Ustnlp16 at SemEval-2025 Task 9: Improving Model Performance through Imbalance Handling and Focal Loss

Zhuoang Cai, Zhenghao Li, Yang Liu, Liyuan Guo, Yangqiu Song

Department of Computer Science and Engineering,

The Hong Kong University of Science and Technology (HKUST), Hong Kong SAR

{zcaiat, zlika, yliuhy, lguoar}@connect.ust.hk, yqsong@cse.ust.hk

Abstract

Classification tasks often suffer from imbalanced data distribution, which presents challenges in food hazard detection due to severe class imbalances, short and unstructured text, and overlapping semantic categories. In this paper, we present our system for SemEval-2025 Task 9: Food Hazard Detection, which addresses these issues by applying data augmentation techniques to improve classification performance. We utilize transformer-based models, BERT and RoBERTa, as backbone classifiers and explore various data balancing strategies, including random oversampling, Easy Data Augmentation (EDA), and focal loss. Our experiments show that EDA effectively mitigates class imbalance, leading to significant improvements in accuracy and F1 scores. Furthermore, combining focal loss with oversampling and EDA further enhances model robustness, particularly for hard-to-classify examples. These findings contribute to the development of more effective NLP-based classification models for food hazard detection.

1 Introduction

The rapid advancement of natural language processing (NLP) has facilitated the development of automated classification systems across various domains, including food hazard detection. Accurately identifying and categorizing food hazards is essential for ensuring food safety and mitigating health risks associated with contaminated or unsafe food products. However, food hazard classification presents several challenges, including severe class imbalances, ambiguous and unstructured textual descriptions, and the need for high predictive accuracy. Traditional approaches to hazard detection have relied on manual inspections and rule-based classification methods, which are often time-consuming and prone to human error. In contrast, recent advancements in machine learning and NLP have enabled the automation of this process, leveraging text classification models to analyze food hazard reports and categorize them into predefined classes.

Transformer-based models, such as BERT (Devlin, 2019) and RoBERTa (Zhuang et al., 2021), have demonstrated state-of-the-art performance in text classification tasks. However, their effectiveness in imbalanced classification settings remains a challenge, as they tend to favor majority classes while underperforming in minority categories. Class imbalance is a common issue where certain categories have significantly fewer instances than others, leading to biased predictions and reduced model performance on underrepresented classes. To address this issue, researchers have explored various techniques, including data augmentation, resampling methods, and modified loss functions. Easy Data Augmentation (EDA) (Wei and Zou, 2019) generates additional training samples for minority classes, enhancing model generalization. Similarly, focal loss (Lin et al., 2017) modifies the traditional cross-entropy loss function to focus more on difficult-to-classify examples, improving performance on underrepresented categories.

In this study, we systematically investigate the impact of data balancing techniques on transformer-based models for food hazard classification. Specifically, we evaluate the effectiveness of oversampling, EDA, and focal loss in mitigating class imbalance and improving classification performance. Through extensive experimentation, we demonstrate that these strategies enhance model robustness, particularly in detecting minority-class hazards. Our findings contribute to the development of more reliable NLP-based classification models for food safety applications, providing valuable insights into optimal approaches for handling class imbalance in text classification tasks.

2 Related Work

2.1 Text Classification

Text classification, a core NLP task, involves assigning predefined labels to text. Traditional methods used rule-based approaches and machine learning models like Naïve Bayes, SVM, and Random Forests. Deep learning, particularly transformer-based models, has significantly improved performance by capturing contextual dependencies (Fan et al., 2024b).

Models like BERT (Devlin, 2019) and RoBERTa (Zhuang et al., 2021) set new benchmarks but struggle with imbalanced datasets, where minority classes are often overlooked. This issue is critical in domains like food hazard detection, where rare cases carry significant risks. To address class imbalance, researchers employ resampling techniques (Lauron and Pabico, 2016), cost-sensitive learning, and modified loss functions like focal loss (Lin et al., 2017). Data augmentation has also proven effective in enhancing classification robustness, especially in low-resource settings.

2.2 Data Augmentation

Data augmentation expands training datasets to improve model generalization (Fan et al., 2024a), particularly in NLP, where it helps mitigate class imbalance in text classification. Traditional methods like synonym replacement, back-translation, and paraphrasing (Wei and Zou, 2019) enhance lexical diversity while preserving meaning. Easy Data Augmentation (EDA) is widely used due to its simplicity, applying synonym replacement, random insertion, swap, and deletion to boost performance on imbalanced datasets. More advanced techniques leverage contextual embeddings (e.g., Word2Vec, FastText) and transformer-based text generation, though excessive alterations risk label noise. Recent studies combine augmentation with resampling strategies (Adegbenjo and Ngadi, 2024), improving accuracy and F1 scores, especially for underrepresented classes. Building on these advancements, we integrate oversampling, EDA, and focal loss to enhance classification in food hazard detection. Our approach effectively mitigates class imbalance and strengthens model robustness, particularly for minority classes.

3 Task Definition and Dataset

SemEval-2025 Task 9 involves classifying short food recall reports into predefined hazard-related
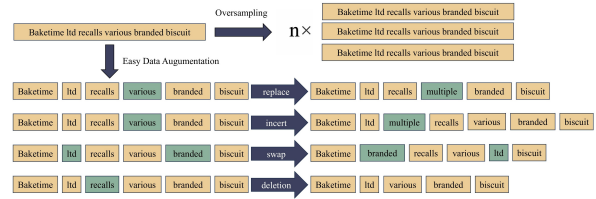

Figure 1: The structured pipeline for food hazard detection.

labels (Randl et al., 2025). The objective is to develop robust NLP models that accurately identify hazard types and product categories despite imbalanced, noisy, and unstructured text.

The dataset comprises thousands of entries obtained from government agencies. Each entry includes a title, typically a brief recall identifier, and a text description that varies in length and format, often containing domain-specific terminology. The data are marked by severe class imbalance, with many hazard categories significantly underrepresented.

4 Methodology

Recent work in food hazard detection highlights the importance of addressing data imbalance, short and unstructured text, and overlapping semantic categories (Adegbenjo and Ngadi, 2024). SemEval-2025 Task 9 intensifies these challenges by providing a real-world dataset where certain hazard classes are severely underrepresented, necessitating specialized techniques to ensure fair and robust classification. In this study, we adopt a transformer-based approach, leveraging BERT (Devlin, 2019) and RoBERTa (Zhuang et al., 2021), and integrate three key strategies—random oversampling, Easy Data Augmentation (EDA), and focal loss—to enhance performance on minority classes while maintaining overall accuracy.

Our methodology follows a structured pipeline, shown in Figure 1, which consists of data preprocessing and augmentation, transformer-based model fine-tuning, and final evaluation using standard classification metrics. The sections below describe how these components are cohesively combined to tackle the real-world complexities of food hazard reports.

4.1 Data Preprocessing and Augmentation

All textual inputs undergo cleaning and tokenization before fine-tuning. We remove stopwords, numerical tokens, and other non-informative el-

Operation	Description
Synonym Replacement	Replace with synonyms
Random Insertion	Insert random words
Random Swap	Swap two words
Random Deletion	Remove words ($p = 0.1$)

Table 1: Easy Data Augmentation (EDA) operations applied to the dataset.

ements, followed by lemmatization to standardize word forms. Outliers are then filtered using the interquartile range (IQR) to reduce extreme text lengths that could bias the model. We adopt Word-Piece tokenization (Song et al., 2020) to handle out-of-vocabulary (OOV) tokens, thereby preserving subword-level information crucial for short and domain-specific texts.

To counteract severe class imbalance, we introduce a combined augmentation strategy that integrates random oversampling with Easy Data Augmentation (EDA) (Wei and Zou, 2019). Oversampling is performed after tokenization to ensure each minority class is represented at a target sample rate. EDA, summarized in Table 1, is then applied to further expand the diversity of minority samples by introducing lexical and structural variations. Rather than applying each augmentation technique independently, we incorporate them into a unified process that consistently enhances minority-class coverage and lexical variety. This integrated augmentation stage aligns with prior findings that emphasize synergy between resampling and data augmentation for imbalanced text classification (Lauron and Pabico, 2016).

4.2 Classification Model and Imbalance Handling

The classification model builds on BERT and RoBERTa, which are fine-tuned for multi-class prediction. While cross-entropy loss remains the baseline choice, we adopt focal loss (Lin et al., 2017) to emphasize hard-to-classify examples in minority classes. The focal loss function is given by:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t), \qquad (1)$$

where α_t balances class contributions, γ focuses on difficult samples, and p_t is the predicted probability for the correct class. We set $\alpha = 1$ and $\gamma = 2$ based on initial experiments indicating improved recall for underrepresented hazards.

Random oversampling is performed using the strategy:

$$\text{sampling\_strategy} = \{\, k : \text{target\_count} \mid$$
$$v < \text{target\_count}, \forall k, v\,\}, \qquad (2)$$

where each minority class is upsampled to match a threshold of the majority class size. By applying oversampling in tandem with EDA, we ensure that minority classes benefit from both quantitative and qualitative increases in training samples.

4.3 System Configurations

We evaluate several configurations to highlight the effect of each balancing technique (Table 2). The *Baseline* employs standard BERT fine-tuning without augmentation, while additional setups incorporate oversampling, EDA, focal loss, or a combination thereof. We also include RoBERTa variants, reflecting the same augmentation and imbalance strategies. This design enables a comprehensive comparison of how each technique—alone or combined—contributes to classification performance on short, imbalanced food hazard reports.

This integrated methodology ensures that each stage—preprocessing, augmentation, model training—cooperates to address the unique challenges posed by SemEval-2025 Task 9, namely short, imbalanced, and domain-specific textual data.

5 Experiments

Our study strategically integrates three techniques—Easy Data Augmentation (EDA), oversampling and focal loss—to address class imbalance in classification tasks. The sequence of application is as follows: EDA is applied before tokenization to enhance data diversity, oversampling is applied after tokenization to balance class distribution, and focal loss is utilized during training to optimize the model's focus on difficult samples. These experiments not only demonstrate the effectiveness of each individual method but also highlight the synergistic benefits of their combination. The results show that this integrated approach enhances data diversity, balances class distribution, and improves model performance by prioritizing challenging samples.

5.1 Experimental Setup

Oversampling The sample rate is set to r. For classes whose size is smaller than r% of the most-frequent class, we perform oversampling to ensure

Configuration	Description
Baseline	Standard BERT fine-tuning with cross-entropy loss
BERT + Oversampling	Resample minority classes after tokenization
BERT + EDA	Apply data augmentation using EDA
BERT + Focal Loss	Replace cross-entropy with focal loss
BERT + EDA + Focal Loss	Combine lexical augmentation and focal loss
RoBERTa Variants	Mirror each configuration using RoBERTa

Table 2: Model configurations used for evaluation.

their size matches that of r% of the most-frequent class.

Easy Data Augmentation The sample rate is set to r. For each instance input, we perform EDA. Each of the four operations—Random Synonym Replacement, Random Insertion, Random Swap, and Random Deletion—has a 50% probability of being applied. For the first three operations, the parameter (n), which indicates the number of times the operation is to be performed, is randomly selected within the range of 1 to the total number of words in the text. In the case of the Random Deletion operation, each word is assigned a 10% probability of being deleted.

Focal Loss Alpha (α), the balance parameter for class imbalance, is set to 1. Gamma (γ), the focusing parameter for hard examples, is set to 2. The method for aggregating the loss values (reduction) is "mean".

5.2 Training Details

We utilized the 'title' and 'text' fields from the dataset released by the organizers. In the data processing phase, categorical labels were encoded into numerical values using the LabelEncoder. The dataset was subsequently split into training and testing sets, with 20% allocated for testing. If Easy Data Augmen- tation (EDA) was enabled, data augmentation techniques were applied specifically to the training subset (train_df). Additionally, if oversampling was employed, data augmentation was conducted after the tokenization process.

For our models, we utilized BERT (Devlin, 2019) and RoBERTa (Zhuang et al., 2021). During training, the total batch size was set to 32. The AdamW optimizer (Kingma, 2014) was used with a learning rate of 5e-5, and dropout was specified at 0.0. The learning rate schedule followed a 'cosine with warmup' strategy, incorporating a warmup phase equivalent to 10% of the total training steps

to gradually adjust the learning rate and enhance model convergence. The default loss function used was CrossEntropyLoss, unless focal loss was selected.

In our study on NLP food hazard classification, we concentrated on addressing the issue of class imbalance, particularly in predicting the most imbalanced label, "product." To evaluate our model's performance, we established BERT, provided by the organizers, as our baseline. The evaluation utilized inputs from both "text" and "title" to enhance the model's effectiveness. We employed several metrics to assess performance, including accuracy, F1-score (macro), and F1-score (weighted). See Table 3 for the results for predicting "product" (data split: training/validation/test).

Oversampling BERT achieved an accuracy of 0.22, with an F1-Macro score of 0.03 and an F1-Weighted score of 0.13, highlighting its limitations in handling the imbalanced dataset. Oversampling with a sample rate at 0.1 yielded the best performance among the oversampling variations, achieving an accuracy of 0.50, an F1-Macro score of 0.25, and an F1-Weighted score of 0.45. This indicates that oversampling can effectively address class imbalance and improve classification performance.

Easy Data Augmentation The experimental results presented in Table 3 also demonstrate the effectiveness of Easy Data Augmentation (EDA) in enhancing the performance of the model for the "product." Compared to the baseline accuracy of 0.22, the application of EDA across multiple configurations consistently improved performance. EDA with a sample rate of 0.2 increased accuracy to a maximum of 0.55. Similarly, the F1 macro score improved significantly from 0.03 to 0.30, while the F1 weighted score rose from 0.13 to 0.52. These findings underscore the potential of EDA as a valuable technique for improving model performance in imbalanced classification tasks.

(ST2) Product Detection

Training Methods	Accuracy	F1-Macro	F1-Weighted
$BERT_{base}$	0.22	0.03	0.13
$Oversampling_{0.1}$	0.50	0.25	0.45
$Oversampling_{0.2}$	0.45	0.22	0.41
$Oversampling_{0.5}$	0.47	0.23	0.42
$Oversampling_{1.0}$	0.29	0.13	0.26
$EDA_{0.1}$	0.54	0.29	0.50
$EDA_{0.2}$	0.55	0.30	0.52
$EDA_{0.5}$	0.55	0.30	0.51
$EDA_{1.0}$	0.54	0.30	0.51
Focal loss + $Oversampling_{0.1}$	0.49	0.24	0.44
Focal loss + $Oversampling_{0.2}$	0.47	0.23	0.41
Focal loss + $Oversampling_{0.5}$	0.48	0.25	0.44
Focal loss + $EDA_{0.1}$	0.53	0.29	0.50
Focal loss + $EDA_{0.2}$	0.54	0.29	0.51
Focal loss + $EDA_{0.5}$	0.54	0.30	0.51
Focal loss + $EDA_{1.0}$	0.53	0.29	0.50
Oversampling + $EDA_{0.1}$	0.49	0.25	0.45

Table 3: BERT model with different strategies for predicting product. The subscript isthe sample rate. For example, 0.1 means to upsample categories that are less than 10% of the maximum sample to 10% of the maximum sample.

(ST2) Hazard Detection

Training Methods	Accuracy	F1-Macro	F1-Weighted
$BERT_{base}$	0.58	0.17	0.53
BERT + Focal loss + $EDA_{0.1}$	0.86	0.59	0.85
RoBERTa + Focal loss + $EDA_{0.1}$	0.86	0.59	0.85

Table 4: BERT and RoBERTa with focal loss and EDA on predicting hazard. The subscript is sample rate. For example, 0.1 means to upsample categories that are less than 10% of the maximum sample to 10% of the maximum sample.

Combination   After observing that handling imbalance can enhance model performance, we explored the effects of combining this strategy with focal loss, which places greater emphasis on harder-to-classify examples. We also examined the combination of EDA with oversampling. While these combinations did result in performance improvements that surpassed the baseline BERT model, they did not achieve the same high levels of effectiveness as EDA alone. For instance, the combination of focal loss with oversampling yielded an accuracy of 0.49 at a sample rate of 0.1, while EDA at the same rate achieved an accuracy of 0.54.

Prediction on Hazard   To ensure that the model predicts well across different tasks, we also made predictions on "hazard." The combination of focal loss and Easy Data Augmentation (EDA) significantly increased performance from an accuracy of 0.58 for BERT to 0.86 at a sample rate of 0.1. We also investigated the RoBERTa model in a similar manner; however, it performed comparably to BERT, leading us to discontinue further exploration of RoBERTa. Table 4 shows some results of these experiments.

In summary, the results show that this integrated approach enhances data diversity, balances class distribution, and improves model performance by prioritizing challenging samples. Our findings indicate that EDA is a particularly effective technique for addressing class imbalance in food hazard classification. While combinations with focal loss and oversampling improve performance, they do not surpass EDA alone. Additionally, these methods may require further fine-tuning and more training steps to optimize their effectiveness.

6 Conclusion

In this paper, we strategically combined three methods - Easy Data Augmentation (EDA), oversampling, and focal loss - to tackle class imbalance in classification tasks. Our model ranked 13th on ST1, and 12th on ST2. It surpassed the baseline in both ST1, predicting product and hazard category, and ST2, predicting specific hazard and product.

Future work will explore strategies to enhance model performance in data-constrained and domain-shift scenarios. We plan to investigate the integration of advanced augmentation techniques and refined loss functions, along with fine-tuning the existing methodology.

Limitations

While our approach demonstrated notable improvements in food hazard classification, several limitations remain. First, despite the effectiveness of EDA and oversampling in mitigating class imbalance, these techniques may introduce synthetic noise into the dataset. Augmented samples, particularly those generated via lexical modifications, may not always accurately preserve the semantic meaning of the original text, potentially leading to misclassification.

Second, our reliance on transformer-based models presents computational challenges. Fine-tuning these models requires significant hardware resources, making it less feasible for real-time applications or deployment in resource-constrained environments. Additionally, while focal loss improves performance on hard-to-classify examples, it requires careful tuning of hyperparameters, which may not generalize well across different datasets or classification tasks.

Another key limitation is the potential lack of generalizability. Our model was trained on the SemEval-2025 Task 9 dataset, which, despite being a real-world dataset, has specific linguistic characteristics and class distributions. This may limit the model's ability to perform well on other food safety-related classification tasks with different text structures, hazard categories, or reporting styles. Future work should explore domain adaptation techniques and evaluate performance across multiple datasets.

Finally, while our approach improves minority class detection, the gap between majority and minority class performance remains. Additional techniques, such as contrastive learning, cost-sensitive training, or adaptive resampling, could be explored to further enhance model fairness and robustness.

References

Adeyemi O. Adegbenjo and Michael O. Ngadi. 2024. Handling the imbalanced problem in agri-food data analysis. *Foods*, 13(20).

Jacob Devlin. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Wei Fan, Haoran Li, Zheye Deng, Weiqi Wang, and Yangqiu Song. 2024a. GoldCoin: Grounding large language models in privacy laws via contextual integrity theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3321–3343, Miami, Florida, USA. Association for Computational Linguistics.

Wei Fan, Weijia Zhang, Weiqi Wang, Yangqiu Song, and Hao Liu. 2024b. Chain-of-choice hierarchical policy learning for conversational recommendation. *Preprint*, arXiv:2310.17922.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Maureen Lyndel C Lauron and Jaderick P Pabico. 2016. Improved sampling techniques for learning an imbalanced data set. *arXiv preprint arXiv:1601.04756*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2020. Fast wordpiece tokenization. *arXiv preprint arXiv:2012.15524*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.