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

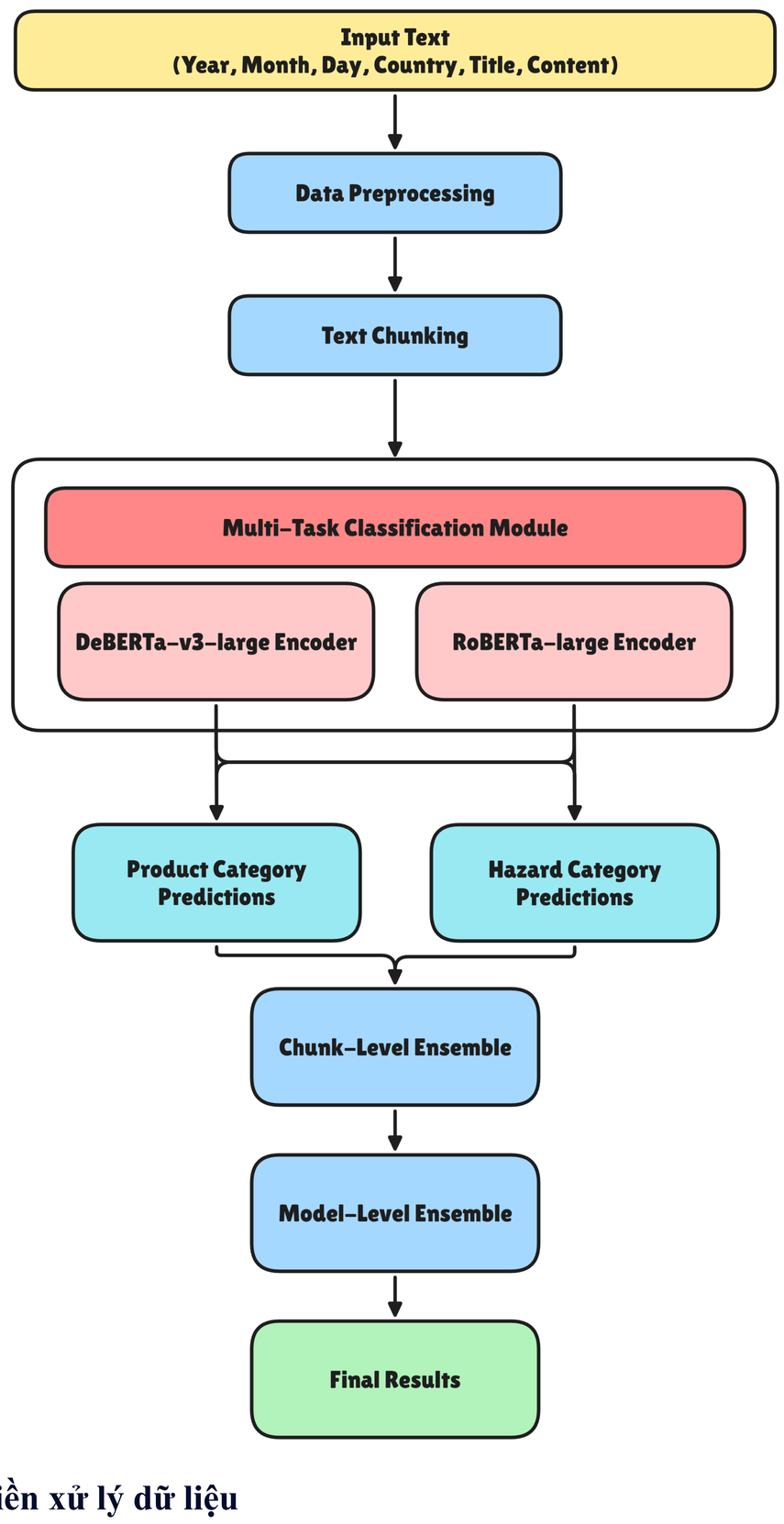
Ngô Minh Trí, Nguyễn Đình Khôi
Trường Đại học Công nghệ Thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh (ĐHQG-HCM)

GIỚI THIỆU

An toàn thực phẩm là vấn đề then chốt, ảnh hưởng trực tiếp đến sức khỏe cộng đồng và kinh tế xã hội. Các báo cáo sự cố an toàn thực phẩm hiện nay được công bố rộng rãi trên web, tạo ra nguồn dữ liệu lớn nhưng chứa nhiều nhiễu, văn bản dài và khó khai thác tự động. Việc phát hiện sớm mối nguy từ các báo cáo này có vai trò quan trọng trong các hệ thống giám sát an toàn thực phẩm hiện đại.

SemEval-2025 Task 9 tập trung vào bài toán phát hiện mối nguy thực phẩm từ văn bản web, yêu cầu đồng thời dự đoán loại mối nguy và nhóm sản phẩm trong bối cảnh dữ liệu dài và mất cân bằng lớp nghiêm trọng. Các hệ thống hàng đầu chủ yếu sử dụng mô hình Transformer lớn kết hợp Focal Loss, tăng cường dữ liệu và ensemble. Tuy nhiên, nhiều phương pháp có pipeline phức tạp và chi phí triển khai cao, đặt ra nhu cầu về một giải pháp tinh gọn nhưng hiệu quả.

PHƯƠNG PHÁP



Tiền xử lý dữ liệu

- Trích xuất văn bản thuần từ HTML, loại bỏ thẻ và định dạng không cần thiết.
- Chuẩn hóa văn bản và loại bỏ boilerplate
- Khử trùng lặp câu bằng so khớp chính xác và so khớp mờ.
- Chuẩn hóa các thực thể mối nguy và sản phẩm (ví dụ: E. coli → Escherichia coli).
- Hợp nhất tiêu đề và nội dung thành một chuỗi văn bản thống nhất.

Chia đoạn văn bản

- Mã hóa văn bản và chia thành các chunk tối đa 512 token.
- Áp dụng chồng lấp ngữ cảnh giữa các chunk để tránh mất thông tin.
- Loại bỏ chunk kém thông tin và gán nhãn theo văn bản gốc.

Học đa nhiệm

Hệ thống áp dụng học đa nhiệm để đồng thời dự đoán hai nhãn: (i) loại mối nguy thực phẩm và (ii) nhóm sản phẩm liên quan.

Một encoder Transformer chung được sử dụng để học biểu diễn ngữ nghĩa dùng chung, sau đó tách thành hai head phân loại độc lập cho hai tác vụ. Hàm mất mát đa nhiệm được tính bằng tổng có trọng số của hai tác vụ:

$$L = \lambda_p \mathcal{L}_{product} + \lambda_h \mathcal{L}_{hazard}$$

Hàm mất mát Focal

Để khắc phục mất cân bằng lớp, hệ thống sử dụng Focal Loss nhằm giảm ảnh hưởng của các mẫu dễ và tập trung vào các mẫu khó.

Công thức Focal Loss:

$$L_{focal} = -\alpha(1 - p_t)^{\gamma} log(p_t)$$

Trong đó:

p_t: xác suất dự đoán cho nhãn đúng

α: hệ số cân bằng lớp

γ: tham số điều chỉnh mức độ tập trung vào mẫu khó

Chiến lược Ensemble

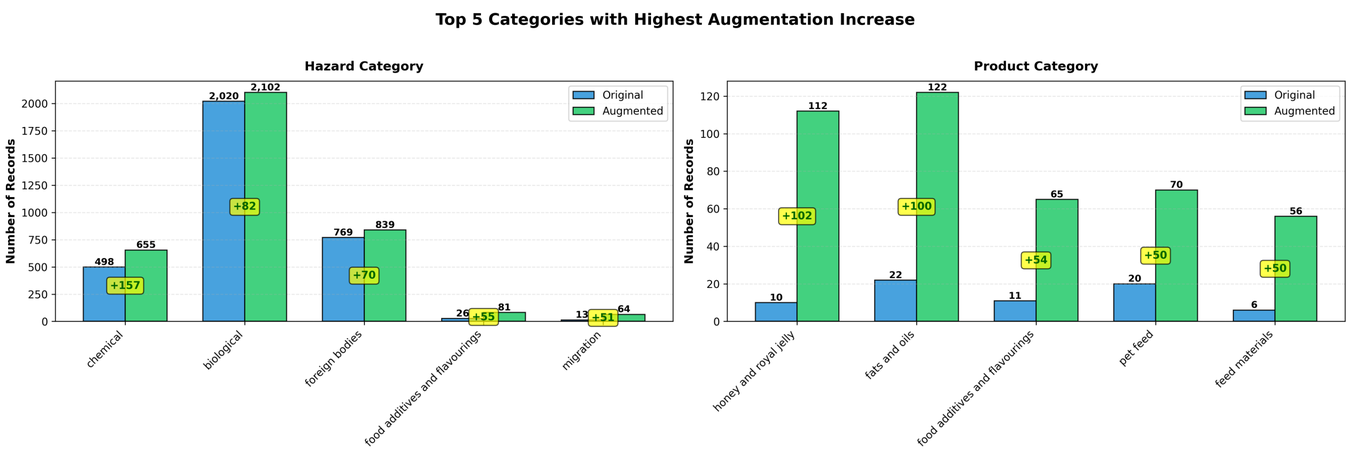
- **Ensemble mức đoạn văn bản:** mỗi báo cáo được chia thành nhiều đoạn và mô hình dự đoán độc lập cho từng đoạn. Xác suất ở mức tài liệu được tính bằng cách trung bình xác suất của các chunk.
- **Ensemble mức mô hình:** dự đoán ở mức tài liệu từ hai mô hình DeBERTa và RoBERTa được kết hợp bằng ensemble mềm với trọng số. Trọng số được lựa chọn thông qua grid search trên tập xác thực nhằm tối đa hóa macro-F1, sau đó được cố định cho giai đoạn suy luận.

BỘ DỮ LIỆU

Dữ liệu gốc

Bộ dữ liệu được cung cấp trong SemEval-2025 Task 9, gồm các báo cáo an toàn thực phẩm thu thập từ web, mỗi mẫu gồm văn bản và hai nhãn: hazard-category (10) và product-category (22). Dữ liệu đa ngôn ngữ, chủ yếu là tiếng Anh (≈82%) và tiếng Đức (≈15%). Tập dữ liệu được chia thành 4787 mẫu huấn luyện và 1197 mẫu xác thực.

Dữ liệu tăng cường



KẾT QUẢ

Top	Team	F1-macro	Feature
1	Anastasia	0.8223	Meta, Title, Text
2	MyMy	0.8112	Meta, Title, Text
3	**Ours (best weight)**	**0.8042**	Title, Text