The Food Hazard Detection using Deep Learning

Sufian Ahmad
Registration: 450554
MS Data Science
sahmad.msds23seecs@seecs.edu.pk

1 Introduction

Thousands of food hazard incidents are reported online daily, threatening consumer health, damaging manufacturers reputations, and impacting the economy. Accurate classification of these hazards is essential for understanding their implications, finding out the root cause and improving web crawler's ability to categorize incidents automatically. To address this challenge, deep learning techniques can be used to develop robust classification models. We can improve the accuracy and efficiency of identifying food hazards by using deep learning. This approach will not only optimize the detection process but also support taking measures to lower risks, ensure consumer safety, and protect reputation.

Classification of the food hazards from text and their explainability is an under-explored area in Deep Learning. Therefore, Deep Learning can be applied to food hazard incident reports to classify the type and category of food hazards. The relevance of the work is given below

- The authors [3] used traditional machine learning with tf-idf, transformer-based LLMs with few-shot prompt approaches (RoBERTa, XLM-R, llama3 and GPT-3) to verify the labels corresponding to the hazard news reported.

- There are numerous methods depending on the type of hazard, and selecting the appropriate one is challenging. Therefore, they [6] labelled the data and applied traditional machine learning models (SVM, RF) and multilayer perceptron (MLP) to classify the most effective methods based on the hazard type.

- The European Union RASFF data was used to compare neural and non-neural machine learning models for food safety risk prediction [2]. In neural models, they applied MLP and a Conv1d. In non-neural models, they used LR, RF, and SVM.

- The researcher [5] used a machine-learning approach to detect and classify food allergens in food. They applied Decision Tree, RF, SVM, and K-Nearest Neighbors models.

- They used a data augmentation technique to evaluate its impact on LLM performance for food hazard and product analysis, as LLMs struggle to provide accurate predictions due to short text [4]. Data was generated using ChatGPT-4o-mini and subsequently used to train two large language models, RoBERTa-base and Flan-T5-base.

2 Methodology

We chose the paper [3] as our baseline. The dataset was divided into 4067 for training and 1017 for testing samples. The number of classes in each category is in table 1.

Categories	Number of Classes
Hazard category	12
Product Category	24
Product	1024
Hazard	130

Table 1: Number of Classes in each Category

Our methodology involves using a BERT language model [1] and its tokenizer to generate tokens and embeddings from input text after removing the special symbols. These embeddings capture the semantic relationships between words and consist of numerical representations, which serve as input for deep learning models along with their corresponding labels. We trained the deep learning models (LSTM, MLP) from scratch and also used Logistic Regression. We fine-tune the BERT models (12 layers & 110M parameters) and compare the results. During the deep learning model training, I configure the different hyperparameters (number of layers, epochs, hidden size, learning rate, batch size) to observe their impact on the output and fine tune the Google BERT model to analyse the fine-tune approach with the training (LSTM, MLP). We used the F-score as the evaluation metric.

The pipeline for BERT in our methodology is shown in Fig. 1.

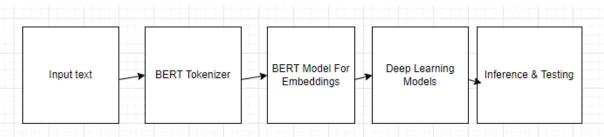

Figure 1: pipeline with BERT Model

3 Results

First, I trained the LSTM and MLP on the embeddings from the BERT model. Secondly, I used the Logistics regression on the same embeddings, and third, I fine-tuned the BERT to make a comparison of results with the baseline paper. The results are summarised in fig 2. The details of results can be found on github.

Model	F1-Score
Hazard-Category	
BERT-LSTM	0.8252
BERT-MLP	0.8124
BERT-LR	0.822
BERT-FineTune	0.8504
Product-Category	
BERT-LSTM	0.44
BERT-MLP	0.4873
BERT-LR	0.5211
BERT-FineTune	0.7372
Hazard	
BERT-LSTM	0.0092
BERT-MLP	0.418
BERT-LR	0.4427
BERT-FineTune	0.506
Product	
BERT-LSTM	0.0614
BERT-MLP	0.003
BERT-LR	0.009
BERT-FineTune	0.3

Figure 2: Results Comparison

The fine-tuned BERT outperform the other approaches. The comparison of the fine-tuned BERT with the baseline paper [3] is given in Fig 3

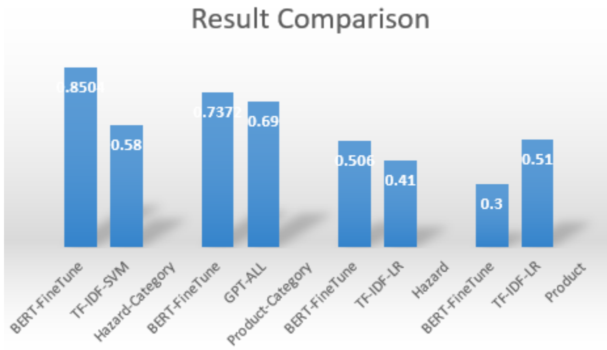

Figure 3: Results Comparison With Baseline Paper

As we can observe from the Figure 3, the BERT fine-tune model has a higher F1 Score compared to the baseline paper, except for one category of products. In the hazard category, the F1 score is 0.85 as compared to (TF-IDF-SVM) 0.58. For product-category, it is 0.735 in contrast to GPT-ALL(0.69) and hazard category it is 0.508 as compared to TF-IDF-LR (0.41). It shows that the Fine-tuned BERT approach is better than all other approaches, and it has significantly improved the F1 score.

4 Discussion and Conclusion

The dataset was highly imbalanced, and I observed the overfitting issue during the training of the LSTM and the MLP. I configure the hidden size, number of layers and epochs starting from lower and gradually increasing to observe the overfitting effect. To overcome overfitting, I used the simpler model, logistic regression, and there was some improvement in the results, but the issue of overfitting was still persistent. At last, I used the pre-trained BERT model and fine-tuned it and observed that fine-tuning has significantly improved the results and overfitting issue was somewhat solved, but still present. To conclude, the fine-tuned BERT approach outperforms the other approaches. The fine-tuned BERT model is highly effective in finding out the hazards from the hazard news reported. Due to an imbalanced class dataset, the overfitting was still present, and the up-sampling and down-sampling were not useful due to only a few samples per class in each category.

5 Future Work and Team Contribution

My Contributions are as follows:

- I used the large language model BERT to generate embeddings for the food hazard news.
- I train the deep learning models (LSTM, MLP) on embeddings generated from the BERT Model and use different hyperparameter configurations that can be found on github.
- To overcome the overfitting issue, I used the simpler model, LR and fine-tuned the BERT model.
- Make a comparison of results, and the fine-tuned BERT model outperforms the other approaches.

I used the only BERT model for embeddings, therefore, other LLMS can be used. The Few-shot, one-shot can also be used with the instruction-tuned LLM. To overcome over overfitting issue, the data can be extended, especially for those classes that have few samples.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Alberto Nogales, Rodrigo Díaz-Morón, and Álvaro J García-Tejedor. A comparison of neural and non-neural machine learning models for food safety risk prediction with european union rasff data. *Food Control*, 134:108697, 2022.

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

[4] Areeg Fahad Rasheed, M Zarkoosh, Shimam Amer Chasib, and Safa F Abbas. Data augmentation to improve large language models in food hazard and product detection. *arXiv preprint arXiv:2502.08687*, 2025.

[5] Hafsah Shaukat, Arooj Sultan, and Humayun Salahuddin. Enhancing food safety: A machine learning approach for accurate detection and classification of food allergens. *Journal of Computing & Biomedical Informatics*, 2024.

[6] Long-yu Zhu, Lijuan Yan, Fang Zhao, Xuewen Guo, Dunming Xu, Jingzhang Lv, Lin Ding, Na Niu, Jun-qin Qiao, Shumian Ma, et al. Evaluation of methods for the detection of hazardous substances in food based on machine learning. *New Journal of Chemistry*, 48(3):1399–1406, 2024.