Habib University at SemEval-2025 Task 9: Using Ensemble Models for Food Hazard Detection

Rabia Shahab
Habib University
rs07528@st.habib.edu.pk

Iqra Azfar
Habib University
ia07614@st.habib.edu.pk

Hammad Sajid
Habib University
hs07606@st.habib.edu.pk

Ayesha Enayat
Habib University
ayesha.enayat@sse.habib.edu.pk

1 Abstract

Food safety incidents cause serious threats to public health, requiring efficient detection systems. This study contributes to SemEval 2025 Task 9: Food Hazard Detection by leveraging insights from existing literature and using multiple BERT-based models for multi-label classification of food hazards and product categories. Using a dataset of food recall notifications, we applied preprocessing techniques to prepare data and address challenges like class imbalance. Experimental results show strong hazard classification performance on ensembled models such as **DistilBERT**, **SciBERT**, and **DeBERTa** but highlight product classification variability. Building on Tyagi et al. (Tyagi et al., 2023) and Madry et al.'s (Rebuffi et al., 2021) work, we explored strategies like ensemble modeling and data augmentation to improve accuracy and explainability, paving the way for scalable food safety solutions.

2 Introduction

The task at hand which lies in the domain of **Food Hazard Detection**(Randl et al., 2025), focuses on developing explainable machine learning systems to classify food hazard-related reports. This task is crucial as food safety incidents pose significant threats to public health and the global economy, leading to foodborne illnesses and product recalls. The challenge involves two key sub-tasks:

- **Sub-task 1 (ST1)**: Classifying food products and hazards categories from textual data.
- **Sub-task 2 (ST2)**: Detecting precise vector representations for product and hazard.

These tasks emphasize both accurate prediction and explainability to enhance trust and usability in real-world applications. The task overview paper provides detailed insights into the structure and objectives of this challenge. Our main strategy involves leveraging multiple BERT-based models for multi-label classification of food hazards and product categories. We employed an ensemble approach, combining models like `DistilBERT`, `DeBERTa`, and `SciBERT` to improve predictive performance and robustness. Additionally, we utilized preprocessing techniques to clean and normalize the dataset, and applied data augmentation to address class imbalance, ensuring a more diverse and representative training set. By participating in this task, we discovered that ensemble learning significantly enhances model performance, achieving a **macro F1 score of 0.7844 on our internal validation set** for Subtask 1 (and **0.4482** on the official test set). For Subtask 2, the ensembled approach yielded **0.442** in the conception-phase internal validation and **0.0315** in the evaluation test phase. According to the official leaderboard, our system ranked **26th for Subtask 1** and **24th for Subtask 2**.

3 Background

Recent advancements in NLP have enabled automated food hazard detection in consumer reviews. Transformer-based models like BERT classify reviews as "safe," "potentially unsafe," or "unsafe," addressing challenges such as class imbalance and limited data. Maharana et al. fine-tuned BERT on expert-validated e-commerce data, achieving a precision of 0.77, recall of 0.71, and F1-score of 0.74 (Maharana et al., 2019). However, small dataset size and linguistic variability hindered generalization.

In foodborne illness detection, encoder-based transformers (RoBERTa, XLM-RoBERTa) and traditional models (SVM, logistic regression) were compared for classifying 7,546 food recall announcements. A novel GPT-CICLe approach combined Conformal Prediction with GPT-3.5-turbo for few-shot learning, reducing large language

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1351–1356

model usage by 60–98% while maintaining accuracy (Randl et al., 2024). TF-IDF-SVM achieved the highest macro F1 scores (0.58 for hazard-category, 0.59 for product-category), outperforming RoBERTa in low-resource settings.

For automating food safety news impact classification, a stacking ensemble integrated classifiers like Naive Bayes, SVM, XGBoost, CNN, LSTM, and BERT. Combining TF-IDF and Word2Vec embeddings enhanced text representation, achieving an F1-score of 0.8052 (Song et al., 2020). Despite its success, computational complexity and dataset constraints limited real-time applications, prompting future efforts to optimize configurations and explore multilingual datasets. In relation to prior work, our method extends Maharana et al.'s fine-tuning of BERT on e-commerce data (Maharana et al., 2019) by tackling **multi-label classification** of both hazards and products simultaneously, rather than single-label safety judgments. Unlike Randl et al.'s GPT-CICLe few-shot prompts (Randl et al., 2024), we employ **deterministic local augmentation** for reproducibility and scale. Compared to Song et al.'s stacking of heterogeneous classifiers (Song et al., 2020), we focus exclusively on **transformer ensembles** to leverage deep contextual embeddings across domain-specific and general models.

4 Data

The dataset used for evaluation consists of food hazard recall notifications collected from various sources, focusing on food safety. It contains a total of 5,966 rows and 10 columns out of which the 6 required columns for our task are shown in Table 1. The dataset was preprocessed to ensure uniformity and relevance for model training. Numerical identifiers, phone numbers, addresses, dates, and unnecessary fields (e.g., Domestic Est. Number, Recall Class) were removed using regex patterns to prevent bias and maintain privacy. Special characters and excessive spaces were eliminated for consistent formatting, and all text was normalized to a uniform format. These steps optimized the dataset for improved machine learning model performance in classification tasks.

5 Experimental Setup

5.1 BERT Baseline

BERT is based on the Transformer architecture introduced in the paper "Attention is All You Need" (Vaswani et al., 2017). BERT was chosen due to its strong contextual understanding and pre-trained knowledge, making it highly effective for text classification tasks with limited training data. Pre-trained models like BERT allow efficient fine-tuning, improving generalization without requiring large datasets.

Table 1: Overview of the dataset fields

Field Name	Description
title	A brief title of the recall notification.
text	Detailed information about the recall.
hazard-category	The category of the hazard (e.g., biological, allergens).
product-category	The category of the product affected (e.g., meat, dairy).
hazard	The specific hazard identified (e.g., listeria monocytogenes).
product	The specific product involved in the recall.

Table 2: Training Parameters for BERT Baseline Model

Parameters	Values
Epochs	3
Batch Size	8
Logging Steps	10
Warmup Steps	500
Weight Decay	0.01

To optimize performance while maintaining efficiency, specific hyperparameters were selected in Table 2. A batch size of 8 was used to balance memory constraints and training stability. Weight decay (0.01) helped prevent overfitting by penalizing large weights. Three training epochs were chosen as BERT fine-tunes effectively with minimal epochs, ensuring stable training, and allowing frequent weight updates without excessive computational costs.

5.2 Ensemble Methods

To improve the efficiency of our tasks, ensembling approach was used that combined multiple individual models to improve overall predictive performance and robustness. The models chosen were:

1. **DistilBERT:** Faster and smaller than standard BERT model so it is efficient in NLP tasks.

2. **DeBERTa:** Good for High-accuracy NLP tasks.

3. **SciBERT:** Scientific domain specific so good in hazards detection for our task.

Model Selection Rationale We chose **SciBERT** for its pre-training on scientific text, which closely matches the technical language of food recall reports; **DeBERTa** for its disentangled attention mechanism that yields richer contextual representations; and **DistilBERT** to balance throughput and accuracy in inference.

The reason for not choosing only the **BERT-Large-uncased** model directly was that large models are well suited for tasks with a significantly larger dataset and more resource-consuming hyperparameters. Based on the findings by Tyagi et al. (Tyagi et al., 2023) ensembled models perform better than large models, as proven by our results. The reason for this is that ensembling combines the predictions of multiple models to produce a more robust and accurate outcome than any individual model, leveraging the strengths and compensating for the weaknesses of different models.

Table 3: Hyperparameter Settings for Ensemble Technique

Hyperparameter	DeBERTa	SciBERT	DistilBERT
Train Epochs	8	8	8
Batch Size	8	16	16
Logging Steps	10	10	10
Warmup Steps	500	500	500
Weight Decay	0.01	0.01	0.01
Parameters	150 M	110 M	67 M

The hyperparameters as shown in Table 3 were chosen to balance training stability and resource constraints. A train and evaluation batch size of 8 was used for DeBERTa due to limited computational resources, while SciBERT and DistilBERT used a batch size of 16 to maximize GPU utilization and training efficiency. The number of training epochs was set to 8 to allow sufficient fine-tuning without risking overfitting. The weight decay value of 0.01 was applied uniformly to regularize the models and prevent overfitting.

Table 4: Model Parameter Comparison

Model	Parameters (Million)
Ensembled Models	327
BERT Large Uncased	336

Table 4 shows that despite BERT having more parameters than the ensemble models combined parameters, the latter performs better. This is because BERT is well suited for tasks with significantly larger datasets, hence ensembling was the suitable approach to our task given our dataset size.

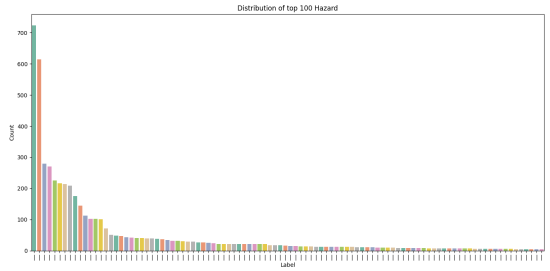
5.3 Data Augmentation

We initially utilized external APIs with structured prompts to generate class-specific, contextually relevant text to address class imbalance and limited sample diversity. While effective for targeted augmentation, this approach was constrained by API rate limits, key exhaustion, and inconsistent latency. To overcome these limitations, we implemented a deterministic local augmentation pipeline using `nlpaug` and BERT-based contextual embeddings.

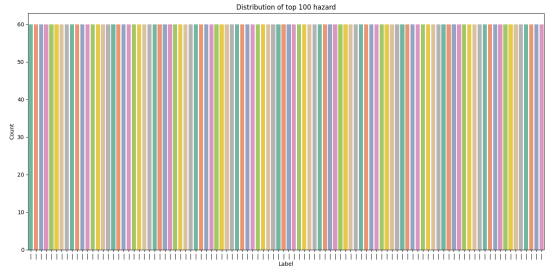
This method employed substitution-based augmentation (top_k = 50) using "ContextualWordEmbsAug" with the "bert-base-uncased" model. Semantic similarity was validated through spaCy's "en_core_web_md" model, using an initial acceptance threshold of 0.85, adaptively reduced to a minimum of 0.70 when necessary. Each sample underwent up to 200 augmentation attempts, generating a maximum of 10 high-similarity, semantically coherent variants.

Each class—Hazard and Product—was augmented independently to preserve class-specific semantics and avoid drift. Augmentation was confined within individual class groups to maintain label integrity. Overrepresented classes were randomly undersampled using a fixed seed to meet a target count, while underrepresented classes were synthetically expanded. This explains the reduction in dominant class frequencies, as illustrated in Fig. 1 (Hazard) and Fig. 2 (Product), which show class distributions before and after augmentation.

Two augmentation passes were applied per class, and metadata fields (*augmentation_pass*, *try_number*, *is_augmented*) were retained for traceability. This method, implemented using `nlpaug`, yielded a measurable macro F1 score increase from 0.10 (baseline) to 0.32 after augmentation, reflecting a consistent improvement across all classes. The method is scalable, reproducible, and independent of external services, offering a robust solution for large-scale augmentation. Additional results are discussed in the later sections.

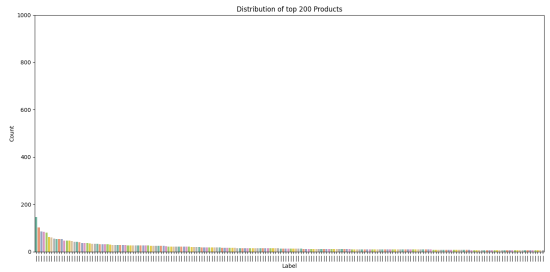


(a) Before Augmentation for Hazard Class (Top 100 Labels)

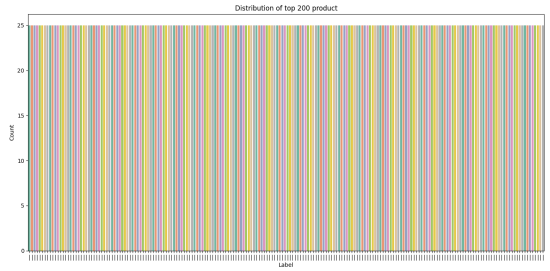


(b) After Augmentation for Hazard Class (Top 100 Labels)

Figure 1: Comparison of Hazard Class Distribution Before and After Augmentation



(a) Before Augmentation for Product Class (Top 200 Labels)



(b) After Augmentation for Product Class (Top 200 Labels)

Figure 2: Comparison of Product Class Distribution Before and After Augmentation

6 Results and Discussion

6.1 Subtask 1: Hazard and Product Category Identification

Initially, we fine-tuned the **BERT-base-uncased** model to establish a baseline for performanFce. During training, we recorded key metrics such as training loss, validation loss, accuracy, and F1 score for each epoch, using an 80-20 training-validation split. The macro F1 score, reported in Table 5, represents the result on the test dataset, as obtained by submitting the model predictions to **CodaBench**. This baseline model achieved a macro F1 score of **0.41**.

Subsequently, we adopted an ensemble learning approach, combining multiple models to improve performance. The models included in the ensemble were **DistilBERT-base-uncased**, **DeBERTa-base-uncased**, and **SciBERT-base**, the latter being fine-tuned specifically on scientific data. By averaging the logits from these models to generate final predictions, we significantly improved the macro F1 score, achieving **0.78**.

This improvement aligns with the findings of Tyagi et al. (Tyagi et al., 2023), which suggest that ensemble approaches, leveraging multiple foundational models, can outperform larger individual models, despite the latter having more parameters. To validate this hypothesis, we trained the **BERT-large-uncased** model on the same dataset; however, it achieved only a macro F1 score of **0.52**, demonstrating that the ensemble method outperforms larger single-model configurations.

Additionally, we implemented a Conformal In-Context Learning (CICLe) approach, as proposed by Randl et al. (Randl et al., 2024), which utilizes logistic regression as a base classifier and prompts **Llama-3.1 B** for hazard and product category classification. This approach resulted in a macro F1 score of **0.51**, further contributing to the evaluation of different strategies for model improvement.

Table 5: Performance of various fine-tuned models employed for identifying Hazard and Product categories for conception phase ST 1.

Model(s)	Accuracy	Loss	Macro F1 Score
BERT-base-uncased (baseline)	0.56	9.27	0.41
Ensembled (DistilBERT, DeBERTa, SciBERT)	0.39	1.45	**0.78**
BERT-large-uncased	0.43	5.67	0.52
Conformal In-Context Learning (CICL)	-	-	0.51

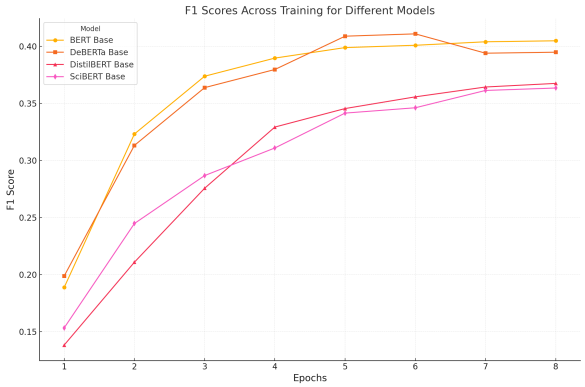


Figure 3: Visual representation of F1 scores across models. The ensembled models show better scores in Macro F1 score as indicated in Table 5.

6.2 Subtask 2: Hazard and Product Identification

For Subtask 2, we directly employed an ensemble of models, including **DistilBERT-base-uncased**, **DeBERTa-base-uncased**, and **SciBERT-base**. Despite the improved model architecture, the extreme class imbalance in this subtask resulted in poor performance, with the models struggling to effectively capture the underlying patterns. Fine-tuning the ensemble on the imbalanced dataset led to a slight improvement, achieving a macro F1 score of **0.12** as indicated in Table 6, which, though better, was still suboptimal for a classification task of this nature.

Table 6: Performance of various fine-tuned models/techniques employed for identifying Hazards and Products for Conception Phase ST2.

Model(s)	Accuracy	Loss	Macro F1 Score
BERT-base-uncased (baseline)	0.32	9.6	0.08
Ensembled (DistilBERT, DeBERTa, SciBERT)	0.53	5.54	0.12
Augmented Ensembled	0.68	4.32	**0.32**

6.3 Quantitative Findings and Analysis

The ensembled models achieved the highest macro F1 score of **0.4482** for Subtask 1 and **0.0315** for Subtask 2 in the evaluation phase. This significantly outperformed the baseline models, demonstrating the effectiveness of model ensembling in handling complex classification tasks.

Ablation studies further confirmed that the ensemble approach consistently outperformed individual models, highlighting the benefits of combining multiple architectures. Additionally, data augmentation played a crucial role in addressing class imbalance, leading to improved model performance.

Phase	Sub-task 1 Score	Sub-task 2 Score
Conception Phase ST1	0.784	0.000
Conception Phase ST2	0.000	0.442
Evaluation Phase	**0.4482**	**0.0315**

Table 7: Scores for Conception and Evaluation Phases

The Conception Phase ST2 score of 0.442 refers to performance on our held-out 20% validation split prior to evaluation-phase tuning. The official evaluation macro F1 (0.0315) was obtained by submitting to the SemEval CodaLab test set, with labels withheld to simulate unseen conditions, which equally weights performance across all label categories irrespective of class distribution. In Task 9, this includes both hazard type and product category labels, requiring robust handling of multi-label imbalance and partial annotations as described in the task formulation (Randl et al., 2025).

7 Conclusion and Future Work

In this paper, we presented our participation in **SemEval 2025 Task 9: Food Hazard Detection**, where we applied advanced BERT-based models to tackle the multi-label classification problem of food hazards and product categories. Although built on established components, our integration of controlled data augmentation with ensemble diversity contributes a reproducible and competitive baseline for food hazard detection under imbalanced conditions. Through extensive experimentation with models such as DistilBERT, SciBERT, and DeBERTa, we achieved strong performance in hazard classification. Our ensemble approach demonstrated promising results, though product classification still exhibited variability, highlighting areas for further improvement. Currently, we are ranked **26** and **24** internationally based on our scores in Subtasks 1 and 2 in the **Submitted** leader-

board.

In the future, the emphasis will be placed on optimizing data augmentations toward mitigating class imbalance issues and building model robustness. Here, extending augmentations over a longer window will diversify and make training data much more representative when considering minority classes. This could include enhancing the diversity of the data with further techniques like domain-specific paraphrasing or leveraging generative approach models for GPT-based synthetic data. Hyperparameters such as learning rate, batch size, and weight decay will also be tuned toward maximizing the model's stability and efficiency. Finally, a study will be performed to discover methods for easing resource consumption while maintaining the performance of the previously underlined concepts to allow for a better level of generalization and scalability in handling a more complex food hazard detection task. The future work will further focus on incorporating multilingual datasets to enhance generalizability, integrating explainability modules such as SHAP or LIME for model transparency, and exploring active learning strategies to iteratively refine the model with user feedback in real-world deployment settings.

8 Limitation

The ensemble approach, combining DistilBERT, DeBERTa, and SciBERT, demonstrated a good performance in hazard classification but faced notable limitations. Despite data augmentation efforts that improved class distributions and boosted overall scores, the models struggled with product classification due to the inherent complexity of diverse product categories and extreme class imbalance. Even after augmentation, synthetic data failed to capture nuanced domain-specific patterns fully. The ensemble's computational complexity also limited real-time deployment, while hyperparameter tuning introduced trade-offs between resource efficiency and training stability. These limitations highlight the need for more advanced augmentation techniques, cost-sensitive learning, or hierarchical classification strategies to address underrepresented classes and improve scalability.

9 Code and Reproducibility

To encourage reproducibility and facilitate future work on food hazard detection, we have made our source code, preprocessing scripts, and configuration files publicly available at:

`https://github.com/HammadxSaj/Sem-Eval-Task09-Dataset`

This repository includes the implementation for all models described in this paper, including BERT fine-tuning, ensemble logic, and data augmentation routines.

References

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.

Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2024. Cicle: Conformal in-context learning for largescale multi-class food risk classification. *ACL*.

Korbinian Randl, John Pavlopoulos, Aron Henriksson, Tony Lindgren, and Juli Bakagianni. 2025. SemEval-2025 task 9: The food hazard detection challenge.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles, and Timothy A. Mann. 2021. Data augmentation can improve robustness. *NeurIPS*, abs/2111.05328.

Bo Song, Kefan Shang, Junliang He, Wei Yan, and Tianjiao Zhang. 2020. Impact assessment of food safety news using stacking ensemble learning. *IOS Press*.

Nancy Tyagi, Aidin Shiri, Surjodeep Sarkar, Abhishek Kumar Umrawal, and Manas Gaur. 2023. Simple is better and large is not enough: Towards ensembling of foundational language models. *MASC-SLL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS*, abs/1706.03762.