ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC VÀ KỸ THUẬT THÔNG TIN

PHÁT HIỆN MỐI NGUY TRONG THỰC PHẨM BẰNG PHƯƠNG PHÁP HỌC ĐA NHIỆM VỚI HÀM MẤT MÁT FOCAL

BÁO CÁO ĐỒ ÁN - DS310.Q11

Ngô Minh Trí – 23521640
Nguyễn Đình Khôi – 23520774

Hồ Chí Minh, 2025

Mục lục

1 Giới thiệu

1.1 Bối cảnh

An toàn thực phẩm là một trong những vấn đề then chốt, có tác động trực tiếp đến sức khỏe cộng đồng và chất lượng cuộc sống của con người. Chỉ một sai sót nhỏ trong chuỗi sản xuất, chế biến hoặc phân phối thực phẩm cũng có thể dẫn đến những hậu quả nghiêm trọng, không chỉ gây ảnh hưởng tiêu cực đến sức khỏe người tiêu dùng mà còn kéo theo các tổn thất đáng kể về kinh tế và uy tín đối với các doanh nghiệp liên quan. Trong bối cảnh toàn cầu hóa cùng với sự phát triển mạnh mẽ của thương mại điện tử, chuỗi cung ứng thực phẩm ngày càng mở rộng và phức tạp, khiến việc kiểm soát và quản lý các rủi ro an toàn thực phẩm trở nên khó khăn hơn.

Song song với đó, các thông tin liên quan đến sự cố và cảnh báo an toàn thực phẩm hiện nay được công bố rộng rãi trên nhiều nguồn trực tuyến khác nhau, bao gồm các trang web của cơ quan quản lý, báo chí điện tử và mạng xã hội. Các hệ thống giám sát chính thức, điển hình như Hệ thống Cảnh báo Nhanh về Thực phẩm và Thức ăn chăn nuôi (RASFF) của Liên minh Châu Âu hay các thông báo thu hồi sản phẩm của Cục Quản lý Thực phẩm và Dược phẩm Hoa Kỳ (FDA), mỗi năm ghi nhận hàng nghìn báo cáo liên quan đến các mối nguy thực phẩm. Lượng thông tin lớn này nhanh chóng được lan truyền trên môi trường web, phản ánh xu hướng ngày càng phổ biến của việc giám sát an toàn thực phẩm dựa trên dữ liệu trực tuyến.

1.2 Động lực và Công trình liên quan

Trước thực trạng nêu trên, việc tự động thu thập, nhận diện và phân loại các mối nguy an toàn thực phẩm từ các báo cáo sự cố trên web trở thành một yêu cầu cấp thiết. Các hệ thống tự động không chỉ góp phần giảm tải khối lượng công việc cho các chuyên gia trong lĩnh vực an toàn thực phẩm mà còn hỗ trợ phát hiện sớm các rủi ro tiềm ẩn, từ đó nâng cao hiệu quả của các cơ chế cảnh báo và phòng ngừa.

SemEval-2025 Task 9 [7] được tổ chức nhằm đáp ứng nhu cầu thực tiễn này, tập trung đánh giá các mô hình trí tuệ nhân tạo trong nhiệm vụ dự đoán đồng thời loại mối nguy thực phẩm và nhóm sản phẩm liên quan, dựa trên tiêu đề và nội dung của các báo cáo sự cố thu thập từ web. Nhiệm vụ đặt ra nhiều thách thức quan trọng, bao gồm sự mất cân bằng lớp nghiêm trọng trong dữ liệu, yêu cầu cao về khả năng tổng quát hóa của mô hình, cũng như nhu cầu nâng cao độ chính xác phân loại để hướng tới triển khai trong các hệ thống giám sát an toàn thực phẩm tự động trong thực tế.

Các hệ thống đạt thứ hạng cao nhất tại SemEval-2025 Task 9 [7] chủ yếu khai thác các mô hình ngôn ngữ tiền huấn luyện dựa trên kiến trúc Transformer, kết hợp với nhiều chiến lược xử lý mất cân bằng lớp và tăng cường dữ liệu. Tiêu biểu, hệ thống của đội *Anastasia* [3] đạt thứ hạng cao nhất nhờ sử dụng các mô hình DeBERTa [2] và RoBERTa [5] kích thước lớn, kết hợp hàm mất mát Focal cùng chiến lược tăng cường dữ liệu và ensemble mềm để cải thiện hiệu năng phân loại. Các công trình tiếp theo như *MyMy* [6], *SRCB* [10], *PATeam* [9] và *HU* [8] tiếp tục mở rộng theo các hướng tiếp cận phức tạp hơn, bao gồm khai thác mô hình ngôn ngữ lớn (LLM), kỹ thuật truy hồi tri thức (RAG), sinh dữ liệu tăng cường bằng LLM, cũng như các pipeline nhiều giai đoạn nhằm xử lý sự mất cân bằng lớp và cải thiện hiệu năng tổng thể.

Mặc dù đạt được hiệu suất cao trong nhiệm vụ, các phương pháp nêu trên thường phụ thuộc vào các mô hình quy mô lớn và nhiều thành phần bổ trợ phức tạp, điều này

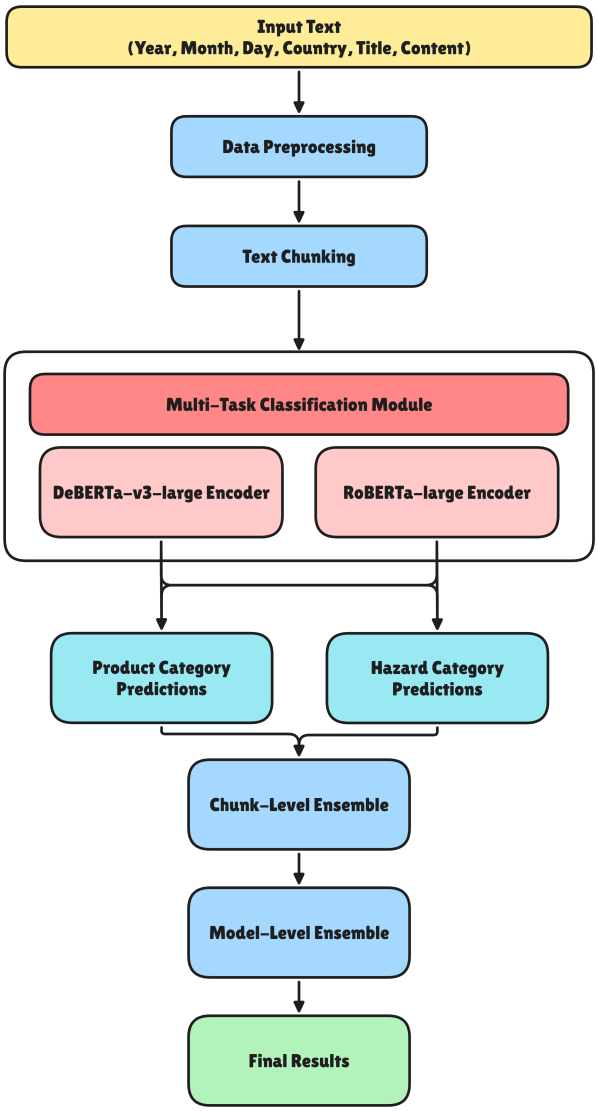
có thể làm gia tăng chi phí huấn luyện và suy luận, đồng thời hạn chế khả năng triển khai trong các hệ thống giám sát an toàn thực phẩm tự động ngoài thực tế. Xuất phát từ nhận định này, nhóm nghiên cứu kế thừa các ý tưởng hiệu quả từ hệ thống Anastasia [3] và phát triển theo hướng tinh gọn, nhằm cân bằng giữa hiệu suất mô hình và tính thực tiễn triển khai.

Cụ thể, nhóm đề xuất một hệ thống dựa trên các mô hình ngôn ngữ tiền huấn luyện thuộc họ Transformer, bao gồm DeBERTa [2] và RoBERTa [5], được tinh chỉnh để giải quyết bài toán phát hiện mối nguy thực phẩm từ văn bản web. Trong giai đoạn huấn luyện, hàm mất mát Focal Loss [4] được áp dụng nhằm giảm ảnh hưởng của các lớp chiếm ưu thế và cải thiện khả năng học của các lớp hiếm. Đồng thời, chiến lược học đa nhiệm (multi-task learning) [1] được triển khai, cho phép mô hình đồng thời dự đoán hai nhãn đầu ra là loại mối nguy và nhóm sản phẩm, qua đó khai thác mối liên hệ ngữ nghĩa giữa hai tác vụ.

Bên cạnh đó, các kỹ thuật tăng cường dữ liệu được thiết kế có chọn lọc nhằm cải thiện độ chính xác phân loại trong khi vẫn hạn chế nguy cơ quá khớp. Cuối cùng, hệ thống áp dụng chiến lược học tổ hợp hai tầng, bao gồm ensemble ở mức đoạn văn bản và ensemble ở mức mô hình. Ở tầng thứ nhất, các dự đoán từ nhiều đoạn văn bản (chunk) thuộc cùng một báo cáo được tổng hợp bằng phương pháp trung bình xác suất nhằm thu được dự đoán ổn định ở mức tài liệu. Ở tầng thứ hai, các phân phối xác suất đầu ra từ hai mô hình nền DeBERTa [2] và RoBERTa [5] được kết hợp thông qua ensemble mềm (soft voting) với trọng số được tinh chỉnh trên tập xác thực. Chiến lược ensemble hai tầng này cho phép tận dụng đồng thời thông tin ngữ cảnh trải dài trong văn bản dài và tính bổ sung giữa các mô hình khác nhau, từ đó nâng cao hiệu năng tổng thể so với từng mô hình đơn lẻ.

Mã nguồn và toàn bộ cấu hình thực nghiệm của hệ thống được công bố công khai tại:
https://github.com/chisngooo/FoodHazarDectection-DS310-FinalProject
Ngoài ra, nhóm cũng cung cấp một phiên bản demo trực quan tại:
https://github.com/NgDinhKhoi0709/FoodHazarDectection-DS310-FinalProjectDemo,
cho phép người dùng dễ dàng quan sát, thực hiện và đánh giá hoạt động của hệ thống.

2 Phương pháp đề xuất



Hình 1: Pipeline tổng quát của hệ thống phát hiện mối nguy thực phẩm

2.1 Tiền xử lý dữ liệu

Trong các bài toán phân loại văn bản thu thập từ web, dữ liệu đầu vào thường có mức độ không đồng nhất cao và chứa nhiều thành phần nhiễu như mã HTML, boilerplate mang tính hành chính, các đoạn nội dung lặp lại, cũng như sự không nhất quán trong cách biểu diễn các thực thể chuyên ngành. Những yếu tố này làm gia tăng độ dư thừa thông tin và có thể gây nhiễu cho quá trình học biểu diễn ngữ nghĩa của mô hình, đặc biệt trong bối cảnh dữ liệu có phân bố nhãn dài đuôi và mức độ chồng lấn ngữ nghĩa cao giữa các lớp.

Xuất phát từ nhận định này, nhóm đề xuất một pipeline tiền xử lý dữ liệu nhiều bước nhằm: (i) loại bỏ các thành phần không mang thông tin ngữ nghĩa, (ii) giảm sự trùng lặp nội dung trong cùng một báo cáo, và (iii) chuẩn hóa biểu diễn của các thực thể quan trọng liên quan đến mối nguy và sản phẩm thực phẩm. Tổng quan pipeline của hệ thống được minh họa trong Hình 1. Để đánh giá định lượng tác động của từng bước tiền xử lý,

nhóm tiến hành một nghiên cứu *ablation* có kiểm soát, trong đó các biến thể tiền xử lý khác nhau được so sánh dựa trên chất lượng biểu diễn ngữ nghĩa của văn bản.

Variant	Score	Avg. words	TTR	Tri-dup	HTML	CosSim
v0_raw	53.69	293.60	0.6088	0.6772	20329	0.4094
v1_html	64.70	290.33	0.6094	0.6766	0	0.4093
v2_basic	67.48	290.33	0.6094	0.6766	0	0.4093
v3_dedup	67.08	266.23	0.6297	0.6497	0	0.4087
v4_entities	67.48	290.35	0.6094	0.6767	0	0.4092
v5_full	66.83	263.36	0.6370	0.6467	0	0.4062

Bảng 1: Kết quả ablation các biến thể tiền xử lý dữ liệu. Score là điểm chất lượng tổng hợp phản ánh mức độ nhiễu, dư thừa và chất lượng biểu diễn ngữ nghĩa (SentenceTransformer-based evaluation, 3 000 mẫu).

2.1.1 Trích xuất nội dung từ HTML

Do dữ liệu được thu thập trực tiếp từ các trang web thông báo sự cố và thu hồi thực phẩm, nội dung văn bản thường chứa nhiều thẻ HTML và thành phần định dạng không cần thiết. Các thành phần này không mang thông tin ngữ nghĩa liên quan đến mối nguy thực phẩm và có thể gây nhiễu cho quá trình mã hóa văn bản của mô hình ngôn ngữ.

Nhóm sử dụng thư viện *BeautifulSoup* để phân tích cú pháp HTML và trích xuất phần văn bản thuần túy (plain text), trong đó các đoạn văn được nối lại bằng khoảng trắng nhằm bảo toàn tính liên tục của nội dung. Kết quả ablation cho thấy đây là bước tiền xử lý có tác động mạnh nhất: so với dữ liệu thô (*v0_raw*), biến thể chỉ loại bỏ HTML (*v1_html*) giúp tăng điểm đánh giá tiền xử lý từ 53.69 lên 64.70, tương ứng mức cải thiện hơn 11 điểm, đồng thời loại bỏ hoàn toàn 20 329 thẻ HTML còn sót lại. Điều này cho thấy mã HTML là một nguồn nhiễu nghiêm trọng và việc trích xuất nội dung thuần túy là bước tiền xử lý bắt buộc.

2.1.2 Chuẩn hóa văn bản cơ bản

Sau khi trích xuất văn bản thuần túy, dữ liệu tiếp tục được chuẩn hóa thông qua các bước làm sạch cơ bản, bao gồm loại bỏ các ký tự không hiển thị (ví dụ `\xa0`), gom các khoảng trắng dư thừa và cắt bỏ khoảng trắng ở đầu và cuối chuỗi. Mục tiêu của bước này là đảm bảo tính nhất quán về mặt định dạng và giảm nhiễu bề mặt trước khi đưa văn bản vào các bước xử lý sâu hơn.

Kết quả thực nghiệm cho thấy bước chuẩn hóa cơ bản giúp cải thiện thêm chất lượng dữ liệu: điểm đánh giá tăng từ 64.70 (*v1_html*) lên 67.48 (*v2_basic*). Mặc dù các chỉ số ngữ nghĩa ở mức embedding (độ tương đồng cosine trung bình xấp xỉ 0.409) gần như không thay đổi, bước này góp phần ổn định cấu trúc văn bản và hạn chế các sai lệch định dạng có thể ảnh hưởng tiêu cực đến tokenizer và mô hình Transformer trong giai đoạn huấn luyện.

2.1.3 Loại bỏ nhiễu và boilerplate

Các báo cáo thu hồi thực phẩm thường chứa nhiều đoạn thông tin mang tính khuôn mẫu như tuyên bố pháp lý, thông tin liên hệ hoặc các mô tả được lặp lại gần như nguyên văn giữa nhiều báo cáo. Những đoạn boilerplate này không đóng góp trực tiếp vào việc xác

định loại mối nguy hoặc nhóm sản phẩm, nhưng lại làm tăng độ dài và mức độ dư thừa của văn bản.

Nhóm sử dụng các biểu thức chính quy được thiết kế theo miền bài toán để loại bỏ những thành phần này. Việc giảm nhiễu và boilerplate giúp mô hình tập trung nhiều hơn vào các đoạn nội dung giàu thông tin ngữ nghĩa liên quan trực tiếp đến an toàn thực phẩm, đồng thời hạn chế hiện tượng học lệch vào các mẫu câu hành chính lặp lại.

2.1.4 Khử trùng lặp câu

Nhằm giảm thiểu sự dư thừa thông tin trong cùng một báo cáo, nhóm áp dụng phương pháp khử trùng lặp ở mức câu, kết hợp giữa so khớp chính xác và so khớp mờ (*fuzzy matching*). Mỗi câu được chuẩn hóa về chữ thường, loại bỏ các chuỗi số dài và ký tự dư thừa trước khi so sánh. Hai câu được xem là trùng lặp nếu độ tương đồng chuỗi, đo bằng thuật toán *SequenceMatcher*, vượt quá một ngưỡng xác định trước.

Kết quả ablation cho thấy bước khử trùng lặp giúp giảm đáng kể mức độ dư thừa n-gram: tỉ lệ trùng lặp trigram giảm từ 0.6766 (*v2_ basic*) xuống 0.6497 (*v3_ dedup*), đồng thời độ dài trung bình của văn bản giảm từ khoảng 290 từ xuống còn 266 từ. Tuy nhiên, điểm đánh giá tổng hợp giảm nhẹ từ 67.48 xuống 67.08, phản ánh sự đánh đổi giữa việc loại bỏ dư thừa và việc giữ lại đầy đủ ngữ cảnh mô tả trong các báo cáo dài. Do đó, khử trùng lặp được xem là một bước hỗ trợ, cần áp dụng thận trọng để tránh làm mất các chi tiết quan trọng.

2.1.5 Chuẩn hóa thực thể chuyên ngành

Dữ liệu web thường tồn tại nhiều biến thể cách viết khác nhau của cùng một thực thể, đặc biệt là tên vi sinh vật và tác nhân gây hại (ví dụ *"E. coli"*, *"e coli"*). Sự không nhất quán này có thể làm phân mảnh biểu diễn ngữ nghĩa và gây bất lợi cho các lớp hiếm trong dữ liệu long-tail.

Nhóm tiến hành chuẩn hóa một số thực thể quan trọng bằng cách ánh xạ các biến thể phổ biến về một dạng chuẩn (ví dụ *"Escherichia coli"*). Kết quả thực nghiệm cho thấy biến thể chuẩn hóa thực thể (*v4_ entities*) đạt điểm đánh giá cao nhất (67.48), tương đương với pipeline làm sạch cơ bản, trong khi độ dài trung bình văn bản (290.35 từ) và mức độ chồng lấn ngữ nghĩa (độ tương đồng cosine trung bình 0.4092) gần như không thay đổi. Điều này cho thấy chuẩn hóa thực thể giúp tăng tính nhất quán ngữ nghĩa mà không gây mất mát thông tin.

2.1.6 Hợp nhất tiêu đề và nội dung

Cuối cùng, tiêu đề và nội dung báo cáo sau khi được làm sạch được nối lại thành một chuỗi văn bản duy nhất và chuyển về chữ thường. Việc hợp nhất này cho phép mô hình khai thác đồng thời thông tin cô đọng từ tiêu đề và ngữ cảnh chi tiết từ nội dung, đặc biệt quan trọng trong các trường hợp mối nguy được nêu rõ ngay trong tiêu đề báo cáo.

2.1.7 Kết luận

Tổng hợp kết quả ablation cho thấy: (i) trích xuất nội dung từ HTML là bước quan trọng nhất, mang lại mức cải thiện lớn nhất về chất lượng dữ liệu; (ii) chuẩn hóa văn bản cơ bản và chuẩn hóa thực thể giúp tăng tính nhất quán ngữ nghĩa mà không gây tác dụng

phụ; và (iii) khử trùng lặp giúp giảm dư thừa nhưng cần áp dụng có chọn lọc để tránh làm mất ngữ cảnh.

Dựa trên các phân tích định lượng trên, nhóm lựa chọn pipeline tiền xử lý đầy đủ làm cấu hình mặc định cho hệ thống, do khả năng cân bằng tốt giữa việc giảm nhiễu, giảm dư thừa và duy trì biểu diễn ngữ nghĩa ổn định. Pipeline này tạo ra dữ liệu đầu vào sạch và nhất quán hơn, từ đó hỗ trợ hiệu quả cho quá trình huấn luyện mô hình đa nhiệm và nâng cao khả năng tổng quát hóa trong nhiệm vụ phát hiện mối nguy an toàn thực phẩm.

2.2 Chia đoạn văn bản

Các mô hình ngôn ngữ tiền huấn luyện dựa trên kiến trúc Transformer như DeBERTa [2] và RoBERTa [5] đều bị giới hạn độ dài đầu vào tối đa, thường là 512 token. Trong khi đó, các báo cáo sự cố an toàn thực phẩm thu thập từ web có thể có độ dài lớn hơn đáng kể. Để tận dụng đầy đủ thông tin từ văn bản gốc mà vẫn tuân thủ ràng buộc của mô hình, nhóm áp dụng chiến lược chia đoạn văn bản (text chunking) ở mức token.

Cụ thể, văn bản sau tiền xử lý được mã hóa thành chuỗi token bằng tokenizer tương ứng với mô hình nền, trong đó các token đặc biệt không được thêm vào nhằm kiểm soát chính xác độ dài đầu vào. Chuỗi token này sau đó được chia thành các đoạn con với độ dài tối đa 512 token. Để giảm thiểu hiện tượng mất ngữ cảnh tại ranh giới giữa các đoạn, các chunk liên tiếp được chồng lấp một số lượng token cố định, giúp bảo toàn mối liên kết ngữ nghĩa xuyên suốt toàn bộ văn bản.

Sau khi giải mã ngược về dạng văn bản, các chunk được chuẩn hóa lại và các đoạn quá ngắn, không mang đủ thông tin ngữ nghĩa, sẽ bị loại bỏ. Mỗi chunk hợp lệ được gán nhãn mối nguy và nhóm sản phẩm giống với văn bản gốc, đồng thời giữ nguyên các siêu dữ liệu liên quan. Tập dữ liệu sau khi chia đoạn được lưu trữ dưới định dạng JSON và được sử dụng làm đầu vào cho mô hình phân loại đa nhiệm trong giai đoạn huấn luyện và suy luận.

Chiến lược chia đoạn dựa trên token kết hợp với cơ chế chồng lấp ngữ cảnh cho phép hệ thống khai thác hiệu quả toàn bộ nội dung của các báo cáo dài, đồng thời nâng cao khả năng học và tổng quát hóa của mô hình trong nhiệm vụ phát hiện mối nguy an toàn thực phẩm.

2.3 Fine-tuning mô hình đa nhiệm với Focal Loss

Sau bước chia đoạn văn bản, mỗi đoạn (chunk) được sử dụng làm đầu vào để tinh chỉnh các mô hình ngôn ngữ tiền huấn luyện dựa trên kiến trúc Transformer, bao gồm DeBERTa-v3-large [2] và RoBERTa-large [5]. Các mô hình encoder-only này đã được chứng minh đạt hiệu năng cao trong nhiệm vụ phát hiện mối nguy thực phẩm tại SemEval-2025 Task 9 [7]. Nhằm khai thác mối liên hệ ngữ nghĩa chặt chẽ giữa các nhãn đầu ra, nhóm áp dụng chiến lược học đa nhiệm (multi-task learning), trong đó mô hình được huấn luyện đồng thời để dự đoán hai nhãn: loại mối nguy thực phẩm và nhóm sản phẩm liên quan, tương tự các hướng tiếp cận hiệu quả được đề xuất trong các hệ thống hàng đầu của cuộc thi [3, 9, 6].

2.3.1 Kiến trúc học đa nhiệm

Cả hai mô hình DeBERTa [2] và RoBERTa [5] đều được sử dụng như bộ mã hóa ngữ cảnh chung (shared encoder). Biểu diễn của token `[CLS]` ở tầng cuối cùng được trích xuất và

đưa qua một lớp dropout nhằm giảm hiện tượng quá khớp. Sau đó, biểu diễn này được chia nhánh thành hai đầu phân loại độc lập, tương ứng với hai tác vụ: phân loại nhóm sản phẩm và phân loại loại mối nguy. Thiết kế này cho phép mô hình học được các biểu diễn dùng chung hiệu quả trong khi vẫn duy trì khả năng phân biệt đặc thù cho từng tác vụ, phù hợp với bản chất liên quan chặt chẽ giữa hai nhãn trong bài toán phát hiện mối nguy thực phẩm

2.3.2 Hàm mất mát Focal

Dữ liệu của bài toán tồn tại hiện tượng mất cân bằng lớp nghiêm trọng, khi một số loại mối nguy và nhóm sản phẩm chỉ xuất hiện với tần suất rất thấp, phản ánh phân bố nhãn dài đuôi (long-tail) đã được ghi nhận trong SemEval-2025 Task 9 [7]. Trong bối cảnh này, việc sử dụng hàm mất mát cross-entropy tiêu chuẩn dễ khiến mô hình bị chi phối bởi các lớp chiếm ưu thế và học kém hiệu quả trên các lớp hiếm. Để khắc phục vấn đề này, nhóm sử dụng hàm mất mát Focal Loss [4], một lựa chọn đã được chứng minh hiệu quả trong các hệ thống đạt thứ hạng cao tại SemEval-2025 Task 9, tiêu biểu là đội *Anastasia* [3].

Focal Loss cho bài toán phân loại đa lớp được định nghĩa như sau:

$$\mathcal{L}_{\text{focal}} = -\alpha(1 - p_t)^{\gamma} \log(p_t), \tag{1}$$

trong đó p_t là xác suất dự đoán của mô hình đối với nhãn đúng, α là hệ số cân bằng lớp, và γ là tham số điều chỉnh mức độ tập trung vào các mẫu khó. Thành phần $(1 - p_t)^{\gamma}$ giúp giảm trọng số của các mẫu dễ (có p_t lớn), từ đó buộc mô hình tập trung nhiều hơn vào các mẫu khó và các lớp hiếm.

Lựa chọn tham số γ. Trong thực nghiệm này, tham số γ được cố định bằng 2.0. Giá trị này được lựa chọn dựa trên hai lý do chính. Thứ nhất, $\gamma = 2.0$ là giá trị đã được đề xuất trong công trình gốc của Focal Loss [4] và được sử nhớp dụng rộng rãi trong các bài toán phân loại có phân bố nhãn dài đuôi. Thứ hai, qua thử nghiệm sơ bộ, giá trị này cho phép giảm đáng kể ảnh hưởng của các mẫu được phân loại đúng với độ tin cậy cao, đồng thời vẫn đảm bảo quá trình huấn luyện ổn định. Các giá trị γ lớn hơn có xu hướng làm giảm tốc độ hội tụ và gây dao động loss, trong khi các giá trị nhỏ hơn không đủ mạnh để xử lý hiệu quả hiện tượng mất cân bằng lớp trong dữ liệu.

Lựa chọn tham số α. Tham số α được sử dụng nhằm điều chỉnh mức độ đóng góp của các lớp trong hàm mất mát. Trong khuôn khổ nghiên cứu này, nhóm đặt $\alpha = 1.0$ cho tất cả các lớp và không sử dụng trọng số lớp riêng biệt. Cách thiết lập này cho phép Focal Loss tập trung hoàn toàn vào cơ chế tái phân phối trọng số dựa trên độ khó của mẫu thông qua tham số γ, đồng thời giữ cho hàm mất mát đơn giản và ổn định. Việc không áp dụng trọng số lớp thủ công cũng giúp tránh hiện tượng quá nhạy cảm với nhiễu và sai lệch thống kê trong dữ liệu web, đặc biệt đối với các lớp hiếm có số lượng mẫu rất nhỏ.

Áp dụng Focal Loss trong huấn luyện. Trong quá trình huấn luyện, Focal Loss được áp dụng trực tiếp trên các logits đầu ra của mô hình cho từng tác vụ phân loại. Việc sử dụng cùng một cấu hình tham số ($\alpha = 1.0$, $\gamma = 2.0$) xuyên suốt quá trình huấn luyện giúp đảm bảo tính nhất quán và khả năng tái lập kết quả. Nhờ đó, mô hình có thể giảm ảnh hưởng của các lớp chiếm ưu thế và tập trung học tốt hơn các mẫu khó, góp phần cải thiện hiệu năng tổng thể trong bài toán phát hiện mối nguy an toàn thực phẩm.

2.3.3 Hàm mất mát đa nhiệm

Trong bối cảnh học đa nhiệm, mô hình được huấn luyện đồng thời trên hai tác vụ liên quan chặt chẽ: phân loại nhóm sản phẩm (*product-category*) và phân loại loại mối nguy (*hazard-category*). Mỗi tác vụ có một hàm mất mát riêng biệt, được tính độc lập trên các logits đầu ra tương ứng. Tổng hàm mất mát của mô hình được xây dựng bằng cách kết hợp tuyến tính hai hàm mất mát thành phần:

$$\mathcal{L} = \lambda_p \mathcal{L}_{\text{product}} + \lambda_h \mathcal{L}_{\text{hazard}}, \tag{2}$$

trong đó $\mathcal{L}_{\text{product}}$ và $\mathcal{L}_{\text{hazard}}$ lần lượt là Focal Loss cho tác vụ phân loại nhóm sản phẩm và loại mối nguy, còn λ_p và λ_h là các hệ số điều chỉnh mức độ đóng góp của từng tác vụ vào quá trình huấn luyện.

Trong thực nghiệm này, hai trọng số được đặt bằng nhau, cụ thể $\lambda_p = \lambda_h = 0.5$. Cách thiết lập này phản ánh giả định rằng hai tác vụ có tầm quan trọng tương đương trong bài toán phát hiện mối nguy thực phẩm, đồng thời giúp đảm bảo sự cân bằng trong quá trình cập nhật tham số của encoder dùng chung. Việc sử dụng trọng số cố định và đối xứng cũng giúp giảm số lượng siêu tham số cần tinh chỉnh, từ đó tăng tính ổn định và khả năng tái lập kết quả của mô hình.

Một hướng tiếp cận phổ biến trong học đa nhiệm là điều chỉnh trọng số các tác vụ dựa trên độ bất định của từng tác vụ, tiêu biểu là phương pháp được đề xuất bởi Cipolla et al. [1]. Theo cách tiếp cận này, tổng hàm mất mát được định nghĩa như sau:

$$\mathcal{L} = \sum_t \left(\frac{1}{2\sigma_t^2} \mathcal{L}_t + \log \sigma_t \right), \tag{3}$$

trong đó σ_t biểu diễn độ bất định (uncertainty) của tác vụ t. Sau khi tái tham số hóa với $s_t = \log \sigma_t^2$, công thức tương đương có dạng:

$$\mathcal{L} = \sum_t \left(\exp(-s_t) \mathcal{L}_t + s_t \right). \tag{4}$$

Cách tiếp cận này cho phép mô hình tự động điều chỉnh mức độ đóng góp của từng tác vụ trong quá trình huấn luyện.

Tuy nhiên, trong khuôn khổ nghiên cứu này, nhóm không áp dụng cơ chế trọng số dựa trên độ bất định mà sử dụng trọng số cố định cho hai tác vụ. Lý do là hai tác vụ có mức độ liên quan ngữ nghĩa cao và được huấn luyện trên cùng tập dữ liệu, do đó không có sự chênh lệch rõ rệt về độ khó hoặc mức độ nhiễu. Thực nghiệm sơ bộ cho thấy việc sử dụng trọng số động không mang lại cải thiện so với cấu hình trọng số cố định mà còn làm giảm khả năng dự đoán, trong khi lại làm tăng độ phức tạp của mô hình và khó tái lập kết quả. Vì vậy, nhóm lựa chọn cách kết hợp đơn giản với trọng số bằng nhau nhằm ưu tiên tính ổn định và hiệu quả thực tiễn.

2.3.4 Chiến lược huấn luyện

Chiến lược huấn luyện và các siêu tham số được thiết lập như sau:

- **Bộ tối ưu**: AdamW với hệ số suy giảm trọng số (weight decay) nhằm giảm hiện tượng quá khớp.
- **Tốc độ học**: 1×10^{-5}, kết hợp với lịch học cosine.

- **Warm-up**: 10% số bước huấn luyện đầu tiên được sử dụng cho giai đoạn làm nóng tốc độ học.
- **Số epoch**: 10.
- **Độ dài đầu vào tối đa**: 512 token cho mỗi đoạn văn bản.
- **Batch size**: 2 mẫu trên mỗi thiết bị, kết hợp với tích lũy gradient trong 4 bước để mô phỏng batch size hiệu dụng lớn hơn.
- **Huấn luyện chính xác hỗn hợp**: sử dụng *mixed precision* (FP16) nhằm giảm chi phí bộ nhớ và tăng tốc quá trình huấn luyện.
- **Chiến lược đánh giá và lưu mô hình**: đánh giá trên tập kiểm tra sau mỗi epoch và lưu mô hình có giá trị hàm mất mát thấp nhất.

Việc kết hợp học đa nhiệm với Focal Loss giúp mô hình học được các biểu diễn ngữ nghĩa dùng chung hiệu quả hơn, đồng thời cải thiện khả năng phân loại các lớp hiếm trong bài toán phát hiện mối nguy an toàn thực phẩm.

2.4 Chiến lược Ensemble

Nhằm khai thác tối đa thông tin từ các báo cáo dài và tận dụng tính bổ sung giữa các mô hình khác nhau, nhóm đề xuất một chiến lược ensemble hai tầng, bao gồm ensemble ở mức đoạn văn bản (chunk-level ensemble) và ensemble ở mức mô hình (model-level ensemble). Cách tiếp cận này cho phép cải thiện độ ổn định và khả năng tổng quát hóa của hệ thống mà không cần huấn luyện thêm mô hình mới.

2.4.1 Ensemble ở mức đoạn văn bản

Do mỗi báo cáo sự cố có thể được chia thành nhiều đoạn văn bản (chunk), mô hình sẽ đưa ra dự đoán độc lập cho từng chunk. Với mỗi chunk, các logits đầu ra cho hai tác vụ phân loại nhóm sản phẩm và loại mối nguy được chuyển đổi thành phân phối xác suất thông qua hàm softmax.

Để thu được dự đoán ở mức tài liệu, các phân phối xác suất của tất cả các chunk thuộc cùng một báo cáo được tổng hợp bằng phép trung bình (mean pooling):

$$\mathbf{p}_{\text{doc}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{p}_i, \tag{5}$$

trong đó $\mathbf{p}_i$ là vector xác suất dự đoán của chunk thứ i và N là số lượng chunk của tài liệu. Cách tổng hợp này giúp làm giảm nhiễu từ các đoạn văn bản kém thông tin và nhấn mạnh các tín hiệu ngữ nghĩa xuất hiện nhất quán xuyên suốt tài liệu.

2.4.2 Ensemble ở mức mô hình

Bên cạnh ensemble ở mức đoạn, nhóm tiếp tục kết hợp dự đoán từ hai mô hình nền khác nhau là DeBERTa và RoBERTa. Mỗi mô hình cung cấp một phân phối xác suất ở mức tài liệu cho từng tác vụ. Các phân phối này được kết hợp bằng phương pháp ensemble mềm (soft voting) thông qua trung bình có trọng số:

$$\mathbf{p}_{\text{ens}} = w \cdot \mathbf{p}_{\text{DeBERTa}} + (1 - w) \cdot \mathbf{p}_{\text{RoBERTa}}, \tag{6}$$

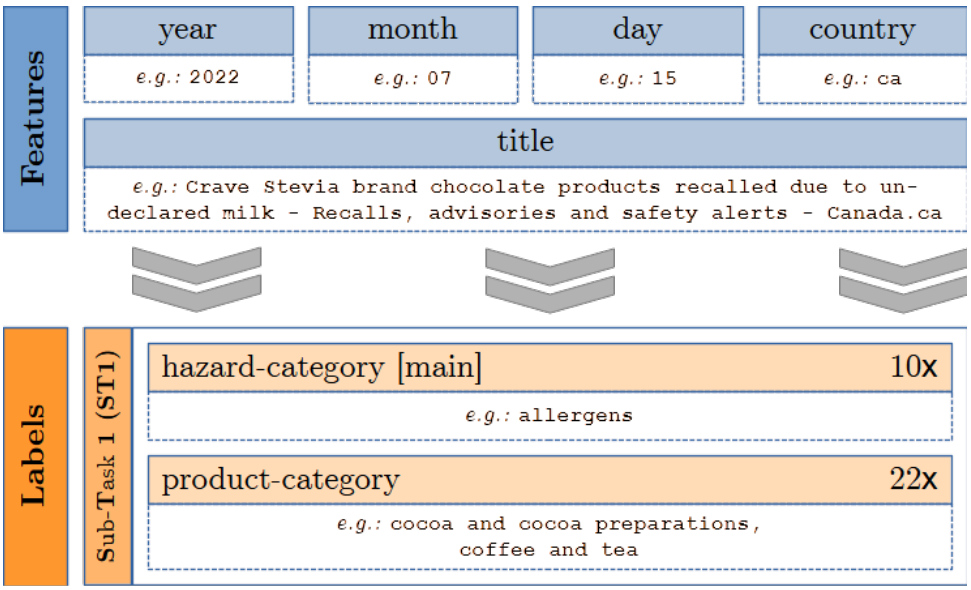
trong đó $w \in [0, 1]$ là trọng số của mô hình DeBERTa và được lựa chọn sao cho tối đa hóa macro-F1 trên tập xác thực.

Trọng số ensemble được tìm kiếm bằng phương pháp grid search trên tập xác thực, đảm bảo rằng quá trình lựa chọn siêu tham số không gây rò rỉ thông tin từ tập kiểm tra. Sau khi xác định trọng số tối ưu, chiến lược ensemble này được áp dụng cố định cho tập kiểm tra và giai đoạn suy luận.

3 Bộ dữ liệu

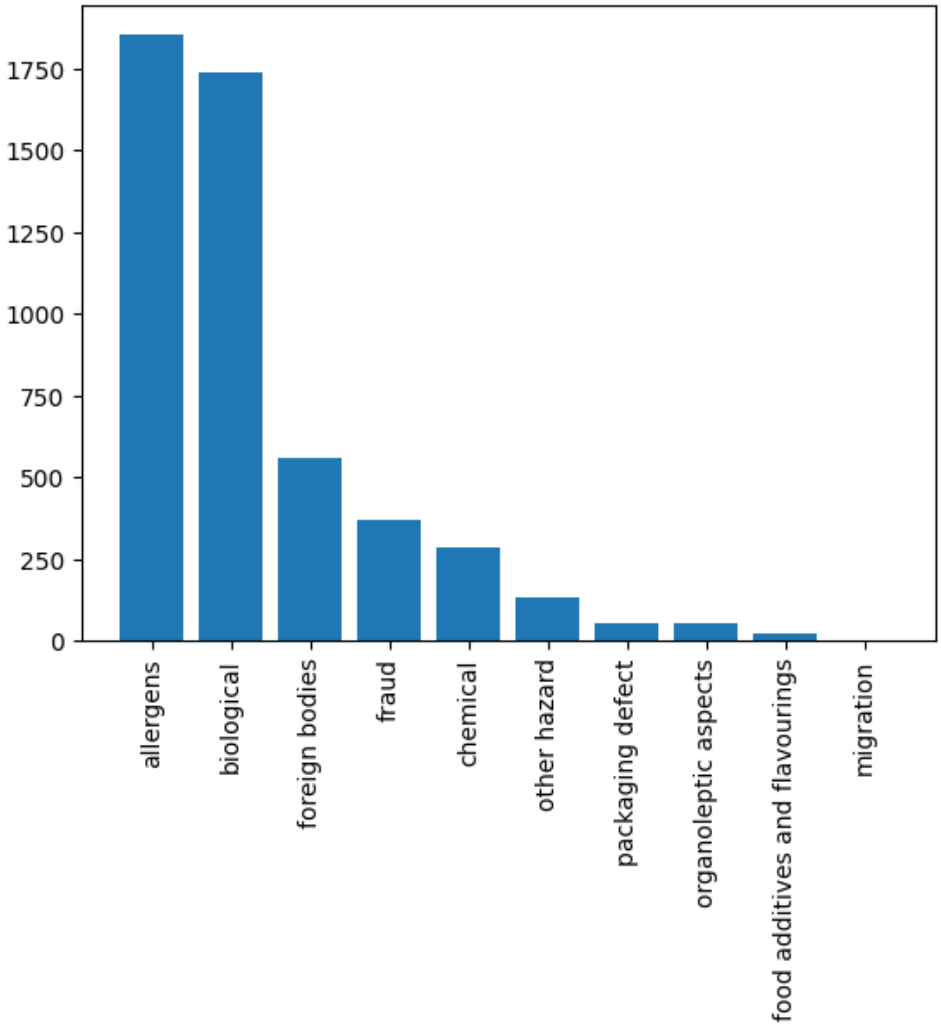
3.1 Dữ liệu gốc (Original Data)

Bộ dữ liệu được sử dụng trong nghiên cứu này được cung cấp trong khuôn khổ SemEval-2025 Task 9 [7], bao gồm các báo cáo sự cố và cảnh báo an toàn thực phẩm được thu thập từ các nguồn web công khai. Mỗi mẫu dữ liệu bao gồm tiêu đề, nội dung văn bản và hai nhãn phân loại tương ứng với hai tác vụ của bài toán: **hazard-category** (loại mối nguy) và **product-category** (nhóm sản phẩm).



Hình 2: Tổng quan cấu trúc bộ dữ liệu SemEval-2025 Task 9. Các khối màu xanh biểu diễn đặc trưng đầu vào (năm, tháng, ngày, quốc gia, tiêu đề và nội dung văn bản), trong khi các khối màu cam biểu diễn nhãn chuẩn cho từng tác vụ. Con số bên phải mỗi nhãn cho biết số lượng giá trị phân biệt tương ứng.

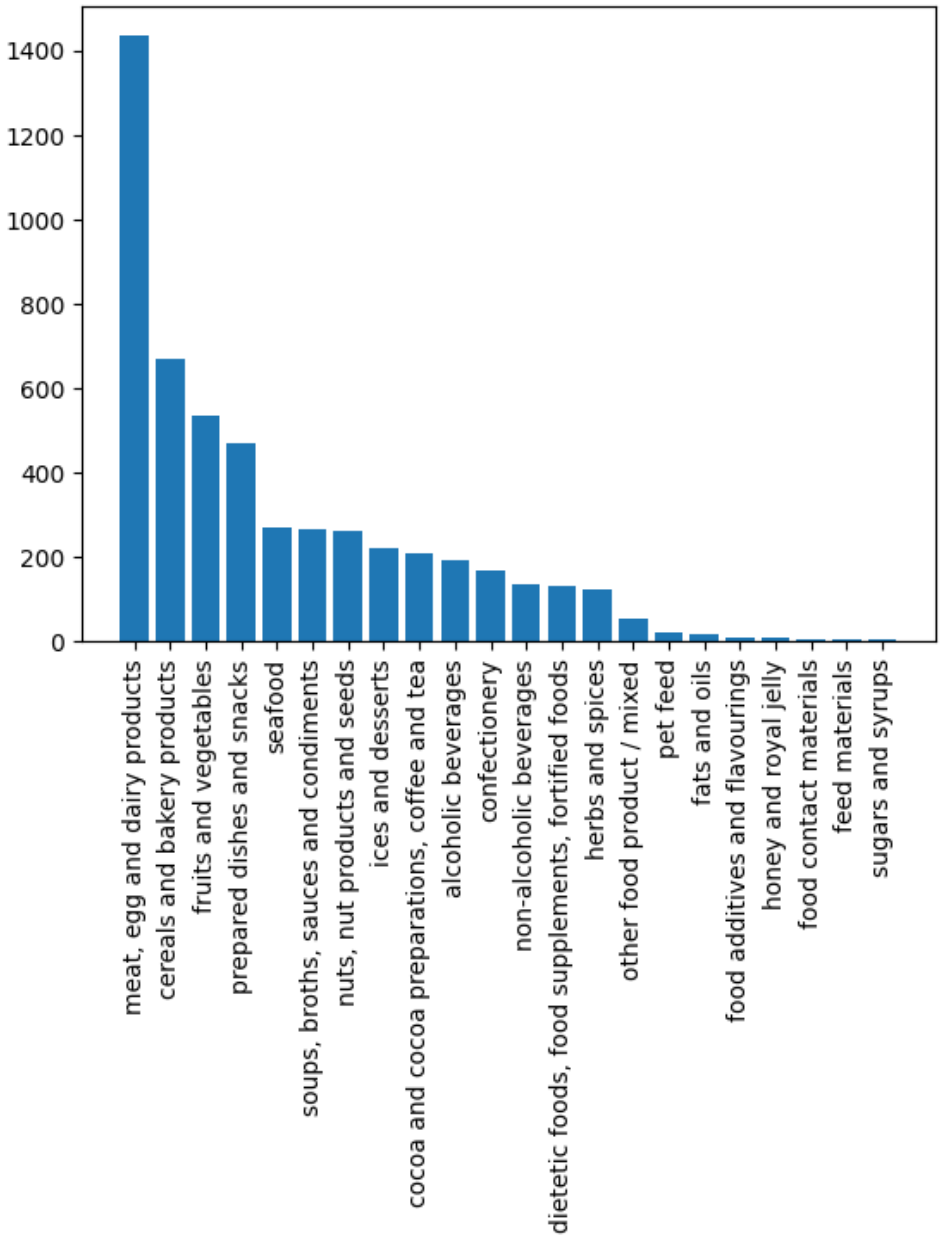
Phân bố nhãn mối nguy (hazard-category). Tập dữ liệu bao gồm 10 loại mối nguy thực phẩm với phân bố nhãn mất cân bằng rõ rệt, phản ánh đặc trưng long-tail của bài toán. Hai lớp chiếm ưu thế là *allergens* và *biological*, trong khi một số lớp hiếm như *migration* và *food additives and flavourings* chỉ xuất hiện với số lượng rất nhỏ. Phân bố cụ thể của các nhãn mối nguy được liệt kê như sau:



Hình 3: Phân bố nhãn mối nguy (hazard-category) trong tập dữ liệu gốc

- allergens: 1854
- biological: 1741
- foreign bodies: 561
- fraud: 371
- chemical: 287
- other hazard: 134
- packaging defect: 54
- organoleptic aspects: 53
- food additives and flavourings: 24
- migration: 3

Phân bố nhãn nhóm sản phẩm (product-category). Dữ liệu được gán nhãn theo 22 nhóm sản phẩm khác nhau, với sự chênh lệch đáng kể về số lượng mẫu giữa các lớp. Các nhóm sản phẩm phổ biến nhất chủ yếu liên quan đến thịt, sữa và các sản phẩm ngũ cốc, trong khi một số nhóm khác chỉ xuất hiện với tần suất thấp. Các nhóm sản phẩm có số lượng mẫu lớn nhất bao gồm:



Hình 4: Phân bố nhãn nhóm sản phẩm (product-category) trong tập dữ liệu gốc

- meat, egg and dairy products: 1434
- cereals and bakery products: 671
- fruits and vegetables: 535
- prepared dishes and snacks: 469

- seafood: 268
- soups, broths, sauces and condiments: 264
- nuts, nut products and seeds: 262
- ices and desserts: 222
- cocoa and cocoa preparations, coffee and tea: 210
- alcoholic beverages: 193
- ...

Sự mất cân bằng trong phân bố nhãn nhóm sản phẩm làm gia tăng độ phức tạp của bài toán phân loại đa lớp và đặt ra yêu cầu cao đối với khả năng tổng quát hóa của mô hình.

Phân bố nhãn mối nguy và nhóm sản phẩm được minh họa lần lượt trong Hình 3 và Hình 4, cho thấy rõ đặc trưng long-tail của tập dữ liệu.

"Randsland brand Super Salad Kit recalled due to Listeria monocytogenes"	
hazard:	listeria monocytogenes
hazard-category:	biological
product:	salads
product-category:	fruits and vegetables
"Create Common Good Recalls Jambalaya Products Due To Misbranding and Undeclared Allergens"	
hazard:	milk and products thereof
hazard-category:	allergens
product:	meat preparations
product-category:	meat, egg and dairy products
"Nestlé Prepared Foods Recalls Lean Cuisine Baked Chicken Meal Products Due to Possible Foreign Matter Contamination"	
hazard:	plastic fragment
hazard-category:	foreign bodies
product:	cooked chicken
product-category:	prepared dishes and snacks

Hình 5: Ví dụ các mẫu văn bản trong bộ dữ liệu SemEval-2025 Task 9 kèm theo nhãn tương ứng. Mỗi mẫu bao gồm tiêu đề báo cáo thu hồi thực phẩm và các nhãn *hazard-category* và *product-category* được gán thủ công.

Phân bố ngôn ngữ. Bộ dữ liệu có tính đa ngôn ngữ, phản ánh nguồn thu thập từ nhiều cơ quan quản lý an toàn thực phẩm khác nhau trên thế giới. Trong đó, tiếng Anh chiếm tỷ lệ lớn nhất với khoảng 82.49% số mẫu, tiếp theo là tiếng Đức với 14.84%. Một số ngôn ngữ khác như tiếng Trung, Afrikaans, Luxembourgish, Hy Lạp, Đan Mạch và Ý chỉ xuất hiện với tỷ lệ rất nhỏ (dưới 2% mỗi ngôn ngữ).

Mặc dù dữ liệu không đồng nhất hoàn toàn về mặt ngôn ngữ, phần lớn các báo cáo sử dụng các ngôn ngữ có tài nguyên phong phú, giúp giảm bớt độ phức tạp trong quá trình tiền xử lý. Tuy nhiên, sự đa dạng ngôn ngữ vẫn đặt ra thách thức nhất định cho mô hình, đặc biệt trong việc học các biểu diễn ngữ nghĩa ổn định và tổng quát hóa tốt trong bối cảnh dữ liệu có phân bố nhãn dài đuôi.

Chia tập dữ liệu. Tập dữ liệu gốc gồm tổng cộng 5984 mẫu và được chia thành hai phần theo tỉ lệ 80/20:

- **Tập huấn luyện**: 4787 mẫu
- **Tập xác thực**: 1197 mẫu

Phân bố nhãn không đồng đều trên cả hai tập tiếp tục phản ánh rõ đặc trưng long-tail của bài toán, đồng thời là động lực chính cho việc áp dụng các chiến lược học đa nhiệm, hàm mất mát Focal và các phương pháp ensemble trong hệ thống đề xuất.

3.2 Dữ liệu tăng cường (Augmented Data)

Trong bài toán phát hiện mối nguy thực phẩm từ văn bản web, dữ liệu huấn luyện không chỉ đối mặt với hiện tượng mất cân bằng lớp nghiêm trọng mà còn chịu ảnh hưởng mạnh từ sự đa dạng và chồng lấn ngữ nghĩa giữa các nhãn. Các báo cáo an toàn thực phẩm thường được thu thập từ nhiều nguồn khác nhau, với cách diễn đạt không đồng nhất và mức độ chi tiết khác nhau, khiến mô hình dễ học lệch vào các mẫu ngôn ngữ phổ biến nếu dữ liệu huấn luyện không đủ phong phú về mặt ngữ nghĩa.

Xuất phát từ thực tế này, nhóm xây dựng và so sánh ba cấu hình dữ liệu huấn luyện khác nhau: *Original* (không tăng cường), *Aug1* (tăng cường có kiểm soát) và *Aug2* (tăng cường mạnh). Mục tiêu của thí nghiệm không chỉ nhằm cải thiện hiệu năng trên các lớp hiếm, mà quan trọng hơn là đánh giá khả năng tổng quát hóa của mô hình khi dữ liệu được mở rộng với các mức độ khác nhau.

3.2.1 Mô tả các cấu hình dữ liệu.

- **Original**: Tập dữ liệu gốc của SemEval-2025 Task 9 gồm 5 984 mẫu. Cấu hình này phản ánh trung thực phân bố dữ liệu thực tế nhưng còn hạn chế về độ bao phủ ngữ nghĩa, đặc biệt đối với các lớp cực hiếm và các trường hợp có nội dung chồng lấn.
- **Aug1**: Tập dữ liệu tăng cường có kiểm soát với 6 496 mẫu. Việc tăng cường tập trung vào (i) các lớp cực hiếm như *migration* và *food additives and flavourings*, và (ii) các lớp có tần suất cao nhưng dễ gây nhầm lẫn về ngữ nghĩa như *biological*, *chemical* và *allergens*. Mức tăng vừa phải giúp mở rộng không gian biểu diễn mà vẫn giữ được phân bố tổng thể gần với dữ liệu gốc.
- **Aug2**: Tập dữ liệu tăng cường mạnh với tổng cộng 8 000 mẫu. Cấu hình này được thiết kế nhằm khảo sát giới hạn của tăng cường dữ liệu khi số lượng mẫu được mở rộng đáng kể.

3.2.2 So sánh hiệu năng giữa các cấu hình dữ liệu.

Bảng 2 trình bày kết quả thực nghiệm của các mô hình đơn lẻ (DeBERTa và RoBERTa) trên ba cấu hình dữ liệu, được đánh giá trên cả tập xác thực và tập kiểm tra.

Dữ liệu	Mô hình	VALID_Prod	VALID_Haz	VALID_Avg	TEST_Prod	TEST_Haz	TEST_Avg
Original	DeBERTa	0.6721	0.8093	0.7407	0.7763	0.7465	0.7614
Original	RoBERTa	0.7023	0.8117	0.7570	0.7731	0.7574	0.7653
Aug1	DeBERTa	0.7340	0.8186	0.7763	0.7453	0.7246	0.7349
Aug1	RoBERTa	**0.7413**	**0.8455**	**0.7934**	**0.7896**	**0.8159**	**0.8027**
Aug2	DeBERTa	0.7820	0.8460	0.8140	0.7687	0.7741	0.7714
Aug2	RoBERTa	0.7733	0.8142	0.7938	0.7682	0.7205	0.7443

Bảng 2: So sánh hiệu năng của các cấu hình dữ liệu tăng cường. VALID và TEST lần lượt là tập xác thực và tập kiểm tra; Avg là trung bình macro-F1 của hai tác vụ.

3.2.3 Phân tích và lựa chọn cấu hình Aug1.

Kết quả trong Bảng 2 cho thấy xu hướng nhất quán. Trên dữ liệu *Original*, cả hai mô hình đạt hiệu năng ở mức khá, tuy nhiên macro-F1 trên tập kiểm tra vẫn còn hạn chế, phản ánh khả năng tổng quát hóa chưa cao do dữ liệu huấn luyện chưa đủ đa dạng về mặt ngữ nghĩa.

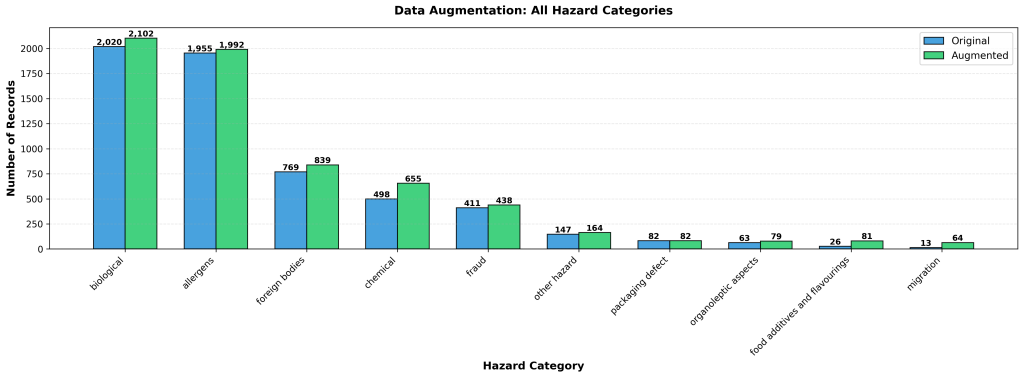
Khi áp dụng tăng cường dữ liệu ở mức vừa phải với *Aug1*, hiệu năng của mô hình RoBERTa được cải thiện rõ rệt và ổn định trên cả tập xác thực và tập kiểm tra, đạt macro-F1 cao nhất (0.8027). Điều này cho thấy *Aug1* đã bổ sung đủ các biến thể ngữ nghĩa cần thiết để mô hình học được các ranh giới quyết định rõ ràng hơn, đặc biệt đối với các lớp hiếm và các trường hợp dễ gây nhầm lẫn, trong khi vẫn duy trì được sự nhất quán với phân bố dữ liệu gốc.

Ngược lại, mặc dù *Aug2* đạt điểm rất cao trên tập xác thực, hiệu năng trên tập kiểm tra lại giảm đáng kể, đặc biệt đối với mô hình RoBERTa. Việc mở rộng dữ liệu lên tới 8 000 mẫu có khả năng đã làm lệch phân bố ngữ nghĩa và đưa vào quá nhiều mẫu tăng cường kém đại diện, khiến mô hình học các tín hiệu không bền vững và làm suy giảm khả năng tổng quát hóa.

Từ các phân tích trên, nhóm lựa chọn *Aug1* (6 496 mẫu) làm cấu hình dữ liệu tăng cường chính cho toàn bộ hệ thống, do đạt được sự cân bằng tối ưu giữa quy mô dữ liệu, độ đa dạng ngữ nghĩa và tính ổn định khi suy luận.

3.2.4 Phân tích chi tiết dữ liệu Aug1.

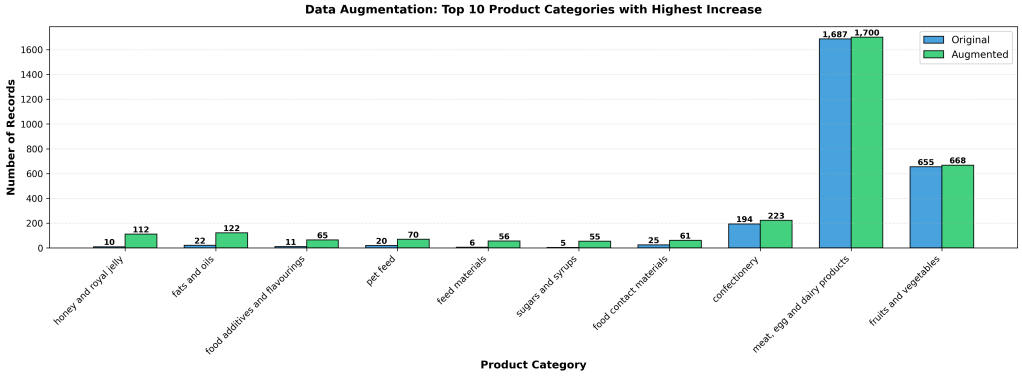
Sau khi xác định *Aug1* là cấu hình tối ưu, phần này tập trung phân tích cấu trúc và đặc điểm của tập dữ liệu tăng cường này nhằm làm rõ cơ chế giúp cải thiện hiệu năng và khả năng tổng quát hóa của mô hình. Aug1 gồm tổng cộng 6 496 mẫu, tăng khoảng 8.5% so với tập *Original*, trong đó việc tăng cường được thiết kế có chọn lọc theo từng nhóm nhãn thay vì cân bằng cơ học về mặt số lượng.



Hình 6: So sánh phân bố nhãn *hazard-category* trước và sau tăng cường dữ liệu (Aug1).

Như minh họa trong Hình 6, chiến lược tăng cường trong Aug1 được áp dụng trên toàn bộ các nhóm mối nguy với mức độ khác nhau tùy theo đặc trưng của từng lớp. Các lớp cực hiếm như *migration* và *food additives and flavourings* được tăng cường đáng kể nhằm giảm tình trạng thiếu mẫu nghiêm trọng, giúp mô hình có đủ tín hiệu để học được các biểu diễn ngữ nghĩa cơ bản thay vì chỉ ghi nhớ một vài mẫu đơn lẻ.

Bên cạnh đó, một điểm quan trọng của Aug1 là việc bổ sung dữ liệu có chọn lọc cho các lớp có tần suất cao như *biological* và *allergens*, dù các lớp này đã chiếm ưu thế trong tập gốc. Nguyên nhân là các lớp này thường chứa các mô tả rất đa dạng và có mức độ chồng lấn ngữ nghĩa cao với các lớp khác, đặc biệt là *chemical* và *foreign bodies*. Trong nhiều báo cáo an toàn thực phẩm, các tác nhân sinh học, hóa chất và dị ứng có thể cùng xuất hiện trong một ngữ cảnh, khiến mô hình dễ dự đoán sai nếu chỉ học dựa trên tần suất xuất hiện. Việc tăng cường các lớp lớn này giúp mô hình tiếp cận nhiều biến thể diễn đạt hơn, từ đó học được ranh giới quyết định ổn định và rõ ràng hơn giữa các nhóm mối nguy.



Hình 7: So sánh phân bố nhãn *product-category* trước và sau tăng cường dữ liệu (Aug1, Top 10 nhóm tăng nhiều nhất).

Đối với tác vụ phân loại *product-category*, chiến lược tăng cường của Aug1 cũng tuân theo nguyên tắc tương tự (Hình 7). Các nhóm sản phẩm có số lượng mẫu rất thấp như *honey and royal jelly*, *fats and oils* và *food additives and flavourings* được tăng cường mạnh nhằm giảm độ chênh lệch giữa các lớp và hạn chế hiện tượng mô hình bỏ qua hoàn toàn các nhóm hiếm trong quá trình huấn luyện.

Song song đó, một số nhóm sản phẩm phổ biến như *meat, egg and dairy products* và *fruits and vegetables* vẫn được bổ sung nhẹ. Việc này không nhằm cân bằng số lượng, mà

nhằm tăng tính đa dạng về ngữ cảnh và cách diễn đạt trong văn bản web, phản ánh tốt hơn thực tế các báo cáo an toàn thực phẩm, nơi cùng một nhóm sản phẩm có thể được mô tả theo nhiều cách khác nhau tùy nguồn và ngôn ngữ.

Đáng chú ý, các nhóm liên quan đến phụ gia và thành phần (*food additives and flavourings*) đóng vai trò cầu nối quan trọng trong bài toán học đa nhiệm, do chúng có thể xuất hiện đồng thời như một tín hiệu ở cả hai nhãn *hazard-category* và *product-category*. Việc tăng cường dữ liệu cho các nhóm này giúp mô hình giảm nhầm lẫn chéo giữa hai tác vụ và khai thác hiệu quả hơn mối liên hệ ngữ nghĩa giữa loại mối nguy và nhóm sản phẩm.

Tổng thể, Aug1 không được thiết kế để cân bằng phân bố nhãn một cách cơ học, mà hướng tới việc mở rộng không gian biểu diễn ngữ nghĩa theo cách có kiểm soát. Với quy mô vừa phải (6 496 mẫu), Aug1 giúp mô hình học tốt hơn các lớp hiếm, đồng thời làm rõ ranh giới giữa các lớp có nội dung chồng lấn, từ đó cải thiện khả năng tổng quát hóa trên dữ liệu chưa từng quan sát. Chiến lược này đóng vai trò quan trọng trong việc nâng cao hiệu năng và độ ổn định của hệ thống phát hiện mối nguy thực phẩm đề xuất.

4 Kết quả

4.1 Thiết lập đánh giá

Trong SemEval-2025 Task 9 [7], hiệu năng hệ thống được đánh giá bằng chỉ số **macro-F1**, được tính trung bình trên hai tác vụ hazard-category và product-category. Chỉ số này đặc biệt phù hợp với bài toán có phân bố nhãn dài đuôi, do đảm bảo mỗi lớp đóng góp ngang nhau vào điểm số tổng thể.

Macro-F1 cho từng tác vụ được định nghĩa như sau:

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{C} \sum_{c=1}^{C} \frac{2 \cdot \mathrm{Precision}_c \cdot \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}, \tag{7}$$

trong đó C là số lượng lớp của tác vụ tương ứng.

Điểm số cuối cùng của hệ thống được tính bằng trung bình cộng macro-F1 của hai tác vụ:

$$\mathrm{Score} = \frac{1}{2} \left(\mathrm{F1}_{\mathrm{hazard}} + \mathrm{F1}_{\mathrm{product}} \right). \tag{8}$$

Đáng chú ý, macro-F1 của tác vụ product-category chỉ được tính trên các mẫu mà mô hình dự đoán đúng nhãn hazard-category, phản ánh cấu trúc phân cấp của bài toán.

4.2 Kết quả xếp hạng

Trước khi so sánh hệ thống đề xuất với các đội tham gia khác trong SemEval-2025 Task 9, nhóm tiến hành phân tích chi tiết hiệu năng của từng mô hình thành phần cũng như tác động của chiến lược ensemble lên kết quả cuối cùng. Việc này nhằm làm rõ đóng góp của từng thành phần trong hệ thống và lý giải một cách minh bạch kết quả đạt được.

4.2.1 Kết quả mô hình đơn lẻ

Bảng 3 trình bày hiệu năng của hai mô hình nền khi được huấn luyện đơn lẻ với cấu hình dữ liệu tăng cường Aug1 và đánh giá trên tập kiểm tra.

Mô hình	TEST_Product_F1	TEST_Hazard_F1	TEST_Avg_F1
DeBERTa (single)	0.7453	0.7246	0.7349
RoBERTa (single)	**0.7896**	**0.8159**	**0.8027**

Bảng 3: Kết quả trên tập kiểm tra của các mô hình đơn lẻ với dữ liệu Aug1.

Kết quả cho thấy RoBERTa vượt trội hơn DeBERTa trên cả hai tác vụ. Đặc biệt, RoBERTa đạt macro-F1 trung bình 0.8027, cho thấy khả năng tổng quát hóa mạnh mẽ trong bối cảnh dữ liệu đa ngôn ngữ và phân bố nhãn dài đuôi. Trong khi đó, DeBERTa tuy có hiệu năng thấp hơn nhưng vẫn học được các đặc trưng ngữ nghĩa bổ sung, tạo tiền đề cho chiến lược ensemble ở mức mô hình.

4.2.2 Ensemble với trọng số suy ra từ tập xác thực

Tiếp theo, nhóm áp dụng chiến lược ensemble mềm giữa DeBERTa và RoBERTa, trong đó trọng số được lựa chọn trực tiếp dựa trên giá trị macro-F1 tối ưu trên tập xác thực (VALID). Kết quả trên tập kiểm tra được trình bày trong Bảng 4.

TEST_Product_F1	TEST_Hazard_F1	TEST_Avg_F1
0.8161	0.7524	0.7842

Bảng 4: Kết quả ensemble khi sử dụng trọng số được chọn trực tiếp từ tập xác thực.

Mặc dù chiến lược này giúp cải thiện đáng kể điểm product-category so với từng mô hình đơn lẻ, macro-F1 tổng thể lại giảm so với RoBERTa single. Nguyên nhân chính là sự lệch phân bố giữa tập xác thực và tập kiểm tra, khiến trọng số tối ưu trên VALID không còn phù hợp khi suy luận trên dữ liệu chưa quan sát. Hiện tượng này cho thấy ensemble nhạy cảm với cách lựa chọn trọng số và cần được tinh chỉnh cẩn trọng để đảm bảo khả năng tổng quát hóa.

4.2.3 Ensemble với trọng số tinh chỉnh

Để khắc phục hạn chế trên, nhóm tiến hành tinh chỉnh trọng số ensemble theo hướng ổn định hóa, thay vì tối ưu cứng theo tập xác thực. Với cấu hình trọng số được điều chỉnh, hệ thống đạt kết quả tốt nhất trên tập kiểm tra như trình bày trong Bảng 5.

TEST_Product_F1	TEST_Hazard_F1	TEST_Avg_F1
0.7909	**0.8174**	**0.8042**

Bảng 5: Kết quả ensemble tốt nhất sau khi tinh chỉnh trọng số.

Kết quả này cho thấy việc điều chỉnh trọng số ensemble theo hướng cân bằng hơn giữa hai mô hình giúp cải thiện rõ rệt tác vụ hazard-category, đồng thời duy trì hiệu năng cao trên product-category. Macro-F1 trung bình đạt 0.8042, cao hơn cả mô hình đơn lẻ tốt nhất, chứng minh tính bổ sung hiệu quả giữa DeBERTa và RoBERTa khi được kết hợp hợp lý.

4.2.4 So sánh với các hệ thống khác

Cuối cùng, Bảng 6 trình bày kết quả xếp hạng Top-6 hệ thống trong SemEval-2025 Task 9. Hệ thống của nhóm đạt macro-F1 0.8042 với cấu hình **TITLE + TEXT**, xếp hạng thứ 3 và vượt qua nhiều hệ thống có pipeline phức tạp hơn.

Hạng	Đội	Macro-F1	Đặc trưng
1	Anastasia	0.8223	META, TITLE, TEXT
2	MyMy	0.8112	META, TITLE, TEXT
3	**Ours (best weight)**	**0.8042**	**TITLE, TEXT**
4	SRCB	0.8039	TITLE, TEXT
5	PATeam	0.8017	TITLE, TEXT
6	HU	0.7882	TITLE, TEXT

Bảng 6: Kết quả xếp hạng SemEval-2025 Task 9 (macro-F1).

Kết quả macro-F1 cao của hệ thống đề xuất đến từ sự kết hợp hiệu quả của bốn thành phần chính trong pipeline. (1) **Tiền xử lý và chia đoạn văn bản** giúp giảm nhiễu đặc trưng của dữ liệu web và cho phép mô hình khai thác đầy đủ thông tin từ các báo cáo dài thông qua cơ chế chia đoạn có chồng lấp, từ đó ổn định hóa đầu vào và cải thiện chất lượng biểu diễn ngữ nghĩa. (2) **Học đa nhiệm** [1] cho phép mô hình đồng thời dự đoán *hazard-category* và *product-category*, qua đó khai thác mối liên hệ ngữ nghĩa chặt chẽ giữa hai tác vụ và giảm hiện tượng nhầm lẫn chéo, đặc biệt trong các trường hợp nội dung mối nguy và nhóm sản phẩm có mức độ chồng lấn cao. (3) **Hàm mất mát Focal Loss** [4] đóng vai trò quan trọng trong việc xử lý phân bố nhãn dài đuôi của SemEval-2025 Task 9 [7], bằng cách giảm trọng số của các mẫu dễ và tập trung học tốt hơn các mẫu khó cũng như các lớp hiếm. (4) **Chiến lược tăng cường dữ liệu có kiểm soát** giúp mở rộng không gian biểu diễn ngữ nghĩa theo cách cân bằng, vừa bổ sung các lớp cực hiếm như *migration* và *food additives and flavourings*, vừa tăng tính đa dạng diễn đạt cho các lớp dễ gây nhầm lẫn như *biological* và *chemical*, cũng như các nhóm sản phẩm phụ thuộc mạnh vào ngữ cảnh văn bản web. Nhờ đó, mô hình đạt khả năng tổng quát hóa tốt hơn trên dữ liệu đa ngôn ngữ và phân bố dài đuôi.

Xét về đóng góp của từng mô hình nền, **RoBERTa** [5] cho thấy năng lực tổng quát hóa vượt trội khi được huấn luyện đơn lẻ với dữ liệu Aug1, đạt macro-F1 0.8027 trên tập kiểm tra. Kết quả này phản ánh khả năng học biểu diễn ngữ nghĩa mạnh mẽ của RoBERTa trong bối cảnh dữ liệu không đồng nhất và đa ngôn ngữ. Trong khi đó, **DeBERTa** [2], dù đạt hiệu năng thấp hơn (macro-F1 0.7349), vẫn học được các đặc trưng ngữ nghĩa bổ sung và đóng vai trò quan trọng trong chiến lược ensemble, giúp cải thiện hiệu năng tổng thể khi được kết hợp hợp lý.

Kết quả ensemble cũng cho thấy tính nhạy cảm của việc lựa chọn trọng số. Khi trọng số được suy ra trực tiếp từ tập xác thực, hệ thống chỉ đạt macro-F1 0.7842 trên tập kiểm tra, phản ánh sự lệch phân bố giữa các tập dữ liệu và nguy cơ quá khớp vào tập xác thực. Ngược lại, khi trọng số ensemble được tinh chỉnh theo hướng ổn định hóa, hệ thống đạt kết quả tốt nhất với macro-F1 0.8042, cao hơn cả mô hình đơn lẻ tốt nhất. Điều này cho thấy việc cân bằng hợp lý giữa các mô hình nền giúp khai thác hiệu quả tính bổ sung về đặc trưng ngữ nghĩa, đồng thời nâng cao khả năng tổng quát hóa của hệ thống trong điều kiện dữ liệu thực tế.

5 Kết luận

Trong nghiên cứu này, nhóm đã đề xuất một hệ thống phát hiện mối nguy thực phẩm từ văn bản web dựa trên các mô hình Transformer, kết hợp học đa nhiệm, hàm mất mát Focal và chiến lược ensemble mềm. Thông qua pipeline tiền xử lý có hệ thống, cơ chế chia đoạn văn bản, cùng chiến lược tăng cường dữ liệu toàn diện, mô hình đạt hiệu năng cao và ổn định trong bối cảnh dữ liệu có phân bố dài đuôi và chồng lấn ngữ nghĩa.

Kết quả thực nghiệm trên SemEval-2025 Task 9 cho thấy hệ thống đề xuất đạt macro-F1 0.8042 với cấu hình TITLE + TEXT, nằm trong nhóm dẫn đầu của cuộc thi. Đặc biệt, RoBERTa đơn lẻ đã cho hiệu năng rất cạnh tranh, cho thấy tiềm năng triển khai các hệ thống gọn nhẹ nhưng hiệu quả cao trong các ứng dụng giám sát an toàn thực phẩm thực tế.

Trong tương lai, các hướng mở rộng có thể bao gồm khai thác thêm thông tin ngữ cảnh thời gian–không gian, áp dụng các mô hình đa ngôn ngữ chuyên biệt, hoặc tích hợp tri thức miền nhằm tiếp tục nâng cao khả năng tổng quát hóa và độ tin cậy của hệ thống.

Thông tin thêm: Bối cảnh nghiên cứu và định hướng triển khai

Công trình này được phát triển dựa trên kinh nghiệm và nền tảng từ hệ thống thi đấu chính thức của chính nhóm tại SemEval-2025 Task 9 [7]. Trong khuôn khổ cuộc thi, nhóm tác giả đã đạt **hạng nhất** và bài báo tương ứng đã được chấp nhận đăng tại ACL Anthology [3]. Hệ thống thi đấu đạt hiệu năng cao nhất nhờ khai thác nhiều thành phần phức tạp, bao gồm mô hình Transformer kích thước lớn, tăng cường dữ liệu mạnh, siêu dữ liệu (META) và chiến lược ensemble nhiều tầng.

Tuy nhiên, các hệ thống tối ưu cho môi trường cạnh tranh thường có chi phí huấn luyện và suy luận cao, cũng như độ phức tạp lớn trong triển khai. Do đó, đồ án này không nhằm tái hiện toàn bộ hệ thống thi đấu, mà tập trung phát triển một hướng tiếp cận cân bằng giữa hiệu năng và tính khả thi triển khai thực tế. Nhóm chủ động đơn giản hóa pipeline, giới hạn đầu vào ở mức **TITLE + TEXT**, đồng thời tối ưu các thành phần cốt lõi như tiền xử lý dữ liệu, học đa nhiệm với Focal Loss và tăng cường dữ liệu có kiểm soát.

Cách tiếp cận này hướng tới việc xây dựng một hệ thống gọn nhẹ, ổn định và dễ mở rộng, phù hợp hơn với các kịch bản ứng dụng thực tế trong giám sát an toàn thực phẩm tự động, trong khi vẫn duy trì hiệu năng cạnh tranh trên bài toán SemEval-2025 Task 9.

Tài liệu

[1] Roberto Cipolla, Yarin Gal, and Alex Kendall. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.

[2] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543, 2021.

[3] Tung Thanh Le, Tri Minh Ngo, and Trung Hieu Dang. Anastasia at SemEval-2025 task 9: Subtask 1, ensemble learning with data augmentation and focal loss for food risk classification. In Sara Rosenthal, Aiala Rosá, Debanjan Ghosh, and Marcos

Zampieri, editors, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 141–147, Vienna, Austria, July 2025. Association for Computational Linguistics.

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020.

[6] Ben Phan and Jung-Hsien Chiang. MyMy at SemEval-2025 task 9: A robust knowledge-augmented data approach for reliable food hazard detection. In Sara Rosenthal, Aiala Rosá, Debanjan Ghosh, and Marcos Zampieri, editors, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 812–822, Vienna, Austria, July 2025. Association for Computational Linguistics.

[7] Korbinian Randl, John Pavlopoulos, Aron Henriksson, Tony Lindgren, and Juli Bakagianni. SemEval-2025 task 9: The food hazard detection challenge. In Sara Rosenthal, Aiala Rosá, Debanjan Ghosh, and Marcos Zampieri, editors, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2523–2534, Vienna, Austria, July 2025. Association for Computational Linguistics.

[8] Muhammad Saad, Meesum Abbas, Sandesh Kumar, and Abdul Samad. HU at SemEval-2025 task 9: Leveraging LLM-based data augmentation for class imbalance. In Sara Rosenthal, Aiala Rosá, Debanjan Ghosh, and Marcos Zampieri, editors, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1593–1601, Vienna, Austria, July 2025. Association for Computational Linguistics.

[9] Xue Wan, Fengping Su, Ling Sun, Yuyang Lin, and Pengfei Chen. PATeam at SemEval-2025 task 9: LLM-augmented fusion for AI-driven food safety hazard detection. In Sara Rosenthal, Aiala Rosá, Debanjan Ghosh, and Marcos Zampieri, editors, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1912–1918, Vienna, Austria, July 2025. Association for Computational Linguistics.

[10] Yuming Zhang, Hongyu Li, Yongwei Zhang, Shanshan Jiang, and Bin Dong. SRCB at SemEval-2025 task 9: LLM finetuning approach based on external attention mechanism in the food hazard detection. In Sara Rosenthal, Aiala Rosá, Debanjan Ghosh, and Marcos Zampieri, editors, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 996–1003, Vienna, Austria, July 2025. Association for Computational Linguistics.