáo **F1 theo từng lớp** (F1 Neg., F1 Neu., F1 Pos.) để quan sát trực tiếp mô hình mạnh/yếu ở lớp nào. Đồng thời, chúng tôi sử dụng **Macro-F1** (Mac F1) để tổng hợp chất lượng phân loại theo cách *đối xử công bằng giữa các lớp*:

$$\text{Macro-F1} = \frac{1}{|C|} \sum_{c \in C} \text{F1}_c, \quad \text{với } C = \{NEG, NEU, POS\}. \tag{2.8}$$

Lý do chọn Macro-F1: Macro-F1 giúp tránh tình trạng mô hình "được điểm" nhờ dự đoán tốt lớp chiếm đa số, tránh dự đoán sai ở một lớp quan trọng (tiêu cực).

2.3.3 Thước đo cho Rationale Generation

Đầu ra rationale là một chuỗi văn bản, do đó việc đánh giá không thể chỉ dựa trên "đúng/sai" như phân loại. Trong báo cáo này, chúng tôi dùng hai nhóm thước đo phổ biến và bổ sung cho nhau: (i) thước đo dựa trên **chồng lấp n-gram** (ROUGE) và (ii) thước đo dựa trên **tương đồng ngữ nghĩa theo embedding** (BERTScore). Với mỗi mẫu, giả sử rationale tham chiếu (do người gán nhãn) là r và rationale mô hình sinh ra là $\hat{r}$.

ROUGE (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum)

ROUGE là nhóm thước đo đánh giá mức độ chồng lấp giữa văn bản sinh và văn bản tham chiếu. Trong đó:

- **ROUGE-1 (R-1)**: chồng lấp unigram (từ đơn),
- **ROUGE-2 (R-2)**: chồng lấp bigram (cặp từ),
- **ROUGE-L (R-L)**: dựa trên độ dài *Longest Common Subsequence* (LCS),
- **ROUGE-Lsum (R-Lsum)**: biến thể thường dùng cho tóm tắt nhiều câu, tính ROUGE-L trên cấu trúc "câu" (hữu ích khi rationale có nhiều mệnh đề).

ROUGE đo "mức giống về mặt từ vựng và cấu trúc" giữa $\hat{r}$ và r; điểm cao thường tương ứng với việc mô hình dùng các từ khóa/cụm từ tương tự người gán nhãn. Tuy nhiên, ROUGE có hạn chế: hai rationale có thể đúng về nghĩa nhưng dùng từ khác vẫn bị điểm thấp.

BERTScore

BERTScore đo độ tương đồng ngữ nghĩa giữa $\hat{r}$ và r bằng cách so khớp các token dựa trên embedding từ các mô hình kiểu BERT. Tóm tắt ý tưởng: mỗi token trong $\hat{r}$ sẽ được so khớp với token "gần nhất" trong r theo cosine similarity trên không gian embedding; từ đó tạo ra các biến thể tương tự Precision/Recall và F1 ở mức token-embedding. Trong bảng, nhóm báo cáo **BERTScore** như một thước đo phản ánh "độ giống về nghĩa" tốt hơn so với chồng lấp n-gram thuần túy.

3 Tổng quan các công trình nghiên cứu liên quan

Bài báo **Sentiment Reasoning for Healthcare** tập trung vào việc phân tích cảm xúc trong hội thoại y tế tiếng Việt và sinh rationale tương ứng giải thích vì sao mô hình đưa ra nhãn đó. Công trình này đóng vai trò quan trọng trong lĩnh vực hiểu ngôn ngữ tự nhiên trong môi trường y tế, đặc biệt khi xử lý dữ liệu đa dạng, nhiều cảm xúc và có tính nhạy cảm cao.

3.1 Bộ dữ liệu

3.1.1 Phân chia dữ liệu

Tập dữ liệu đầy đủ gồm 30,000 mẫu với 5 ngôn ngữ (tiếng Anh, tiếng Trung giản thể và phồn thể, tiếng Đức, tiếng Pháp). Bài báo tập trung vào phần dữ liệu tiếng Việt với tổng cộng 8,086 mẫu, được chia thành hai phần:

- **Training set**: 5,695 mẫu
- **Test set**: 2,183 mẫu

Bên cạnh đó, nhóm tác giả sử dụng các văn bản (transcript) được dich từ hệ thống nhận dạng tiếng nói (ASR) trong môi trường bệnh viện nhằm đánh giá hiệu quả của mô hình trong các ngữ cảnh thực tế.

3.1.2 Các đặc trưng của bộ dữ liệu

a. human_justification

Đây là lời giải thích/bình luận của annotator về lý do tại sao câu text mang nhãn cảm xúc đó (reasoning). Dùng để *giải thích mô hình* (explainability), là nguồn dữ liệu huấn luyện cho rationale-based models(training with rationales).

b. label

Có 3 label với phân bố giữa các lớp:

- **Positive:** 30%.
- **Neutral:** 50%,
- **Negative:** 20%,

Sự thiên lệch này là đặc trưng của các cuộc trò chuyện y tế thực tế, nơi mà lời khuyên và giải thích là rất phổ biến.

c. text

Lời của cả bác sĩ và bệnh nhân bằng tiếng Việt. Đây là feature input chính cho Sentiment Analysis.

d. audio

Đoạn ghi âm cuộc trò chuyện giữa bác sĩ và bệnh nhân, có độ dài trung bình từ 7-8 giây.

Bảng 3.1: Một số mẫu dữ liệu từ tập Sentiment Reasoning (minh họa)

human_justification	label	text	duration (s)	audio
nỗi đau tinh thần	negative	giận nó hay đau đớn cả ngày, không ngủ được và cảm giác căng thẳng kéo dài...	6.0	▶ _____
mô tả khách quan	neutral	khớp tại Việt Nam dễ sản xuất, nên lựa chọn phương án phù hợp với điều kiện hiện có...	5.0	▶ _____
khen ngợi xương và khớp chắc khỏe	positive	cái hệ cơ xương khớp thật là khỏe, vận động nhẹ nhàng và cảm giác cải thiện rõ...	6.0	▶ _____

3.2 Phương pháp thực nghiệm của tác giả

3.2.1 Automatic Speech Recognition (ASR)

Nhóm tác giả sử dụng mô hình wav2vec 2.0 (118M tham số) để chuyển đổi lời nói thành văn bản, đạt WER 29.6% trên tập kiểm tra.

End-to-end Sentiment Classification:
PhoWhisper (phiên bản base, huấn luyện trên 844 giờ tiếng Việt) được dùng cho phân loại cảm xúc bằng cách gắn thêm một tầng phân loại lên encoder.

End-to-end Sentiment Reasoning:
Mô hình Qwen2-Audio (7B tham số) được fine-tune trong cả hai thiết lập *Label Only* và *Label + Rationale* để thực hiện suy luận cảm xúc.

3.2.2 Phân loại cảm xúc

Kiến trúc **Encoder** rất phù hợp cho nhiệm vụ phân loại cảm xúc, nên một tầng phân loại tuyến tính được gắn trực tiếp vào đầu ra của mô hình. Tuy nhiên, các Encoder không thể sinh *rationale*, vì vậy chúng được sử dụng như các mô hình **baseline** trong bài báo.

- **PhoBERT-base** (110M tham số)
- **RoBERTa-base** (110M tham số)
- **ViHealthBERT** (110M tham số)

Các Encoder này được tiền huấn luyện trên tập dữ liệu tiếng Việt quy mô lớn.

3.2.3 Mô hình sinh

Các mô hình sinh (*generative models*) được sử dụng cho cả hai chế độ:

- **Label Only**: mô hình chỉ sinh ra nhãn cảm xúc.
- **Label + Rationale**: mô hình đồng thời sinh ra nhãn cảm xúc và đoạn giải thích (*rationale*) dưới dạng văn bản tự nhiên.

Trong bài báo, nhóm tác giả sử dụng cả hai loại mô hình sinh:

Encoder–Decoder

- **BARTPho** (139M tham số) tiền huấn luyện trên 20GB tiếng Việt từ Wikipedia và kho dữ liệu báo chí.
- **ViT5-base** (223M tham số) tiền huấn luyện trên 71GB tiếng Việt.

Decoder-only

- **Vistral-7B-Chat** (7B tham số)
- **vmlu-llm1** (7B tham số)

Các mô hình sinh phù hợp với nhiệm vụ tạo giải thích nhờ khả năng mô hình hóa chuỗi mạnh và khả năng sinh văn bản tự nhiên.

3.3 Kết quả

Model	Acc.	F1 Neg.	F1 Neu.	F1 Pos.	Mac F1	R-1	R-2	R-L	R-Lsum	BERTScore
Encoder (Label Only)										
PhoBERT	0.6674	0.6969	0.6607	0.6377	0.6651	–	–	–	–	–
ViHealthBERT	0.6752	0.6970	0.6718	0.6535	0.6741	–	–	–	–	–
Encoder–Decoder (Label Only)										
ViT5	0.6628	0.6922	0.6687	0.6007	0.6545	–	–	–	–	–
BARTPho	0.6523	0.6870	0.6571	0.5841	0.6427	–	–	–	–	–
Decoder (Label Only)										
vmlu-llm	0.6592	0.6768	0.6769	0.5911	0.6483	–	–	–	–	–
Vistral7B	0.6716	0.6858	0.6771	0.6398	0.6676	–	–	–	–	–
Encoder–Decoder (Label + Rationale)										
ViT5	0.6633	0.6936	0.6572	0.6335	0.6615	0.3910	0.2668	0.3653	0.3660	0.8093
BARTPho	0.6619	0.7029	0.6460	0.6265	0.6585	0.3871	0.2613	0.3658	0.3683	0.8077
Decoder (Label + Rationale)										
vmlu-llm	0.6729	0.7039	0.6714	0.6307	0.6687	0.3947	0.2467	0.3789	0.3796	0.8086
Vistral7B	0.6812	0.7152	0.6765	0.6425	0.6781	0.4155	0.2788	0.3880	0.3900	0.8101

Bảng 3.2: Kết quả đánh giá các mô hình với input là human transcript

Model	Acc.	F1 Neg.	F1 Neu.	F1 Pos.	Mac F1	R-1	R-2	R-L	R-Lsum	BERTScore
Encoder (Label Only)										
PhoBERT	0.6166	0.6418	0.6231	0.5658	0.6102	–	–	–	–	–
ViHealthBERT	0.6198	0.6307	0.6261	0.5934	0.6167	–	–	–	–	–
Encoder–Decoder (Label Only)										
ViT5	0.6157	0.6412	0.6258	0.5523	0.6064	–	–	–	–	–
BARTPho	0.6056	0.6364	0.6156	0.5311	0.5944	–	–	–	–	–
Decoder (Label Only)										
vmlu-llm	0.6216	0.6551	0.5186	0.5186	0.6011	–	–	–	–	–
Vistral7B	0.6299	0.6377	0.6537	0.5609	0.6174	–	–	–	–	–
Encoder–Decoder (Label + Rationale)										
ViT5	0.6189	0.6305	0.6286	0.5837	0.6143	0.3571	0.2202	0.3350	0.3366	0.8044
BARTPho	0.6129	0.6523	0.6028	0.5665	0.6072	0.3956	0.2652	0.3728	0.3774	0.8106
Decoder (Label + Rationale)										
vmlu-llm	0.6395	0.6585	0.6557	0.5723	0.6289	0.3853	0.2386	0.3663	0.3671	0.8092
Vistral7B	0.6354	0.6485	0.6479	0.5892	0.6285	0.3558	0.2337	0.3343	0.3394	0.7994

Bảng 3.3: Kết quả đánh giá các mô hình với input ASR transcript

Nhận xét

Dựa trên kết quả thực hiện thực nghiệm giữa hai loại dữ liệu đầu vào là văn bản gốc (Human Transcript - Bảng 3.1) và văn bản từ hệ thống nhận dạng giọng nói (ASR Transcript - Bảng 3.2), nhóm đưa ra các nhận định sau:

Nhóm mô hình **Encoder** đạt hiệu năng cao và ổn định nhất trong thiết lập *Label Only*. Đặc biệt, ViHealthBERT vượt trội hơn PhoBERT trên cả hai chỉ số vì có tiền huấn luyện trên dữ liệu y tế chuyên ngành khi xử lý hội thoại lâm sàng.

Ngược lại, các mô hình **Encoder-Decoder** (ViT5, BARTPho) trong thiết lập *Label Only* cho kết quả thấp hơn đáng kể so với nhóm Encoder. Điều này cho thấy mô hình sinh (generative) không phải là lựa chọn tối ưu khi chỉ thực hiện phân loại nhãn, do cơ chế decoder có thể làm suy giảm độ chính xác trong các tác vụ thuần phân biệt.

Các mô hình **Decoder-only** (vmlu-llm, Vistral7B) đạt hiệu năng trung gian trong chế độ *Label Only*, cao hơn nhóm Encoder-Decoder nhưng vẫn chưa vượt qua nhóm Encoder. Trong nhóm này, Vistral7B cho kết quả tốt nhất, cho thấy tiềm năng của các mô hình ngôn ngữ lớn được huấn luyện chuyên biệt cho tiếng Việt ngay cả khi chỉ thực hiện phân loại nhãn.

Khi mở rộng sang thiết lập *Label + Rationale*, hiệu năng của các mô hình **Encoder-Decoder** cải thiện nhẹ và ổn định hơn so với chế độ Label Only, tuy nhiên vẫn không đạt mức của nhóm Encoder trong bài toán phân loại.

Ngược lại, các mô hình **Decoder-only** thể hiện ưu thế rõ rệt trong thiết lập này, đạt Accuracy và Macro-F1 cao nhất, đồng thời vượt trội về các chỉ số ROUGE và BERTScore. Kết quả này khẳng định khả năng sinh rationale mạch lạc và giàu ngữ nghĩa của kiến trúc decoder-only, phù hợp với mục tiêu giải thích quyết định trong bài toán *Sentiment Reasoning*.

Cuối cùng, kết quả BERTScore xấp xỉ 0.8 trong cả hai trường hợp *human transcript* và *ASR transcript* cho thấy chất lượng ngữ nghĩa của rationale hầu như không bị suy giảm khi sử dụng transcript sinh tự động, qua đó củng cố tính khả thi của hệ thống trong các kịch bản triển khai thực tế.

4 Phương pháp nghiên cứu thực nghiệm

Phần này mô tả toàn bộ quy trình mà nhóm đã thực hiện để xây dựng hệ thống *Sentiment Reasoning* theo hướng: (i) tích hợp và so sánh hai mô hình ASR nhằm chọn ra transcript tốt nhất; (ii) tái lập và huấn luyện các nhóm mô hình (Encoder / Encoder–Decoder / Decoder-only LLM) cho hai thiết lập *Label Only* và *Label + Rationale*; và (iii) **tối ưu tham số huấn luyện** bằng cách chạy nhiều cấu hình để chọn cấu hình tốt nhất.

4.1 Thiết lập bài toán và định dạng Input–Output

4.1.1 Đầu vào

Đầu vào của hệ thống là dữ liệu hội thoại y tế dưới dạng âm thanh (audio) hoặc transcript. Với dữ liệu âm thanh, hệ thống sử dụng một mô hình nhận dạng tiếng nói tự động (ASR) để chuyển đổi tín hiệu âm thanh thành văn bản (*ASR transcript*). Với dữ liệu văn bản thuần (*human transcript*), đầu vào là bản ghi do con người tạo. Như vậy, cùng một nội dung hội thoại có thể tồn tại dưới hai dạng transcript khác nhau: *human transcript* (ít nhiễu) và *ASR transcript* (có nhiễu do lỗi nhận dạng).

4.1.2 Đầu ra và hai chế độ huấn luyện

Trong báo cáo này, nhóm xét hai chế độ bài toán:

(1) Label Only. Ở chế độ *Label Only*, mô hình chỉ dự đoán nhãn cảm xúc:

$$y \in \{\mathit{POSITIVE},\ \mathit{NEUTRAL},\ \mathit{NEGATIVE}\}.$$

Với các mô hình *Encoder*, nhãn được suy ra từ lớp phân loại (classification head). Với các mô hình sinh chuỗi (Encoder–Decoder, Decoder-only), nhãn được sinh ra dưới dạng chuỗi token.

(2) Label + Rationale. Ở chế độ *Label + Rationale*, mô hình dự đoán nhãn và sinh đồng thời rationale:

$$(y, r), \quad \text{trong đó } r \text{ là văn bản tự do mô tả lý do của nhãn.}$$

Để đồng nhất đầu ra giữa các kiến trúc sinh chuỗi và giảm lỗi parse nhãn khi suy luận, nhóm chuẩn hóa định dạng đầu ra theo mẫu cố định sau:

```
Label: <POSITIVE|NEUTRAL|NEGATIVE>
Rationale: <text>
```

Thiết kế này giúp tách bạch rõ phần "quyết định" (label) và phần "giải thích" (rationale), đồng thời hỗ trợ tự động đánh giá rationale bằng ROUGE/BERTScore.

4.2 Thiết lập môi trường huấn luyện

Vì các mô hình *Decoder-only LLM* có số lượng tham số lớn và tiêu tốn đáng kể bộ nhớ GPU, nhóm áp dụng chiến lược fine-tuning tiết kiệm tài nguyên nhằm tránh hiện tượng *out-of-memory* (OOM) và đảm bảo khả năng tái lập

thực nghiệm. Cụ thể, các mô hình Decoder-only được fine-tune bằng kỹ thuật **Low-Rank Adaptation (LoRA)** kết hợp với **4-bit quantization**, cho phép cập nhật một tập con tham số có kích thước nhỏ trong khi giữ nguyên trọng số gốc của mô hình.

Thiết lập này giúp giảm đáng kể yêu cầu bộ nhớ và chi phí tính toán, đồng thời vẫn duy trì khả năng thích nghi với dữ liệu hội thoại y tế và phong cách sinh rationale của bài toán. Đối với các mô hình Encoder và Encoder–Decoder có quy mô nhỏ hơn, nhóm thực hiện fine-tuning trực tiếp trên GPU CUDA của máy chủ mà không cần áp dụng các kỹ thuật nén tham số bổ sung.

Việc sử dụng các chiến lược huấn luyện khác nhau phản ánh đặc thù tài nguyên của từng họ kiến trúc, đồng thời đảm bảo so sánh hiệu năng được thực hiện trong điều kiện khả thi và công bằng.

4.3 Thiết lập ASR và quy trình lựa chọn mô hình cho pipeline

4.3.1 Quy trình chạy ASR trên dataset

Nhóm chuẩn hoá quy trình chạy ASR để đảm bảo hai mô hình được so sánh công bằng và transcript đầu ra có thể sử dụng trực tiếp cho các mô hình sentiment.

Bước 1: Chuẩn hoá dữ liệu âm thanh và cách giải mã. Nhóm sử dụng `datasets` để tải dataset `leduckhai/Sentiment-Reasoning` và giải mã audio về waveform với tần số lấy mẫu thống nhất (16kHz). Việc chuẩn hoá sampling rate giúp giảm sai khác do tiền xử lý và đảm bảo đầu vào ASR nhất quán.

Bước 2: Suy luận ASR theo batch và lưu transcript. Với PhoWhisper, nhóm triển khai suy luận bằng `transformers.pipeline` cho tác vụ `automatic-speech-recognition` và chạy trên GPU. Với Sherpa, nhóm chạy suy luận bằng engine ONNX tương ứng để thu được transcript. Đầu ra của mỗi mô hình được lưu vào các trường riêng (ví dụ `asr_text_phowhisper`, `asr_text_sherpa`) nhằm phục vụ so sánh và các thí nghiệm downstream.

Bước 3: Hậu xử lý tối thiểu để giữ nguyên "nhiễu ASR". Nhóm chỉ thực hiện các bước hậu xử lý tối thiểu như chuẩn hoá khoảng trắng và ký tự điều khiển. Nhóm **không tự động sửa chính tả** hoặc thay đổi nội dung từ vựng, vì mục tiêu là phản ánh trung thực lỗi ASR trong điều kiện triển khai thực tế.

4.3.2 Quy trình đánh giá và lựa chọn ASR

Khi thực hiện đánh giá độ chính xác của mô hình ASR, nhóm phát hiện đoạn text không đầy đủ nội dung so với audio. Chính vì vậy, ta không dùng các metric thông thường như WER, CER để đánh giá được mà phải thực hiện đánh giá dựa trên output của mô hình một cách thủ công và lấy mẫu trên tập kết quả và chia ra để kiểm tra. Các tiêu chí kiểm tra như sau:

- **Hallucination/lặp từ** (đặc trưng quan sát với PhoWhisper): mô hình có thể tự lặp một cụm từ nhiều lần, làm loãng tín hiệu cảm xúc và khiến rationale downstream dễ lan man.
- **Thiếu từ ở đầu hoặc cuối câu** (quan sát với PhoWhisper): bỏ sót biên câu có thể làm mất thông tin ngữ cảnh hoặc phủ định.
- **Token không xác định** (`unk`): làm giảm chất lượng tokenization và ảnh hưởng đến embedding ở các mô hình encoder.

- **Trộn ngôn ngữ** (quan sát với Sherpa): xuất hiện một số từ bị nhận dạng sang tiếng Anh, có thể gây nhiễu cho mô hình sentiment tiếng Việt.
- **Độ đầy đủ câu** (quan sát với Sherpa): Sherpa có xu hướng "dịch đủ câu" hơn, có thể hữu ích cho việc giữ cấu trúc phát ngôn.

Sau vòng kiểm tra này, nhóm tiến hành **đánh giá lại sau tinh chỉnh thiết lập suy luận** để chọn cấu hình tối ưu, chuẩn hoá bước hậu xử lý nhằm giảm lặp từ và hạn chế `unk`. Kết quả cuối cùng là lựa chọn một mô hình ASR và một cấu hình suy luận ổn định để tạo transcript dùng xuyên suốt các thí nghiệm downstream ở các mục sau.

4.4 Mô hình Sentiment Reasoning và cách triển khai theo kiến trúc

Sau khi xác định được pipeline ASR phù hợp và chuẩn hoá định dạng transcript đầu vào, nhóm tiến hành triển khai, huấn luyện và đánh giá các mô hình theo ba họ kiến trúc phổ biến trong xử lý ngôn ngữ tự nhiên hiện đại. Việc lựa chọn này bám sát các baseline trong paper gốc, đồng thời cho phép kiểm chứng một cách hệ thống ảnh hưởng của *bias kiến trúc* lên hai mục tiêu cốt lõi của bài toán: **hiệu năng phân loại cảm xúc** và **khả năng sinh lý do (rationale)**. Trên mỗi họ kiến trúc, nhóm thực hiện hai thiết lập: *Label Only* (chỉ dự đoán nhãn cảm xúc) và *Label + Rationale* (dự đoán nhãn kèm giải thích) nhằm đánh giá đầy đủ mức độ đánh đổi giữa độ chính xác và tính minh bạch.

4.4.1 Nhóm Encoder: PhoBERT và ViHealthBERT

Với các mô hình thuộc họ *Encoder* (PhoBERT, ViHealthBERT), bài toán được mô hình hoá theo khuôn khổ phân loại đa lớp chuẩn. Cụ thể, transcript đầu vào được mã hoá thành các biểu diễn ngữ cảnh (*contextual embeddings*) thông qua cơ chế self-attention, sau đó một lớp phân loại tuyến tính (*classification head*) ánh xạ biểu diễn tổng hợp (thường là token `[CLS]` hoặc phép pooling) sang phân phối xác suất trên ba lớp *Positive/Neutral/Negative*. Quá trình huấn luyện tối ưu hàm mất mát cross-entropy, qua đó mô hình học ranh giới phân lớp dựa trên tín hiệu ngôn ngữ trong transcript.

4.4.2 Nhóm Encoder–Decoder: ViT5 và BARTPho

Đối với các mô hình *Encoder–Decoder* (ViT5, BARTPho), nhóm đưa bài toán về dạng *text-to-text* nhằm thống nhất cách triển khai giữa hai nhiệm vụ phân loại và sinh giải thích. Ở thiết lập *Label Only*, mô hình được huấn luyện để sinh một chuỗi ngắn biểu diễn nhãn cảm xúc. Ở thiết lập *Label + Rationale*, mô hình được huấn luyện sinh chuỗi có cấu trúc theo template `Label | Rationale`, trong đó phần nhãn được đặt ở đầu để dễ trích xuất và phần rationale theo sau như một đoạn văn bản tự do.

Lợi thế cốt lõi của kiến trúc Encoder–Decoder nằm ở sự phân tách chức năng: Encoder học biểu diễn ngữ cảnh giàu thông tin từ transcript, trong khi Decoder thực hiện sinh chuỗi tự hồi quy để diễn đạt rationale mạch lạc và có tính giải thích. Do đó, nhóm này đặc biệt phù hợp với bài toán *Sentiment Reasoning*, nơi yêu cầu đồng thời một quyết định phân lớp và một diễn giải ngôn ngữ hoá cho quyết định đó.

4.4.3 Nhóm Decoder: Qwen3-8B, Vistral-7B và Llama-7B

Về nhóm mô hình Decoder, nhóm chọn mô hình có kết quả cao nhất trong paper là Vistral7B và mở rộng triển khai thêm 2 model mới là Llama3-8b và Qwen3-8B, kỳ vọng rằng 2 mô hình này sẽ cho kết quả tốt hơn. Với các mô

hình *Decoder* quy mô lớn (Qwen3-8B, Vistral-7B, Llama-7B), nhóm khai thác năng lực sinh văn bản và suy luận theo ngữ cảnh của LLM thông qua cơ chế prompt. Cụ thể, transcript được đặt trong một prompt hướng dẫn rõ ràng yêu cầu mô hình trả về nhãn cảm xúc và rationale theo đúng định dạng chuẩn hoá. Thiết lập này giúp đảm bảo đầu ra nhất quán và cho phép so sánh công bằng giữa các mô hình sinh chuỗi.

Tuỳ theo tài nguyên tính toán và mục tiêu thực nghiệm, nhóm cân nhắc **hai hướng triển khai**: (i) *zero-shot/few-shot prompting* để thiết lập baseline nhanh, và (ii) fine-tuning (LoRA) để thích nghi tốt hơn với đặc thù dữ liệu y tế và phong cách rationale trong dataset. Trong báo cáo này, nhóm Decoder được đánh giá như một hướng tiếp cận giàu tiềm năng cho rationale chất lượng cao, đồng thời được phân tích về độ ổn định nhãn khi đầu vào chứa nhiễu từ ASR.

4.5 Tối ưu tham số

Một điểm nhấn trong phương pháp của nhóm là thực hiện tối ưu siêu tham số theo cách có hệ thống thay vì cố định cấu hình mặc định. Trong thực nghiệm NLP, các mô hình (đặc biệt khi fine-tune) rất nhạy với learning rate, batch size và schedule; do đó, việc không tối ưu tham số có thể dẫn đến kết luận sai lệch về "mô hình nào tốt hơn" khi thực chất khác biệt đến từ cấu hình huấn luyện. Bởi vậy, nhóm tiến hành chạy nhiều cấu hình và lựa chọn cấu hình tối ưu theo tiêu chí đánh giá trên tập validation, nhằm tăng tính công bằng và độ tin cậy của so sánh.

4.5.1 Không gian tham số khảo sát

Các siêu tham số được khảo sát bao gồm: learning rate; batch size và gradient accumulation (để phù hợp giới hạn bộ nhớ); max sequence length (điều khiển khả năng giữ ngữ cảnh hội thoại); warmup ratio/steps và weight decay (ổn định quá trình fine-tune); cùng với các tham số decoding cho mô hình sinh chuỗi như beam size, max new tokens, temperature và top-p (nếu sử dụng sampling). Trong các tham số trên, learning rate được ưu tiên khảo sát theo lưới do có ảnh hưởng mạnh nhất đến hội tụ và tổng quát hoá.

4.5.2 Giao thức chọn cấu hình tối ưu

Mỗi cấu hình được huấn luyện trên cùng một chiến lược chia tập train/validation và được lựa chọn dựa trên metric mục tiêu. Với thiết lập *Label Only*, tiêu chí ưu tiên là Macro-F1 trên validation nhằm phản ánh công bằng hiệu năng giữa các lớp. Với thiết lập *Label + Rationale*, nhóm ưu tiên Macro-F1 (không suy giảm đáng kể) đồng thời các chỉ số ROUGE/BERTScore đạt mức hợp lý, qua đó đảm bảo mô hình vừa duy trì chất lượng phân loại vừa tạo được rationale hữu ích. Sau khi xác định cấu hình tốt nhất trên validation, nhóm cố định cấu hình này để báo cáo kết quả trên test nhằm hạn chế nguy cơ overfitting vào tập test.

4.6 Các cải tiến chính của nhóm

Từ quá trình tái lập baseline và phân tích đặc thù dữ liệu y tế (đặc biệt khi transcript được tạo bởi ASR), nhóm tập trung vào **bốn hướng cải tiến**: (i) chuẩn hoá định dạng đầu ra và bổ sung hậu xử lý để trích nhãn ổn định đối với mô hình sinh chuỗi; (ii) kiểm soát độ dài và mức độ tập trung của rationale để tăng tính hữu ích và cải thiện ROUGE/BERTScore; (iii) áp dụng chiến lược huấn luyện theo giai đoạn (*curriculum*) nhằm cân bằng giữa hiệu năng phân loại và khả năng giải thích; và (iv) thiết kế đánh giá song song trên human transcript và ASR transcript để kiểm tra độ bền vững của mô hình trước nhiễu ASR trong điều kiện triển khai thực tế.

5 Kết quả thực nghiệm

Nội dung dưới đây trình bày sự đối chiếu giữa kết quả công bố trong bài báo gốc (*P: Paper*) và kết quả thực nghiệm của nhóm nghiên cứu (*E: Experiment*).
Đối với cấu trúc Decoder-only, nhóm chỉ tái thực hiện chạy mô hình Vistral-7B (mô hình đạt hiệu năng tối ưu nhất theo báo cáo gốc), đồng thời mở rộng thử nghiệm trên hai mô hình khác là Llama3-8B và Qwen-3-8B.

Model	Acc.		F1 Neg.		F1 Neu.		F1 Pos.		Macro F1		R-1		R-2		R-L		R-Lsum		BERTScore	
	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E	P	E
Encoder (Label)																				
PhoBERT	0.6674	0.6590	0.6969	0.6886	0.6607	0.6500	0.6377	0.6351	0.6651	0.6579	–	–	–	–	–	–	–	–	–	–
ViHealthBERT	0.6752	0.6696	0.6970	0.6965	0.6718	0.6499	0.6535	0.6673	0.6741	0.6712	–	–	–	–	–	–	–	–	–	–
Encoder–Decoder (Label + Rationale)																				
ViT5	0.6633	0.6609	0.6936	0.6842	0.6572	0.6686	0.6335	0.6093	0.6615	0.6540	0.3910	0.3548	0.2668	0.2122	0.3653	0.3291	0.3660	0.3289	0.8093	0.7243
BARTpho	0.6619	0.6610	0.7029	0.6884	0.6460	0.6583	0.6265	0.6288	0.6585	0.6585	0.3871	0.3730	0.2613	0.2227	0.3658	0.3487	0.3683	0.3481	0.8077	0.8025
Decoder (Label + Rationale)																				
Vistral7B	0.6812	0.6833	0.7152	0.6998	0.6765	0.6887	0.6425	0.6498	0.6781	0.6794	0.4155	0.4288	0.2788	0.2835	0.3880	0.3981	0.3900	0.3984	0.8101	0.8153
LLaMA3-8B	–	0.6812	–	0.7152	–	0.6765	–	0.6425	–	0.6781	–	0.4155	–	0.2788	–	0.3880	–	0.3900	–	0.8101
Qwen3-8B	–	0.6841	–	0.6975	–	0.6860	–	0.6626	–	0.6821	–	0.3767	–	0.2071	–	0.3368	–	0.3368	–	0.7639

Bảng 5.1: Bảng đánh giá kết quả thực nghiệm

Dựa trên bảng kết quả thực nghiệm (Bảng 5.1), nhóm đưa ra các nhận xét về hiệu năng của các mô hình như sau:

a) Nhận xét chung kết quả

- Nhìn chung, các mô hình *PhoBERT, ViHealthBERT, ViT5, BARTpho, Vistral7B* đều cho kết quả sát sao với báo cáo gốc.
- Đối với các mô hình Encoder và Encoder-Decoder, độ lệch về chỉ số Accuracy và Macro F1 dao động trong khoảng nhỏ (dưới 1%) → quy trình thực nghiệm của nhóm đảm bảo tính khách quan và độ tin cậy.
- Riêng mô hình Vistral7B, kết quả thực hiện bởi nhóm (E) thậm chí có sự cải thiện nhẹ ở một số chỉ số như Acc (0.6833 so với 0.6812) và BERTScore (0.8153 so với 0.8101) so với bài báo gốc.

b) So sánh hiệu năng giữa các kiến trúc (Encoder vs. Decoder)

- Nhóm Decoder (Vistral, LLaMA3, Qwen3): Thể hiện ưu thế vượt trội ở hầu hết các chỉ số. Đặc biệt, chỉ số Macro F1 của nhóm này đều đạt ngưỡng trên 0.678, cao hơn hẳn so với nhóm Encoder (khoảng 0.65-0.67).
- Nhóm Encoder-Decoder (ViT5, BARTpho): Hiệu năng phân loại (Acc, F1) thấp hơn so với các mô hình Decoder. Tuy nhiên, nhóm này duy trì được sự ổn định giữa các chỉ số ROUGE (R-1, R-2, R-L).

c) Phân tích các mô hình Decoder mở rộng (LLaMA3-8B và Qwen3-8B)

- Qwen3-8B: Đạt kết quả ấn tượng nhất về khả năng phân loại với Acc (0.6841) và Macro F1 (0.6821), vượt qua Vistral7B (mô hình tốt nhất trong paper gốc). Tuy nhiên, khả năng sinh Rationale của Qwen3 (thể hiện qua các chỉ số ROUGE và BERTScore) lại thấp hơn so với Vistral7B và LLaMA3.
- LLaMA3-8B: Cho kết quả cực kỳ ổn định, tương đương với Vistral7B ở các chỉ số phân loại và có điểm số ROUGE tương đối cao, cho thấy khả năng hiểu và tóm tắt ngữ cảnh tốt.

⇒ **Mô hình Vistral7B vẫn cho thấy sự cân bằng tốt nhất** giữa việc phân loại (Label) và giải thích (Rationale). Trong khi đó, **Qwen3-8B là lựa chọn tối ưu nếu ưu tiên độ chính xác của nhãn dự đoán**. Các kết quả thực nghiệm này cho thấy các mô hình Large Language Models (LLMs) với kích thước tham số lớn hơn (8B) đang mang lại hiệu quả tích cực cho bài toán thực hiện.

6 Kết luận

6.1 Thành tựu và hạn chế

6.1.1 Thành tựu

Dựa trên quá trình triển khai và so sánh với bài báo gốc, nhóm đã đạt được các kết quả sau:

- **Tái lập và xác thực kết quả:** Nhóm đã triển khai thành công các mô hình tiêu biểu từ nghiên cứu gốc (PhoBERT, ViT5, Vistral7B) với kết quả thực nghiệm sát sao và cả vượt trội hơn so với số liệu công bố trong bài báo. Đặc biệt, mô hình Vistral7B do nhóm chạy đạt chỉ số Acc (0.6833) và BERTScore (0.8153) cao hơn mức báo cáo gốc.
- **Cải tiến hệ thống ASR:** Khác với nghiên cứu gốc chỉ sử dụng PhoWhisper, nhóm đã thử nghiệm thêm framework Sherpa-ONNX.
- **Thử nghiệm thêm với các LLM tiên tiến:** Nhóm đã thử nghiệm thành công hai mô hình Llama3-8B và Qwen3-8B. Trong đó, Qwen3-8B thiết lập một cột mốc mới về độ chính xác phân loại với Acc (0.6841) và Macro F1 (0.6821).
- **Đảm bảo tính toàn diện:** Kết quả thực nghiệm đã bao quát đầy đủ cả 3 dạng kiến trúc (Encoder, Encoder-Decoder, Decoder), cung cấp cái nhìn đa chiều về hiệu năng giữa việc chỉ phân loại nhãn (Label) và việc kết hợp giải thích (Rationale).

6.1.2 Hạn chế

Mặc dù đạt được những kết quả khả quan, nghiên cứu vẫn tồn tại một số điểm cần cải thiện:

- **Sự khác nhau giữa transcript của audio và human_transcript:** Điều này gây khó khăn khi thực hiện đánh giá các mô hình ASR, không dùng các metrics thông thường như WER, CER để đánh giá được mà phải đánh giá thủ công.
- **Sự sụt giảm chất lượng Rationale trên các mô hình mới:** Dù Qwen3-8B có độ chính xác phân loại cao nhất, nhưng khả năng sinh lời giải thích (Rationale) lại kém hơn đáng kể so với Vistral7B.
- **Chi phí tài nguyên:** Việc vận hành các mô hình Decoder-only với kích thước tham số lớn (8B) đòi hỏi tài nguyên tính toán cao dễ dẫn đến OOM. .

6.2 Định hướng phát triển trong tương lai

Từ những thành tựu và hạn chế nêu trên, nhóm đề xuất các hướng phát triển tiếp theo:

- **Fine-tuning prompt cho Rationale:** Thực hiện tinh chỉnh (Fine-tuning) mô hình Llama3 và Qwen3 trên tập dữ liệu tiếng Việt lớn hơn để cải thiện chất lượng sinh lời giải thích hoặc sử dụng prompt CoT, thay vì chỉ sử dụng phương pháp Zero-shot.
- **Tối ưu hóa Pipeline ASR-LLM:** Trong hướng nghiên cứu tiếp theo, nhóm dự kiến tối ưu pipeline ASR bằng cách thử nghiệm các kiến trúc mới như Conformer, đặc biệt là mô hình Parakeet-CTC thế hệ mới. Conformer kết hợp hiệu quả giữa self-attention và convolution, cho phép mô hình nắm bắt đồng thời ngữ cảnh dài hạn và đặc trưng âm học cục bộ, vốn rất quan trọng trong hội thoại y tế. Việc tích hợp mô hình này kỳ vọng giúp cải thiện độ đầy đủ transcript và tăng độ ổn định của các mô hình phân tích cảm xúc phía sau.

Tài liệu tham khảo

[1] A. Author *et al.*, "AI-Supported Shared Decision-Making (AI-SDM): Conceptual Framework," *JMIR AI*, vol. 1, e75866, 2025. [Online]. Available: https://ai.jmir.org/2025/1/e75866/PDF

[2] A. Author *et al.*, "Evaluating Large Language Models for Sentiment Analysis: A Comparative Study of Accuracy and Scalability," *IEEE*, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/11154211

[3] A. Author *et al.*, "MultiMed-ST: Large-scale Many-to-many Multilingual Medical Speech Translation," *arXiv preprint arXiv:2504.03546*, 2025. [Online]. Available: https://arxiv.org/pdf/2504.03546

[4] A. M. Antoniadi *et al.*, "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review," *Applied Sciences*, vol. 11, no. 11, Art. 5088, 2021, doi: 10.3390/app11115088.

[5] Coderivers, "Python Sentiment Analysis: Unveiling the Emotional Tone of Text," [Online]. Available: https://coderivers.org/blog/python-perform-sentiment-analysis/

[6] K. Denecke and Y. Deng, "Sentiment analysis in medical settings: New opportunities and challenges," *Artificial Intelligence in Medicine*, vol. 64, no. 1, pp. 17–27, 2015, doi: 10.1016/j.artmed.2015.03.006.

[7] R. Guidotti *et al.*, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, Art. 93, pp. 1–42, 2018, doi: 10.1145/3236009.

[8] L. D. Khai, "Sentiment Reasoning Dataset," *HuggingFace*, 2024. [Online]. Available: https://huggingface.co/datasets/leduckhai/Sentiment-Reasoning

[9] L. D. Khai, "Sentiment Reasoning GitHub Repository," *GitHub*, 2024. [Online]. Available: https://github.com/leduckhai/Sentiment-Reasoning

[10] D. Le, K. Nguyen, *et al.*, "Sentiment Reasoning for Healthcare," *arXiv preprint arXiv:2407.21054*, 2024. [Online]. Available: https://arxiv.org/pdf/2407.21054

[11] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[12] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, pp. 74–81, 2004.

[13] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.

[14] W.-Y. Loh, C.-H. Ooi, *et al.*, "Explainable Artificial Intelligence in Healthcare: A Survey," *Computer Methods and Programs in Biomedicine*, 2022.

[15] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008, doi: 10.1561/1500000011.

[16] S. Author *et al.*, "Evaluating Large Language Models: A Comprehensive Survey," *arXiv preprint arXiv:2310.19736*, 2023. [Online]. Available: https://arxiv.org/pdf/2310.19736

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://arxiv.org/pdf/1706.03762

[18] Y. Author *et al.*, "Emotion helps sentiment: A multi-task model for sentiment and emotion analysis," *arXiv preprint arXiv:1911.12569*, 2019. [Online]. Available: https://arxiv.org/pdf/1911.12569

[19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[20] L. Zunic, P. Corcoran, and I. Spasic, "Sentiment Analysis in Health and Well-Being: Systematic Review," *JMIR Medical Informatics*, vol. 8, no. 1, e16023, 2020, doi: 10.2196/16023.