HU at SemEval-2025 Task 9: Leveraging LLM-Based Data Augmentation for Class Imbalance

Muhammad Saad[†], **Meesum Abbas**[†], **Sandesh Kumar**[†], **Abdul Samad**[†]
[†]Habib University, Dhanani School of Science & Engineering, Pakistan
{ms08063, ma08056}@st.habib.edu.pk
{sandesh.kumar, abdul.samad}@sse.habib.edu.pk

Abstract

Food safety is a critical global concern, and automating the detection of food hazards from recall reports can improve public health monitoring and regulatory compliance. This paper presents our submission for SemEval-2025 Task 9: The Food Hazard Detection Challenge. We tackle the inherent class imbalance in this task by leveraging advanced data augmentation techniques, including LLM-based synthetic data generation, synonym replacement, and back-translation.

We employ transformer-based models such as DistilBERT, fine-tuned with these augmented datasets, to enhance performance. Our system achieves significant improvements, obtaining a Macro-F1 score of **0.7882** in ST1 and **0.5099** in ST2.[1] Additionally, we analyze the impact of augmentation strategies and compare multiple architectures, highlighting challenges in handling implicit hazards. Our findings underscore the effectiveness of LLM-based augmentation in addressing extreme class imbalance while demonstrating the strengths and limitations of transformer models in food safety applications.

1 Introduction

Ensuring food safety is a critical challenge in public health, requiring timely detection of hazards leading to product recalls. Traditional methods rely on manual expert analysis, which is time-consuming and lacks scalability. Recent advances in natural language processing (NLP) have enabled automated food hazard detection from recall reports, improving regulatory oversight. However, class imbalance in real-world datasets, where some classes are overrepresented, remains a challenge for accurate predictions (Gao, 2020).

SemEval-2025 Task 9: *The Food Hazard Detection Challenge* focuses on classifying food hazards and products from textual reports (Randl et al., 2025). This task is crucial for enhancing food security and public health interventions. It consists of two subtasks:

- **ST1**: Classifying the hazard category and product category.
- **ST2**: Identifying the exact hazard and exact product mentioned in the report.

The primary challenge of this task is the extreme class imbalance, where certain classes appear far more frequently than others. Figure 1 illustrates the severity of class imbalance through a probability distribution.

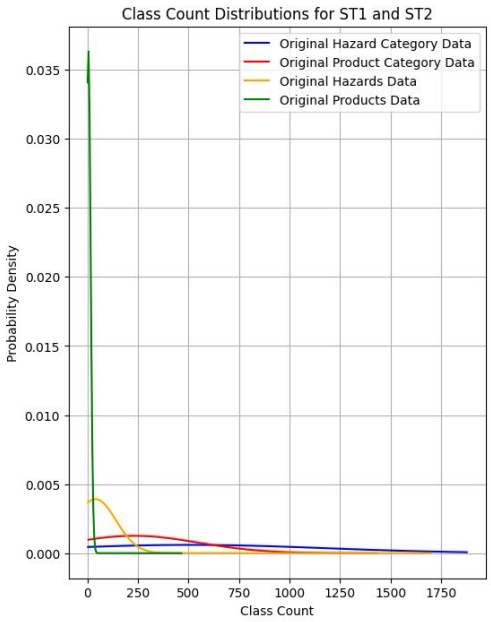

Figure 1: Distribution of classes by hazard-category, product-category, hazard, and product

We propose a transformer-based model augmented with synthetic data from LLMs like GPT-4o, Gemini Flash 1.5, and T5 to address this imbalance. Our approach leverages DistilBERT, which has proven effective in handling class imbalance,

[1]Our Code: https://github.com/msaadg/hu_semeval_task9

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2200–2208

and fine-tunes it on both original and augmented datasets.

Our system ranked 5th in ST1 and 4th in ST2. Despite these achievements, challenges persist in identifying implicit hazards and dealing with highly imbalanced categories, which require further refinement (Henning et al., 2023). The results confirm that data augmentation plays a key role in overcoming class imbalance.

The dataset provided by the SemEval-2025 organizers contains structured recall reports sourced from food regulatory bodies. It consists of:

- **Training Data:** 5,082 samples.
- **Validation Data:** 565 samples.
- **Test Data:** 997 samples.
- **Hazard Categories:** 10
- **Product Categories:** 22
- **Hazards:** 128
- **Products:** 1,142

Each report includes 'year', 'month', 'day', 'country', 'title', 'text', 'hazard-category', 'product-category', hazard, and product. During preprocessing, we merged 'title' and 'text' into a unified field 'title_text' to enhance contextual representation.

2 Related Work

The challenge of food hazard detection has been extensively studied, with recent advancements leveraging Natural Language Processing (NLP) for automated risk assessment. Traditional approaches have primarily relied on rule-based methods and handcrafted feature extraction, which often fail to generalize across diverse recall reports. According to (Gao, 2020), while these methods have been widely used, they struggle with the complexity and scale of modern datasets. More recently, deep learning models, particularly transformers, have demonstrated superior performance in food safety classification tasks, significantly outperforming traditional methods (Buyuktepe et al., 2025).

2.1 Food Hazard Detection Using NLP

Food safety monitoring requires extracting key hazard-related information from unstructured recall reports. Earlier work focused on keyword-based extraction and ontology-driven approaches. However, with the advent of deep learning, transformer-based architectures like BERT and DistilBERT have enabled more accurate hazard classification, as seen in recent studies (Zhou et al., 2020). These models offer improved flexibility and performance over rule-based methods, as they can capture semantic nuances in text. Our work builds upon these advancements by tackling the extreme class imbalance inherent in food recall datasets, an issue that has been discussed in the context of NLP-based food safety applications (Gao, 2020).

2.2 Handling Class Imbalance in NLP

Class imbalance poses a significant challenge in multi-class NLP classification, particularly when dealing with underrepresented categories. According to (Henning et al., 2023), traditional methods like oversampling and undersampling often lead to overfitting and loss of information, which can negatively impact model performance. More recently, synthetic data augmentation has emerged as a promising solution to this issue. Studies such as (Meng et al., 2020) and (Gao, 2020) have demonstrated the efficacy of techniques like contextual augmentation, back-translation, and paraphrasing in mitigating class imbalance. Our approach extends this by leveraging Gemini Flash 1.5 and GPT-4o for targeted LLM-based augmentation, generating diverse synthetic data to improve model generalization, particularly for rare hazard categories.

2.3 Explainability in Food Safety NLP

Explainability is critical in automated food safety monitoring, ensuring transparency and trustworthiness. According to (Ribeiro et al., 2016), techniques like LIME and SHAP provide insights into how machine learning models make predictions. However, these techniques often struggle with implicit hazard detection, particularly for rare or underrepresented classes. Our experiments in ST2 confirm the said limitations, with (Pavlopoulos et al., 2022) highlighting similar challenges when interpreting complex models in the food hazard domain. These findings emphasize the need for alternative interpretability methods, which we explore further in our work.

3 System Overview

In this section, we present our methodology for tackling the task. As mentioned earlier, the primary challenge of this task lies in the severe class

imbalance in both hazard and product categories, which hinders model generalization (Gao, 2020). Additionally, implicit relationships between hazards and products pose an extra layer of complexity (Henning et al., 2023). To address these issues, we integrated transformer-based models with diverse data augmentation strategies and applied a series of training optimizations to enhance model performance.

3.1 Data Augmentation Strategies

Given the highly skewed distribution of hazard and product categories, where several classes appear fewer than ten times in the dataset (see Appendix A), we employed multiple augmentation techniques to enrich data diversity and improve classification robustness. As noted by (Gao, 2020), augmentation techniques like synonym replacement and back-translation have been proven to alleviate class imbalance. The augmentation strategies included synonym & contextual replacement, back-translation, paraphrasing and large language model based synthetic data generation. Synonym & contextual replacement was implemented using the NLTK WordNet, where at most 5 words in the text were replaced with contextually appropriate synonyms (Meng et al., 2020). Back-translation was performed using French and German translations to generate alternative textual representations while preserving semantic consistency, on texts where the class count was under 50 for ST1 and under 20 for ST2. T5-base was used to paraphrase sentences for classes with less than 30 entries in the original training dataset, so that alternative formulations of the text were generated while maintaining the original meaning and retaining the classes. The augmented dataset had at least 30 entries for each class of hazard and product. Furthermore, we employed state-of-the-art LLMs like Gemini Flash 1.5, GPT-4o & o1-mini for their diverse text-generation capabilities to synthesize recall reports for classes with less than 50 entries in the original dataset such that each class has at least 50 entries in the augmented dataset. This significantly expanded the training dataset while also diversifying the kind of texts for each class. This approach aligns with studies where LLM-based data augmentation has been shown to improve generalization in imbalanced datasets (Zhou et al., 2020).

To quantify the contribution of each augmentation technique, we report the distribution of augmented samples generated for ST1 and ST2. For ST1, synonym and contextual replacement, back-translation, and LLM-based synthetic data generation were applied to address class imbalance in hazard and product categories. For ST2, similar techniques were used, with additional emphasis on paraphrasing via T5-base to enhance fine-grained hazard and product identification. The total number of samples after augmentation was 15,570 for ST1 and 63,082 for ST2, expanding the original 5,082 training samples. Table 1 summarizes the count-wise distribution of samples from each augmentation technique for both subtasks.

To ensure the integrity of the augmented dataset, as detailed in Table 1, all augmented samples underwent human verification to eliminate label leakage and maintain data consistency. The augmented data was carefully merged with the original dataset, following a structured approach to avoid overfitting on artificial samples, which is a common issue when synthetic data is introduced (Shorten et al., 2021). Figure 2 shows the balanced class distribution after data augmentation.

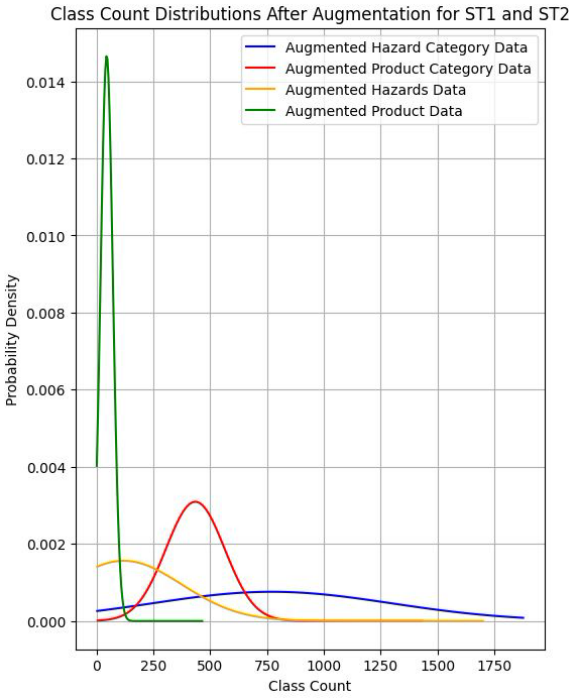

Figure 2: Final Distribution of Classes after Augmentation

3.2 Model Architectures

We experimented with multiple transformer architectures to determine the most effective model for both subtasks. Initially, we used 'bert-base-

Table 1: Distribution of Augmented Samples for ST1 and ST2 by Augmentation Technique

Augmentation Technique	**ST1 Samples**	**ST2 Samples**
Synonym & Contextual Replacement	4150	20,167
Back-translation	3210	1,130
Paraphrasing	0	15,036
LLM-based Synthetic Data Generation	2564	10,823
Total Augmented Samples	9924	47,156

uncased' as our baseline, which performed very poorly on low sample classes. Recognizing the need for a more effective model, we explored various transformer-based alternatives, including RoBERTa, XLM-R, ALBERT, and Sci-BERT. However, 'distilbert-base-uncased' emerged as the most effective model across both subtasks, significantly outperforming other models. This aligns with findings that DistilBERT has shown to outperform other transformer variants in tasks with class imbalance due to its reduced computational demands and ability to retain performance comparable to larger transformer models (Zhou et al., 2020).

3.3 Training Pipeline and Optimization

Training pipelines for both subtasks incorporated 3 major steps: augmenting the data, merging augmented data with original data, and finally training the model. To further enhance model performance, we also incorporated optimization techniques into our pipeline, such as early stopping, class weights, and learning rate scheduling.

Furthermore, for ST2 we experimented with ensemble modeling by training two DistilBERT models on different random seeds and aggregating their logits using a max-logit selection strategy (Gao, 2020). This approach enhanced model robustness, particularly in identifying low-resource hazard and product categories. See Appendix B to get the complete overview of the pipelines for both subtasks, respectively.

3.4 Explainability Methods

Explainability techniques such as LIME and SHAP presented significant limitations. According to (Ribeiro et al., 2016), LIME and SHAP are valuable for interpreting model predictions, but we faced two major challenges in ST2:

- These models are resource-intensive, requiring significant computational power. Preliminary tests indicated that completing one epoch on Google Colab's T4 GPU would take approximately 18 days, making them impractical.
- LIME and SHAP primarily identify explicit features contributing to predictions but struggle with implicit hazards. For instance, the term "Latvian" was highlighted as the most significant contributor to a hazard category, but it wasn't the actual hazard, illustrating the models' limitations in predicting exact hazards and products.

4 Experimental Setup

The experimental setup aimed to address challenges such as class imbalance and computational efficiency while ensuring robust training and evaluation. We used Google Colab's T4 GPU and Kaggle's T4x2 GPUs for efficient fine-tuning. Our models were trained using the AdamW optimizer with a learning rate of $5e^{-5}$, a batch size of 8, and a maximum of 5 epochs. Early stopping was applied to avoid overfitting, halting training when validation loss plateaued. To address class imbalance, class weights were computed based on inverse class frequency (see Appendix C), which helped the model focus on underrepresented hazard and product categories (Gao, 2020). Learning rate scheduling was applied using a linear decay schedule with a warm-up phase, allowing for stable training convergence.

Text preprocessing involved merging the 'title' and 'text' fields into a single 'title_text' feature to maximize contextual representation. Tokenization was handled using model-specific tokenizers, like 'AutoTokenizer' for BERT and DistilBERT, with input sequences truncated to 512 tokens for computational feasibility. We utilized nlpaug 1.1.11 for synonym replacement, contextual augmentation, and back-translation, and Gemini 1.5 Flash and GPT-4o for synthetic data generation, which helped mitigate class imbalance (Meng et al., 2020).

Evaluation metrics included Macro-F1, with scores calculated on both ST1 and ST2.

5 Results & Analysis

The results of our experiments reveal a clear trend in the impact of data augmentation and model selection on performance improvements. Our final model, 'distilbert-base-uncased', demonstrated significant gains in Macro-F1 scores over the baseline model, highlighting the effectiveness of augmentation techniques in addressing severe class imbalance and enhancing classification accuracy across underrepresented classes.

5.1 Performance Gains

The baseline model, 'bert-base-uncased', trained for five epochs, achieved a Macro-F1 score of 0.4965 for ST1 and 0.009 for ST2. This stark contrast between the two subtasks underscores the challenge of exact hazard and product detection due to the large number of low-frequency categories. The poor performance in ST2 highlights the difficulty of predicting fine-grained hazard and product labels without sufficient examples of each class. Event extraction models, like those proposed by (Harrag and Gueliani, 2020), face similar challenges in detecting rare event categories, particularly when there is insufficient training data.

By implementing synonym replacement through NLTK WordNet, ST1 saw a notable improvement to 0.701, but ST2 remained largely unaffected. This suggests that simple lexical augmentation is effective for coarse-grained classification but does not introduce sufficient diversity for granular hazard-product identification. The need for more diverse augmentation strategies for fine-grained predictions has been observed in previous studies as well (Meng et al., 2020).

Then in order to augment the dataset that improves score in ST-2, we made use of the infamous encoder-decoder transformer model, t5-base, in order to paraphrase the texts with more contextual information and grammatical correctness. The score, after using this, jumped to 0.43, which suggests that attention based mechanisms are good at retaining information, while adding diversity to the texts, which helped the model learn its characteristics more effectively.

The introduction of LLM-based augmentation using Gemini Flash 1.5 and GPT-4o significantly improved performance, increasing ST1 to 0.779 and ST2 to 0.47. The synthetic samples generated by Gemini provided more varied examples for rare categories, reducing model bias towards majority classes. This improvement was also evident in ST2, which benefited from the increased exposure to underrepresented hazard-product pairs. The effectiveness of LLM-based data augmentation for imbalanced datasets has been demonstrated in similar domains. (Assael et al., 2022).

Further augmentation using 'nlpaug' techniques, including back-translation (French and German) and contextual synonym replacement, further enhanced classification performance, bringing ST1 to 0.811 and ST2 to 0.49. The introduction of sentence-level diversity allowed the model to better generalize beyond the original training samples, mitigating the imbalance problem further, as seen in studies utilizing back-translation for food safety tasks (Shorten et al., 2021).

The final transition to 'distilbert-base-uncased', trained on the fully augmented dataset, resulted in a Macro-F1 of 0.7882 securing 5th place for ST1 and 0.5099 securing 4th place for ST2 on the test dataset. Notably, the improvements in ST2 suggest that increasing the diversity of samples was crucial for extracting implicit hazard-product relationships, reinforcing the necessity of extensive augmentation for fine-grained classification (Zhou et al., 2020).

5.2 Error Analysis

An in-depth analysis of misclassifications revealed that the model performed well on high-frequency categories but struggled with extremely rare hazards and products. The class imbalance led to instances where certain hazards or products, despite their unique nature, were mapped to broader, more frequently occurring categories. For example, rare food contaminants were often misclassified into broader chemical hazard categories, indicating a lack of precise decision boundaries for low-sample classes. Similar misclassifications have been discussed in food hazard detection tasks, where rare instances are misclassified into broader categories (Harrag and Gueliani, 2020).

Additionally, implicit hazards presented significant challenges. Many instances of food recall reports describe contamination or issues without explicitly stating the hazard category. The model struggled to infer implicit relationships between food safety incidents and their corresponding hazard types. This aligns with our earlier observations regarding the limitations of LIME and SHAP in capturing implicit relationships (Pavlopoulos et al., 2022).

Another common misclassification pattern was observed in ST2, where highly specific hazard-product pairs were mislabeled due to insufficient positive training samples. Although augmentation improved classification, the model still exhibited difficulty in capturing the nuance of rare pairings, reinforcing the need for further improvements in dataset balancing strategies (Ozyegen et al., 2022).

5.3 Comparison with Leaderboard Results

While our augmentation strategies and model optimizations narrowed the performance gap compared to top systems (0.8223 for ST1, 0.5473 for ST2), further improvements are still possible. Similar gaps have been observed in other NLP tasks, where augmentation and optimization were key factors (Gao, 2020)

Key differences in approaches of top teams include:

- **LLM-Enhanced Augmentation:** Top teams used DeBERTa v3 Large and RoBERTa Large with fine-tuned LLMs (e.g., Gemini, RAG), while our pipeline focused on Gemini Flash 1.5 and 'nlpaug', significantly improving minority class detection.
- **Ensemble Learning:** High-ranking teams used multiple transformer models with soft voting, while we used two ensembled DistilBERT models for ST2, improving robustness but lacking the power of multiple model ensembles.
- **Chunking and Data Representation:** Some teams experimented with chunking input data into various token sizes, but we used a fixed title + text representation, optimizing classification but possibly limiting generalization.
- **Fine-Tuning Strategies:** Leading teams used LoRA fine-tuning on RoBERTa-based models, whereas we focused on DistilBERT and augmentation-centric enhancements.

While our system performed well under constraints, future iterations could benefit from LLM-based retrieval mechanisms (e.g., RAG), soft voting across multiple LLMs, and chunking strategies to improve hazard-product representation. As noted in Assael (2022), LLM-based reasoning could provide richer context for rare classes and implicit relationships (Assael et al., 2022).

6 Conclusion

This work tackled the challenges of food hazard detection in SemEval-2025 Task 9, focusing on extreme class imbalance and enhancing model generalization through LLM-based augmentation (Gao, 2020). By employing a combination of contextualized synonym replacement, paraphrasing, back-translation, and synthetic data generation, we significantly improved classification performance, particularly for low-frequency categories (Shorten et al., 2021). Among various transformer models, 'distilbert-base-uncased' provided the best trade-off between efficiency and accuracy, achieving a final Macro-F1 of 0.7882 (ST1) and 0.5099 (ST2). While these improvements are noteworthy, further refinements are necessary for addressing ongoing challenges, especially implicit hazard detection.

7 Limitations

Despite the advancements made in this work, several limitations remain. Class imbalance continues to be a major issue, particularly for rare hazard and product categories, where the model still struggles with suboptimal performance (Henning et al., 2023). While data augmentation techniques have alleviated some of the imbalance, further enhancement is needed (Meng et al., 2020). Implicit hazard detection remains an ongoing challenge, especially when hazards are inferred rather than explicitly stated. This underlines the need for more advanced interpretability techniques to handle such implicit relationships (Ribeiro et al., 2016). Additionally, while our approach has shown improvements, incorporating strategies like contrastive data augmentation, hierarchical classification, and ensemble learning (Ozyegen et al., 2022) could further boost model robustness and generalization.

Acknowledgments

We thank the SemEval-2025 Task 9 organizers for providing the dataset and challenge framework (Randl et al., 2025). We acknowledge computational support from Google Colab and Kaggle, which enabled large-scale training. Contributions from the open-source NLP community, including augmentation libraries and transformer models, were instrumental in our research.

References

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.

Okan Buyuktepe, Cagatay Catal, Gorkem Kar, Yamine Bouzembrak, Hans Marvin, and Anand Gavai. 2025. Food fraud detection using explainable artificial intelligence. *Expert Systems*, 42(1):e13387.

Jie Gao. 2020. Data augmentation in solving data imbalance problems. Master's thesis, KTH Royal Institute of Technology.

Fouzi Harrag and Selmene Gueliani. 2020. Event extraction based on deep learning in food hazard arabic texts. *arXiv preprint arXiv:2008.05014*. Accessed: 2023-02-28.

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, Dubrovnik, Croatia. Association for Computational Linguistics.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017, Online. Association for Computational Linguistics.

Ozan Ozyegen, Hadi Jahanshahi, Mucahit Cevik, Beste Bulut, Deniz Yigit, Fahrettin F Gonen, and Ayşe Başar. 2022. Classifying multi-level product categories using dynamic masking and transformer models. *Journal of Data, Information and Management*, 4(1):71–85. © The Author(s), under exclusive licence to Springer Nature Switzerland AG 2022.

John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. 2022. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3721–3734, Dublin, Ireland. Association for Computational Linguistics.

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.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Connor Shorten, Taghi M. Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of Big Data*, 8(1):101.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.

A Class Imbalance Illustration

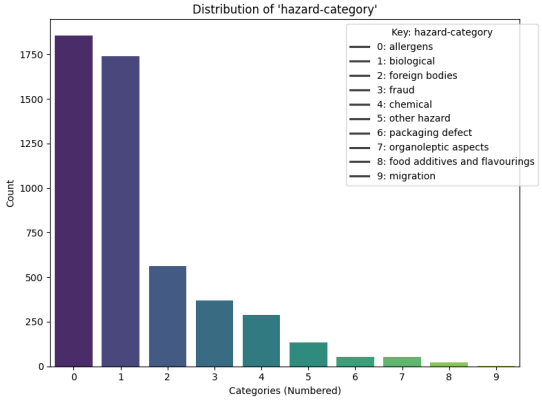

Figure 3: Hazard Category Distribution

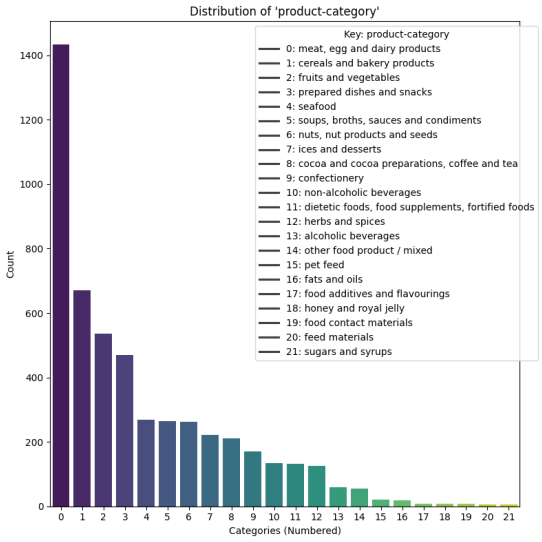

Figure 4: Product Category Distribution

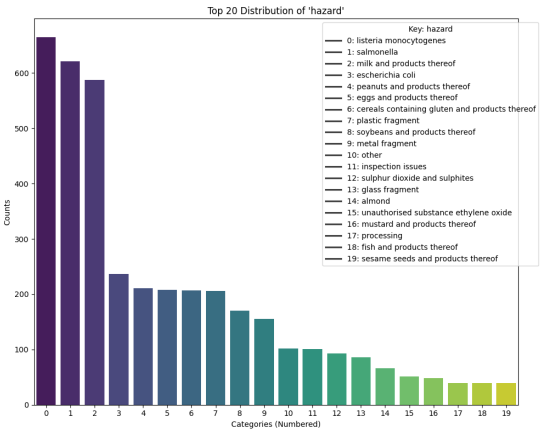

Figure 5: Top 20 Hazards Distribution

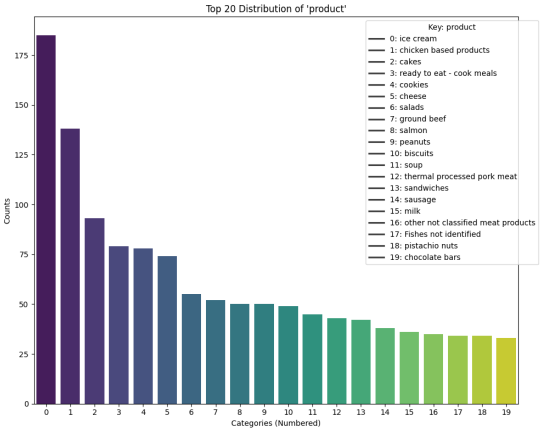

Figure 6: Top 20 Products Distribution

B Pipelines for ST1 and ST2

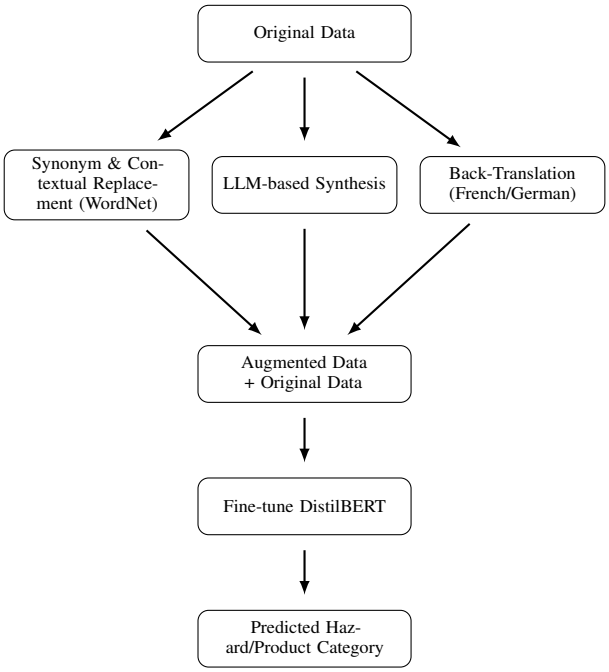

Figure 7: Pipeline for ST1

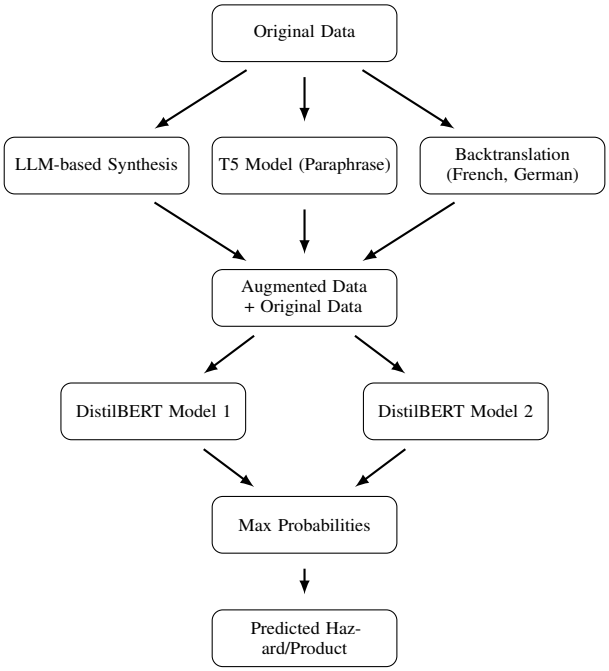

Figure 8: Pipeline for ST2

C Class Weights Implementation

To address the class imbalance in our dataset, we incorporated class weights into the loss function during training. The primary goal was to ensure that underrepresented classes, which occur far less frequently, were given more importance in the loss

computation, thereby guiding the model to better focus on these classes. This was particularly important in tasks like hazard and product category classification, where some categories were significantly more frequent than others.

The class weights were calculated based on the inverse frequency of each class in the training dataset. Specifically, for each class, the weight was computed as the inverse of its frequency relative to the total number of samples. The weight for class i is given by:

$$w_i = \frac{N}{f_i}$$

Where:

- w_i is the weight for class i,
- N is the total number of samples in the training dataset,
- f_i is the frequency of class i.

These computed weights were then integrated into the loss function, ensuring that the model penalized misclassifications of rare classes more than those of more frequent ones. By doing so, we mitigated the effect of class imbalance and helped the model focus on learning from the underrepresented classes, which would otherwise be overshadowed by the more frequently occurring classes.

This approach aligns with the findings of (Henning et al., 2023), where the use of class weights has been shown to improve model performance in imbalanced classification tasks by preventing the model from being biased towards the majority class.

By adjusting the loss function in this manner, we were able to improve the model's ability to detect and classify rare hazard and product categories, ultimately leading to better performance on both subtasks ST1 and ST2.