WITM at #SMM4H-HeaRD 2025: Detection and Extraction of Food Recalls and Foodborne Disease Outbreaks using linguistically-Informed Language Models

Suchandra Chakraborty, Diya Saha, Sudeshna Jana

Research and Innovation Lab, Tata Consultancy Services
West Bengal, India
{suchandra.c, diya.saha, sudeshna.jana}@tcs.com

Abstract

This paper presents our contribution to Task 5 of the Social Media Mining for Health (#SMM4H) 2025 shared task, which addresses the classification of foodborne disease outbreaks and recall events from news articles, with a primary focus on FDA press releases. Task 5 is divided into two subtasks: (1) sentence-level classification of news articles, and (2) extraction of key entities and events from text. To tackle these challenges, we employed transformer-based architectures and large language models, leveraging both few-shot and zero-shot prompt engineering strategies. Additionally, we explored various prompt formulation and data augmentation techniques to assess their influence on system performance. Our proposed systems achieved an F1 score of 92.3% in Subtask 1 and an average score of 48% in Subtask 2 on the test dataset. On the test set, our models attained the second-highest accuracy in both subtasks.

Introduction

The rising prevalence of food safety incidents continues to raise serious public health and economic concerns globally (Amico et al. 2018; Kase, Zhang, and Chen 2017; Boatemaa et al. 2019; Nerín, Aznar, and Carrizo 2016). Foodborne diseases remain a leading cause of illness and mortality (Bouzembrak and Marvin 2016; Potter et al. 2012; Pádua et al. 2019; Lüth et al. 2019), often necessitating large-scale recalls that impose significant financial burdens on industries and governments (Deng, den Bakker, and Hendriksen 2016). Identifying the root causes of such events—including contamination during production, packaging, or storage—is essential for effective prevention strategies (Zhou et al. 2020; Scallan and Mahon 2012; Gupta et al. 2004; Hall et al. 2013).

Despite the urgency, computational models for detecting and analyzing food safety incidents remain limited, largely due to the scarcity of structured data in this domain (Allard et al. 2016; Moumni Abdou et al. 2019). To address this gap, Task 5 of the Social Media Mining for Health (#SMM4H) 2025 shared task—Detection and Extraction of Food Recalls and Foodborne Disease Outbreaks in Online News Articles—aims to foster the development of automated systems for identifying such events from unstructured textual sources. The task centers on news articles and press releases, with a particular focus on FDA-issued statements. The dataset comprises 3,172 training articles, 357 development articles, and 1,005 test articles, sourced from FDA press releases, news websites, and scientific literature.

Recent advancements in language technologies and data-driven methodologies have shown promise in addressing foodborne risk assessment (Harris et al. 2017; Altenburger and Ho 2019; Maharana et al. 2019; Deng, Cao, and Horn 2021). National systems such as PulseNet (Swaminathan et al. 2001), NARMS (Gupta et al. 2004), FoodNet (Scallan and Mahon 2012), and the National Outbreak Reporting System (Hall et al. 2013) illustrate the importance of data integration in surveillance. Moreover, the adoption of whole-genome sequencing (WGS) in platforms like GenomeTrakr and EnteroBase has enabled real-time pathogen monitoring at scale (Allard et al. 2016; Zhou et al. 2020). Recently, (Jana, Sinha, and Dasgupta 2024) demonstrated the utility of large language models and transformer-based architectures for event detection in food safety contexts using a multi-task sequence labeling framework.

Motivated by these developments, our work in the #SMM4H 2025 Task 5 explores transformer-based models and LLMs for sentence-level classification and entity/event extraction in food safety-related articles, with an emphasis on FDA press releases.

Methodology

The shared task comprises two subtasks: (1) Multi-class classification of news articles related to food safety and (2) entity and event extraction from the same. We use both zero-shot/few-shot and supervised approaches to tackle these subtasks effectively.

Subtask 1: News Article Classification

(a) Classification with LLaMA 3.1 8B To classify news articles into one of three categories—1) Food Recall, 2) Foodborne Disease Outbreak, or 3) Neither—we employ the LLaMA 3.1 8B Instruct model (Touvron et al. 2023) in both zero-shot and few-shot settings. We construct various prompts incorporating definitions and representative examples for each category to guide the model's classification (Results section reports the best-performing configurations).

(b) Supervised Fine-Tuning with DistilBERT We also fine-tuned a pre-trained `DistilBERT base` (Sanh et al.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

2020) for the same classification task. The model, originally trained on large-scale general corpora, was adapted to the domain-specific task using the provided labeled dataset. Fine-tuning was performed using a batch size of 32, a maximum sequence length of 128, a learning rate of 2×10^{-5}, and early stopping after 800 steps to avoid overfitting.

Subtask 2: Entity and Event Extraction For extracting structured information from the articles, specifically six entity types such as Target Organization, Product Name, Cause of Incident, Disease Caused, Number of People Affected, and Location, we explore two strategies:

(a) Entity Extraction via Prompting and Linguistically Informed Bi-LSTM We utilize the LLaMA 3.1 8B Instruct model (Touvron et al. 2023) to extract entities directly from article text in both zero-shot and few-shot settings. Prompts are formulated in natural language, instructing the model to identify and list key fields such as *organization, product, disease, number of people affected*, and *location* (Ref: Appendix A.2). For extracting the *cause* field, we adopt a separate approach using a linguistically informed Bi-LSTM model (Dasgupta et al. 2018).

(b) Supervised NER with RoBERTa-Large For this approach, we modeled the extraction task as token-level sequence labeling, targeting required entities. A supervised NER model based on `RoBERTa-large` (Liu et al. 2019) was fine-tuned using IOB-tagged given training dataset. Entity mentions were identified via exact string matching and aligned to tokens using a character-to-token mapping that accounted for subword tokenization. Training was conducted using cross-entropy loss with a learning rate of 2×10^{-5}, batch size 8, and 10 epochs. A sliding window was applied at inference to process long texts, and overlapping predictions were merged to form final entity spans.

Results

Table 1: Sub-1 Classification Results

Model	Acc	F1	P	R
LLaMA Z-S (Test)	0.824	0.824	0.867	0.824
LLaMA Z-S (Val)	0.919	0.927	0.944	0.919
LLaMA F-S (Val)	0.77	0.84	0.77	0.76
DistilBERT (Val)	0.922	0.909	0.922	0.915
DistilBERT (Test)	0.931	0.923	0.916	0.931
SMM4H Task5-Mean (Test)	0.855	0.854	0.887	0.855

Table 2: Sub-2 Entity and Event Extraction Results (Pt-1)

Model	Avg	Org	Prdt	Cau
LLaMA Z-S + BiLSTM(Test)	0.444	0.810	0.400	0.220
LLaMA F-S (Test)	0.48	0.89	0.43	0.22
LLaMA Z-S (Val)	0.429	0.800	0.420	0.040
LLaMA F-S (Val)	0.64	0.93	0.67	0.34
RoBERTa (Test)	0.208	0.370	0.210	0.070
RoBERTa (Val)	0.277	0.539	0.425	0.163
SMM4H Task5-Mean (Test)	0.336	0.56	0.343	0.135

We evaluated our models on both Subtask-1 and Subtask-2 using the official validation set and test set. The following tables summarize the performance of our submissions. In Tables 1, 2 and 3, we have used the following abbreviations: *Acc: Accuracy, F1: F1-Score, P: Precision, R: Recall, Avg: Macro-averaged F1 score across entity/event types, Org: Organization, Prdt: Product, Cau: Cause, Dis: Disaster, Num_Aff: Number of Affected Individuals, Loc: Location, Z-S: Zero shot, F-S: Few shot*. Our experiments show that LLaMA achieved superior performance in both tasks, though with key differences in model effectiveness. For Subtask-1 (classification), LLaMA's zero-shot approach attained strong results (82.4% F1 on test, 92.7% on validation), while DistilBERT surprisingly matched this performance (90.9% F1) despite its smaller size (Ref: Table 1). This parity stems from DistilBERT's fine-tuned specialization for classification, its efficient tokenization, and focused learning on labeled FDA recall patterns, making it competitive with LLMs for this simpler task. In Subtask-2, however, LLaMA significantly outperformed RoBERTa (48% vs. 20.8% avg F1), excelling at organization identification (81.0% F1) but struggling with abstract concepts like causes (22.0% F1) (Ref: Table 2 and 3). RoBERTa's poorer performance highlights the limitations of standard NER architectures for complex, relationship-dependent extraction, where LLaMA's generative flexibility and world knowledge proved critical. Comparing with SMM4H Task-5 Mean results in the evaluation phase, our model demonstrates a significantly robust performance in both subtasks of Task 5. While fine-tuned compact models (DistilBERT) can rival LLMs in narrow classification tasks, LLMs dominate when tasks require nuanced understanding or multi-entity reasoning, revealing a clear trade-off between specialization and generalization.

Table 3: Sub-2 Entity and Event Extraction Results (Pt-2)

Model	Dis	Num_Aff	Loc
LLaMA Z-S + BiLSTM(Test)	0.500	0.560	0.430
LLaMA F-S (Test)	0.60	0.560	0.420
LLaMA Z-S (Val)	0.580	0.570	0.460
LLaMA F-S (Val)	0.75	0.69	0.63
RoBERTa (Test)	0.180	0.300	0.320
RoBERTa (Val)	0.174	0.153	0.369
SMM4H Task5-Mean (Test)	0.33	0.487	0.395

Conclusion and Future Work

Our current approach faces key limitations in identifying primary causes among multiple candidate causes, tracking multi-region outbreaks, and extracting standardized disease terminology, primarily due to challenges in entity prioritization and domain-specific normalization. To address these, in future work, we would develop multi-stage prompting to better handle competing entities, integrate public health databases for disease name standardization, and implement cross-verification against source texts to improve extraction reliability. We would also refine location-specific processing for multi-region outbreaks while maintaining the model's computational efficiency through these targeted enhancements. These targeted enhancements will address current weaknesses while maintaining the model's efficiency.

References

Allard, M. W.; Strain, E.; Melka, D.; Bunning, K.; Musser, S. M.; Brown, E. W.; and Timme, R. 2016. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *Journal of clinical microbiology*, 54(8): 1975–1983.

Altenburger, K. M.; and Ho, D. E. 2019. Is Yelp actually cleaning up the restaurant industry? A re-analysis on the relative usefulness of consumer reviews. In *The World Wide Web Conference*, 2543–2550.

Amico, P. D.; Nucera, D.; Guardone, L.; Mariotti, M.; Nuvoloni, R.; and Armani, A. 2018. Seafood products notifications in the EU Rapid Alert System for Food and Feed (RASFF) database: Data analysis during the period 2011–2015. *Food Control*, 93: 241–250.

Boatemaa, S.; Barney, M.; Drimie, S.; Harper, J.; Korsten, L.; and Pereira, L. 2019. Awakening from the listeriosis crisis: Food safety challenges, practices and governance in the food retail sector in South Africa. *Food Control*, 104: 333–342.

Bouzembrak, Y.; and Marvin, H. J. 2016. Prediction of food fraud type using data from Rapid Alert System for Food and Feed (RASFF) and Bayesian network modelling. *Food Control*, 61: 180–187.

Dasgupta, T.; Saha, R.; Dey, L.; and Naskar, A. 2018. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue*, 306–316.

Deng, X.; Cao, S.; and Horn, A. L. 2021. Emerging applications of machine learning in food safety. *Annual Review of Food Science and Technology*, 12: 513–538.

Deng, X.; den Bakker, H. C.; and Hendriksen, R. S. 2016. Genomic epidemiology: whole-genome-sequencing–powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annual review of food science and technology*, 7: 353–374.

Gupta, A.; Nelson, J. M.; Barrett, T. J.; Tauxe, R. V.; Rossiter, S. P.; Friedman, C. R.; Joyce, K. W.; Smith, K. E.; Jones, T. F.; Hawkins, M. A.; et al. 2004. Antimicrobial resistance among campylobacter strains, United States, 1997–2001. *Emerging infectious diseases*, 10(6): 1102.

Hall, A. J.; Wikswo, M. E.; Manikonda, K.; Roberts, V. A.; Yoder, J. S.; and Gould, L. H. 2013. Acute gastroenteritis surveillance through the national outbreak reporting system, United States. *Emerging infectious diseases*, 19(8): 1305.

Harris, J. K.; Hawkins, J. B.; Nguyen, L.; Nsoesie, E. O.; Tuli, G.; Mansour, R.; and Brownstein, J. S. 2017. Research brief report: Using twitter to identify and respond to food poisoning: The food safety stl project. *Journal of Public Health Management and Practice*, 23(6): 577.

Jana, S.; Sinha, M.; and Dasgupta, T. 2024. FORCE: A Benchmark Dataset for Foodborne Disease Outbreak and Recall Event Extraction from News. In *Proceedings of The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, 163–169.

Kase, J. A.; Zhang, G.; and Chen, Y. 2017. Recent foodborne outbreaks in the United States linked to atypical vehicles—lessons learned. *Current Opinion in Food Science*, 18: 56–63.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Lüth, S.; Boone, I.; Kleta, S.; and Al Dahouk, S. 2019. Analysis of RASFF notifications on food products contaminated with Listeria monocytogenes reveals options for improvement in the rapid alert system for food and feed. *Food Control*, 96: 479–487.

Maharana, A.; Cai, K.; Hellerstein, J.; Hswen, Y.; Munsell, M.; Staneva, V.; Verma, M.; Vint, C.; Wijaya, D.; and Nsoesie, E. O. 2019. Detecting reports of unsafe foods in consumer product reviews. *JAMIA open*, 2(3): 330–338.

Moumni Abdou, H.; Dahbi, I.; Akrim, M.; Meski, F. Z.; Khader, Y.; Lakranbi, M.; Ezzine, H.; and Khattabi, A. 2019. Outbreak investigation of a multipathogen foodborne disease in a training institute in Rabat, Morocco: case-control study. *JMIR public health and surveillance*, 5(3): e14227.

Nerín, C.; Aznar, M.; and Carrizo, D. 2016. Food contamination during food process. *Trends in food science & technology*, 48: 63–68.

Pádua, I.; Moreira, A.; Moreira, P.; de Vasconcelos, F. M.; and Barros, R. 2019. Impact of the regulation (EU) 1169/2011: Allergen-related recalls in the rapid alert system for food and feed (RASFF) portal. *Food control*, 98: 389–398.

Potter, A.; Murray, J.; Lawson, B.; and Graham, S. 2012. Trends in product recalls within the agri-food industry: Empirical evidence from the USA, UK and the Republic of Ireland. *Trends in food science & technology*, 28(2): 77–86.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108.

Scallan, E.; and Mahon, B. E. 2012. Foodborne Diseases Active Surveillance Network (FoodNet) in 2012: a foundation for food safety in the United States. *Clinical Infectious Diseases*, 54(suppl_5): S381–S384.

Swaminathan, B.; Barrett, T. J.; Hunter, S. B.; Tauxe, R. V.; and Force, C. P. T. 2001. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerging infectious diseases*, 7(3): 382.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Zhou, Z.; Alikhan, N.-F.; Mohamed, K.; Fan, Y.; Achtman, M.; Brown, D.; Chattaway, M.; Dallman, T.; Delahay, R.; Kornschober, C.; et al. 2020. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome research*, 30(1): 138–152.

System Prompt Template

Here, we have shown zero-shot and few-shot prompt templates used for two different subtasks. In our experience, using few-shot template for classification in a biased dataset does not return output. So, for Subtask-1, we have used the zero-shot template provided in A.1. But from the experiments, we have derived that providing examples in prompt gives more context to LLMs in entity event extraction, resulting similar to gold-standard outputs. So, for Subtask-2, we have used few-shot prompt template (Ref: A.2).

So, if the given input document `(doc id: 5002 in the Test Dataset)` is: *Oct. 28 update: As of Tuesday night, Oct. 27, the number of outbreak-associated cases of Shigella reported to the Santa Clara County Public Health Department stands at 190. There are 151 Santa Clara County residents and 39 people who live in other counties among those sickened. Of the 190 total cases, 92 are lab-confirmed (72 of whom are Santa Clara County residents). There are 20 confirmed cases from other jurisdictions, including the counties of San Mateo, Alameda, Santa Cruz, Marin and Merced. Nearly all of those sickened have reported that they ate at Mariscos San Juan 3 restaurant in downtown San Jose, CA, on Friday, Oct. 16, or Saturday, Oct. 17, 2015. Oct. 26 update: The Shigella outbreak linked to the Mariscos San Juan restaurant in San Jose, CA, is now estimated to involve 182 people, with new cases being reported from Marin and Merced counties. Seventy-two of the cases have been laboratory-confirmed. The downtown seafood restaurant is still closed, and the investigation is said to be focusing on an employee who may have been responsible for the situation. Oct. 22 update: According to the Santa Clara County Health Department, there are now 110 people with illnesses linked to the Mariscos San Juan seafood restaurant in downtown San Jose, CA. Santa Clara County Director of Public Health Dr. Sara Cody said Thursday that the illnesses are beginning to level off. Previous coverage follows: Mariscos San Juan at 205 N. 4th St. in downtown San Jose, CA, was closed Oct. 18 after the Santa Clara County Public Health Department connected the seafood restaurant with an outbreak of Shigella that reportedly may have sickened at least 80 people. The restaurant remains closed, and Santa Clara health officials say 11 of the Shigella victims have been treated in intensive care units at area hospitals. All of those stricken with the intestinal infection that causes fever, abdominal pain and diarrhea report dining at the San Jose restaurant on either the previous Friday or Saturday. Shigella is a pathogen that usually can be treated successfully with antibiotics. Local health officials said Monday that they expect the number of those sickened will grow, and they issued a request for action by all clinicians in the area. Clinicians treating suspected Shigella patients are being asked to test stool cultures and order antimicrobial susceptibility testing and blood cultures if the person is hospitalized. Doctors were also asked to "tailor therapy based on results of susceptibility testing, recognizing that routine antimicrobial susceptibility tests for Shigella may not include some commonly available oral antibiotics." Area emergency rooms were reporting they were treating multiple patients with vomiting and fevers as high as 104 degrees F. The downtown Mariscos San Juan is one of the restaurant chain's three locations in San Jose. The Willow Street restaurant had its permit suspended in August. The third location on Senter Road in San Jose remains open. Santa Clara County has suspended the permits of 81 restaurants during the past six months for a variety of code violations.* According to our prompt, the output for LLaMA 3.1 8B for Subtask 1 will be: *Foodborne Disease Outbreak*. For Subtask 2, the extracted entities will be: 1) Location: *Nationwide*; 2) Organization: *Mariscos San Juan*; 3) Product: *N/A*; 4) Cause: *An employee who may have been responsible for the situation*; 5) Disease: *Shigella infection* and 6) Number of People Affected: *190*.

A.1) Zero-shot Prompt Template given to LLaMA 3.1 for Subtask 1

To illustrate the prompting strategy used for multi-class classification for Subtask 1, we provide the following prompt template.

> **You are a helpful assistant, you always only answer FROM THE GIVEN CONTEXT.**
> Foodborne diseases and food recalls are major public health concerns, with outbreaks causing significant illness, and recalls aiming to prevent further harm to consumers.
> A document is labeled as "Food Recall" if the document reported events of a product recall; "Foodborne Disease Outbreak" if the document mentions events that report an foodborne disease outbreak or "NEUTRAL"- in case none of the above factors hold. Detect foodborne disease outbreaks and recall events from news articles, particularly those in the form of FDA press releases.
> Classify the news article into one of three categories: Food Recall, Foodborne Disease Outbreak, or Neither.

A.2) Few-shot Prompt Template given to LLaMA 3.1 for Subtask 2

To illustrate the few-shot prompting strategy used for entity and event extraction, we provide the following prompt template.

> Foodborne diseases and food recalls are major public health concerns, with outbreaks causing significant illness, and recalls aiming to prevent further harm to consumers. Extract *required entity* from the given text.
> For example if the given news is: `EXAMPLE NEWS`.
> Then the *extracted entity* will be e_1.
> The given news is:
> $< NEWS >$
> Return the extracted *extracted entity* from the given news document. Do not return anything else. If the information is not available say: "N/A".