CSECU-Learners at SemEval-2025 Task 9: Enhancing Transformer Model for Explainable Food Hazard Detection in Text

Monir Ahmad[1], Md. Akram Hossain[2], and Abu Nowshed Chy[1]
[1]University of Chittagong, Chattogram-4331, Bangladesh
[2]Shanto-Mariam University of Creative Technology, Dhaka-1230, Bangladesh
{ahmad.csecu, akram.hossain.cse.cu}@gmail.com, nowshed@cu.ac.bd

Abstract

Food contamination and associated illnesses represent significant global health challenges, leading to thousands of deaths worldwide. As the volume of food-related incident reports on web platforms continues to grow, there is a pressing demand for systems capable of detecting food hazards effectively. Furthermore, explainability in food risk detection is crucial for building trust in automated systems, allowing humans to validate predictions. SemEval-2025 Task 9 proposes a food hazard detection challenge to address this issue, utilizing content extracted from websites. This task is divided into two sub-tasks. Sub-task 1 involves classifying the type of hazard and product, while sub-task 2 focuses on identifying precise hazard and product "vectors" to offer detailed explanations for the predictions. This paper presents our participation in this task, where we introduce a transformer-based method. We fine-tune an enhanced version of the BERT transformer to process lengthy food incident reports. Additionally, we combine the transformer's contextual embeddings to enhance its contextual representation for hazard and product "vectors" prediction. The experimental results reveal the competitive performance of our proposed method in this task, which achieved 7th place in both sub-tasks. We have released our code at https://github.com/AhmadMonir-CSECU/SemEval-2025_Task9.

1 Introduction

Ensuring food safety is a growing concern; identifying and explaining food risks from online text-based sources could help mitigate this issue. However, the explainability of decision mechanisms related to food risk classification remains underexplored. Enhancing this understanding could help humans quickly assess the validity of predictions and utilize meta-learning approaches, such as clustering or pre-sorting examples. To address these challenges, SemEval-2025 Task 9 (Randl et al., 2025) proposed two sub-tasks: i) Text classification for predicting food hazards, which predicts the type of hazard and product, and ii) Detection of food hazards and product "vectors", which aims to identify the specific hazard and product. To demonstrate a clear view of the task definition, we articulate an example in Table 1.

Prior research (de Noordhout et al., 2014; Marvin et al., 2017) showed that developing early detection methods through compiling epidemiological data and evaluating cases may help us prevent foodborne illness outbreaks. To automate food safety detection, Maharana et al. (2019) investigated several machine learning (ML) models, including linear support vector machines, multinomial Naive Bayes, and weighted logistic regression along with over-sampling and under-sampling techniques on Amazon.com food reviews and FDA food recalls linked data. However, ML-based approaches are being limited to learning complex global contextual information resulting in poor performance. To address this limitation, several studies have explored probabilistic models (Wang et al., 2023) and transformer-based approaches (Xiong et al., 2023; Randl et al., 2024). Nevertheless, transformer-based models exhibit superior performance compared to other methods.

In this work, we have proposed a method based on an enhanced Bidirectional Encoder Representations from Transformers (BERT). We utilize the contextual embedding from the transformer for downstream purposes. To predict hazard and product "vectors", we duplicate and concatenate the embedding, then pass the combined representation into a classifier for final predictions.

We organize the rest of the paper as follows: Section 2 describes our proposed system for the SemEval-2025 food hazard detection task. In Section 3, we detail the system design, including parameter configurations, and present the experimental results along with a performance analysis. Fi-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1263–1268

Title	Labels			
	Sub-task 1		Sub-task 2	
Allan Reeder recalls Cocovite Liquid Whole Egg due to Fipronil	**Hazard-category**	**Product-category**	**Hazard**	**Product**
	chemical	meat, egg, and dairy products	phenylpyrazole	eggs

Table 1: An example of SemEval-2025 Task 9.

nally, we conclude with potential future directions and the limitations of our system in Section 4.

2 Food Hazard Detection Framework

In this section, we introduce our proposed framework for the food hazard detection task. The task consists of two distinct sub-tasks. Sub-task 1 involves predicting the categories of food hazards and products. The sub-task 2 focuses on predicting the exact hazards and products. Both of these are structured as multi-class classification problems. Figure 1 illustrates the overview diagram of our proposed framework.

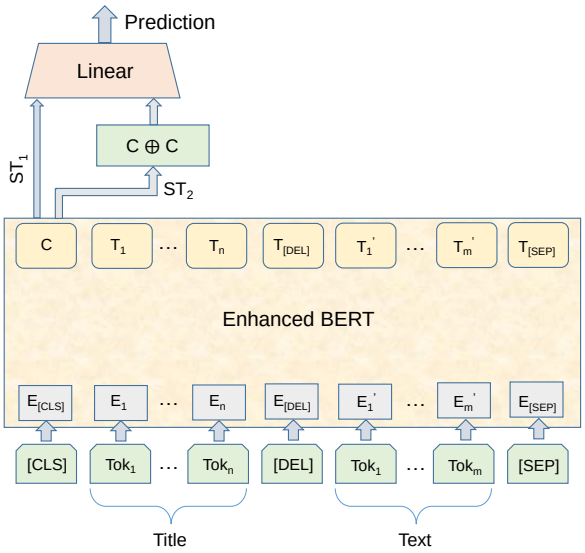

Figure 1: Overview diagram for our proposed method for SemEval-2025 Task 9: Food Hazard Detection Challenge. Here ⊕ indicates concatenation operation.

Our approach incorporates the "Title" and "Text" fields from the dataset. We represent the sequence as: "Title" + [DEL] + "Text" as the input to the transformer where [DEL] indicates a delimiter. We embed a '#' token between "Title" and "Text" as [DEL] to mark the boundary between them. Following (Zhou et al., 2021), we leverage the Enhanced BERT transformer to capture contextual embedding of the sequence. For sub-task 1, the [CLS] token representation is directly fed into the classification layer. For sub-task 2, we replicate the [CLS] representation, concatenate the copies, and then pass the aggregated embedding to the classifier. Finally, the model predicts based on the unnormalized scores (logits) computed by the Linear layer (Paszke et al., 2019).

2.1 Encoder Model

Unlike traditional sequence-based models such as LSTM (Schuster and Paliwal, 1997) and CNN (Goodfellow et al., 2016), transformer models can capture long-term dependencies and enhance the relationships between tokens in a sequence by leveraging multi-head attention and positional embedding mechanisms. To obtain contextualized feature representations of food hazard contexts, we fine-tuned the BERT transformer model as the encoder.

2.1.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), developed by Google's research team in 2018, is a language model that leverages the transformer (Vaswani et al., 2017) architecture. It was pre-trained using passages from BooksCorpus (Zhu et al., 2015) and English Wikipedia. Unlike traditional unidirectional models, BERT performs bidirectional training of transformers, allowing it to understand the context of sentences more deeply. This two-way method has allowed BERT to attain top performance on different natural language processing (NLP) tasks. BERT's pre-training involves two tasks. The first one is Masked Language Modeling (MLM). In this task, BERT randomly masks certain tokens in the input and trains the model to predict these masked tokens using the surrounding context. The second one is the Next Sentence Prediction (NSP). Here, BERT determines whether a pair of sentences are consecutive in the original text.

We utilize the base uncased version of BERT[1] in

[1] https://huggingface.co/google-bert/bert-base-uncased

our task. It comprises 12 transformer blocks (i.e., hidden layers) with 12 attention heads and contains 110M parameters. The hidden size is 768 and the vocabulary size is 30,522.

2.1.2 Extension of BERT

The original BERT encoder supports up to 512 sequence lengths. To handle longer sequences, we utilize the implementation by (Zhou et al., 2021) of Enhanced BERT. Unlike other transformer models that support longer sequences like ModernBERT (Warner et al., 2024), Longformer (Beltagy et al., 2020), it keep the original architecture of the BERT encoder. Let L be the sequence length that is greater than 512, the Enhanced BERT segments the sequence into two overlapping sub-segments which can be represented as follows:

- **Segment 1**: [CLS] Token_1, Token_2, ..., Token_{510}, [SEP]
- **Segment 2**: [CLS] Token_{L-511}, Token_{L-510}, ..., Token_{L-1}, [SEP]

Both of them are then forward-passed to the original BERT encoder. Then we obtain a merged representation of the sub-segments by:

$$
\begin{aligned}
H_1 &= \text{Pad}(\text{BERT}(\text{Segment}_1), \text{bottom padding}) \\
H_2 &= \text{Pad}(\text{BERT}(\text{Segment}_2), \text{top padding}) \\
T &= [T_0, T_1, ..., T_{L-1}] = \frac{H_1 + H_2}{M_1 + M_2 + \epsilon}
\end{aligned} \tag{1}
$$

Here, $H_1, H_2, T \in \mathbb{R}^{L \times d}$. M_1 and M_2 are attention masks for segment 1 and segment 2 respectively. The d indicates the hidden size of the encoder (e.g., 768 for $\text{BERT}_{\text{BASE}}$). ϵ is a small constant to prevent division by zero.

2.2 Classification

We utilize the [CLS] token embedding, c, which corresponds to T_0 in Equation 1, from the transformer for classification purposes. For sub-task 1, we directly feed the embedding into a linear feed-forward layer. For sub-task 2, we duplicate and concatenate the embedding before passing the concatenated embedding into the linear layer. The logits, y, are obtained as follows:

$$y = cW^T + b, \tag{2}$$

Here, $W \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$ are the model's parameters. n indicates the number of classes to be predicted. Finally, the model predicts the class corresponding to the maximum logit.

3 Experiments and Evaluation

3.1 Dataset Overview

To assess the performance of the proposed methods, the organizers of SemEval-2025 Task 9 introduced a benchmark dataset (Randl et al., 2025), derived from CICLe (Randl et al., 2024). This dataset comprises manually annotated English food recall reports sourced from official food agency websites, such as the FDA. Each instance includes six attributes: "year", "month", "day", "country", "title", and "text". The dataset is partitioned into three subsets, as detailed in Table 2.

The competition is structured into two sub-tasks. In sub-task 1, a model is expected to predict the hazard category and product category associated with a given instance. Sub-task 2 extends this challenge by requiring the identification of the exact hazard and product labels. The dataset covers 1,142 distinct products (e.g., "ice cream," "chicken-based products," "cakes"), which are grouped into 22 product categories (e.g., "meat, egg, and dairy products," "cereals and bakery products," "fruits and vegetables"). Additionally, the dataset contains 128 unique hazard labels (e.g., "salmonella," "listeria monocytogenes," "milk and products thereof"), categorized into 10 broader hazard categories. Notably, the dataset exhibits a significant class imbalance (Randl et al., 2024). In the evaluation phase, we merge the train and validation set to train our model and evaluate it with the unseen test set in the CodaLab competition[2].

Fold	Samples
Train	5082
Validation	565
Test	997
Total	6644

Table 2: The statistics of the SemEval-2025 Task 9 dataset.

3.2 Evaluation Measures

To evaluate the performance of the participant's proposed system, the organizers use the macro-averaged F1 score (Sokolova and Lapalme, 2009), which is essential for datasets with a long tail distribution problem. Given a set of true labels y

[2] https://codalab.lisn.upsaclay.fr/competitions/19955

and predicted labels $\hat{y}$, the performance score for a sub-task aggregates the performance on two classification tasks by:

$$F1_h = F1_{\text{macro}}(y_h, \hat{y}_h) \quad (3)$$

$$F1_p = F1_{\text{macro}}(y_p \mid \hat{y}_h = y_h, \hat{y}_p) \quad (4)$$

$$S = \frac{F1_h + F1_p}{2} \quad (5)$$

Where $F1_h$ is the macro F1-score for hazard labels and $F1_p$ is the macro F1-score for product labels, conditioned on correct hazard classification. The evaluation considers both hazard and product classifications, ensuring a balanced assessment across different levels of granularity.

3.3 Parameter Settings

In this section, we outline the parameter configurations for our proposed method. Our model is implemented using PyTorch (Paszke et al., 2019) and the Hugging Face Transformers library (Wolf et al., 2019). We fine-tune the uncased $\text{BERT}_{\text{BASE}}$ pre-trained language model, employing mixed-precision training (Micikevicius et al., 2017) through the Apex library[3] to enhance computational efficiency. Optimization is performed using the AdamW optimizer (Loshchilov and Hutter, 2017). The maximum sequence length is fixed at 1024 tokens. The optimal hyperparameters, as determined by validation set performance, are detailed in Table 3, while default values are maintained for all other parameters. Training is carried out on a T4 GPU utilizing Google Colab (Bisong, 2019).

Hyper-parameters	Optimal Value
Training batch size	8
Encoder learning rate	3e-5
Classifier learning rate	1e-4
Number of epochs	7
Manual seed	66

Table 3: Hyperparameter settings for our method.

3.4 Results and Analysis

In this section, we present a comparative analysis of our proposed system against selected methods for food hazard detection. Following the benchmark set by SemEval-2025 Task 9, system rankings are determined based on the macro-F1 score. Tables 4 and 5 summarize the performance comparisons for sub-task 1 and sub-task 2, respectively. Our system demonstrates competitive performance across both sub-tasks, highlighting its effectiveness in identifying food hazard categories, product categories, specific hazards, and specific products. Upon analyzing Tables 4 and 5, it is evident that sub-task 2 presents greater challenges compared to sub-task 1. This is primarily due to the increased number of target classes and the pronounced class imbalance, making accurate predictions more complex.

Team	Macro-F_1	Features
Baseline (BERT)	0.6670	title
Baseline (TFIDF + LR)	0.4980	title
Anastasia (1st)	0.8223	year, month, day, country, title, text
MyMy (2nd)	0.8112	year, month, day, country, title, text
HU (5th)	0.7882	title, text
BitsAndBites (6th)	0.7873	title, text
CSECU-Learners (7th)	0.7863	title, text
ABCD (8th)	0.7860	title, text
MINDS (9th)	0.7857	title, text
Habib University (26th)	0.4482	N/A
Howard University-AI4PC (27th)	0.1426	text

Table 4: Performance comparison of our proposed method (Team CSECU-Learners) with other selected participants' methods for sub-task 1.

3.5 Ablation Study

In this section, we evaluate the contribution of various components in our model by selectively turning off them. Our findings indicate that each component plays a crucial role in overall performance. The "text" feature, in particular, has a significant impact, as removing it leads to a performance drop

[3] https://github.com/NVIDIA/apex

Team	Macro-F_1	Features
Baseline (BERT)	0.1650	title
Baseline (TFIDF + LR)	0.1830	title
SRCB (1st)	0.5473	title, text
MyMy (2nd)	0.5278	year, month, day, country, title, text
MINDS (5th)	0.4862	title, text
Fossils (6th)	0.4848	title, text
CSECU-Learners (7th)	0.4797	title, text
PuerAI (8th)	0.4783	N/A
Zuifeng (9th)	0.4712	N/A
JU-NLP (25th)	0.0126	title, text
Anaselka (26th)	0.0049	title, text

Table 5: Performance comparison of our proposed method (Team CSECU-Learners) with other selected participants' methods for sub-task 2.

of 6.59% and 8.95% in macro-F1 scores on the test set for sub-task 1 and sub-task 2, respectively. Additionally, the "title" feature also proves beneficial, with its removal resulting in a slight decrease in performance 0.21% for sub-task 1 and 0.18% for sub-task 2. For sub-task 2, we observe that concatenating the [CLS] token embedding enhances the macro-F1 score by 2.12%. In contrast, that strategy reduces the macro-F1 by 0.96% for the sub-task 1. This might be because of the larger number of classes to be predicted for sub-task 2 (128 hazards and 1142 products) than for sub-task 1 (10 hazard categories and 22 product categories).

Method	ST-1	ST-2
CSECU-Learners	0.7863	0.4797
- Title	0.7842	0.4779
- ($c \oplus c$)	-	0.4585
- Text	0.7204	0.3902
+ ($c \oplus c$)	0.7767	-

Table 6: Ablation study results for sub-task 1 and sub-task 2.

Therefore, we can hypothesize that the impact of feature concatenation on model performance is not universal; it depends heavily on the scale and nature of the classification problem. This approach tends to be advantageous in tasks involving a large number of classes, where greater representational power is beneficial. However, in tasks with relatively few classes, increasing the input dimensionality may introduce unnecessary complexity, potentially leading to overfitting. In such cases, the model may learn to rely on spurious patterns in the data rather than focusing on the core discriminative features.

4 Conclusion and Future Direction

In this work, we addressed the challenge of food hazard detection by participating in SemEval-2025 Task 9. We proposed a transformer-based approach, leveraging an enhanced version of the BERT model to handle the complexities of lengthy food incident reports. By combining the transformer's embeddings, our method enhances contextual representations for accurate hazard and product vector prediction. Our approach demonstrated competitive performance in this task, highlighting its effectiveness in classifying hazards and providing precise explanations for predictions.

In the future, we intend to explore other state-of-the-art transformer models pre-trained on biomedical datasets, as they may offer enhanced performance for this task. Due to the imbalanced nature of the dataset, we also aim to apply augmentation techniques that could improve learning across all classes.

Limitations

Our system can process sequences with a maximum length of 1024 tokens. However, many food incident reports exceed this limit, and incorporating the full text could improve the model's contextual understanding. Furthermore, while we utilized the base version of transformer models, their larger variants have shown superior performance in various downstream tasks, an aspect not explored in this study. Additionally, the issue of class imbalance remains unaddressed, which may limit the model's ability to generalize effectively across underrepresented classes, potentially impacting overall prediction accuracy.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Ekaba Bisong. 2019. Google colaboratory. In *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pages 59–64. Springer.

Charline Maertens de Noordhout, Brecht Devleesschauwer, Frederick J Angulo, Geert Verbeke, Juanita Haagsma, Martyn Kirk, Arie Havelaar, and Niko Speybroeck. 2014. The global burden of listeriosis: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, 14(11):1073–1082.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. Deep learning, volume 1.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

Hans JP Marvin, Esmée M Janssen, Yamine Bouzembrak, Peter JM Hendriksen, and Martijn Staats. 2017. Big data in food safety: An overview. *Critical reviews in food science and nutrition*, 57(11):2286–2295.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

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.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xinxin Wang, Yamine Bouzembrak, AGJM Oude Lansink, and HJ van der Fels-Klerx. 2023. Weighted bayesian network for the classification of unbalanced food safety data: Case study of risk-based monitoring of heavy metals. *Risk Analysis*, 43(12):2549–2561.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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(7).

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14612–14620.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.