Anastasia at SemEval-2025 Task 9: Subtask 1, Ensemble Learning with Data Augmentation and Focal Loss for Food Risk Classification.

Tung Thanh Le, Tri Minh Ngo, Trung Hieu Dang
VNUHCM - University of Information Technology
23521740@gm.uit.edu.vn, 23521640@gm.uit.edu.vn, 23521672@gm.uit.edu.vn

Abstract

This paper describes our system for SemEval-2025 Task 9, Subtask 1: The Food Hazard Detection Challenge, which focuses on predicting the type of food hazard and product from incident report titles collected from the web. We employed an ensemble learning approach, combining models trained with various data augmentation techniques to enhance performance on this text classification task. To address class imbalance, we fine-tuned the models using focal loss. Our system achieved Top 1 with a score of 0.8223, demonstrating the effectiveness of ensemble methods and data augmentation in improving classification accuracy for food safety risk assessment.

1 Introduction

Food safety is a growing global concern, with food-related hazards posing risks to public health and the economy. Identifying and categorizing these hazards from online incident reports is crucial for early detection and prevention. The SemEval-2025 Task 9, Subtask 1 (Randl et al., 2025) addresses this issue by evaluating AI models for classifying food hazards and associated products based on web-sourced report titles. This task presents challenges such as handling imbalanced data, ensuring model explainability, and improving classification accuracy to support automated food risk monitoring systems.

Our approach to this task involved employing an ensemble learning method that integrates multiple BERT (Devlin et al., 2019) models, including RoBERTa-large (Liu et al., 2019) and DeBERTa-v3-large (He et al., 2023), trained with various data augmentation strategies. To address the class imbalance commonly found in food hazard classification tasks, we fine-tuned these models using focal loss (Lin et al., 2020). This approach not only helped improve performance but also ensured our system's ability to generalize well across diverse hazard categories. By leveraging both lightweight and intensive data augmentation techniques, we crafted a solution that maintained high accuracy while prioritizing transparency, which is essential in explainable AI.

You can access our system's code through the following GitHub repository: Semeval-Task9-The-Food-Hazard-Detection-Challenge-2025.

2 Related Work

Food safety risk classification is crucial for protecting public health and ensuring regulatory compliance. Traditional approaches relied on rule-based systems and expert knowledge, but advances in machine learning and natural language processing have significantly improved classification accuracy and scalability.

(Nogales et al., 2022) introduced a deep learning framework that incorporates categorical embeddings to predict food safety risks using European Union data. Their model demonstrated high accuracy in predicting product categories, hazard types, and appropriate actions, laying the foundation for large-scale food safety classification using neural architectures.

(Randl et al., 2024) proposed CICLe, a conformal in-context learning approach for large-scale multi-class food risk classification. By integrating conformal prediction, CICLe provides reliable uncertainty estimates, enhancing decision-making in high-risk scenarios. Additionally, they introduced a dataset of 7,546 labeled food recall announcements, serving as a benchmark for future studies.

Recent advances in AI-driven text classification have demonstrated significant potential in regulatory and news analysis. (Hassani et al., 2025) conducted an empirical study utilizing large language models (LLMs) to classify requirements-related provisions in food safety regulations. In a related effort, (Xiong et al., 2023) proposed a hierarchical Transformer-based model for food safety news

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 141–147
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

classification, addressing the challenge of long-text processing.

Morever, (Maharana et al., 2019) used BERT to detect unsafe food reports in Amazon reviews, linking them to FDA recalls (2012–2014). Their model achieved an F1 score of 0.74 and identified potential underreporting of food safety issues. Similarly, (Wang et al., 2022) reviewed machine learning applications in food safety, highlighting improvements in monitoring, detection, and prediction.

These studies collectively demonstrate the progress in food risk classification. Building upon this foundation, our work explores strategies to enhance both classification accuracy and explainability, with a focus on real-world applicability.

3 System Description

We performed Exploratory Data Analysis (Rao et al., 2021) and discovered that the data suffers from severe class imbalance. To address this issue, we augmented the data by creating multiple different datasets and chunking them into various sizes. We trained different variants of BERT models using Focal Loss to mitigate the impact of the imbalance in the classes.

To further improve performance, we applied an ensemble method using soft voting on the probabilities of each label, combining the results from multiple models to optimize accuracy and minimize classification errors.

3.1 System Overview

Our system is structured as shown in Figure 1 and consists of the following stages: a) **Data:** Preprocessing, augmentation to create two additional datasets, and chunking the data into different sizes; b) **Training:** Training models using both multitask learning (Zhang and Yang, 2017) and single-task learning approaches; c) **Ensemble:** Combining model predictions using soft voting (Manconi et al., 2022) based on the probabilities of each label.

3.2 Training models

3.2.1 Focal Loss

Focal Loss is used to minimize the effect of easily classified examples and emphasize harder-to-classify ones. We apply Focal Loss for both multitask and single-task scenarios.

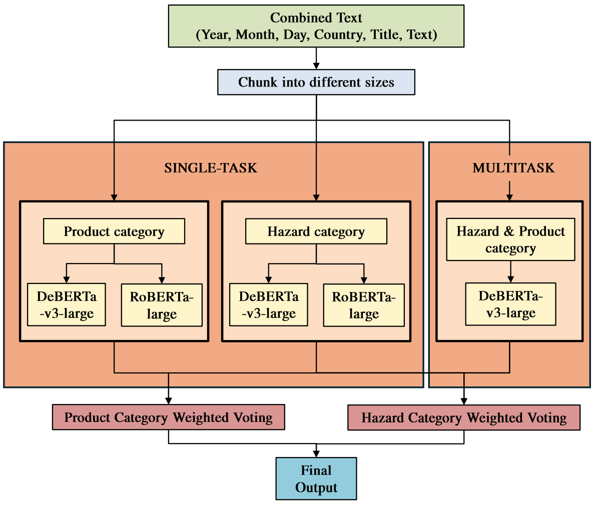
Combined Text
(Year, Month, Day, Country, Title, Text)
Chunk into different sizes
SINGLE-TASK
MULTITASK
Product category
Hazard category
Hazard & Product category
DeBERTa -v3-large
RoBERTa-large
DeBERTa -v3-large
RoBERTa-large
DeBERTa-v3-large
Product Category Weighted Voting
Hazard Category Weighted Voting
Final Output

Figure 1: The weight voting ensemble architecture based on the combination of fine-tuning multilingual contextual language models.

$$\mathcal{L}_{FL}(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t)$$

In this formula, p_t is the probability of the true class based on the model's prediction, α_t is a balancing factor for each class, used to adjust the impact between classes, especially when dealing with imbalanced datasets, and γ is the focusing parameter that helps adjust the focus on hard examples. When $\gamma = 0$, Focal Loss becomes the standard Cross-Entropy loss. As γ increases, the impact of easy examples decreases, and the model focuses more on the hard-to-classify examples.

3.2.2 Multitask Learning

Multitask learning is a type of machine learning approach in which multiple related tasks are learned simultaneously, sharing representations to improve performance on each task. In this study, we leverage multitask learning to train a model that simultaneously predicts two types of labels: *product category* and *hazard category*. By training the model on both tasks at once, the shared knowledge between the tasks can enhance the overall model's generalization.

To implement this, we use a transformer-based architecture (Vaswani et al., 2017) as shown in Figure 2, specifically the DeBERTa-v3-large model, which is fine-tuned on both classification tasks. The model consists of a pre-trained BERT-based encoder that captures the contextualized representation of text and two separate classifiers: one for the product category and another for the hazard category.

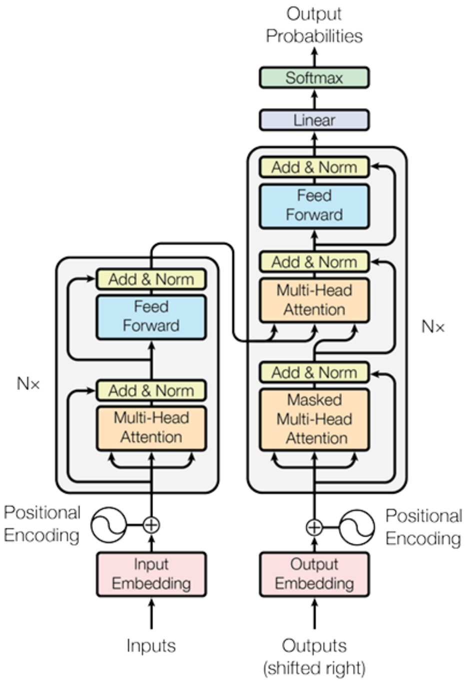

Figure 2: The architecture of the transformer model used in this work.

The multitask model is optimized using a custom loss function called *Focal Loss*, which helps to address class imbalance in the training data. Focal Loss is designed to reduce the impact of easy-to-classify examples and focus more on hard-to-classify instances, thereby improving model performance on imbalanced datasets. Specifically, we use Focal Loss for both product and hazard classification tasks. The model computes the final loss as the weighted average of the individual losses for each task:

$$\text{Loss} = 0.5 \times \text{Product Loss} + 0.5 \times \text{Hazard Loss}$$

The individual task losses are computed using Focal Loss, where the loss for each task is calculated as:

$$\text{Focal Loss} = \alpha(1 - p_t)^{\gamma} \times \text{CrossEntropyLoss}$$

We apply data balancing techniques, such as oversampling and undersampling (Yang et al., 2024), to address the class distribution issues in both tasks. Oversampling is applied to the least frequent categories, while undersampling is applied to the most frequent ones, leading to a more balanced distribution of the classes on the original dataset.

We also calculate class weights (Xu et al., 2020) based on the frequency of each class in the dataset. These weights are used in the model's loss function to give more importance to minority classes, further improving the model's ability to classify rare categories effectively.

3.2.3 Single-task Learning

In this approach, we train two separate models, DeBERTa-v3-large (He et al., 2023) and RoBERTa-large (Liu et al., 2019), each focusing on a specific classification task: *product category* and *hazard category*. Each model is fine-tuned independently for its respective task without any shared learning between them.

To address class imbalance within the dataset, we employ data augmentation instead of traditional oversampling or undersampling techniques (Gao, 2020). For each task, we first augment a dataset to ensure that less frequent labels are represented sufficiently in both the training and validation splits. This step ensures that no label is underrepresented in the validation set. then, we perform additional augmentation to increase the overall volume of data while maintaining the original distribution of classes. We prioritize preserving the natural class distribution, as artificially balancing the data could lead to the loss of important patterns, which would degrade the performance of the model.

Focal loss is also applied for each of the tasks to further help address the class imbalance. For evaluation, we utilize the macro F1 (Opitz and Burst, 2021) score for each label. The macro F1 score calculates the F1 score for each class individually and then averages them, ensuring that each label is treated equally regardless of frequency.

Through this approach, we leverage the benefits of data augmentation to ensure balanced representation across tasks, while focusing on preserving the class distribution to optimize model performance.

3.3 Ensemble

In our model, the *Ensemble* method is implemented using the **soft voting** technique, where the probabilities from multiple models are aggregated as follows:

$$P(y=c)=\sum_{i=1}^{N} w_i P_i(y=c) \qquad (1)$$

In equation (1), $P_i(y=c)$ represents the probability of class c predicted by model i, while w_i is the weight assigned to that model. The weights w_i are optimized using grid search on the validation set during the Conception Phase.

The weight optimization process follows these steps:

- Define a grid of possible weight values w_i, ensuring that $\sum w_i = 1$.
- Evaluate each set of weights using the validation set and compute the ensemble model's performance.
- Select the optimal set of weights based on evaluation metrics.

The results show that the *Ensemble* model significantly improves performance compared to individual models, as it leverages weighted aggregation instead of relying on a single model's prediction.

4 Experiment

4.1 Datasets

We used three different datasets for this experiment: the original dataset, a lightly augmented version, and a heavily augmented version.

4.1.1 Data Augmentation

Light Augmentation: In this phase, we focused specifically on the most underrepresented classes in the dataset. We generated additional synthetic samples for the following categories: 9 product categories with the lowest representation and 4 hazard categories with the lowest representation. This targeted approach aimed to ensure that the model receives more examples from these underrepresented classes, which helps to mitigate the bias toward the majority classes and improve overall model performance.

Heavy Augmentation: In the heavy augmentation phase, we applied extensive modifications to the dataset, generating a larger volume of synthetic samples separately for hazard categories and product categories. This approach enhanced the representation of minority classes, improving the model's ability to generalize. Additionally, the dataset was split into two separate subsets: one for hazard classification and another for product classification, as this dataset is used for single-task learning.

All three datasets were split using an 80:20 ratio for training and validation. The dataset statistics after augmentation are summarized in Table 1.

Dataset	Train Samples	Validation Samples
Original	4787	1197
Light Augmentation	5187	1297
Heavy Augmentation - Hazard	8224	2057
Heavy Augmentation - Product	13417	3355

Table 1: Dataset statistics after augmentation

4.1.2 Preprocessing

The data preprocessing follows a systematic approach applied to all datasets. Special characters (excluding punctuation) are removed, newlines are replaced with spaces, and consecutive spaces are consolidated. Punctuation is standardized for readability.

After cleaning, the text is segmented into sentence-based chunks of 512, 768, 1024, and 1280 tokens, approximately 400, 650, 900, and 1150 words, to preserve contextual coherence while adhering to model constraints.

For the heavily augmented dataset, additional preprocessing steps are applied. Non-English text is translated into English to ensure consistency across all the data, allowing the model to process it uniformly. Additionally, HTML tags, which might have been included in the original dataset, are removed using BeautifulSoup (Pant et al., 2024), ensuring that only the relevant textual content is retained and improving the quality of the data used for model training.

4.2 Experiment Environment

We used RoBERTa-large and DeBERTa-v3-large models for classification, trained on NVIDIA P100 and T4 GPUs via the Kaggle platform. RoBERTa-large was trained for 8 hours, while DeBERTa-v3-large took 12 hours per model. The training used a learning rate of 2×10^{-5}, batch sizes of 4 for training and 2 for evaluation, 10 epochs, weight decay of 0.01, logging every 10 steps, and a warm-up ratio of 0.1.

4.3 Results

Table 2 presents a comparison of different model configurations across various methods, datasets, model types, token sizes, and weight voting scores.

METHOD	DATA	MODEL NAME	TOKEN CHUNK	HAZARD SCORE	PRODUCT SCORE	SCORE	WEIGHT HAZARD	WEIGHT PRODUCT
Single-Task	Light	DeBERTa-v3-large	512	0.7861	0.7486	0.7673	0.3500	0.1842
			768	0.7990	0.7640	0.7815	0.3500	0.2632
			1024	0.7789	0.7960	0.7874	0.0000	0.0000
			1280	0.7819	0.7875	0.7847	0.2000	0.0000
		RoBERTa-large	512	0.7680	0.7515	0.7598	0.0500	0.1842
			768	0.7691	0.8292	0.7991	0.0000	0.0000
			1024	0.7719	0.7522	0.7621	0.0000	0.0000
			1280	0.7839	0.7869	0.7854	0.0000	0.0000
	Heavy	DeBERTa-v3-large	512	0.7613	0.7945	0.7779	0.0500	0.2632
			768	0.7712	0.7984	0.7848	0.0000	0.0000
			1024	0.7599	0.7490	0.7544	0.0000	0.0000
		RoBERTa-large	512	0.7775	0.7837	0.7806	0.0000	0.0000
MultiTask	Original	DeBERTa-v3-large	512	0.7291	0.7963	0.7627	0.0000	0.1053

Table 2: Result comparison based on method, data, model type, token size, and weight voting

For single-task learning on the Light dataset, DeBERTa-v3-large with 768 tokens achieves the highest overall score of 0.7815, while RoBERTa-large with 768 tokens achieves a slightly higher product score of 0.8292. On the Heavy dataset, DeBERTa-v3-large with 512 tokens achieves the best overall score of 0.7779.

In multi-task learning with the Original dataset, DeBERTa-v3-large with 512 tokens performs with an overall score of 0.7627. Weight voting scores indicate the influence of hazard and product classification, where certain models receive higher weights in hazard or product recognition, such as DeBERTa-v3-large (512 tokens, Light dataset) with a weight hazard score of 0.35.

By using grid search, we optimized the weight voting scheme to obtain the final model combination. The optimized weight allocation, as shown in Table 2, resulted in a final overall score of 0.8223.

5 Conclusion

In summary, we presented an ensemble-based approach for the food hazard detection task in SemEval 2025 Task 9, Subtask 1. By combining DeBERTa-v3-large and RoBERTa-large models with data augmentation and focal loss, we achieved a top performance with a macro F1 score of 0.8223. Our results highlight the importance of model ensembling, data augmentation, and addressing class imbalance for multi-class classification tasks.

Future work will focus on improving the model's ability to distinguish between similar hazard types by incorporating advanced techniques such as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), which combines information retrieval and generation to enhance context and reduce ambiguity. Additionally, we plan to explore few-shot learning and GAN-based data augmentation (Wang and Wan, 2020) to generate more realistic data, addressing class imbalance and boosting performance with limited labeled data. These methods are expected to improve model generalization and enhance its ability to handle complex hazard detection tasks.

6 Limitations

Although our system achieved strong results, several limitations remain. First and most notably, we submitted multiple test runs, violating SemEval's single-submission rule. This may have led to an unfair advantage, and we take full responsibility. We are committed to strictly following submission policies in future shared tasks to ensure fairness. Second, while data augmentation helped address class imbalance, we did not apply rigorous quality control to synthetic samples, which risks propagating label noise—especially in underrepresented classes. Third, our work lacks inference latency metrics and comparisons with modern large language models (e.g., GPT-4), limiting insight into real-world deployment and performance against current state-of-the-art systems. Future work should incorporate human-validated augmentation, efficiency benchmarks, and LLM-based baselines.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jie Gao. 2020. Data augmentation in solving data imbalance problems. Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS).

Shabnam Hassani, Mehrdad Sabetzadeh, and Daniel Amyot. 2025. An empirical study on llm-based classification of requirements-related provisions in food-safety regulations. *Empirical Software Engineering*, 30.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Adyasha Maharana, Kunlin Cai, Joseph Hellerstein, Yulin Hswen, Michael Munsell, Valentina Staneva, Miki Verma, Cynthia Vint, Derry Wijaya, and Elaine O Nsoesie. 2019. Detecting reports of unsafe foods in consumer product reviews. *JAMIA Open*, 2(3):330–338.

Andrea Manconi, Giuliano Armano, Matteo Gnocchi, and Luciano Milanesi. 2022. A soft-voting ensemble classifier for detecting patients affected by covid-19. *Applied Sciences*, 12(15).

Alberto Nogales, Rodrigo Díaz-Morón, and Álvaro J. García-Tejedor. 2022. A comparison of neural and non-neural machine learning models for food safety risk prediction with european union rasff data. *Food Control*, 134:108697.

Juri Opitz and Sebastian Burst. 2021. Macro f1 and macro f1.

Sakshi Pant, Er. Narinder Yadav, Milan, Monnie Sharma, Yash Bedi, and Anshuman Raturi. 2024. Web scraping using beautiful soup. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, volume 1, pages 1–6.

Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2024. CICLe: Conformal in-context learning for largescale multi-class food risk classification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7695–7715, Bangkok, Thailand. Association for Computational Linguistics.

Korbinian Randl, John Pavlopoulos, Aron Henriksson, Tony Lindgren, and Juli Bakagianni. 2025. SemEval-2025 task 9: The food hazard detection challenge. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

A Suresh Rao, B. Vishnu Vardhan, and Hafeezuddin Shaik. 2021. Role of exploratory data analysis in data science. In *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, pages 1457–1461.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ke Wang and Xiaojun Wan. 2020. Adversarial text generation via sequence contrast discrimination. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 47–53, Online. Association for Computational Linguistics.

Xinxin Wang, Yamine Bouzembrak, AGJM Oude Lansink, and H. J. van der Fels-Klerx. 2022. Application of machine learning to the monitoring and prediction of food safety: A review. *Comprehensive Reviews in Food Science and Food Safety*, 21(1):416–434.

Shufeng Xiong, Wenjie Tian, Vishwash Batra, Xiaobo Fan, Lei Xi, Hebing Liu, and Liangliang Liu. 2023. Food safety news events classification via a hierarchical transformer model. *Heliyon*, 9(12):e17806.

Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar. 2020. Class-weighted classification: Tradeoffs and robust approaches. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10544–10554. PMLR.

Cynthia Yang, Egill A. Fridgeirsson, Jan A. Kors, Jenna M. Reps, and Peter R. Rijnbeek. 2024. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *Journal of Big Data*, 11(1):7.

Yu Zhang and Qiang Yang. 2017. An overview of multitask learning. *National Science Review*, 5(1):30–43.