Towards Explainable Food Hazard Detection: A Neuro-Symbolic Approach

Neelima Prasad
CU Boulder
nepr1244@colorado.edu

Karthik Sairam
CU Boulder
kasa6776@colorado.edu

Advait Deshmukh
CU Boulder
adde1214@colorado.edu

1 Introduction

Task / Research Question Description Food safety is becoming an increasingly important issue worldwide. As our food systems grow more complex and interconnected, the risks of contamination and food borne illnesses rise. Moreover, with the rise of social media, there are a myriad of food safety reports flooding the web that is difficult to sort through. We propose a classification model that would be able to identify and explain food-related risks as well as particular hazards from online sources. For any text source, our model would be able to perform text classification for food hazard prediction to the type of hazard and product, as well as food hazard and product vector detection, to predict the exact hazard and product.

Motivation and Limitations of existing work Food risk classification based on texts is currently under explored. Previous methods, like (Tao et al., 2021) struggle with explainability, since there are many diverse approaches for it. Due to the potential high economic impact, transparency is crucial for this task. We aim to create a high explainable model that can not only predict hazards but is also understandable. This transparency is crucial for trust and practical application in food safety. Given the need for explainability with this task, we turn to a neuro-symbolic approach to leverage human reasoning and discretization which results in a more interpretable model.

Likely challenges and Mitigations This task is difficult as it involves balancing the model's ability to generalize across related tasks while specializing in a particular domain. The main challenge to overcome in solving this task would be to create an architecture to improve upon the baseline that BERT provides, which is an F1 score of 0.83. In class, we have studied techniques that leverage modular approaches to improve upon neural methods. Our biggest struggle will be to exploit domain specific knowledge in a way that will boost performance and generate results that are more explainable and generalizable. To make the overall task of identifying the product and the hazard associated with it for each text, we break the problem down into two subtasks. The first subtask would be text classification for food hazard prediction, or predicting the type of hazard and product. The second subtask would be food hazard and product vector detection, or predicting the exact hazard and product. By breaking up the problem like this, we ensure that even if we are unable to produce quality results for the second subtask, that our model would at least be able to classify the type of hazards and products associated with it.

2 Related Works

(Zini and Awad, 2022) present a survey on the explainability of deep models in NLP by underlining the importance of explainability in domains where understanding the decision-making process is critical, which directly corresponds to our task of building an explainable food hazard detection model. The authors focus on elucidating why explainability is especially tricky when it comes to textual data, supporting their claim with reasons such as the opaque nature of word embeddings and the inherent interpretability of the attention mechanism in transformers. The specific avenue that we would like to consider building upon from this paper is that of quantitatively assessing the explainability of our model and its textual data.

The authors of (Assael et al., 2022) propose Ithaca, a deep neural network architecture that can perform the tasks of restoring ancient texts from Greek inscriptions, in addition to also attributing a place of origin and date of writing of the inscription. We think that this is relevant because of its model-specific explanation, in which the authors'

claim that a "high-level of generalization is often involved" (in epigraphy), resonates with the fact that food risk classification often is not accompanied by transparency. Their use of models that perfectly fit the three tasks Ithaca excels at (text restoration, geographical attribution and geographical attribution), suggests that such an approach intends to enable the readers and the authors to not only have a deeper understanding about the solution proposed, but also to reason about the structure of the model. We will be building on a similar approach, albeit incorporating a symbolic approach for the food hazard detection solution, where we will leverage model-specific explanations.

In (Ribeiro et al., 2016), the main focus is to develop a system that explains why a classifier made a certain prediction, which is done by identifying the important parts of the input that contribute towards the decision making and presenting the visual artifacts that establish a relationship between the input and the prediction. Since model-agnostic methods do not take into account the model's structure since they work on a black-box approach, the authors intend to develop Local Interpretable Model-agnostic Explanations (LIME). The outlined drawbacks of evaluating only the accuracy of the model (dataset shift, cross validation overestimation and data leakage) to explain it's performance are interesting, since they can also be applied directly towards the explainable food hazard detection model we will be working on. Our proposed methodology takes a slightly different approach, which is in being explainable right from the start, hence not needing an approach to generate explanations using a separate methods like the one the authors have proposed.

An attention based mechanism was deployed by (Pavlopoulos et al., 2022) in context of toxicity detection. One of the approaches they experimented with in their work was a systematic application of attention as a rationale extraction mechanism which is applied at inference. This added a layer of explainability to the problem of toxic label identification for text. Moreover, by applying a probability threshold to the attention scores for each token of the post, they achieved impressive results in the task of toxic sequence detection. While the idea of analysing the last layer of the models adds some explainability to the approach, it doesn't overcome the black box nature of the internals of the model. We intend to have more explicit symbolic components that would be employed earlier in the approach.

In an effort to build a system for early detection of food hazards, a framework proposed by (Ihm et al., 2017) aims at extracting information from social media and online news. The authors propose a multi step framework to extract, filter and process data from multiple online sources. They employ neural methods at the document filtering stage followed by rule based methods to fill the food hazard event templates prescribed by the Korean government. This task that authors address is similar to our proposed task. However, the strictly rule-based approach for extracting information comes with its own set of challenges. This can often cause missed fields leading to incomplete information. A hybrid approach will attempt to overcome this limitation.

(Tao et al., 2021) propose an approach in extracting entities related to food-borne outbreaks from twitter posts. They develop a dual-task BERTweet model to a) classify tweets and b) extract entities related to the outbreak. They modify the architecture of BERTweet model for the proposed tasks and achieve state of the art performance on the first task and a high precision on the second task. Though this seems impressive, their approach remains purely neural and lacks explainability. We intend to use a similar BERT-based baseline for our work, with our symbolic component remaining a key differentiating factor.

References

Yannis Assael, Thea Sommerschield, Brendan Shillingford, et al. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603:280–283.

Hwon Ihm, Kyoungrok Jang, Kangwook Lee, Gwan Jang, Min-Gwan Seo, Kyoungah Han, and Sung-Hyon Myaeng. 2017. Multi-source food hazard event extraction for public health. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 414–417.

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.

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.

Dandan Tao, Danying Zhang, Ruoyu Hu, et al. 2021. Crowdsourcing and machine learning approaches for extracting entities indicating potential foodborne outbreaks from social media. *Scientific Reports*, 11(1):21678.

Julia El Zini and Mariette Awad. 2022. On the explainability of natural language processing deep models. *ACM Comput. Surv.*, 55(5).