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

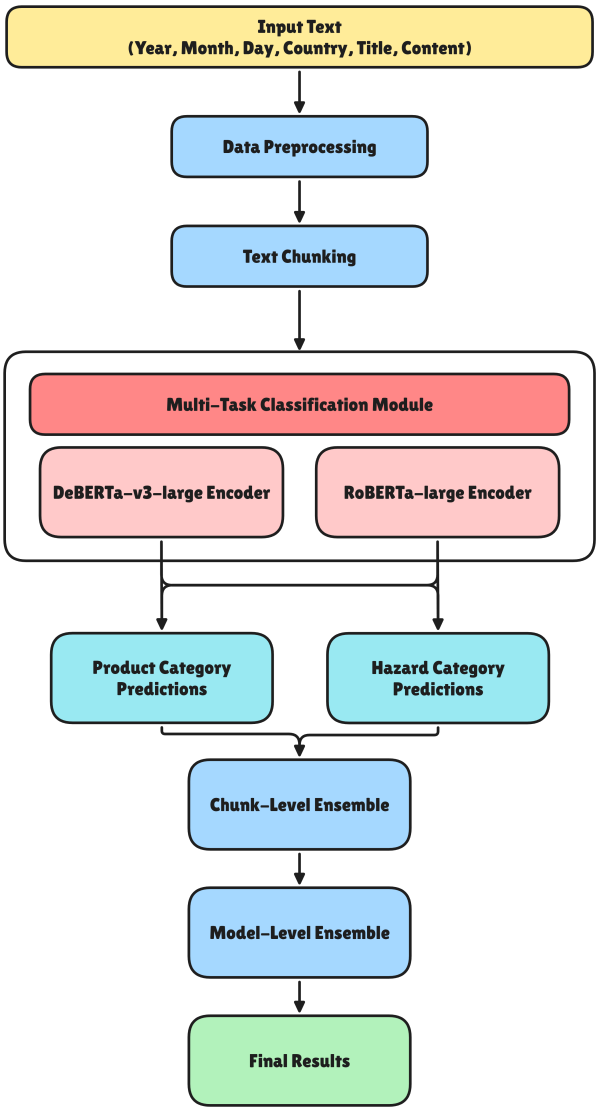
có thể làm gia tăng chi phí huấn luyện và suy luận, đồng thời hạn chế khả năng triển khai trong các hệ thống giám sát an toàn thực phẩm tự động ngoài thực tế. Xuất phát từ nhận định này, nhóm nghiên cứu kế thừa các ý tưởng hiệu quả từ hệ thống Anastasia [3] và phát triển theo hướng tinh gọn, nhằm cân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

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 chứa nhiều nhiễu như mã HTML, thông tin lặp lại, boilerplate mang tính hành chính, cũng như sự không nhất quán trong cách biểu diễn các thực thể chuyên ngành. Những yếu tố này có thể ảnh hưởng tiêu cực đến quá trình học và khả năng tổng quát hóa của mô hình. Do đó, nhóm đề xuất một pipeline tiền xử lý dữ liệu nhiều bước nhằm làm sạch, chuẩn hóa và tăng tính nhất quán ngữ nghĩa của văn bản trước khi đưa vào mô hình học sâu. Tổng quan pipeline của hệ thống được minh họa trong Hình 1.

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 các thẻ HTML và thành phần định dạng không cần

thiết. 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.

2.1.2 Chuẩn hóa văn bản cơ bản

Sau khi trích xuất, văn bản 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. Bước này giúp đảm bảo tính nhất quán về mặt định dạng và giảm nhiễu không cần thiết trước khi đưa văn bản vào các bước xử lý sâu hơn.

2.1.3 Loại bỏ nhiễu và boilerplate

Các báo cáo thu hồi thực phẩm thường chứa những đoạn thông tin mang tính khuôn mẫu như tuyên bố pháp lý, thông tin liên hệ hoặc các đoạn mô tả lặp lại giữa nhiều báo cáo. Những thành phần này không đóng góp trực tiếp vào việc xác định loại mối nguy hoặc nhóm sản phẩm, do đó được loại bỏ bằng các biểu thức chính quy được thiết kế theo miền bài toán. Việc loại bỏ boilerplate giúp mô hình tập trung vào các đoạn nội dung giàu thông tin ngữ nghĩa liên quan đến an toàn thực phẩm.

2.1.4 Khử trùng lặp câu

Nhằm giảm thiểu sự dư thừa thông tin trong cùng một văn bản, nhóm áp dụng phương pháp khử trùng lặp ở mức câu, kết hợp giữa so khớp chính xác và so khớp mờ (*fuzzy matching*). Cụ thể, 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. Cách tiếp cận này cho phép loại bỏ hiệu quả các câu có nội dung gần giống nhau trong khi vẫn bảo toàn các thông tin quan trọng.

2.1.5 Chuẩn hóa thực thể chuyên ngành

Để giảm sự phân mảnh ngữ nghĩa do các biến thể cách viết khác nhau của cùng một thực thể, nhóm tiến hành chuẩn hóa một số thực thể quan trọng liên quan đến mối nguy và sản phẩm thực phẩm. Ví dụ, các biến thể như *"E. coli"* hoặc *"e coli"* được ánh xạ về dạng chuẩn *"Escherichia coli"*. Việc chuẩn hóa này giúp mô hình học được các biểu diễn ngữ nghĩa nhất quán hơn, đặc biệt trong bối cảnh dữ liệu có sự mất cân bằng lớp nghiêm trọng.

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. Văn bản hợp nhất này được sử dụng làm đầu vào cho các bước tiếp theo trong pipeline, bao gồm chia đoạn văn bản và huấn luyện mô hình phân loại đa nhiệm.

Nhờ pipeline tiền xử lý được thiết kế có hệ thống, dữ liệu đầu vào được làm sạch và chuẩn hóa một cách hiệu quả, góp phần cải thiện độ ổn định trong quá trình huấn luyện và nâng cao hiệu năng tổng thể của mô hình 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 xuất hiện với tần suất rất thấp, tương tự phân bố nhãn dài đuôi (long-tail) được ghi nhận trong SemEval-2025 Task 9 [7]. Để giảm ảnh hưởng của các lớp chiếm ưu thế và tập trung hơn vào các mẫu khó, nhóm sử dụng hàm mất mát Focal Loss [4] thay cho hàm cross-entropy truyền thống, một lựa chọn đã được chứng minh hiệu quả trong hệ thống của đội*Anastasia* [3] đạt hạng nhất.

Công thức 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. Khi $\gamma > 0$, các mẫu được phân loại đúng với độ tin cậy cao sẽ bị giảm trọng số, từ đó giúp mô hình tập trung học tốt hơn các mẫu khó và các lớp hiếm.

2.3.3 Hàm mất mát đa nhiệm

Trong bối cảnh học đa nhiệm, mỗi tác vụ có một hàm mất mát riêng biệt. Tổng hàm mất mát của mô hình được tính bằng cách kết hợp 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. Trong thực nghiệm này, hai trọng số λ_p và λ_h được đặt bằng nhau nhằm đảm bảo sự cân bằng giữa hai tác vụ và giữ kiến trúc mô hình đơn giản, ổn định.

Một hướng tiếp cận phổ biến trong học đa nhiệm là sử dụng trọng số dựa trên độ bất định của từng tác vụ, như được đề xuất bởi Cipolla et al. [1]. Công thức tổng quát của hàm mất mát đa nhiệm dựa trên độ bất định được viết 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 trở thành:

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

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 sử dụng trọng số cố định cho hai tác vụ nhằm ưu tiên tính đơn giản và khả năng tái lập kết quả.

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:

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

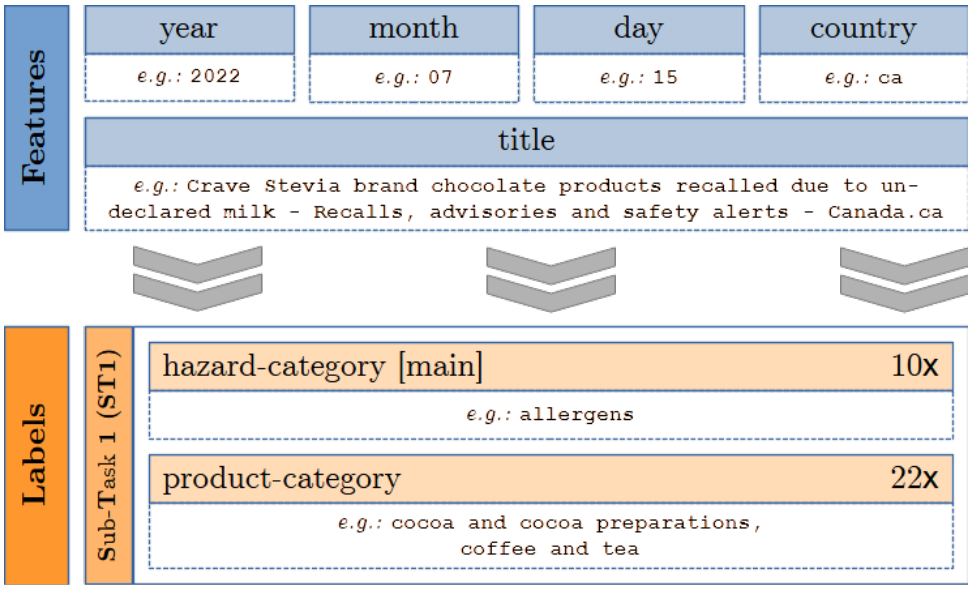
trong đó $w \in [0, 1]$ là trọng số của mô hình DeBERTa và được lựa chọn sao cho tối đa hóa macro-F1 trên tập xác thực.

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)

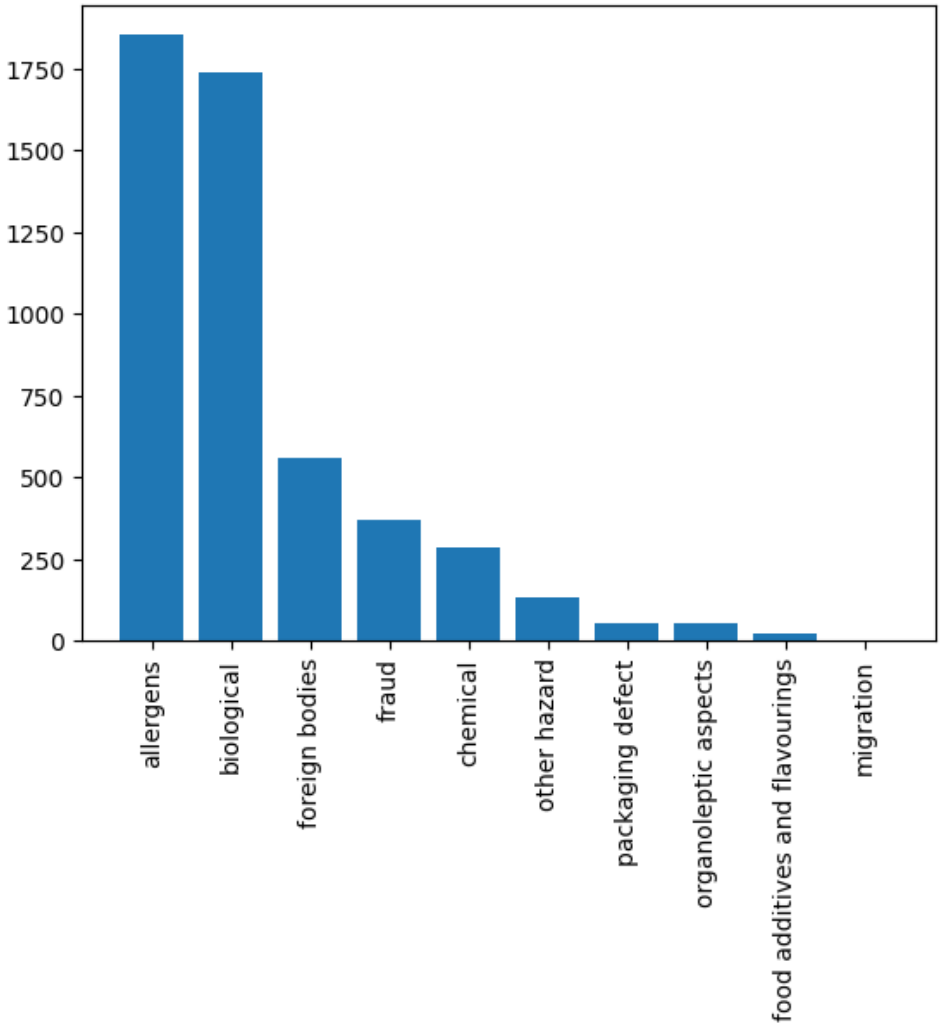
Bộ dữ liệu được sử dụng trong nghiên cứu này được cung cấp trong khuôn khổ SemEval-2025 Task 9 [7], bao gồm các báo cáo sự cố và cảnh báo an toàn thực phẩm được thu thập từ các nguồn web công khai. Mỗi mẫu dữ liệu bao gồm tiêu đề, nội dung văn bản và hai nhãn phân loại tương ứng với hai tác vụ của bài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

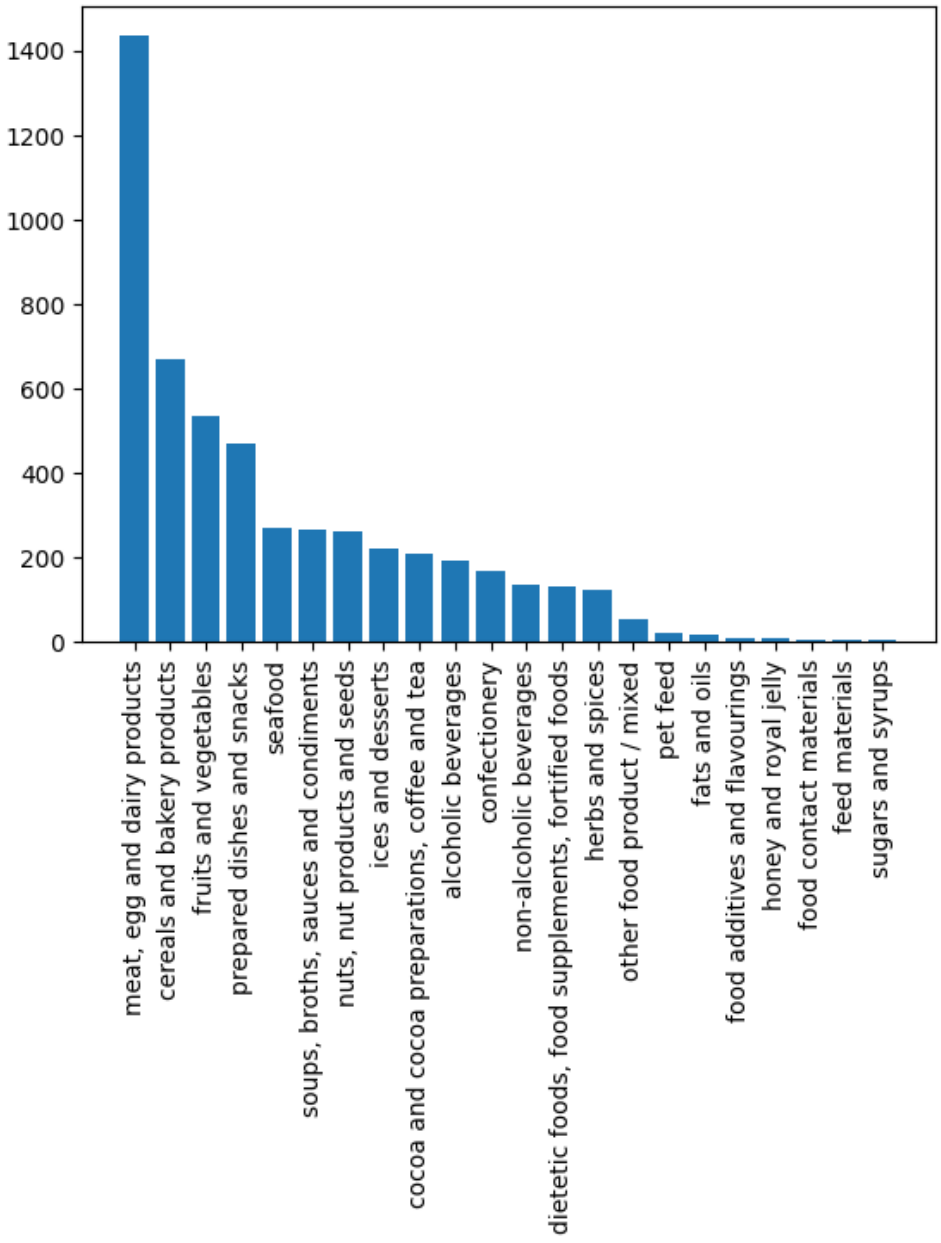


Hình 3: Phân bố nhãn mối nguy (hazard-category) trong tập dữ liệu gốc

- 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.

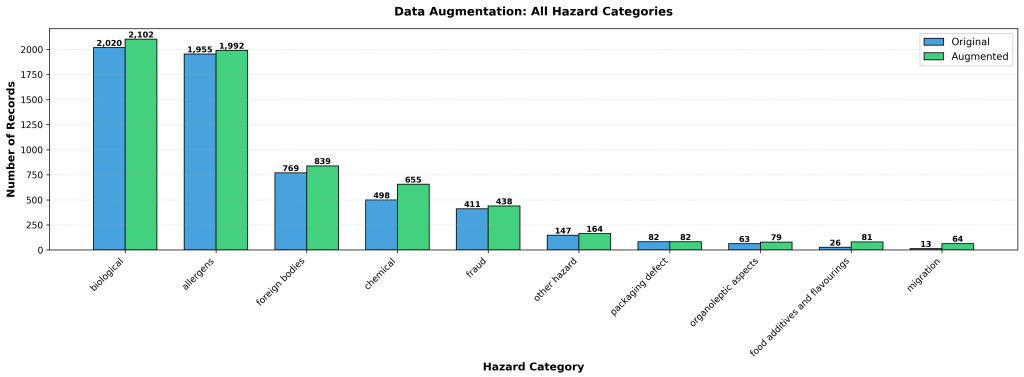
Chia tập dữ liệu. Tập dữ liệu gốc gồm tổng cộng 5984 mẫu và được chia thành hai phần theo tỉ lệ 80/20:

- **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)

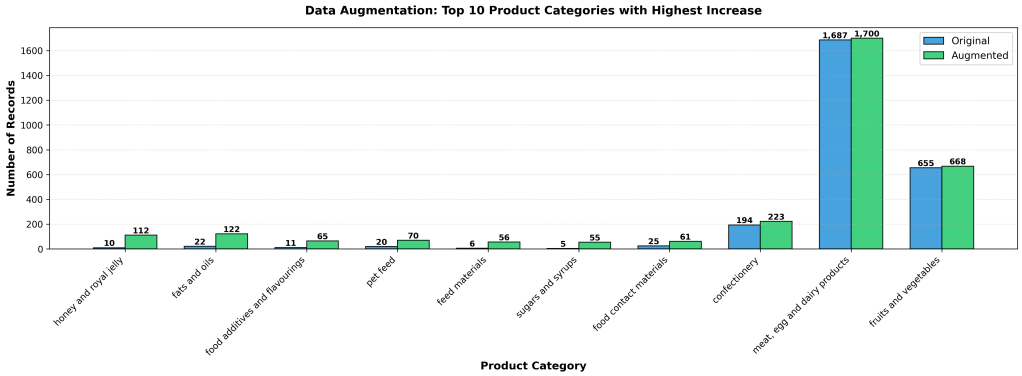
Để giải quyết đồng thời hai thách thức chính của bài toán là mất cân bằng lớp và chồng lấn ngữ nghĩa giữa các nhãn, nhóm xây dựng thêm một tập dữ liệu tăng cường bên cạnh tập dữ liệu gốc. Khác với các phương pháp tăng cường chỉ tập trung vào các lớp hiếm, chiến lược của nhóm được thiết kế theo hướng toàn diện, nhằm cải thiện cả khả năng nhận diện các nhãn ít xuất hiện lẫn độ ổn định dự đoán đối với các nhãn có tần suất cao nhưng dễ gây nhầm lẫn.



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

Như minh họa trong Hình 6, việc tăng cường dữ liệu được áp dụng trên toàn bộ các nhóm mối nguy, với mức độ bổ sung 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 hiện tượng thiếu mẫu nghiêm trọng và giúp mô hình có đủ tín hiệu để học được các biểu diễn ngữ nghĩa cơ bản cho các lớp này.

Bên cạnh đó, các lớp có tần suất cao như *biological* và *allergens* cũng được bổ sung thêm dữ liệu, dù đã chiếm ưu thế trong tập gốc. Lý do là các lớp này thường chứa nhiều mô tả phức tạp và đa dạng, với ranh giới ngữ nghĩa không hoàn toàn tách biệt so với các lớp khác như *chemical* hoặc *foreign bodies*. Trong thực tế, nhiều báo cáo an toàn thực phẩm đề cập đồng thời đến tác nhân sinh học, hóa chất hoặc các biểu hiện liên quan đến dị ứng, khiến mô hình dễ dự đoán sai nếu chỉ dựa vào tần suất xuất hiện. Việc tăng cường có chọn lọc các lớp lớn này giúp mô hình học được ranh giới quyết định rõ ràng và ổn định hơn.



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

Đối với tác vụ phân loại *product-category* (Hình 7), chiến lược tăng cường cũng được áp dụng theo nguyên tắc tương tự. Các nhóm sản phẩm có số lượng mẫu rất thấp, chẳng hạn như *honey and royal jelly*, *fats and oils* hay *food additives and flavourings*, được tăng cường mạnh nhằm giảm độ chênh lệch giữa các lớp. Đồng thời, một số nhóm sản phẩm phổ biến như *meat, egg and dairy products* và *fruits and vegetables* vẫn được bổ sung nhẹ để tăng tính đa dạng về ngữ cảnh và cách diễn đạt trong văn bản web.

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

Tổng thể, chiến lược tăng cường dữ liệu được thiết kế không chỉ nhằm cân bằng phân bố nhãn về mặt số lượng, mà còn hướng tới việc cải thiện chất lượng biểu diễn ngữ nghĩa và khả năng phân biệt giữa các lớp có nội dung chồng lấn. Điều này đặc biệt quan trọng trong bối cảnh dữ liệu có phân bố dài đuôi và được thu thập từ nhiều nguồn web với cách diễn đạt không đồng nhất, góp phần nâng cao hiệu năng và độ ổn định của hệ thống trong cả giai đoạn huấn luyện và suy luận.

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:

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

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

Bảng 1 trình bày kết quả xếp hạng Top-6 hệ thống có điểm số cao nhất trong cuộc thi. Hệ thống của nhóm chúng tôi đạt macro-F1 **0.8042** với cấu hình **TITLE + TEXT**, nằm trong nhóm dẫn đầu và vượt qua nhiều hệ thống sử dụng 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 1: Kết quả xếp hạng SemEval-2025 Task 9 (macro-F1).

Kết quả macro-F1 cao của hệ thống đến từ **4 yếu tố chính**. (1) **Tiền xử lý + Chia đoạn văn bản** giúp giảm nhiễu từ văn bản web và tận dụng tốt thông tin trong các báo cáo dài bằng cơ chế chia đoạn có chồng lấp, từ đó ổn định hóa đầu vào cho mô hình. (2) **Học đa nhiệm** [1] khai thác mối liên hệ giữa hazard-category và product-category, giúp mô hình học biểu diễn dùng chung tốt hơn và giảm nhầm lẫn chéo giữa hai nhãn. (3) **Focal Loss** [4] tăng trọng số cho các mẫu khó và lớp hiếm, đặc biệt hiệu quả với dữ liệu long-tail của SemEval-2025 Task 9 [7]. (4) **Dữ liệu tăng cường (augmentation)** được thiết kế theo hướng toàn diện: vừa bổ sung các lớp hiếm (ví dụ *migration*, *food additives and flavourings*), vừa tăng thêm cho các lớp dễ nhầm lẫn như *biological* và *chemical* (do chồng lấn ngữ nghĩa), cũng như các nhóm sản phẩm dễ gây nhiễu theo ngữ cảnh văn bản web. Việc tăng đa dạng diễn đạt giúp mô hình tổng quát hóa tốt hơn trên dữ liệu đa ngôn ngữ.

Về đóng góp của mô hình nền, **RoBERTa** [5] đạt macro-F1 **0.8027** khi huấn luyện đơn lẻ, cho thấy năng lực xử lý đa ngôn ngữ và tổng quát hóa mạnh giữa hai tác vụ. Trong khi đó, **DeBERTa** [2] đạt **0.7349** nhưng vẫn bổ trợ tích cực trong ensemble, giúp tăng nhẹ điểm tổng nhờ tính bổ sung về đặc trưng ngữ nghĩa. Khi tinh chỉnh trọng số hợp lý, hệ thống ensemble đạt kết quả tốt nhất **0.8042** với cấu hình **TITLE + TEXT**.

Một hạn chế đáng chú ý là khi **grid search trọng số ensemble trực tiếp trên tập *public test***, điểm số chỉ đạt **0.7842** (Top-10). Điều này cho thấy trọng số tối ưu trên public test có thể không ổn định do lệch phân bố giữa các split; vì vậy, lựa chọn trọng số dựa trên tập xác thực (hoặc điều chỉnh nhẹ theo hướng ổn định hóa) mang lại kết quả tổng quát tốt hơn trên đánh giá cuối cùng.

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.

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.