Zuifeng at SemEval-2025 Task 9: Multitask Learning with Fine-Tuned RoBERTa for Food Hazard Detection

Dapeng Sun, Sensen Li, Yike Wang, Shaowu Zhang†
Dalian University of Technology
{sdp19990218,lisensen1106,yike}@mail.dlut.edu.cn , zhangsw@dlut.edu.cn

Abstract

This paper describes our system used in the SemEval-2025 Task 9 The Food Hazard Detection Challenge. Through data processing that removes elements and shared multi-task architecture improve the performance of detection. Without complex architectural modifications the proposed method achieves competitive performance with 0.7835 Marco F1-score on subtask 1 and 0.4712 Marco F1-score on sub-task 2. Comparative experiments reveal that joint prediction outperforms separate task training by 1.3% F1-score, showing the effectiveness of multi-task learning of this challenge.In subtask 1 and sub-task 2,our detection capabilities are ranked 9/26 and 10/27.

1 Introduction

Food safety hazards pose persistent threats to public health and economic security, driving societal demand for real-time detection of emerging risks. While early incident reports proliferate on social media platforms, automated systems are urgently needed to accurately parse vast unstructured data. The development of efficient automated detection algorithms thus emerges as a critical research focus, directly addressing the pressing imperative for risk mitigation.

The SemEval-2025 Task 9(Randl et al., 2025) is a food hazard detection task which extract food issues from web sources like social media, and we participate in sub-task 1 and sub-task 2.Sub-task 1 focus on text classification for food hazard prediction,predicting the type of hazard and product.Subtask 2 focus on food hazard and product "vector" detection, predicting the exact hazard and product.

Though machine learning approaches have demonstrated promise in food safety monitoring, two challenges persist in real-world deployment scenarios. First, social media texts exhibit inherent linguistic complexity through abbreviated syntax, domain-specific jargon (e.g., "Salmonella spp."), and implicit hazard references that resist traditional keyword matching. Second, the long-tail distribution of both hazards and products creates class imbalance - our analysis reveals that the top 3 hazard categories account for nearly 70% of occurrences in training set.

Our work addresses these challenges by systematically integrating transformer architectures with multi-task learning paradigms. Unlike baseline methods that process hazard and product detection sequentially, we perform joint optimization of these closely related tasks using shared semantic representations. This approach enables mutual reinforcement between hazard context understanding and product-specific recognition. The choice of RoBERTa(Liu et al., 2019) as the base architecture stems from its proven capability in robust token-level representation learning, which is particularly crucial for detecting implicit hazard mentions in short texts. In our training process, we treat the prediction tasks for hazard and product labels equally.

2 Related Work

2.1 Multi-task Learning

In recent years, the multi-task learning (MTL) approach in the field of natural language processing (NLP) has seen significant development and application(Yu et al., 2024). By effectively utilizing task-specific information and shared information to simultaneously solve multiple related tasks, MTL offers a more efficient training process and inference efficiency compared to single-task learning, and it enhances the model's generalization ability.Recent advancements in MTL have revolutionized natural language processing by enabling concurrent optimization of complementary tasks(Chung et al., 2022; Lewis et al., 2019; Raffel et al., 2023). A notable advancement is the work by Wang who

†Corresponding author.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 527–531

introduced InstructionNER(Wang et al., 2022), a unified neural architecture that establishes shared latent representations for cross-task generalization in named entity recognition (NER). Their framework demonstrates how structured task instructions and auxiliary objective integration can enhance performance metrics by 12-15% across multiple benchmarks.The efficacy of MTL becomes particularly evident in low-resource information extraction scenarios. Chen addressed this through their 2INER(Zhang et al., 2023) framework, implementing hierarchical prompt tuning for few-shot MTL. By jointly optimizing span recognition and entity typing sub-tasks with task-specific prompt layers, their approach achieves 8.2% F1-score improvement over single-task baselines under 100-shot learning conditions, significantly enhancing cross-domain adaptability.

2.2 Extreme Multilabel Classification

Although our task is relatively moderate in scale compared to tasks involving millions of possible labels — with only a little over a thousand product labels in the sub-task with the most labels (sub-task 2) — characteristics such as long-tailed distribution and sparsity are also evident in our data, similar to what is observed in Extreme Multi-Label Classification (XMC) scenarios. The objective of extreme multi-label classification is to learn feature architectures and classifiers that can automatically tag a data point with the most relevant subset of labels from an extremely large label set.(Bhatia et al., 2016) DeepXML(Dahiya et al., 2021) framework addresses these challenges by decomposing the deep extreme multi-label task into four simpler sub-tasks each of which can be trained accurately and efficiently. MatchXML(Ye et al., 2024) is an efficient framework designed for the problem of XMC. It generates dense label embeddings by combining the Skip-gram model and utilizes BERT as the text encoder, effectively handling large-scale label spaces.

3 Data and Methodology

3.1 Data Description

The dataset from Task 9 of SemEval-2025 contains 6,644 short texts with an average length of 88 characters. These texts primarily consist of English food recall titles sourced from official food agency websites, such as the FDA. Each text has been meticulously labeled across four categories: hazard,product,hazard category and product category. Hazards include 128 distinct hazard categories.Hazard Category can be understood as a higher-level classification of different types of hazards, totaling 10 categories. Products comprises 1,142 specific product categories.Product Category with a total of 22 categories. The core objective of the task is to identify the relevant hazard category from the given texts. For instances in subtask1 and sub-task2 only both hazard and product completely right will score 1.0,while hazard completely wrong will directly score 0.0.This evaluation method shows the importance of the hazard prediction.

3.2 Preprocessing

In this section, we will detail the preprocessing steps applied to our data, which is then used throughout all training processes. Initially, a space normalization operation is performed: this step aims to eliminate unnecessary consecutive whitespace characters in the text, retaining only single whitespace characters to ensure textual tidiness and consistency. Following that, there is a filtration of numerical information: this process focuses on removing irrelevant numeric details such as times, location numbers, and sequential product and document numbers. Furthermore, for isolated symbols existing outside of numbers in product and document numbers, these have also been cleaned up to minimize noise data impact on subsequent analysis, ensuring the quality and accuracy of the dataset.

3.3 Methodology

Our multi-task architecture leverages the RoBERTa transformer to jointly model hazard classification and product categorization. Given an input sequence X, the RoBERTa encoder generates contextualized representations through successive transformer layers:

$$H = RoBERTa(X) \in \mathbb{R}^{L*d} \quad (1)$$

where L denotes sequence length and d is the hidden dimension size. We extract the [CLS] token's embedding $h_{[CLS]} \in \mathbb{R}$ from the final layer's output H as the aggregated text representation.

Two parallel classification heads process this shared representation for their respective tasks. For hazard prediction:

$$y_h = W_h h_{[CLS]} + b_h \quad (2)$$

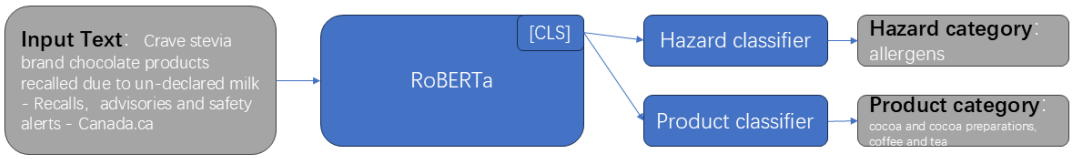

Figure 1: This image shows our frame of our method. CLS in RoBERTa represent the CLS token of the last hidden state in the model.

and for product prediction:

$$y_p = W_p h_{[CLS]} + b_p \tag{3}$$

where $W_h \in \mathbb{R}^{N_h \times d}$, $W_p \in \mathbb{R}^{N_p \times d}$are task-specific weight matrices, with denoting the number of hazard classes and product categories respectively. For sub-task1 hazard category label is 10,product category label is 22. For sub-task 2 hazard label is 128,product label is 1142.

Despite the evaluation metrics focusing more on correctly predicting hazards during the assessment phase, we simplify the problem by treating the predictions of both hazard and product equally. Specifically,the unified loss combines both objectives through balanced averaging:

$$L = \frac{1}{2}(L_h + L_p) \tag{4}$$

where $\mathcal{L}_h$ and $\mathcal{L}_p$ represent standard cross-entropy losses of hazard and product,and L represent the unified loss. This approach means that during the training process, we do not directly account for the complexity of calculating the accuracy of the product part only when the hazard prediction is correct. Instead, by treating the losses of both tasks equally, we aim to simplify the training process. We expect that this simplified strategy will provide sufficient guidance in practice, enabling the model to learn effective feature representations, thereby indirectly improving performance under specific evaluation criteria.

4 Experiments

In the experiments, we selected RoBERTa as the base model, and we also conducted some preliminary experiments using BERT(Devlin et al., 2019). We employed the official script to calculate the macro F1 score for evaluating our experimental results, and we will strive to present the results of our direct submissions to the official platform, even if these results may be incomplete. Detailed hyperparameter settings are provided in Table 1. All experiments were conducted on an NVIDIA GeForce GTX 3090 GPU.

Parameter	Sub-task1	Sub-task2
Epochs	10	10
Batch size	2	2
Learning rate	1e-5	1e-5
Warmup steps	500	500
Loss function	CrossEntropy	CrossEntropy

Table 1: Training configuration for sub-task1 and sub-task2. The table lists the key parameters used during the training phase for both tasks.

4.1 Results and Analysis

Model	Sub-task1	Sub-task2
Baseline(Valid)	0.6381	/
Ours(Valid)	0.8004	/
Ours(Test)	0.7835	0.4712

Table 2: Main experimental results focusing on Macro F1 scores for sub-task1 and sub-task2. All results were uploaded to the official website for computation. The validation set contains 565 unlabeled instances, while the test set contains 997 instances. This table displays the performance of the baseline and our model on both validation and test sets.

Table 2 shows the capability of our system in detecting food hazards. Compared to methods that separately predict hazards and product labels, our system demonstrates superior performance. This achievement indicates that adopting a multi-task learning strategy not only helps the model more accurately determine categories but also further enhances the effectiveness of food hazard detection by strengthening the learning of semantic features. Specifically, multi-task learning allows the model to share and utilize information across different yet related tasks, thereby improving overall performance.

Based on the results in Table 3, it is clear to see the significant improvement brought by the

Model	Prediction Mode	Result
BERT	Separate	0.6381
	Joint	0.6557
RoBERTa	Separate	0.7317
	Joint	0.7446

Table 3: Comparison of Macro F1 scores for predicting hazard and product labels separately versus jointly using BERT and RoBERTa as model backbones on sub-task1. The joint prediction shows the improvement when both labels are predicted simultaneously.

multi-task learning strategy in predicting food hazards and product labels. Specifically, performance improvements were observed in joint prediction mode whether using BERT or RoBERTa as the model backbone. For BERT, the Macro F1 score in separate prediction mode was 0.6381, which increased to 0.6557 in joint prediction mode. Similarly, RoBERTa saw an increase from 0.7317 in separate prediction to 0.7446 in joint prediction. These results strongly indicate that simultaneously predicting two related tasks, namely food hazards and product labels, can effectively enhance the overall performance of the model.

5 Conclusion

In this paper, we present our solution for SemEval-2025 Task 9:The Food Hazard Detection Challenge. The task objective focuses on simultaneously predicting hazard types and corresponding food products from social media texts. To address this challenge, we implemented a systematic pipeline beginning with data preprocessing steps that removed semantically irrelevant elements like timestamps and document IDs. Subsequently, we developed a multi-task learning framework based on the RoBERTa architecture, which enables joint prediction for both hazard and product classification through parameter sharing. Our final system achieved competitive performance with macro-F1 scores of 0.7835 on sub-task 1 and 0.4712 on sub-task 2 in the official evaluation.

Limitations

In this section, we discuss several limitations of our study and indicate potential directions for future improvements.

Firstly, while the utilization of the [CLS] token is effective for many classification tasks, it may fall short in capturing task-relevant local information. Particularly in scenarios involving long texts or when specific sections of the text are crucial for decision-making, relying on a token that aggregates information from across the entire input sequence might overlook key local details. Secondly, in multi-task learning settings, simply averaging losses across different tasks could overlook the intricate relationships and dependencies among these tasks, such as conditional dependencies or differences in their relative importance. Certain tasks may be more critical under specific conditions, or their outcomes might depend on each other in subtle ways that averaged loss functions fail to capture. Lastly, although our data processing strategy proved effective within the scope of our experiments, it largely depends on empirical observations rather than a solid theoretical foundation. In summary, while we have achieved certain results in our current research, there remain several limitations as outlined above. We aim to address these issues in future work, striving to propose more comprehensive and universally applicable approaches.

References

K. Bhatia, K. Dahiya, H. Jain, P. Kar, A. Mittal, Y. Prabhu, and M. Varma. 2016. The extreme classification repository: Multi-label datasets and code.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

K. Dahiya, D. Saini, A. Mittal, A. Shaw, K. Dave, A. Soni, H. Jain, S. Agarwal, and M. Varma. 2021. Deepxml: A deep extreme multi-label learning framework applied to short text documents. In *Proceedings of the ACM International Conference on Web Search and Data Mining*.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural

language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

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. *Preprint*, arXiv:1907.11692.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

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.

Liwen Wang, Rumei Li, Yang Yan, Yuanmeng Yan, Sirui Wang, Wei Wu, and Weiran Xu. 2022. Instructionner: A multi-task instruction-based generative framework for few-shot ner. *Preprint*, arXiv:2203.03903.

Hui Ye, Rajshekhar Sunderraman, and Shihao Ji. 2024. Matchxml: An efficient text-label matching framework for extreme multi-label text classification. *Preprint*, arXiv:2308.13139.

Jun Yu, Yutong Dai, Xiaokang Liu, Jin Huang, Yishan Shen, Ke Zhang, Rong Zhou, Eashan Adhikarla, Wenxuan Ye, Yixin Liu, Zhaoming Kong, Kai Zhang, Yilong Yin, Vinod Namboodiri, Brian D. Davison, Jason H. Moore, and Yong Chen. 2024. Unleashing the power of multi-task learning: A comprehensive survey spanning traditional, deep, and pretrained foundation model eras. *Preprint*, arXiv:2404.18961.

Jiasheng Zhang, Xikai Liu, Xinyi Lai, Yan Gao, Shusen Wang, Yao Hu, and Yiqing Lin. 2023. 2INER: Instructive and in-context learning on few-shot named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3940–3951, Singapore. Association for Computational Linguistics.