Howard University-AI4PC at SemEval-2025 Task 9: Using Open-weight BART-MNLI for Zero Shot Classification of Food Recall Documents

Saurav K. Aryal
Howard University
saurav.aryal@howard.edu

Kritika Pant
Howard University
kritika.pant@bison.howard.edu

Abstract

We present our system for SemEval-2025 Task 9: Food Hazard Detection, a shared task focused on the explainable classification of food-incident reports. The task involves predicting hazard and product categories (ST1) and their exact vectors (ST2) from short texts. Our approach leverages zero-shot classification using the BART-large-MNLI model, which allows classification without task-specific fine-tuning. Our model achieves competitive performance, emphasizing hazard prediction accuracy, as evaluated by the macro-F1 score.

1 Introduction

Food safety is a critical global concern, with food-borne illnesses affecting millions annually and causing significant economic losses. Rapid and accurate detection of food hazards is essential to mitigate these risks, but the sheer volume of food-incident reports makes manual monitoring challenging. Automated systems that identify and classify food hazards from textual data, such as recall notices or social media posts, offer a promising solution. However, these systems must be accurate since it is crucial for building trust and enabling human oversight in food safety applications.

Recent advancements have enabled the development of systems that can automatically detect food hazards from textual data. For example, previous work has focused on identifying foodborne illnesses from social media posts or analyzing food recall reports from official agencies. (Poisoned, 2025)

Despite the progress in automated food safety systems, there is a lack of systems that provide exact predictions for food hazards. Additionally, the imbalanced and ambiguous nature of food-incident data poses significant challenges for traditional supervised learning approaches.In this paper, we aim to address these challenges by developing a system for food hazard detection using zero-shot classification. Our approach leverages the BART-large-MNLI model to predict hazard and product categories (ST1) and their exact vectors (ST2) from short food recall related texts.

The key contribution of our work evaluate the effectiveness of zero-shot classification using a well-studied pre-trained open-weight model as part of SemEval-2025 Task 9 (Randl et al., 2025).

2 Related Work

Utilizing advancements in deep learning and large language models to categorize and classify text is standard practice in research literature across multiple domains and languages (Prioleau and Aryal, 2025; Aryal and Prioleau, 2024; Aryal et al., 2023a,b; Prioleau and Aryal, 2023). Recently, the detection of food hazards from textual data, such as food incidents reports, has gained significant attention due to its potential to improve public health and reduce economic losses. Traditional methods for identifying food hazards often rely on manual analysis of reports from official sources, which can be time-consuming and prone to delays. To address these limitations, recent research has focused on developing automated systems that can efficiently process and classify food-incident reports collected from the web. For example, studies have explored the use of machine learning (ML) techniques to analyze social media posts and web-based reports for early detection of food borne illnesses, demonstrating the potential of these systems to identify unreported events and enhance outbreak response (Radford et al., 2018; Tao et al., 2023). Additionally, systems like FINDER have leveraged aggregated web search and location data to detect food borne illnesses in real-time, showcasing the value of integrating diverse data sources for timely hazard identification (Tao et al., 2023). However, many existing systems lack explainability, which is crucial for ensuring transparency and enabling human

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1919–1923
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

oversight in food safety applications. Our work addresses this gap by introducing classification system that uses zero-shot classification to predict hazard and product categories (ST1) and their exact vectors (ST2) from food-incident reports which supports the development of automated crawlers that can efficiently extract and classify food hazards from web sources like social media. (Poisoned, 2025)

3 Methodology

The task involves two subtasks:

- **ST1:** Hazard and Product Category Classification – Systems are required to predict the general category of the hazard (e.g., "biological hazards") and the product category (e.g., "meat, egg, and dairy products") from the input text.
- **ST2:** Hazard and Product Vector Detection – Systems must predict the exact hazard (e.g., "salmonella") and product (e.g., "ice cream") from the input text.

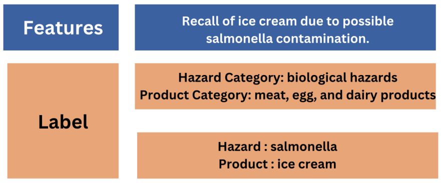

Features
Recall of ice cream due to possible salmonella contamination.
Label
Hazard Category: biological hazards
Product Category: meat, egg, and dairy products
Hazard : salmonella
Product : ice cream

Figure 1: Model inputs in Blue and ground truth in Orange.

The dataset is primarily in English and includes a variety of genres, such as recall notices, social media posts, and news articles. The size and diversity of the dataset make it a challenging benchmark for evaluating the performance of automated systems in real-world scenarios. (Randl et al., 2025)

3.1 Key Algorithms and Modeling Decisions

Our approach focuses on classifying input text into categories by identifying exacts hazards and products along with their respective category. To achieve this, we leverage zero-shot learning using a powerful language model, combined with labels derived from the training dataset. The overall process consists of four key components:

- **Data Preprocessing:** We start with cleaning and formatting the input text from the testing dataset which involves removing additional whitespace and newline characters to form one continuous input text.
- **Label Extraction:**From the training dataset, we extract candidate labels from the datafield —such as categories like hazards and products—along with their corresponding classifications. These labels are then passed to the model alongside the text field from the test dataset. Number of hazards and products extracted:

```
Number of hazards and products extracted:
hazards = 128
products = 1022
hazards category = 10
products category = 22
```

- **Zero-Shot Classification:** To classify the input text, we employ the BART-large-MNLI model, a powerful transformer-based architecture designed for natural language inference. This model allows us to predict hazard and product along their categories by:
  - Assessing whether the input text of testing dataset aligns with a set of predefined candidate labels retrieved from the training dataset.
  - Each label is reformatted as a natural language hypothesis whereas the input text is treated as a premise.
  - Evaluating the relationship between the text and each label in terms of entailment (true), neutral, or contradiction (false).
- **Prediction Generation:** Selecting the label with the highest probability as the final classification result and saving it in a file.

3.2 Core Model:Why BART-large MNLI for zero-shot classification

We selected the BART-large MNLI model as the core of our system because it has shown strong performance in zero-shot classification tasks, particularly in scenarios where labeled training data is limited or unbalanced. The model is pre-trained on the Multi-Genre Natural Language Inference (MNLI) dataset, which enables it to generalize well to unseen tasks by leveraging its understanding of textual entailment and semantic relationships. This makes it highly suitable for our task, where we need

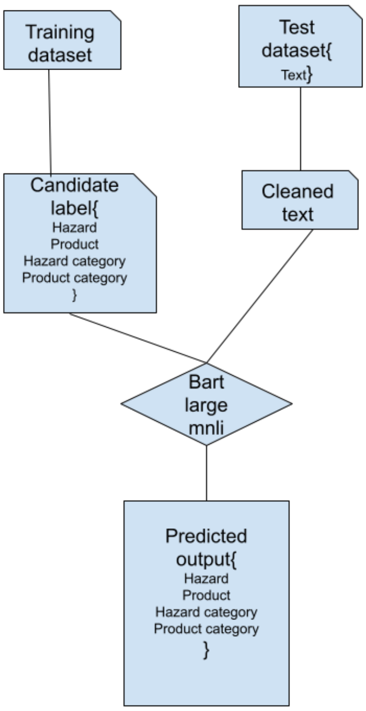

Training dataset
Test dataset{ Text}
Candidate label{ Hazard Product Hazard category Product category }
Cleaned text
Bart large mnli
Predicted output{ Hazard Product Hazard category Product category }

Figure 2: Proposed model approach

to classify food hazards and products from short, unstructured texts without task-specific fine-tuning.

Unlike T5, which is optimized for text-to-text generation and excels in tasks where output generation is necessary (Face, 2025), BART-MNLI is directly aligned with classification tasks, allowing us to leverage its structure without additional task-specific formatting or training. While GPT-based classifiers are powerful, they often require complex prompt engineering and tend to be more computationally intensive.(Hugging Face, 2024) In contrast, BART-MNLI offers a strong balance between performance, interpretability, and resource efficiency.

Furthermore, its zero-shot capability allows it to handle unbalanced datasets effectively without requiring training, making it a robust and scalable choice for food-hazard detection in real-world applications.

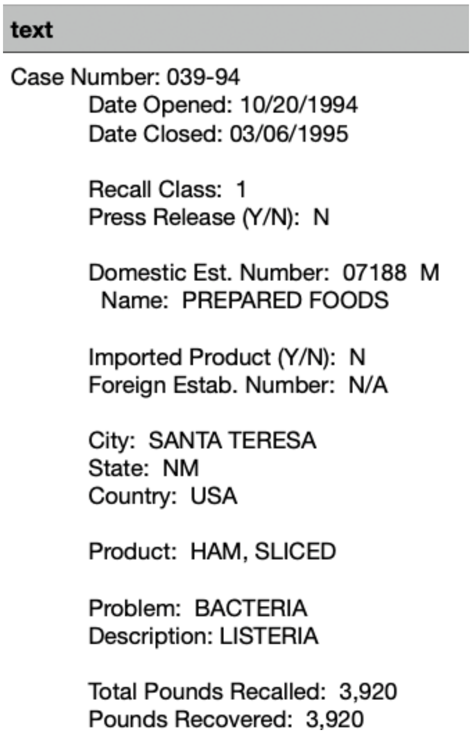
4 Experimental Setup

4.1 Data Splits

The dataset for the Food Hazard Detection Task is divided into three main splits: training, development, and test sets. Each split serves a specific purpose in the development and evaluation of our system.

- **Training Set:** This set is used to extract candidate labels for hazards and products. These labels are essential for the zero-shot classification process, as they provide the model with a predefined set of options to choose from.

- **Test set:** This set is used for the final evaluation of the system. It provides an unbiased measure of the system's performance on completely unseen data, simulating real-world conditions.



text

Case Number: 039-94
Date Opened: 10/20/1994
Date Closed: 03/06/1995

Recall Class: 1
Press Release (Y/N): N

Domestic Est. Number: 07188 M
Name: PREPARED FOODS

Imported Product (Y/N): N
Foreign Estab. Number: N/A

City: SANTA TERESA
State: NM
Country: USA

Product: HAM, SLICED

Problem: BACTERIA
Description: LISTERIA

Total Pounds Recalled: 3,920
Pounds Recovered: 3,920

Figure 3: Example of test dataset

4.2 External Tools and Libraries

Our system relies on several external tools and libraries for data manipulation, text pre-processing, and model inference. Below, we provide a summary of these resources.

- **Transformers:** Used for loading the BART-large-MNLI model and performing zero-shot classification. (Wolf et al., 2020)

- **Pandas:** Used for data manipulation, such as loading the dataset and extracting candidate labels. (Paszke et al., 2019)

- **Torch:** Used for GPU acceleration during model inference. (pandas development team, 2020)

4.3 Evaluation Measures

The task is evaluated using the macro-F1 score, which is calculated separately for hazard and product predictions. The final score is the average of the two F1 scores, with a higher weight given to the accuracy of the prediction of hazards. This evaluation metric ensures that the system performs well on both tasks while prioritizing the correct identification of hazards.

5 Limitation

A key limitation of our approach is that relying on candidate labels extracted from the training set guarantees we will miss unseen or novel labels, limiting the system's ability to generalize to new hazards or products. Additionally, while our approach is preliminary, there may be better-performing models or techniques, such as fine-tuning, few-shot learning, or testing alternative architectures, that could yield improved results. Furthermore, our system only utilizes text data, which could overlook valuable context from other data fields that could enhance performance. Finally, the model is only English, which restricts its applicability to multilingual contexts, highlighting the need for future work to address language diversity and incorporate additional data sources for more robust and generalizable food hazard detection.

Our system underperforms compared to other participants, but we lack a thorough error analysis to understand why. Future work should include a detailed breakdown of performance across different subcategories (e.g., biological vs. chemical hazards) and confusion matrices to identify specific areas of model weakness and common failure patterns in hazard versus product classification.

6 Result

The intention behind the submission was to evaluate the feasibility of using a zero-shot model for food-hazard and product classification from short, unstructured texts—without relying on task-specific fine-tuning or annotated training data, which is often scarce or unbalanced in real-world applications.Our model performed better than random chance, indicating that it captures some meaningful semantic distinctions between hazards and products and score between two of them has been mentioned below:

Model	**Naive Random Classifier**	**BART-MNLI (Zero-shot)**
Task 1	0.0043	0.2426
Task 2	0.07275	0.1380

However, we agree that the performance is not sufficient for deployment and that more extensive work is needed.

For all the tasks where the shared task organizers released a test data and , we used the test data for the results reported in this section (for the and the score was evaluated by calculating the macro-F1-score on the participants' predicted labels using the annotated labels as ground truth.The leaderboard can be found below:

ID	Username	Organization	Score
87	HammadxSajid		0.4482
88	mdalam		0.4455
89	hanguanghui	QF_CS	0.4435
90	kritikapant2003	**"ours"**	0.1426

Table 1: Leaderboard results for ST1

ID	Username	Organization	Score
47	king001	PA14	0.2062
48	sushovit21	IISERB-03	0.2055
49	hanguanghui	QF_CS	0.1529
50	kritikapant2003	**"ours"**	0.1380

Table 2: Leaderboard results for ST2

7 Conclusion

Key contributions of our work include the use of zero-shot classification to handle imbalanced data and the integration of multi-task learning and ensemble methods to improve robustness. We recognize that a more thorough error analysis would provide deeper insights, and we plan to include this in future iterations of our work.We also acknowledge the importance of comparing our zero-shot baseline to fine-tuned transformer models or few-shot learning approaches, which are likely to yield better results by leveraging task-specific supervision. While our current focus was to establish a simple yet meaningful baseline, we plan to explore and benchmark fine-tuned models and other machine learning baselines in future work to better contextualize performance.

Acknowledgements

This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714. The work is solely the responsibility of the authors and does not necessarily represent the official view of the Office of Naval Research.

References

Saurav K Aryal and Howard Prioleau. 2024. Ad-hoc ensemble approach for detecting adverse drug events in electronic health records. *Journal of Computing Sciences in Colleges*, 40(3):238–249.

Saurav K Aryal, Howard Prioleau, Gloria Washington, and Legand Burge. 2023a. Evaluating ensembled transformers for multilingual code-switched sentiment analysis. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 165–173. IEEE.

Saurav K Aryal, Ujjawal Shah, Howard Prioleau, and Legand Burge. 2023b. Ensembling and modeling approaches for enhancing alzheimer's disease scoring and severity assessment. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1364–1370. IEEE.

Hugging Face. 2025. T5 - transformers documentation. Accessed: April 28, 2025.

Hugging Face. 2024. Openai gpt - hugging face transformers documentation. Accessed: 2025-04-28.

The pandas development team. 2020. *pandas: powerful Python data analysis toolkit*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*.

I Was Poisoned. 2025. Crowdsourced foodborne illness reporting platform.

Howard Prioleau and Saurav Aryal. 2025. Entity only vs. inline approaches: Evaluating llms for adverse drug event detection in clinical text (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29469–29471.

Howard Prioleau and Saurav K Aryal. 2023. Benchmarking current state-of-the-art transformer models on token level language identification and language pair identification. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 193–199. IEEE.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *arXiv preprint arXiv:1812.01813*.

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.

Dandan Tao, Ruofan Hu, Dongyu Zhang, Jasmine Laber, Anne Lapsley, Timothy Kwan, Liam Rathke, Elke Rundensteiner, and Hao Feng. 2023. A novel foodborne illness detection and web application tool based on social media. *Foods*, 12(14):2769.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.