BitsAndBites at SemEval-2025 Task 9: Improving Food Hazard Detection with Sequential Multitask Learning and Large Language Models

BitsAndBites at SemEval-2025 Task 9: Improving Food Hazard Detection with Sequential Multitask Learning and Large Language Models

Aurora Gensale[1,2], **Irene Benedetto**[1,3], **Luca Gioacchini**[3],
Luca Cagliero[1], **Alessio Bosca**[3]
[1] Politecnico di Torino, Italy, name.surname@polito.it
[2] Drivesec S.r.l., Italy, agensale@drivesec.com
[3] JAKALA S.p.A, Italy, name.surname@jakala.com

Abstract

Automatic and early detection of foodborne hazards is crucial for preventing foodborne outbreaks. Existing AI-based solutions often cannot handle complexity and noise in food recall reports and they struggle to overcome the dependency between product and hazard labels. We introduce a methodology for classifying reports on food-related incidents that addresses these challenges. Our approach leverages LLM-based information extraction, to minimize report variability, along with a two-stage classification pipeline. The first model assigns coarse-grained labels that narrow the space of eligible fine-grained labels for the second model. This sequential process allows us to capture hierarchical label dependencies between products and hazards and between their respective categories. Additionally, we designed each model with two classification heads that rely on the inherent relations between food products and associated hazards. We validate our approach on two multi-label classification sub-tasks. Experimental results demonstrate the effectiveness of our approach, which achieves an improvement of +30% and +40% in classification performance compared to the baseline.

1 Introduction

Food hazard detection — identifying potential risks associated with food products — is pivotal for public health. In this context, researchers have started exploring solutions based on traditional machine learning (ML) and deep learning (DL) techniques to automate food hazard detection tasks. This can help mitigate foodborne outbreaks and improve food safety measures (Zhou et al., 2019; Qian et al., 2023).

Albeit promising, such approaches leverage structured data extracted from incomplete or misleading information collected from social media (Maharana et al., 2019; Tao et al., 2023). More interestingly, natural language processing (NLP) techniques powered by large language models (LLMs) have unlocked new possibilities—especially in scenarios with limited labeled data (such as low-resource languages (Perak et al., 2024; Koudounas et al., 2023) or specific context (Pal et al., 2024; Benedetto et al., 2023)).

They enable the extraction of more robust and context-enhanced information from unstructured data when it relies on authoritative sources such as public reports from government agencies (Özen et al., 2025). This allows for more reliable and comprehensive analysis of food hazard trends and their potential impact.

Among others, Randl et al. (2024) have introduced a dataset of publicly available food recall announcements, annotated at two hierarchical levels — i.e., high-level categories and fine-grained labels for both food products and associated hazards. The authors use this dataset to benchmark a food hazard detection methodology for addressing a multi-label report classification task. While the reports included in the dataset provide authoritative insights, they inherently present noise and span thousands of classes, which poses significant analytical challenges.

Relying on the work of Randl et al. (2024), we hypothesize that information about hierarchical structure can enhance a model's classification performance. Specifically, classifying broader product categories first can facilitate subsequent identification of specific food items, and the same applies to hazard classification. Moreover, while the relationship between food products and associated hazards is highly correlated, existing approaches fail to explicitly model these dependencies (Randl et al., 2024).

In this paper, we address the report classification task in a multitask fashion. We propose a novel approach based on sequential multi-head classification. We present three key advances in multi-label food hazard detection:

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 718–725

1. *Multi-Head Architecture:* We decouple product and hazard prediction by leveraging two classification heads, which enables specialized feature learning for each label type.

2. *Hierarchical Constraint Mechanism:* We first predict macro-categories to dynamically restrict fine-grained class probabilities, leveraging label hierarchy to improve accuracy.

3. *LLM-Driven Corpus Normalization:* We apply LLM information extraction to standardize report texts, thus reducing variability and noise prior to classification.

The classification results on The Food Hazard Detection Challenge (Randl et al., 2025) dataset validate the effectiveness of our approach. Our pipeline achieves an F-score of 0.80 for product classification and an F-score of 0.47 for hazard classification. The Multi-Head approach accounts for the largest improvement in performance, adding an absolute F1 of +0.30 on product classification and an absolute F1 of +0.46 on hazard classification to the single-head baselines. Corpus normalization contributes an additional +0.01 F1 improvement by reducing text variability. Enforcing the hierarchical constraints at the Sequential Classification stage yields a marginal +0.005 F1 gain. Additionally, our approach ranks in the top 15 of the public leaderboard ("title and text" tracks), reaching 6th place in ST1 and 13th place in ST2[1].

2 Background

In the last decade, researchers have focused their efforts on exploring AI-based solutions for food hazard detection (Zhou et al., 2019; Qian et al., 2023). Most of the existing research rely on traditional ML (Kumar et al., 2024) and DL techniques (Xiong et al., 2023), from the detection of zoonotic disease sources (Lupolova et al., 2017) to microbial risk assessment (Njage et al., 2019), to name but a few. Nevertheless, the recent development of LLMs and the advancement of NLP techniques (Zhao et al., 2024) have pushed the boundaries towards more sophisticated approaches (Özen et al., 2025; Prabhune et al., 2025; Randl et al., 2024).

Although recent studies have started leveraging insights from the scientific literature (Xiong et al., 2023; Özen et al., 2025), the majority of food risk detection approaches rely on corpora consisting of news or social media posts (Tao et al., 2023; Maharana et al., 2019). Such sources often provide incomplete information and lack precision from both a taxonomical and scientific perspective, making it challenging to extract structured and reliable data for AI-driven food risk assessment (Randl et al., 2024).

The majority of existing approaches frame the problem as a binary classification task where the goal is to detect the presence of an incident from tabular or image data (Wang et al., 2022). Such approaches are promising but are often too simplistic for real-world scenarios (Hu et al., 2022) where food risk assessment requires complete interpretation of textual data, consideration of context, and distinguishing between different levels of risk severity (Danezis et al., 2016; Prache et al., 2022).

To overcome these issues, Randl et al. (2024) created a new dataset of $>$ 6000 food recall announcements from 24 public food safety authority websites spanning 28 years from 1994 to 2022. They formulated the food hazard detection problem as two supervised multi-label classification tasks and organized the collected reports accordingly: (i) the aim of subtask ST_1 is the classification of each report into macro categories of food products (22 labels) and related hazards (10 labels); (ii) the aim of subtask ST_2 is identification of the specific products (1 142 labels) and hazards (128 labels) mentioned in the reports.

Randl et al. (2024) relied on that dataset to validate a methodology based on LLMs and conformal prediction (Vovk et al., 2022). They addressed the two classification subtasks separately by training two different classifiers, each designed with a single classification head that simultaneously processed products and their associated hazards. While promising, this design may limit the model's ability to distinguish between the different aspects of each target.

In contrast, we propose a modification to this architecture by splitting the classification head into two distinct heads, which allows the model to better capture the relationship between food products and potential hazards.

Moreover, while Randl et al. (2024) addressed the two subtasks independently, we introduce a sequential classification approach (see Section 3), where we constrain the classification probabilities of the detailed labels (i.e., subtask ST_2) to the probabilities of the category labels (i.e., subtask ST_1).

[1]Code available at https://github.com/auro736/BitsAndBites_SemEval2025_Task9.

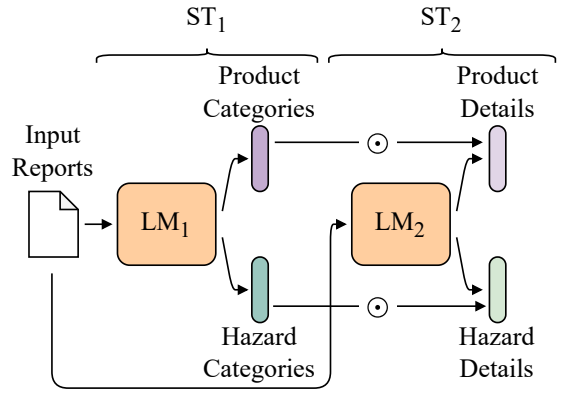


Figure 1: Overview of the adopted methodology. For each subtask (ST_1 and ST_2), we train one classification model with two classification heads: one for classifying the product labels, and one for classifying the food hazard. Then, we constrain the probabilities of the ST_2 detailed labels based on the probabilities of the ST_1 generic categories.

This enables the model to refine its predictions by leveraging the hierarchical dependency between the two tasks, ultimately improving both the accuracy and robustness of food hazard detection.

3 Methodology

In this section, we present our food hazard detection methodology. In Section 3.1, we address the overall problem by relying on a multi-head architecture. This results in two models — one for high-level categories (ST_1) and one for fine-grained labels (ST_2). Then, in Section 3.2, we introduce the sequential classification approach, where fine-grained predictions are guided and constrained by the predictions of the broader categories. Finally, in Section 3.3, we describe how we leverage LLMs to normalize the noisy, unstructured reports through summarization and information extraction, which also enhances classification performance.

3.1 Multi-Head Architecture (MH)

Given the possible correlations between food products and their associated hazards (Randl et al., 2024), we have opted for a double classification head approach.

For each subtask, we split the classification layer of the LM into two classification heads. This way, part of the model parameters are shared across the two subtasks while each classification head is specialized in a different classification subtask (see the green and purple blocks of Figure 1).

As a consequence, for a single subtask we obtain two loss functions, i.e., L_P for the product classification head and L_H for the hazard classification head. We jointly train the two classification heads with a linear combination of the two loss functions. The resulting loss function is $L = \lambda_P \cdot L_P + \lambda_H \cdot L_H$, where $\lambda_P, \lambda_H \in \mathbb{R}$ are multiplicative coefficients to balance the contributions of the head-specific losses.

3.2 Sequential Classification (SC)

We define the set of hazard/product categories $\mathcal{C} = \{c_1, \ldots, c_C\}$ from ST_1 and the set of detailed hazards/products as $\mathcal{D} = \{d_1, \ldots, d_D\}$ from ST_2. As mentioned in Section 2, the dataset exhibits a hierarchical structure between ST_1 and ST_2, i.e., given a hazard/product category $c_i \in \mathcal{C}$, there exists a subset of details $\mathcal{D}_i \in \mathcal{D}$ associated with c_i.

We train two classifiers independently, each tailored to their respective subtask — i.e., LM_1 for ST_1 and LM_2 for ST_2, as highlighted in Figure 1.

First, in ST_1 we leverage LM_1 to predict the probabilities of all the hazard/product categories of an input report. Hence, we assign the report to the hazard/product category with the highest probability, formally $\arg\max_{c_i \in \mathcal{C}} p_{c_i}$, where p_{c_i} is the probability of category c_i.

Then, in ST_2, we exploit the hierarchical relationship between categories and their associated details by weighting the probability of each detail (p_{d_j}) by the probability of its corresponding category (p_{c_i}). Rather than considering a single category, we propagate all probabilities from ST_1 into their corresponding detailed hazard/product probabilities in ST_2:

$$\hat{p}_{d_j} = p_{d_j} \cdot p_{c_i}, \ \forall d_j \in \mathcal{D}_i \ , \ \forall c_i \in \mathcal{C}$$

Hence, we consider the final detailed prediction with the maximum probability, formally $\arg\max_{d_j \in \mathcal{D}} \hat{p}_{d_j}$.

This ensures that the predictions for detailed hazards/products are influenced by the category-level classification, thus maintaining hierarchical consistency between ST_1 and ST_2. As a result, this sequential classification approach should enhance the accuracy and consistency of predictions by restricting LM_2 to only relevant details.

3.3 Corpus Normalization (CN)

The texts included in the dataset follow different formats and structures depending on the type of

	ST_1		ST_2	
	Validation	**Test**	**Validation**	**Test**
Baseline	0.4932	0.4722	0.0031	0.0037
MH	0.7893	**0.7998**	0.4777	0.4644
MH + SC	–	–	**0.4825**	0.4693
MH + CN	**0.8020**	0.7817	0.4813	**0.4700**
MH + CN + SC	–	–	0.4802	0.4681

Table 1: Comparison of classification results across the proposed approaches: sequential classification (SC) and corpus normalization (CN). The experiments are conducted using RoBERTa-large, identified as the best-performing model. The best results are highlighted in **boldface**.

report, the country, the government agency, or the website from which they were extracted. This poses significant challenges for a classifier based on an LM.

To address this issue, we reduce the reports' variability and noise by leveraging another LLM in a zero-shot setting to extract specific information in a uniform and fixed format. Hence, we obtain the final report by prepending the extracted text to the original report.

Below, we provide the prompt template we used to normalize the report:

```
You are an expert in analyzing food-related incident
reports. For the given text, identify the recalled food
product and the motivation for the recall. Add also the
categories that you can infer of the food product and the
motivation.
Provide the output in the following format:
PRODUCT: <food product and its category extracted>
HAZARD: <motivation and its category extracted>
Do not include any additional explanation or output. Follow
the format strictly.
```

4 Experimental Setup

When not explicitly stated, we ran our experiments on the training, validation and test datasets[2] released by the "Food Hazard Detection Challenge" (Randl et al., 2025).

We used RoBERTa-large as the best model among BERT-uncased-large, DeBERTa-v3-large and ModernBERT-large evaluated during a model selection stage (see Appendix A). We used sequence cross-entropy as the loss function as well as the AdamW optimizer (Loshchilov and Hutter, 2017) with a weight decay of 0.01 and a learning rate of 10^{-5}. To prevent overfitting, we remove connections in the last classification layer with a probability of 0.1.

After a hyperparameters tuning stage, we set the coefficients of the multi-head architecture to $\lambda_P = 0.5$ and $\lambda_H = 0.5$ for subtask ST_1, whereas we set $\lambda_P = 1.0$ and $\lambda_H = 1.5$ for subtask ST_2 (please refer to Appendix B for further details on loss coefficient tuning). We extracted the structured information from the reports (CN) with Meta-Llama-3.1-8B-Instruct in zero-shot fashion. We ran all the experiments on a Ubuntu 22.04 Long Term Support (LTS) machine equipped with Intel® Xeon™ Gold 6126 CPU, 1 × Nvidia® RTX 6000 graphics processing unit (GPU), 24 gigabyte (GB) of random access memory (RAM).

Evaluation Metric We evaluated our approach through the *task F1-score* evaluation metric proposed by the organizers of the "SemEval" challenge (Randl et al., 2025). This metric is a customized version of the traditional F1-score accounting for the relation between food products and associated hazards. We provide its implementation in the following code snippet in Python:

```python
from sklearn.metrics import f1_score

def task_f1_score(H_true, P_true, H_pred, P_pred):
  # Compute F1 for hazards:
  H_f1 = f1_score(H_true, H_pred, average='macro')
  # Constraint the products on the predicted hazards
  P_true = P_true[H_pred == H_true]
  P_pred = P_pred[H_pred == H_true]
  # Compute F1 for products:
  P_f1 = f1_score(P_true, P_pred, average='macro')
  # Compute the final task F1-score
  return (H_f1 + P_f1) / 2
```

In a nutshell, we first evaluated the F1-score of the predicted hazards with macro averaging to account for labeling unbalances. Then, we constrained the evaluation to only those instances where the predicted and ground truth hazards align. Within this subset, we then computed the macro average F1-score for the associated food products. Finally, we computed the final task F1-score by averaging the product and hazard scores. This ensures that both hazard detection and product association are jointly considered in the evaluation.

[2]We use the *title* and *text* features from the datasets to create a single concatenated text.

5 Experimental Results

In Table 1, we provide the task F1-scores for the two classification tasks obtained with our approach[3].

We use as baseline a standard classification model based on RoBERTa-large and a single classification head for each task — i.e., ST_1 and ST_2.

Firstly, using a single classification head limits the baseline performance, with a task F1-score of $< 50\%$ in ST_1. This limitation is even more evident in ST_2, where the task F1-score is $< 1\%$. Here, further analysis reveals that merging the product and hazard labels in a unique classification head leads to a strong bias for the predominant class — i.e., the products. As a consequence, 89% of the hazards are misclassified as products.

The multi-head (MH) classification brings a substantial improvement over the baseline in both ST_1 and ST_2, with improvements in the task F1-scores of $> 39\%$ and $> 46\%$ respectively. The classification results confirm our assumption that splitting the hazard and product classification heads allows the classifier to achieve high specialization while accounting for label unbalancing.

Overall, corpus normalization (CN) applied along with MH leads to a performance comparable to simple MH, except for a slight decrease in the test task F1-score of $\sim 2\%$ for ST_1. Apart from this specific case, the information extracted by the LLM follows a uniform and fixed structure, effectively reducing the reports' variability and facilitating the classifier's handling of heterogeneous data. As a result, with CN the classifiers can correctly assign over 100 more reports to the correct products and hazards compared to using the MH alone.

Leveraging the hierarchical structure of labels (i.e., categories and details) through sequential classification (SC) applied alongside simple MH pays off, resulting in a slight task F1-score improvement of $\sim 0.5\%$. On the other hand, SC seems to slightly limit the improvement of the CN approach. Despite the classifier performing comparably with a task F1-score of around 48%, combining the three approaches leads to a slight decrease in performance of $\sim 0.1\%$.

[3]Note that the results presented in this table differ from those in the public competition leaderboard. Here, we have revised and refined the methodology to achieve greater stability and robustness across different model configurations.

6 Conclusions

In this paper, we proposed a novel sequential multi-head classification approach to classify food-related incident reports. We introduced a classification pipeline that integrates (i) multi-head classifiers to split food products and associated hazard labels, (ii) a sequential classification strategy leveraging hierarchical labels, and (iii) LLM information extraction for normalizing reports that exhibit high variability.

Experimental results demonstrate the efficacy of our approach, which yields significant performance improvements over the baseline single-head classifier. The multi-head approach substantially enhances the classifier's performance, mitigating the biases caused by labeling imbalances observed in the baseline model. Corpus normalization reduces report variability and provides a common structure to the texts, thereby slightly improving the classification performance. Finally, sequential classification marginally boosts performance by constraining predictions based on hierarchical label dependencies.

Future work could explore alternative approaches to hierarchical classification constraints, further optimizing the balance between interpretability and performance. Additionally, integrating external knowledge could enhance model robustness and generalization across different food safety scenarios. We also plan to evaluate the use of the embeddings produced by our multi-head, hierarchy-based approach to index incident reports in a retrieval-augmented generation (RAG) framework to enable more efficient retrieval and richer contextualization of historical cases.

Limitations and Ethical Statement

The dataset used in this study, to the best of our knowledge, does not contain any personal information. However, it may include potentially harmful or inappropriate content. This consideration also extends to the model employed, which may generate incorrect responses. The use of this particular dataset and models is subject to the limitations outlined in their respective technical reports and licenses.

Our approach depends on the quality and comprehensiveness of the dataset used. Although it consists of authoritative food recall reports, the dataset may still contain inconsistencies or outdated information, and the performance of the model may vary

due to the presence of other kinds of data. As such, the generalizability of our approach to other food safety datasets or real-world scenarios remains an open question requiring further validation on different food safety records.

Acknowledgments

This research has been carried out by the SmartData@PoliTO center for big data technologies, the HPC@POLITO academic computing center and JAKALA S.p.A. This study was partially carried out within Future Artificial Intelligence Research (FAIR) and received funding from the European Union's NextGenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013), as well as from the European Union's Horizon Europe research and innovation program Extreme Food Risk Analytics (EFRA) (Grant Agreement Number 101093026). This manuscript reflects only the authors' views and opinions, and neither the European Union nor the European Commission can be considered responsible for them.

References

Irene Benedetto, Luca Cagliero, Francesco Tarasconi, Giuseppe Giacalone, and Claudia Bernini. 2023. Benchmarking abstractive models for italian legal news summarization. In *International Conference on Legal Knowledge and Information Systems*.

Georgios P. Danezis, Aristidis S. Tsagkaris, Federica Camin, Vladimir Brusic, and Constantinos A. Georgiou. 2016. Food authentication: Techniques, trends & emerging approaches. *TrAC Trends in Analytical Chemistry*, 85:123–132.

Ruofan Hu, Dongyu Zhang, Dandan Tao, Thomas Hartvigsen, Hao Feng, and Elke Rundensteiner. 2022. TWEET-FID: An Annotated Dataset for Multiple Foodborne Illness Detection Tasks. *arXiv preprint*. ArXiv:2205.10726 [cs].

Alkis Koudounas, Flavio Giobergia, Irene Benedetto, Simone Monaco, Luca Cagliero, Daniele Apiletti, Elena Baralis, et al. 2023. baρtti at geolingit: Beyond boundaries, enhancing geolocation prediction and dialect classification on social media in italy. In *CEUR Workshop Proceedings*. CEUR.

Yogesh Kumar, Inderpreet Kaur, and Shakti Mishra. 2024. Foodborne Disease Symptoms, Diagnostics, and Predictions Using Artificial Intelligence-Based Learning Approaches: A Systematic Review. *Archives of Computational Methods in Engineering*, 31(2):553–578.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization.

Nadejda Lupolova, Tim J. Dallman, Nicola J. Holden, and David L. Gally. 2017. Patchy promiscuity: machine learning applied to predict the host specificity of Salmonella enterica and Escherichia coli. *Microbial Genomics*, 3(10):e000135. Publisher: Microbiology Society,.

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.

Patrick Murigu Kamau Njage, Clementine Henri, Pimlapas Leekitcharoenphon, Michel-Yves Mistou, Rene S. Hendriksen, and Tine Hald. 2019. Machine Learning Methods as a Tool for Predicting Risk of Illness Applying Next-Generation Sequencing Data. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, 39(6):1397–1413.

Ankit Pal, Pasquale Minervini, Andreas Geert Motzfeldt, Aryo Pradipta Gema, and Beatrice Alex. 2024. Open medical llm leaderboard. `https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard`.

Benedikt Perak, Slobodan Beliga, and Ana Meštrović. 2024. Incorporating dialect understanding into LLM using RAG and prompt engineering techniques for causal commonsense reasoning. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 220–229, Mexico City, Mexico. Association for Computational Linguistics.

Akash Prabhune, Vinay Sri Hari, Neeraj Kumar Sethiya, and Mansi Gauniyal. 2025. Utilising consumer reviews for passive surveillance of foodborne illnesses: insights and challenges from the Indian restaurant. *International Journal of Public Health*, 14(1):479–492.

S. Prache, C. Adamiec, T. Astruc, E. Baéza-Campone, P. E. Bouillot, A. Clinquart, C. Feidt, E. Fourat, J. Gautron, A. Girard, L. Guillier, E. Kesse-Guyot, B. Lebret, F. Lefèvre, S. Le Perchec, B. Martin, P. S. Mirade, F. Pierre, M. Raulet, D. Rémond, P. Sans, I. Souchon, C. Donnars, and V. Santé-Lhoutellier. 2022. Review: Quality of animal-source foods. *Animal*, 16:100376.

C. Qian, S. I. Murphy, R. H. Orsi, and M. Wiedmann. 2023. How Can AI Help Improve Food Safety? *Annual Review of Food Science and Technology*, 14(Volume 14, 2023):517–538. Publisher: Annual Reviews.

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. ArXiv:2403.11904 [cs].

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. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2022. *Algorithmic Learning in a Random World*. Springer International Publishing, Cham.

Xinxin Wang, Yamine Bouzembrak, AGJM Oude Lansink, and H. J. van der Fels-Klerx. 2022. Application of machine learning to the monitoring and prediction of food safety: A review. *Comprehensive Reviews in Food Science and Food Safety*, 21(1):416–434. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1541-4337.12868.

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):e17806.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A Survey of Large Language Models. *arXiv preprint*. ArXiv:2303.18223 [cs].

Lei Zhou, Chu Zhang, Fei Liu, Zhengjun Qiu, and Yong He. 2019. Application of Deep Learning in Food: A Review. *Comprehensive Reviews in Food Science and Food Safety*, 18(6):1793–1811. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1541-4337.12492.

Neris Özen, Wenjuan Mu, Esther D. van Asselt, and Leonieke M. van den Bulk. 2025. Extracting chemical food safety hazards from the scientific literature automatically using large language models. *Applied Food Research*, 5(1):100679.

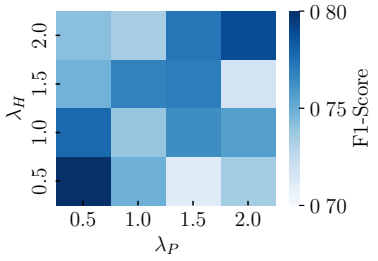
A Model Selection

Here, we complement our experimental results by reporting the model selection task F1-Scores. We organize our dataset relying only on the challenge training data, properly split into training (70%), validation (15%) and test sets (15%).

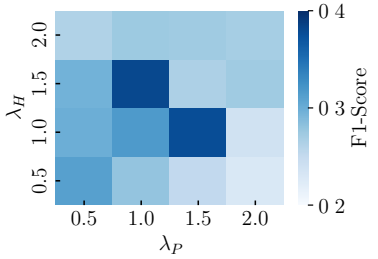
We choose the best model among BERT-uncased-large, RoBERTa-large, DeBERTa-v3-large, ModernBERT-large. We train each model for a maximum of 10 epochs with early stopping and the same experimental settings of Section 4. Appendix A showcases the classification results.

	Task ST_1		Task ST_2	
	Validation	**Test**	**Validation**	**Test**
ModernBERT	0.6379	0.6591	0.2368	0.2223
BERT-uncased	0.6735	0.6866	0.2634	0.2472
DeBERTa-v3	0.7393	0.6741	0.2291	0.2022
RoBERTa	**0.7487**	**0.7394**	**0.3315**	**0.3175**

Table 2: Model selection classification scores for the two subtasks ST_1, ST_2 with sampled data. Best results are in **bold**.



(a) Task ST_1: Classification of categories



(b) Task ST_2: Classification of details

Figure 2: Task F1-Scores with different values of λ_P and λ_H.

.

B Choice of λ_P and λ_H

Here, we complement our experimental results by reporting the results of the hyperparameter tuning stage for λ_P and λ_H introduced in Section 3.1. As for the model selection, we organize our dataset relying only on the challenge training data, properly split into training (70%), validation (15%) and test sets (15%).

We use RoBERTa-large as the best model resulting from the model selection stage and train it for a maximum of 10 epochs with early stopping. We test λ_P and λ_H values in the range from 0.5 to 2.0 and report in Figure 2 the task F1-Scores.