MINDS at SemEval-2025 Task 9: Multi-Task Transformers for Food Hazard Coarse-Fine Classification

Flavio Giobergia
Politecnico di Torino
Turin, Italy
`flavio.giobergia@polito.it`

Abstract

Food safety is a critical concern: hazardous incident reports need to be classified to be able to take appropriate measures in a timely manner. The SemEval-2025 Task 9 on Food Hazard Detection aims to classify food-related incident reports by identifying both the type of hazard and the product involved, at both coarse and fine levels of granularity. In this paper, we present our solution that approaches the problem by leveraging two independent encoder-only transformer models, each fine-tuned separately to classify hazards and food products, at the two levels of granularity of interest. Experimental results show that our approach effectively addresses the classification task, achieving high-quality performance on both subtasks. We additionally include a discussion on potential improvements for future iterations, and a brief description of failed attempts. We make the code available at `https://github.com/fgiobergia/SemEval2025-Task9`.

1 Introduction

The success of Language Models has made it possible to annotate datasets with very limited human intervention. This is the case for a wide variety of tasks, including some with peculiar domains that make it difficult to obtain high-quality labels manually (e.g., classification of legal documents (Shaheen et al., 2020), or dialect detection (Koudounas et al., 2023)). This trend has enabled a thorough analysis of documents that could not be reasonably processed in acceptable times. Among these documents, there are life-critical ones such as the analysis of food-related incident reports (Randl et al., 2024). The Food Hazard Detection task (Randl et al., 2025) from SemEval 2025 focuses specifically on this challenge, with the goal of helping classify food incident reports collected from the web.

The proper classification of these incidents is a vital task, as it provides potentially life-saving insights. These insights are typically in the form of structured labels that indicate the type and severity of the hazard, such as contamination, mislabeling, or adulteration; as well as the specific, or category of food involved. Accurate classification enables regulatory bodies, food safety organizations, and the public to respond effectively by issuing warnings, recalling products, or implementing stricter safety measures.

The large quantities of incidents available online makes manual processing generally infeasible. As such, automated crawlers can be used to find food issues from web sources; whereas Natural Language Processing techniques can be adopted to correctly classify these documents based on (1) the type of food involved, and (2) the type of hazard described.

The task is focused on classifying incidents based on these two targets, at two different levels of granularity (coarse and fine). As a part of this paper, we note that there is limited correlation between the hazard and the type of food – as such, we address the task by proposing a solution based on the fine-tuning of two pretrained, encoder-only transformer-based models, that focus on different aspects of the problem. We show that the proposed solution achieves competitive performance, with a final ranking of 9th place for subtask 1, 5th place for subtask 2.

The rest of the paper is organized as follows. Section 2 describes the dataset and the task under study. We present the proposed approach, and the rationale for it, as a part of Section 3. The main results obtained are shown in Section 4, with some additional considerations on the choices made. We cover some of the failed attempted in Section 5 and the main limitations of the proposed approach in Section 6. Finally, Section 7 wraps up the paper with considerations on the task and solution.

2 Problem description

The goal of the task is to correctly classify incidents available on the web. The dataset is provided in three splits: training data (5,082 samples), validation data (565 samples) and test data (997 samples). Each sample corresponds to the description of an incident. These incidents are taken from official food agency websites (e.g., FDA). For each incident, a set of attributes is known: some are structured (date and country); whereas others are textual (title and text). For training and validation data, the correct classification is also available. This classification consists of two attributes: the *hazard category* (10 classes) and the *product category* (22 classes). Each of these categories is further refined into a specific *hazard* (128 classes) and *product* (1,142 classes). Each text is labeled by two food science or food technology experts. The first sub-task (ST1) of the challenge consists in predicting the hazard and product categories, whereas the second one (ST2) aims to predict the specific hazard and product. For convenience, we report in Table 1 an instance of an incident, with all available information.

For ST2, the main metric of interest is $F_1^{(ST2)} = (F_1^{(h)} + F_1^{(p|h)})/2$. Here, $F_1^{(h)}$ is the macro F_1 score for the hazard classification problem; whereas $F_1^{(p|h)}$ is the macro F_1 score, computed *only* on samples whose hazard has been correctly predicted. For ST1, we adopt a metric computed in a similar way, $F_1^{(ST1)} = (F_1^{(hc)} + F_1^{(pc|hc)})/2$, using the F_1 scores for the hazard and product categories. This choice of metrics places additional importance on the identification of the correct hazard: failure to do so results in a 0 score being achieved, regardless of the performance on the product identification problem.

3 Proposed methodology

The proposed solution consists of two separate encoder-only, pretrained transformers fine-tuned on the fine-coarse dual prediction problem.

We first discuss the rationale behind using two models, each working on a dual granularity. Next, we present the main details of the adopted solution.

3.1 Targets correlations

Multi-task learning allows to exploit useful information from related learning tasks (Zhang and Yang, 2018). Based on this knowledge, a promising
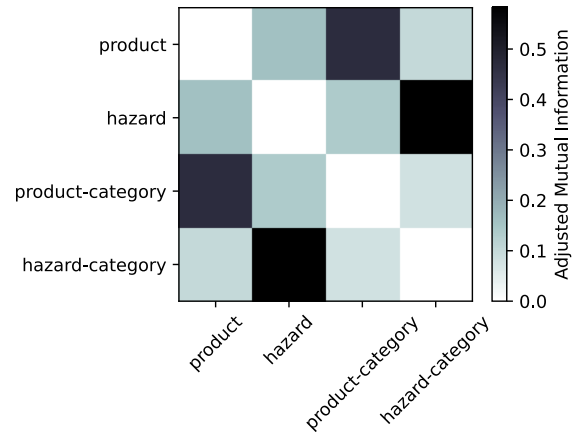

Figure 1: Adjusted Mutual Information between pairs of targets. Fine-coarse categories have high AMI, whereas different categories show lower correlation.

approach to improve a classic transformer-based classifier is to introduce a single backbone, with multiple tasks being aggregated into a single loss function. However, this approach only works if the four target categories are somewhat related. To quantify the relationships between the four targets, we compute the pair-wise Adjusted Mutual Information (AMI) between the targets. The AMI is a version of Mutual Information, which quantifies how mutually dependent (correlated) two variables are (i.e., how informative knowing one variable is in predicting the other one). We remark that some of the targets have a large number of possible classes (up to hundreds, or thousands). As a consequence, we make use of the Adjusted verison of MI, which accounts for random chance and the fact that a large number of clusters tends to produce a higher MI score. We report the pair-wise AMI as a part of Figure 1.

It is clear from this result that the strongest correlation exists between the fine and coarse versions of each target. This is to be expected, because of the hierarchical nature of the relationship. However, the figure also highlights a low correlation between the two targets (product, hazard). This is true for both fine-grained and coarse results. From a domain-agnostic perspective, this may be considered, in many cases, reasonable: a hazard (e.g., the presence of a foreign piece of plastic) is not necessarily related to the type of food affected. Based on these insights, we adopt two separate models: one addressing the coarse-fine prediction of products, the other addressing the coarse-fine prediction of hazards.

Table 1: Example for an incident report. In blue are the time/location metadata, in orange are the text information, in green are the four target classes. <u>Underlined</u> is the information useful to identify the product and product category, in *italic* is the information useful to identify the hazard and hazard category.

year	2014	**hazard-category**	allergens
month	5	**product-category**	ices and desserts
day	4	**hazard**	eggs and products thereof
country	us	**product**	ice cream
title	2013 - Blue Bunny Premium Bordeaux Cherry Chocolate <u>Ice Cream</u> Recalled for *Undeclared Allergen*		
text	Wells Enterprises, Inc., maker of Blue Bunny ice cream said today it has recalled Blue Bunny Premium Bordeaux <u>Cherry Chocolate Ice Cream</u> sold at retail grocery stores in Kansas, Indiana and Iowa because the product may contain *egg* not declared on the label.		

3.2 Adopted architecture

We adopt the same architecture for both models. We rely on a pretrained encoder-only transformer model (e.g., BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019)). We use, as the representation of each incident, the $title$ and the $text$. We tokenize the two as two distinct [SEP]-separated sentences (i.e., making use of Sentence A and B in BERT-like models). We define d as the encoder's hidden size, and n_f and n_c as the fine-grained and the coarse number of classes, respectively. We adopt the output for the [CLS] token as a summary vector $v \in \mathbb{R}^d$ of the entire incident,

$$v = encoder([CLS] \,||\, title \,||\, [SEP] \,||\, text) ,$$

where $\cdot||\cdot$ represents a concatenation operation. We use v as the input to two classifications heads: one for the fine-grained task, the other for the coarse task. Both classification heads are characterized by an initial $d \times d$ layer, followed by a non-linearity (ReLU) and a linear layer that projects the results into n_f- and n_c-dimensional outputs, thus producing the logits $\hat{y}_f \in \mathbb{R}^{n_f}$ and $\hat{y}_c \in \mathbb{R}^{n_c}$ for the two tasks. Assuming a ground truth y_c and y_f for the two problems, we define the multi-task loss function as the cross-entropies for the two granularities, with a scaling factor λ to regulate the weight between the two targets:

$$\mathcal{L} = \sum_i y_{n_c,i} log(\hat{y}_{n_c,i}) + \lambda \sum_j y_{n_f,j} log(\hat{y}_{n_f,j}) .$$

We separately build one model to predict hazards, and the other to predict products involved.

4 Experimental results

We report, as a part of this section, the main results obtained. First, we present an initial overview of the metrics reported. Then, we study the performance of the proposed pipeline, using different backbone models. We further study the results that would be obtained by framing the problem in different ways. Finally, we mention the main hyperparameters adopted for the solution.

4.1 Metrics

We use, as the main metric of interest, the average macro F_1 score – as reported in Section 2, i.e., $F_1^{(ST1)}$ and $F_1^{(ST2)}$. Although these metrics summarize the quality of the solution on the entire task, we additionally report the performance for each task, separately. Based on the large number of classes, and the heavy class imbalance, we choose the macro F_1 score, for each subtask, as the most suitable metric. For ST1, we report the F_1 score for the product and the hazard categories ($F_1^{(pc)}$ and $F_1^{(hc)}$, respectively); whereas for ST2, we report the F_1 score for the specific product and the hazard ($F_1^{(p)}$ and $F_1^{(c)}$, respectively).

4.2 Backbone selection

The proposed approach heavily relies on a valid selection of the backbone model used for the encoding of the incident text and title. A wide variety of encoder-only transformers exist in literature. Based on their popularity, we adopted three possible encoders: BERT (base, large) (Devlin et al., 2019) and RoBERTa (large) (Liu et al., 2019). We report the results obtained in Table 2.

2215

	Subtask 1 (coarse)			Subtask 2 (fine)		
	$F_1^{(hc)}$	$F_1^{(pc)}$	$F_1^{(ST1)}$	$F_1^{(h)}$	$F_1^{(p)}$	$F_1^{(ST2)}$
BERT base	**0.789 ± 0.004**	0.653 ± 0.004	0.721 ± 0.001	0.569 ± 0.013	0.241 ± 0.005	0.405 ± 0.007
BERT large	0.777 ± 0.009	<u>0.708 ± 0.002</u>	<u>0.743 ± 0.005</u>	**0.626 ± 0.006**	<u>0.305 ± 0.003</u>	<u>0.461 ± 0.006</u>
RoBERTa large	<u>0.783 ± 0.007</u>	**0.723 ± 0.005**	**0.754 ± 0.005**	<u>0.625 ± 0.010</u>	**0.337 ± 0.009**	**0.479 ± 0.003**

Table 2: Performance on the various classification problems, for the main solution proposed. Best results for each metric highlighted in **bold**, second best is <u>underlined</u>.

The results clearly show that RoBERTa achieves the best performance in terms of task-related metrics of interest; as well as the best performance for the product-related metrics. Interestingly, RoBERTa is the second-best performer for the hazard categories; with BERT obtaining better results.

4.3 Alternative tasks & baselines

In Section 3, we argue that the most promising multi-task approach appears to be the one with two separate models on fine-coarse targets. Producing a single 4-task solution did not appear to be promising, given the low Mutual Information between products and hazards. We empirically verify this claim, showing that building a single model, trained on 4 tasks, does provide any particular benefit.

For fairness of comparisons, all solutions are trained with the same computing budget, evenly distributed across models. The proposed solution uses 7 + 7 training epochs (7 for each model). As such, we train the single 4-task model for 14 epochs. We adopt the same 1:5 ratio of scaling factors between coarse and fine tasks, as it has provided the best results for the proposed solution.

The results are reported in Table 3. The results obtained are mostly comparable with those achieved by the RoBERTa-based proposed solution. Interestingly, RoBERTa achieves better performance on the fine-grained versions of the problem. The four-task version generally achieves slightly better performance on the coarse problems. Although the "best" result is obtained by using the dual-task to solve Subtask 2, and the four-task version to solve Subtask 1, we propose, for consistency, a single solution based on two dual-task models.

We also report results obtained for a baseline method, namely a Random Forest, trained on (1) the TF-IDF representation (Sparck Jones, 1972) of each document, or (2) the average word embeddings for each word contained in the title and text, using FastText (Bojanowski et al., 2017). Comput-

ing the average word vector (i.e., using distributed bags of words) is a commonly adopted approach with traditional word embeddings, as done in several works (Le and Mikolov, 2014; Giobergia et al., 2020; Reimers and Gurevych, 2019), despite losing the order among words. These baselines provide better context for the difficulty of the problem. The proposed approach significantly outperforms both. Interestingly, the TF-IDF version shows better performance than FT. We expect this to be the case due to the technical nature of the problem: without proper fine-tuning, the word embeddings cannot capture the domain-specific nuances of the problem.

4.4 Hyperparameters

We conducted a tuning phase to identify the best configuration of hyperparameters, by making use of the development set available. The best set of hyperparameters is reported in Table 4. In the interest of limiting the computing cost of this operation, we only tuned a subset of all reported hyperparameters; using well-established values for the others.

5 Failed attempts

In this section, we present some of the attempts that have been considered, but that did not yield promising results.

Hierarchical knowledge injection The fine-coarse labels follow a well-defined hierarchy. In literature, several approaches have been proposed to address hierarchical multi-label classification problems coherently (Giunchiglia and Lukasiewicz, 2020). Based on intuition and experimental results, we additionally acknowledge that predicting the coarse label is an easier task, w.r.t. the prediction of the fine-grained version of the same label. It stands to reason, therefore, that the fine label should be conditioned by the predicted coarse label. Conditioning the fine label choice provides an advantage in early training stages (when the model has not yet learned the relationship between fine

	Subtask 1 (coarse)			Subtask 2 (fine)		
	$F_1^{(hc)}$	$F_1^{(pc)}$	$F_1^{(ST1)}$	$F_1^{(h)}$	$F_1^{(p)}$	$F_1^{(ST2)}$
BERT base (4-task)	**0.785 ± 0.007**	0.703 ± 0.028	0.746 ± 0.013	0.609 ± 0.017	0.288 ± 0.015	0.451 ± 0.012
BERT large (4-task)	0.778 ± 0.001	**0.768 ± 0.026**	**0.773 ± 0.012**	0.610 ± 0.004	<u>0.319 ± 0.015</u>	<u>0.468 ± 0.003</u>
RoBERTa large (4-task)	0.777 ± 0.012	<u>0.729 ± 0.010</u>	<u>0.755 ± 0.003</u>	<u>0.617 ± 0.004</u>	0.289 ± 0.022	0.456 ± 0.010
RF (250 est.) + TF-IDF	0.566 ± 0.010	0.458 ± 0.007	0.528 ± 0.010	0.294 ± 0.008	0.186 ± 0.004	0.256 ± 0.007
RF (250 est.) + FT	0.389 ± 0.042	0.348 ± 0.011	0.367 ± 0.036	0.197 ± 0.015	0.112 ± 0.008	0.185 ± 0.011
Proposed (RoBERTa)	<u>0.783 ± 0.007</u>	0.723 ± 0.005	0.754 ± 0.005	**0.625 ± 0.010**	**0.337 ± 0.009**	**0.479 ± 0.003**

Table 3: Performance on the various classification problems, for other baselines models. Best results for each metric highlighted in **bold**. Second best is <u>underlined</u>.

Hyperparameter	Value
number of epochs	7
batch size	8
warmup	500 steps
learning rate	$5 \cdot 10^{-5}$
weight decay	0.01
λ	5

Table 4: Main hyperparameters used for the proposed solution.

and coarse labels), but the improvement wanes as the training continues, resulting in no substantial advantage over the base solution.

In-Context Learning (ICL) Large Language Models can easily be used for the labelling of documents (e.g., social media posts (Tan et al., 2024), scientific papers (Giobergia et al., 2024), or news articles(Li et al., 2024)). It stands to reason, thus, that these models should be able to perform competitively in this classification task as well. We attempted various few-shot prompt engineering approaches, using LLMs on the small end of the scale (e.g., Llama 3.1 8B (Dubey et al., 2024)). However, as is well-known in literature, ICL can be outperformed by task-specific, fine-tuned SOTA models (Brown et al., 2020). This was the case for this challenge, where the available training data was sufficient to produce an adequately fine-tuned classifier.

6 Limitations

We acknowledge several limitations in the proposed approach, as indicated by the average results obtained in the public leaderboard (9th place for subtask 1, 5th place for subtask 2). Among them, there is the usage of only the textual information, without considering temporal and spatial information

available. In addition, we make a rather strong assumption of independence between products and hazards. While initial results pointed in that direction, we may assume that further explorations could potentially reveal that a single multi-task solution, if properly defined, may yield even better performance. Finally, we note that the problem is characterized by a heavy class imbalance: simple attempts to mitigate this problem (e.g., introducing different weighting schemes for different classes) did not produce promising results. However, more sophisticated approaches (e.g., with data augmentation to increase dataset size and variety (Bayer et al., 2022), or contrastive learning to reduce model biases (Koudounas et al., 2024)) may still be explored to provide additional benefits.

7 Discussion and conclusions

In this paper we discussed a solution to the Food Hazard Detection task of SemEval 2025. The task, framed as a multi-task learning problem, highlighted how hierarchical labels can benefit from being predicted together. We show that we observed no clear benefit in simultaneously predicting results across the two targets (product, hazard). This can be explained given the low Mutual Information observed between the two labels, at all levels of granularity. We presented experimental results that corroborate the claims made and that allow to identify a candidate solution. We finally covered some of the attempts that appeared to be promising, but that did not yield any meaningful improvement.

Acknowledgements

References

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Flavio Giobergia, Luca Cagliero, Paolo Garza, Elena Baralis, et al. 2020. Cross-lingual propagation of sentiment information based on bilingual vector space alignment. In *EDBT/ICDT Workshops*, pages 8–10.

Flavio Giobergia, Alkis Koudounas, and Elena Baralis. 2024. Large language models-aided literature reviews: A study on few-shot relevance classification. In *2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–5. IEEE.

Eleonora Giunchiglia and Thomas Lukasiewicz. 2020. Coherent hierarchical multi-label classification networks. *Advances in neural information processing systems*, 33:9662–9673.

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.

Alkis Koudounas, Flavio Giobergia, Eliana Pastor, and Elena Baralis. 2024. A contrastive learning approach to mitigate bias in speech models. In *Interspeech 2024*, pages 827–831.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024. Large language model agent for fake news detection. *arXiv preprint arXiv:2405.01593*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2024. Cicle: Conformal in-context learning for largescale multi-class food risk classification. *arXiv preprint arXiv:2403.11904*.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Zein Shaheen, Gerhard Wohlgenannt, and Erwin Filtz. 2020. Large scale legal text classification using transformer models. *arXiv preprint arXiv:2010.12871*.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Hanzhuo Tan, Chunpu Xu, Jing Li, Yuqun Zhang, Zeyang Fang, Zeyu Chen, and Baohua Lai. 2024. Hicl: Hashtag-driven in-context learning for social media natural language understanding. *IEEE Transactions on Neural Networks and Learning Systems*.

Yu Zhang and Qiang Yang. 2018. An overview of multi-task learning. *National Science Review*, 5(1):30–43.