Food Hazard Detection

Muhammad Saad
Dhanani School of Science and Engineering
Habib University
Karachi, Pakistan
Email: ms08063@st.habib.edu.pk

Meesum Abbas
Dhanani School of Science and Engineering
Habib University
Karachi, Pakistan
Email: ma08056@st.habib.edu.pk

Abdul Samad
Dhanani School of Science and Engineering
Habib University
Karachi, Pakistan
Email: abdul.samad@sse.habib.edu.pk

Sandesh Kumar
Dhanani School of Science and Engineering
Habib University
Karachi, Pakistan
Email: sandesh.kumar@sse.habib.edu.pk

***Abstract*—Food Hazard Detection, SemEval 2025 Task 9, focuses on identifying and categorizing hazards and product types from food-incident reports sourced online. The goal is to provide explanations for the categorization based on the associated hazards and products mentioned. This task aims to enhance the ability to detect food hazards from social media platforms and other online articles, contributing to improved food safety monitoring and analysis across diverse web sources.**

***Index Terms*—hazard category(main), food category, hazard(main), product**

I. Introduction

Food safety is a critical concern in today's globalized world, where food products traverse vast supply chains before reaching consumers. Ensuring the safety of these products is paramount to preventing foodborne illnesses and safeguarding public health. In the Food Hazard Detection challenge, we aim to address this pressing issue by developing explainable classification systems that analyze food-incident reports collected from diverse web sources, including social media platforms.

II. Research Question

Our research is mainly focused on 'How can deep learning models enhance explainability and accuracy in food hazard and product-category detection from web-sourced reports?'

As automated crawlers become increasingly important in detecting food safety issues, it is crucial to develop systems that can accurately and transparently classify incidents related to food hazards. Through this research, we have the opportunity to create algorithms that effectively identify and categorize potential hazards in food products from incident reports.

The dataset we are working with comprises detailed features such as "year," "month," "day," "language," "country," "title," and "text." It includes 1,142 distinct products, grouped into 22 product categories, and 128 potential hazard values, arranged into 10 hazard categories. This presents a complex and multifaceted problem, further complicated by the inherent class imbalance in the data, which requires innovative approaches to ensure accurate predictions. We will discuss more about the dataset in Section III.

This challenge is comprised of two sub-tasks: (ST1) text classification for food hazard prediction, where we focus on identifying the types of products and hazards, and (ST2) food hazard and product vector detection, which aims to determine the probabilities of each product and hazard associated with each report, and hence also find the product(s) and hazard(s) for them. The task will be evaluated primarily using the macro F1 score, highlighting the importance of both precision and recall in hazard detection across these sub-tasks.

By developing explainable classification systems, we aim not only to improve food safety monitoring but also to contribute to advancements in natural language processing and its applications in public health.

III. Problem Statement

Analyzing our dataset and the sub-tasks, we identify two primary focus areas; classification and explainability. Based on these insights, we define our problem statement as follows:

Develop a system that utilizes deep learning models to accurately and transparently identify food hazards and product categories from unstructured, web-sourced reports. This involves tackling two specific tasks: (1) text classification for food hazard prediction to identify relevant hazards and product types, and (2) vector detection to determine the probabilities of each product and hazard associated with a report.

By addressing these tasks, we aim to overcome the limitations of existing methods, ensuring precise and interpretable predictions while advancing food safety monitoring and contributing to natural language processing in public health applications.

IV. Literature Review

We divided our literature review into two main parts: classification of texts into one of multiple categories for ST1, and then explaining why a specific category was chosen, or more specifically, finding how much each word contributed to the selection of that category for ST2.

One study proposed LOTClass [1], i.e. a weakly-supervised text classification model that relies solely on label names for training, avoiding the need for human-labeled documents. This approach leverages pre-trained neural language models, such as BERT, to predict class categories by learning semantically related words from unlabeled data. Through a self-training process, LOTClass refines the model by iteratively predicting document-level categories. The model achieves approximately 90% accuracy across four benchmark datasets, including AG News, DBPedia, IMDB, and Amazon, without the need for labeled data.

Following the LOTClass study, another paper [2] explored classifying multi-level grocery product categories using transformer models, such as BERT, XLM, and RoBERTa. The research introduces Dynamic Masking to improve classification, focusing on predicting categories at different hierarchical levels (e.g., category, subcategory, segment). The transformer models were compared against traditional methods like SVM and XGBoost. BERT with Turkish embeddings achieved the best performance, recording accuracy and F1-scores exceeding 90% for both categories and subcategories, demonstrating its effectiveness in short text scenarios, such as product titles. This study is directly relevant to our project, as it provides insight into how transformer-based models outperform traditional models, especially for short text classifications. The use of fine-tuning techniques and the focus on multi-level classification can inform our approach to hazard and product categorization in food-incident reports. The paper also highlights common misclassification challenges, such as ambiguous product titles and category overlap, offering potential solutions like improved category definitions and incorporating contextual data, which can be adapted for our task to enhance model accuracy.

Focusing on explainability, Ribeiro et al. [3] introduced LIME (Local Interpretable Model-agnostic Explanations), a tool designed to explain individual predictions made by black-box models, such as SVMs, random forests, and neural networks. The paper also proposes SP-LIME, which selects non-redundant representative explanations to offer a global understanding of a model's behavior. LIME operates by approximating a model's local behavior around a single prediction using a linear model, providing insights into how each feature contributed to the final decision. Both LIME and SP-LIME are model agnostic which makes it particularly relevant for our project as it aligns with our goal of explaining category selection for food hazard reports. The ability to generate interpretable explanations can aid in understanding the contribution of specific words or phrases within food-incident titles, thus enhancing the transparency and reliability of our model's predictions. The study demonstrated the effectiveness of LIME across several NLP tasks, including document classification and sentiment analysis, showing that explanations could reveal flaws in model reasoning, such as predictions based on irrelevant features. This insight is crucial for improving our hazard detection model's interpretability and producing viable results for ST2.

Expanding on the application of Explainable AI (XAI) in food safety, Buyuktepe et al. [4] investigated the use of XAI techniques for predicting and interpreting food fraud cases in the global supply chain using a Deep Neural Network (DNN) model. The study employed tools like LIME, SHAP, and WIT to explain model predictions and identify key features influencing food fraud detection. The DNN model, designed to predict seven types of food fraud (e.g., tampering, illegal importation, expiration date fraud), achieved an accuracy of 81.4%. This research is highly relevant to our project as it showcases the effectiveness of XAI tools in enhancing the interpretability of models, providing both local and global insights into the model's behavior. For instance, SHAP was used to highlight that the feature 'data source' (e.g., RASFF or EMA) had the highest impact on the model's predictions, demonstrating how certain contextual features are pivotal in classification tasks. These insights can be directly applied to improve our hazard detection model by incorporating similar features and XAI methods to ensure that critical contextual data, such as origin country or product name, are effectively leveraged to enhance classification accuracy and interpretability.

Another study [5] explored toxic span detection, which is a task that aims to identify specific toxic spans within texts rather than simply classifying entire posts as toxic or non-toxic. The researchers introduced a new dataset called TOXICSPANS, which annotates English posts with toxic spans, and proposed several methods for the task, including sequence labeling models and binary classifiers enhanced with rationale extraction mechanisms. Among these, Span-BERT sequence labeling emerged as the best performer, achieving an F1 score of 63.0%. Other models, such as BiLSTM-SEQ and CNN-SEQ, were also evaluated. The dataset featured 11,035 toxic posts, with each post's toxic span annotated by multiple crowdworkers. The paper also highlighted the challenges of fine-grained toxicity detection. This study is relevant to our task of food hazard detection, as both involve text classification with an emphasis on explainability, though here the focus is on identifying toxic spans rather than product or hazard categories.

Other studies [6] explored the use of deep neural networks for restoring, attributing, and dating ancient Greek inscriptions. The dataset used was the Packard Humanities Institute's Greek inscriptions, consisting of 78,608 inscriptions. The model, named Ithaca, was trained to perform three tasks: text restoration, geographical attribution, and chronological attribution. Ithaca achieved 61.8% top-1 accuracy for restoration and attributed inscriptions to 84 regions with 70.8% accuracy. In chronological attribution, it predicted dates with an average error of 29.3 years from the ground-truth range. Additionally, the model was designed to enhance human-machine collaboration, allowing historians to improve their restoration accuracy from 25% to 71.7% when combined with Ithaca's predictions. Visualization aids such as saliency maps were used to increase the interpretability of the results. (While this specific study might not be very relevant to our tasks, it was recommended by SemEval and hence we have included it).

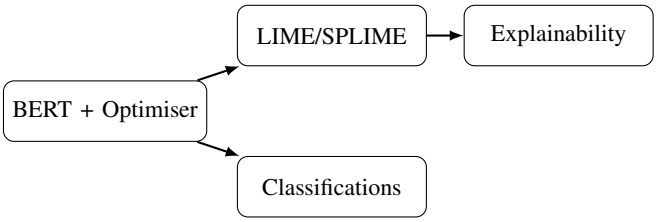

Fig. 1. Overview of the text classification and explainability workflow using BERT and XAI techniques.

Concluding the literature review, we feel that transformer models like BERT would be more suitable for the text classification task (ST1). Meanwhile, XAI techniques like LIME and SP-LIME applied on BERT might offer better explainability for the classifications (ST2). "Fig. 1" gives a comprehensive overview of this flow.

V. Dataset

The provided dataset contains detailed features such as "year," "month," "day," "language," "country," "title," and "text," providing both context and metadata. The "title" consists of short excerpts extracted from official food agency websites such as FDA and social media sites while "text" consists of more detailed descriptions about the hazard report. These are manually labeled by experts in food science or technology. The dataset is further divided into:

- **Training Data:** 5,082 labeled samples.
- **Unlabeled Validation Data:** 565 samples.
- **Unlabeled Test Data:** 997 samples.

In total, there are 6,644 short texts, with lengths varying from a minimum of 5 characters to a maximum of 277 characters, and an average length of 88 characters.

The dataset includes 1,142 different products, which are grouped into 22 product categories (e.g., "meat, egg and dairy products," "cereals and bakery products," "fruits and vegetables"). Similarly, 128 possible hazard values are also sorted into 10 hazard categories (e.g., "biological," "chemical," "fraud").

The distribution of these categories is highly imbalanced, with certain hazard categories like "allergens" and "biological" dominating the dataset, while others such as "migration" and "organoleptic aspects" appear less frequently. Similarly, within product categories, some groups like "meat, egg and dairy products" and "cereals and bakery products" are significantly more common compared to others like "food additives and flavourings" and "feed materials."

Figure 2 shows the distribution of hazard categories, highlighting this imbalance. In Figure 3, the distribution of product categories also reveals the prevalence of certain groups over others.

Furthermore, the texts are concise and vary greatly in length, making the task of classification challenging as it needs to handle a wide range of short text inputs.

For the task, the output for **ST1** involves predicting the appropriate hazard and product categories for each text instance

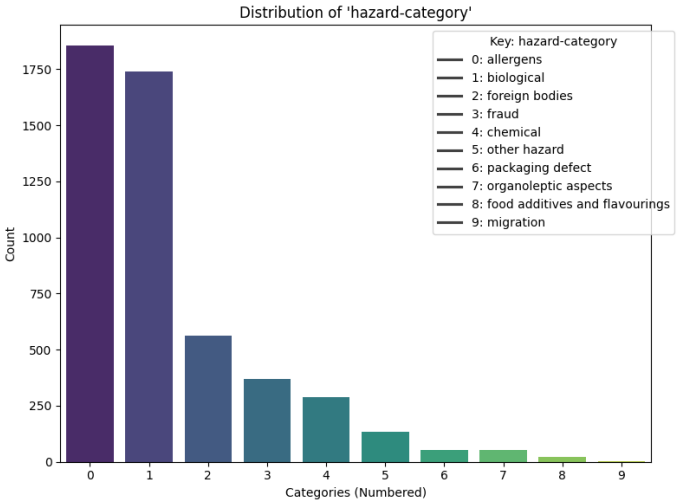

Fig. 2. Hazard Category Distribution

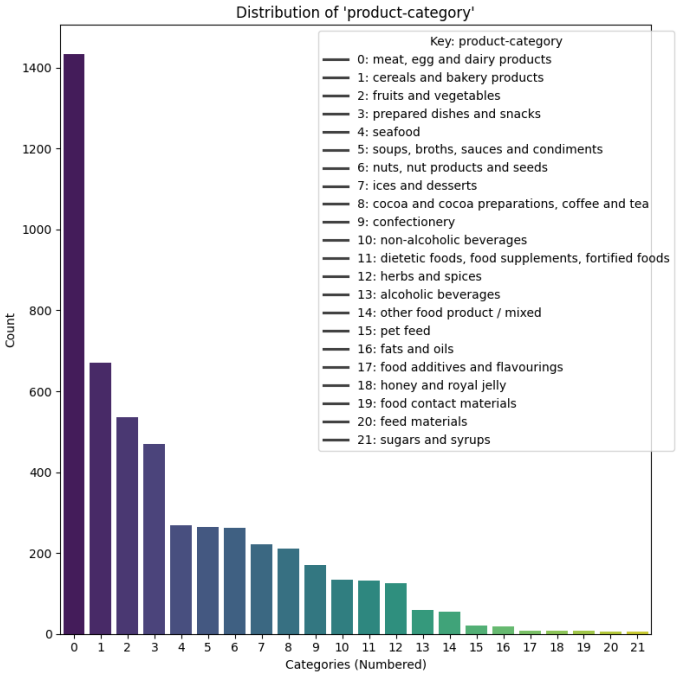

Fig. 3. Product Category Distribution

based on the information provided in the dataset. For **ST2**, the output requires the model to provide "vector"-labels or in other words, the probability of each hazard/product. The actual hazard/product is the one that has the highest probability in the vector.

VI. Evaluation

The evaluation of our model is based on the Macro F1 score, computed separately for two main components: 'hazards_pred' and 'products_pred'. We assess the model's performance by comparing the predicted labels ('hazards_pred' and 'prod-

TABLE I
LITERATURE REVIEW SUMMARY

Year	Model/Technique	Dataset	Accuracy
[1] 2020	LOTClass (BERT-based)	AG News, DBPedia, IMDB, Amazon	≈90%
[2] 2022	BERT, XLM, RoBERTa with Dynamic Masking	Grocery products	>90%
[3] 2016	LIME, SP-LIME	Various NLP tasks	N/A
[4] 2023	DNN with XAI Tools (LIME, SHAP, WIT)	Global Food Supply Chain	81.4%
[5] 2022	Span-BERT, BiLSTM-SEQ	TOXICSPANS dataset	63.0% (Span-BERT)
[6] 2022	Ithaca (Deep Neural Network)	Ancient Greek Inscriptions	70.8% (attribution)

ucts_pred') against the annotated labels ('hazards_true' and 'products_true'), which serve as the ground truth.

The scoring function primarily evaluates the accuracy of the hazard predictions ('hazards_pred'), as this is the focus of the task. The Macro F1 score for hazards is calculated first, and then the Macro F1 score for products is computed only for instances where the hazard prediction matches the ground truth. The final score is the average of the two Macro F1 scores.

The logic behind this scoring system is as follows:

- A submission where both 'hazards_pred' and 'products_pred' are completely correct will receive a score of **1.0**, indicating full accuracy.
- If 'hazards_pred' is entirely correct but 'products_pred' is wrong, the score is set to **0.5**.
- If 'hazards_pred' is completely incorrect, the score is **0.0**, regardless of the correctness of 'products_pred'.

This highlights the critical emphasis on the hazard classification as the primary evaluation metric.

VII. ST1: PREDICTING THE TYPE OF HAZARD AND PRODUCT.

A. Models and Experiments

ST1 required classifying both hazard categories and product categories. However, the models and experiments were largely designed around hazard category as it heavily influenced the final macro F1 score. The same ideas were then replicated to product category classification. Overall, we employed a series of techniques, each building upon the outcomes of the previous one to iteratively refine and improve performance.

Starting with baseline evaluations, we progressively incorporated advanced techniques to address challenges like class imbalance and improve the macro F1 score. This section provides an overview of the methodologies applied and their outcomes.

1) Initial Model Testing: Even though BERT was recommended by the challenge organizers and demonstrated strong results in the literature, we explored alternative models such as RoBERTa and DistilBERT to evaluate their potential. These models, however, performed poorly, yielding low macro F1 scores, particularly for minority classes. After these results, we adopted **BERT Base Uncased**, which consistently outperformed the alternatives. Its robust performance in text classification tasks and its suitability for the challenge requirements made it the definitive choice for further experimentation.

2) BERT Baseline: The BERT Baseline model was fine-tuned on the raw dataset without additional adjustments or optimizations. The dataset presented a significant imbalance, with some classes such as *migration* and *organoleptic aspects* having very few examples. The baseline model achieved a macro F1 score of **0.51**, performing well on majority classes like *allergens* and *biological*, while struggling on minority classes, many of which received F1 scores close to **0.00**. These results highlighted the need for interventions to address the inherent imbalance in the dataset and improve performance for underrepresented classes.

3) Early Stopping: To improve generalization and prevent overfitting, we implemented early stopping, which halted training when the validation loss plateaued for a predefined number of epochs. This method stabilized the training process and slightly improved the macro F1 score to **0.57**. The optimal number of epochs was around 5 and was used in all subsequent experiments. However, the challenges with minority classes persisted, with many still achieving low or zero F1 scores. While early stopping helped prevent overfitting, it did not directly address the imbalance in the dataset, indicating the need for more targeted approaches.

4) Class Weights: To address the imbalance in class representation, we incorporated class weights into the loss function. These weights were computed based on the inverse frequency of each class, ensuring that underrepresented classes contributed more to the loss calculation. This approach marginally improved the macro F1 score to **0.58**, with notable gains for some minority classes, such as *chemical* (F1 = **0.70**) and *fraud* (F1 = **0.64**). However, extremely low-support classes like *migration* and *food additives and flavourings* remained poorly classified. While class weights improved the model's sensitivity to minority classes, their effectiveness was limited by the small size of these classes.

5) Learning Rate Scheduling: A linear learning rate scheduler with a warm-up phase was applied to improve optimization stability and training convergence. This technique gradually increased the learning rate during the warm-up phase and then decreased it linearly for the rest of the training. While the macro F1 score remained at **0.58**, this approach provided better training stability and consistent performance improvements. Given these benefits, learning rate scheduling was incorporated into all subsequent experiments to ensure effective optimization.

6) BERT Model Variants: To explore whether larger or domain-specific pre-trained models could enhance perfor-

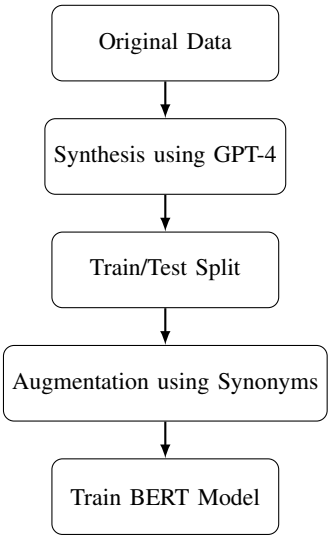

Fig. 4. Pipeline for Data Synthesis and Augmentation in Hazard Classification.

mance, we experimented with variants such as **BERT Large Uncased** and **BioBERT**. These models were fine-tuned with class weights and learning rate scheduling, mirroring the setup of the baseline BERT model. However, no significant improvement was observed, with the macro F1 score remaining at **0.58**. The additional computational cost and complexity of these models did not justify their use, and we concluded that BERT Base Uncased was the most practical and efficient choice for this task.

7) Data Synthesis and Augmentation: To address the persistent issue of class imbalance, we implemented a data synthesis and augmentation pipeline, as illustrated in Figure 4. Minority classes with extremely low support, such as *migration* and *food additives and flavourings*, were supplemented with synthetic examples generated using GPT-4. These examples added valuable diversity and context to the training data. Additionally, synonym replacement techniques were applied to classes with fewer than 500 examples. This augmentation process involved replacing up to five words per sentence with synonyms using WordNet, further diversifying the dataset.

The augmented training set expanded to 4,943 samples, with balanced representation across all classes. This significantly improved the model's performance, achieving a macro F1 score of **0.84**. Minority classes such as *migration* (F1 = **0.62**) and *packaging defect* (F1 = **0.58**) showed notable improvement. Combining the *title* and *text* fields into a single input further enriched the context provided to the model. While this approach required substantial computational resources and careful validation of the synthetic data, it effectively addressed class imbalance and enhanced the model's ability to generalize across all classes.

The progression from baseline evaluations to advanced techniques like data synthesis and augmentation resulted in substantial improvements in the macro F1 score, particularly for minority classes. While early experiments such as class weights and learning rate scheduling provided incremental gains, the synthesis and augmentation pipeline proved to be the most impactful solution, addressing the core challenges of class imbalance and underrepresented categories.

B. Results and Discussion

The results of these experiments are summarized in Table II. The highest macro F1 score achieved was **0.84** for hazard classification, using a combination of data synthesis and augmentation techniques. By applying the same techniques to product classification, we achieved a final macro F1 score of **0.74**. When submitted to the SemEval Codalab leaderboard, our approach yielded a score of **0.7006**, compared to the highest leaderboard score of **0.8249**.

Our results demonstrate notable improvements over the baseline approaches but highlight certain limitations compared to the state-of-the-art methods described in the literature. Previous studies, such as LOTClass [1], achieved approximately 90% accuracy by leveraging weak supervision and self-training, which allowed for effective label prediction without requiring human-annotated datasets. Similarly, the use of hierarchical classification and fine-tuned BERT models in [2] resulted in F1 scores exceeding 90% for multi-level grocery product categorization. These models benefited from balanced datasets and well-defined hierarchical structures, which significantly contributed to their superior performance.

In contrast, the dataset used in our task presented several challenges, including extreme class imbalance and ambiguous class definitions. While our data synthesis and augmentation pipeline effectively mitigated class imbalance, the limitations of synonym-based augmentation and the reliance on synthetic data may have constrained the model's ability to generalize as effectively as models trained on real-world, balanced datasets. For example, minority classes such as *migration* and *food additives and flavourings* showed significant improvement (macro F1 scores of 0.62 and 0.58, respectively), but certain edge cases and semantic overlaps between categories persisted.

Another advantage of prior work, such as [2], was the incorporation of contextual metadata, such as hierarchical labels and domain-specific embeddings, which enhanced model interpretability and accuracy. Our approach, while incorporating combined input fields (title and text), lacked domain-specific enhancements such as dynamic masking or embeddings tailored to food safety datasets, which could be explored in future work.

C. Future Directions

Future research could address the limitations of our approach by incorporating domain-specific knowledge and advanced augmentation techniques. Potential avenues include:

- Implementing more advanced data augmentation strategies, such as back-translation or contextual word replacement, to improve the quality and diversity of synthetic data.
- Exploring data synthesis using advanced models like T5, which could generate contextually rich and diverse

TABLE II
RESULTS SUMMARY OF MODELS AND EXPERIMENTS FOR ST1

Technique	Hazard Macro F1 Score	Product Macro F1 Score
BERT Baseline	0.51	0.49
BERT + Early Stopping	0.57	0.55
BERT + Class Weights	0.58	0.56
BERT + Learning Rate Scheduling	0.58	0.57
BERT Variants (e.g., BERT Large, BioBERT)	0.58	0.57
Data Synthesis and Augmentation	**0.84**	**0.78**
SemEval Codalab Leaderboard Submission	0.7006	

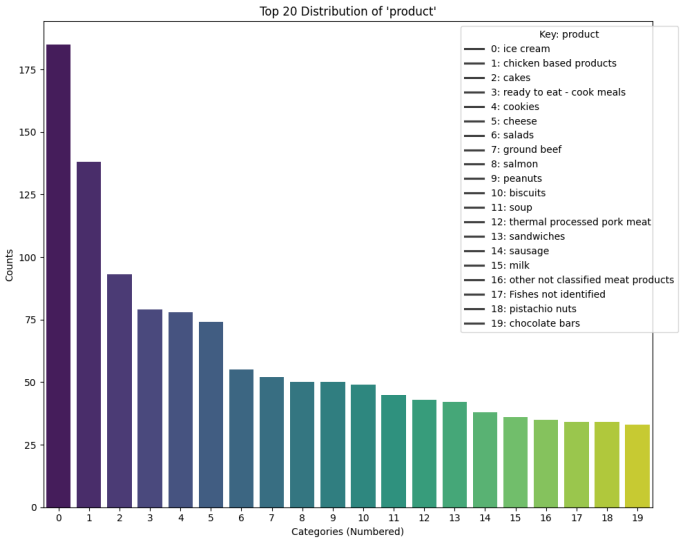

Fig. 5. Top 20 Products Distribution

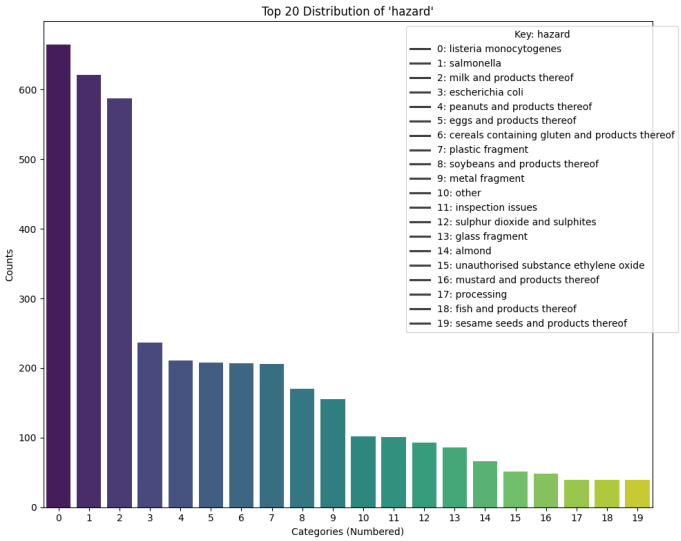

Fig. 6. Top 20 Hazards Distribution

examples tailored to specific minority classes, further reducing the class imbalance.

- Employing ensemble learning approaches by combining predictions from multiple models, such as BERT, RoBERTa, and BioBERT, to improve overall classification robustness and performance.
- Developing domain-specific embeddings fine-tuned for food safety data, which could capture the nuances of hazard and product classifications more effectively.
- Utilizing knowledge graphs or external data sources to enhance the model's understanding of relationships between hazards and products, providing richer contextual information for classification.

These directions aim to build upon the current methodology, addressing its limitations while integrating advancements in NLP and data augmentation to enhance performance and generalization.

VIII. ST2: PREDICTING EXACT HAZARDS AND PRODUCTS

A. Models and Experiments

The detection of exact hazards and products in food-related contexts was a challenging task, particularly when leveraging explainability models such as LIME [3], [4], SPLIME [3], and SHAP [4]. While these models provide insights into feature contributions, their practical application is limited due to several factors:

- **Computational Expense:** These models are resource-intensive, requiring significant computational power. For instance, preliminary testing suggested that completing a single epoch would require approximately 18 days with the available resources, rendering them impractical.
- **Limited Interpretability for Implicit Hazards/Products:** The explainability techniques predominantly identify explicit features contributing to classification categories. This approach struggles with implicit hazards, often necessitating further downstream classification for accurate detection.

For example, in one case, the term *Latvian* was highlighted as the most significant contributor to a hazard category, which might be correct in the context of the text given, but is not the actual hazard, illustrating the model's limitations in predicting exact hazards and products.

To address these challenges, we adopted a novel approach combining data augmentation and ensemble modeling to en-

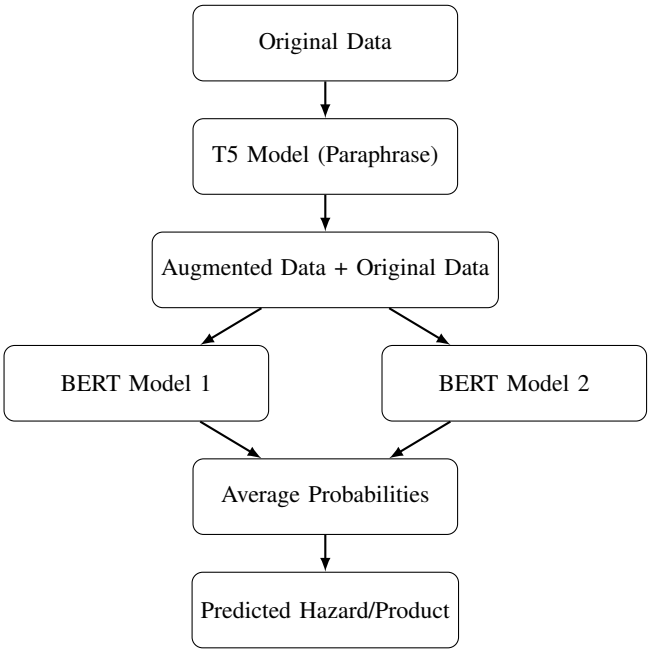

Fig. 7. Pipeline for Data Synthesis, Augmentation, and Hazard/Product Prediction.

hance model performance and generalizability.

Data Augmentation: We utilized the T5 library to paraphrase both titles and body text in the dataset, generating diverse linguistic variations while preserving semantic meaning. This process expanded the dataset by 2,120 rows for hazards, ensuring a minimum of 30 instances per hazard, and by 5,100 rows for products, with at least eight instances per product. This augmentation mitigated data scarcity and imbalance, providing a richer and more representative training set.

Ensemble Modeling: We trained multiple BERT-based classification models to leverage their contextual understanding of text. Each model, fine-tuned with varying configurations, contributed unique insights to the ensemble. Predictions from these models were averaged to produce final outputs, reducing variance and increasing robustness. The ensemble approach smoothed out individual model inconsistencies, improving overall prediction accuracy.

This method effectively addresses the dataset's limitations and improves classification outcomes, as evidenced by significant performance gains in macro F1-scores.

B. Results and Discussions

Baseline BERT models, without data augmentation or ensemble techniques, achieved disappointing macro F1-scores of 0.09 for hazards and 0.0013 for products on the test split. However, following the integration of data augmentation and ensemble modeling, the performance showed a substantial improvement. After augmentation and ensembling, the models reached macro F1-scores of 0.84 for hazards and 0.77 for products, marking a significant boost in prediction accuracy.

These results highlight the effectiveness of the adopted strategy in overcoming the initial limitations of the dataset. The improvement is particularly notable given the challenging nature of the task. Hazards and products often require nuanced context understanding, and our models were able to better capture these complexities with the augmented dataset and diverse ensemble models. This improvement not only reflects the robustness of our method but also underscores the importance of combining data augmentation with ensemble learning to enhance performance in real-world, noisy tasks like food hazard detection.

However, it is important to note that while the local F1-scores are strong, they may be slightly inflated due to overlap between the augmented data and the test split. In particular, proper nouns in paraphrased data occasionally mirrored those in the original input, introducing redundancy. Furthermore, the train-validation-test split was conducted after data augmentation, which increased the likelihood of similar or even identical entries appearing across these splits, potentially leading to overfitting on some instances. While this redundancy likely boosted performance on the test set, it does not fully represent the model's ability to generalize to unseen data.

In comparison to the LIME and SHAP models discussed in earlier studies, our results are notably superior. While LIME [3], SPLIME [3], and SHAP [4] are powerful explainability tools, they focus primarily on interpreting the contributions of individual features to model predictions. These models do not directly provide performance metrics such as accuracy or F1-score, making it difficult to quantitatively compare their results to traditional classification models.

On the official SemEval Codalab leaderboard, our submission secured 4th place with an overall F1-score of 0.4510, just 0.0005 behind the 3rd-ranked team. The highest score achieved in the competition was 0.51, underscoring the competitive nature of the task and validating the significance of our results. While the leaderboard score reflects the relative performance of our solution, it also emphasizes the difficulty of the task and the high quality of competing models. This competitive performance demonstrates that our approach, though not perfect, is highly effective in this challenging domain and holds considerable promise for future improvements.

C. Future Directions

To further enhance model performance, we propose the following improvements:

- **Data Synthesis:** Utilize Gemini's open-source APIs to generate a larger, more diverse dataset that better reflects real-world distributions and reduces redundancy.
- **Constraint-Based Classification:** Leverage patterns between hazard/product categories and their corresponding labels to constrain the output space, thereby reducing class ambiguity and improving prediction accuracy.

These refinements aim to improve both model generalizability and precision, advancing progress in food hazard detection.

REFERENCES

[1] Y. Meng, Y. Zhang, J. Huang, C. Xiong, H. Ji, C. Zhang, and J. Han, "Text classification using label names only: A language model self-training approach," Proc. of the Conf. on Empirical Methods in Natural Language Processing, 2020, pp. 90-100.

[2] O. Ozyegen, H. Jahanshahi, and M. Cevik, "Classifying multi-level product categories using dynamic masking and transformer models," J. of Data, Inf. and Manag., vol. 4, pp. 71–85, 2022. https://doi.org/10.1007/s42488-022-00066-6.

[3] Marco Ribeiro, Sameer Singh, and Carlos Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier," in Proc. 2016 Conf. of the North American Chapter of the Assoc. for Computational Linguistics: Demonstrations, San Diego, CA, 2016, pp. 97–101.

[4] O. Buyuktepe, C. Catal, G. Kar, Y. Bouzembrak, H. Marvin, and A. Gavai, "Food fraud detection using explainable artificial intelligence," Expert Systems, vol. 2023, e13387. https://doi.org/10.1111/exsy.13387.

[5] J. Pavlopoulos, L. Laugier, A. Xenos, J. Sorensen, and I. Androutsopoulos, "From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer," Proc. 60th Annu. Meet. Assoc. Comput. Linguistics (ACL), pp. 3721–3734, May 2022.

[6] Y. Assael et al., "Restoring and attributing ancient texts using deep neural networks," Nature, vol. 603, no. 7900, pp. 280-284, Mar. 2022.