Towards Explainable Food Hazard Detection: A Neuro-Symbolic Approach

Karthik Sairam
CU Boulder
kasa6776@colorado.edu

Neelima Prasad
CU Boulder
nepr1244@colorado.edu

Advait Deshmukh
CU Boulder
adde1214@colorado.edu

Abstract

As food systems become increasingly complex, ensuring food safety is paramount. Traditional methods for detecting food hazards often lack transparency and interpretability, which are crucial for practical applications. In this paper, we present a novel neuro-symbolic approach to food hazard detection, leveraging a combination of neural and symbolic models for enhanced explainability. Our method employs CoCo-Ex to extract concepts from food recall titles and maps them to ConceptNet nodes, followed by filtering irrelevant nodes using the Llama-3.1-8B-Instruct model. A streamlined, context-specific sub-graph of ConceptNet is constructed to establish relationships between hazard/product labels and recall titles. By employing measures like distances between concepts in ConceptNet, we build a hazard classification system which is inherently explainable. While initial results show room for improvement compared to neural-only baselines, this approach, with its superior performance in low-data setting highlights the potential of integrating symbolic reasoning with neural models to improve model transparency and performance in food safety applications.

The code for our pipeline is available at: https://github.com/karthiksairam01/SemEval-Task9

1 Introduction

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. Since this is a complicated, unsolved problem, SemEval released a task called The Food Hazard Detection Challenge, which is designed to evaluate explainable classification systems for titles of food-incident reports collected from the web. To solve this task, we propose a classification model that would be able to predict specific label categories, i.e. hazard(s), product(s) involved in food recall titles from online sources. We aim to create a highly explainable model that can not only predict hazards and products that caused those hazards, but a model that is also understandable. This transparency is crucial for trust and practical application in food safety. Given this need, we turn to a neuro-symbolic approach to leverage human reasoning and discretization which results in a more interpretable model. The approach that we propose for this uses a neural model, Coco Ex, to extract meaningful concepts from the input food recall titles and map them to nodes in ConceptNet. We then filter out the irrelevant nodes using Llama-3.1-8B. Lastly, we use the relevant nodes to algorithmically generate a sub-network of ConceptNet, a popular semantic network, which would allow the parsed labels to be matched to the titles.

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 analyzing 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.

(Becker et al., 2021) introduces a concept extraction tool for ConceptNet called CoCo-Ex, which identifies and extracts concepts from natural language texts and maps them to ConceptNet. It can be used as a way to detect and classify knowledge relations instantiated within texts. We use CoCo-Ex for our task to extract relevant keywords from the input recall titles and labels.

3 Dataset

The training dataset (Randl et al., 2024) comprises of 5082 instances of textual data, ranging from 5 to 277 characters in the food recall titles. It also contains the text from these articles. The food recall titles are manually labeled from food safety authority agencies with the relevant hazard, hazard-category, product and product-category.

On analyzing the data, some keen observations that will shape how we build the framework moving

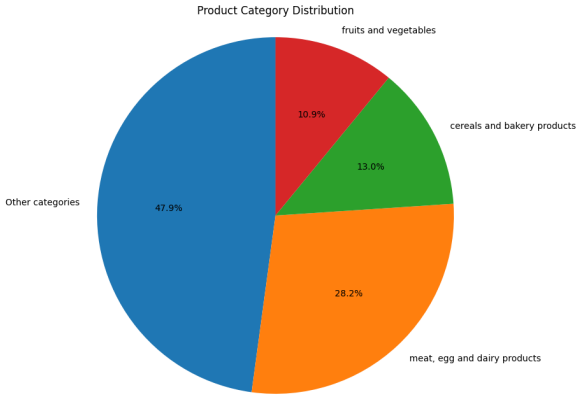

Figure 1: Distribution of product categories in the dataset. The 'Other categories' is purely indicative of other named product categories.

forward are presented below:

1. The data also includes the features such as the release date of the food recall article, and 'hazard-title'/'product-title' which are the spans of the food recall title that are relevant to the classification at hand. These spans are generated via a Logistic Regression classifier based on feature importance.

2. Out of the 127 hazards in total, it contains 2 hazards which only have one title associated with them, meaning that the dataset has only one title example for these specific hazards.

3. The dataset also contains titles that are repeated (one title is repeated 11 times)

4. A popular product-category is meat, egg and dairy products, as can be seen in Figure 1.

Presence of hazards with single/few examples is a challenge for neural approaches as it doesn't provide the model with enough information for training. The neural approach loses on the information present in the label name and simply uses it as a category which is something knowledge-enriching approaches can leverage to their advantage. Hence, we propose a Neuro-symbolic framework that leverages knowledge graphs to reason about the classification task.

4 Background Information : Models

The modules used for our framework are CoCo-Ex, Llama, and ConceptNet. We outline the overview for each of these modules prior to detailing about the pipeline of our framework.

4.1 CoCo-Ex

CoCo-Ex (Becker et al., 2021) is a tool for extracting concepts from texts and linking them to the ConceptNet knowledge graph. Their methodology begins by extracting candidate phrases from the given text using the Stanford Constituency Parser, which are then preprocessed with Spacy, where lemmatization is applied. Following preprocessing, the types of the candidate phrases are matched against a dictionary based on ConceptNet, utilizing word embeddings for semantic similarity. We use CoCo-Ex to create candidate nodes for each concept (title/label) in our pipeline.

4.2 LLaMA

LLaMa (Grattafiori et al., 2024) is a collection of foundation language models with various amounts of parameters. For this task, we use the Llama-3.1-8B-Instruct model which has 8 billion parameters. We primarily use it to filter the candidate nodes generated by CoCo-Ex.

4.3 ConceptNet

ConceptNet (Speer et al., 2018) is an open, multilingual knowledge graph that contains nodes that model concepts (single or multiple words) and labeled edges that model relations. Some example relations are "is a", "is used for" and "part of". ConceptNet 5.7 contains a total of 36 relations, which include both symmetric and asymmetric relations. We primarily use ConceptNet to judge the distances between different "Concept clusters" and assign a category to each title.

5 Methodology

Our method to addressing the classification problem is comprised of three components, as shown in Figure 2. The goal is to map each of the food recall titles with the correct product/product category and hazard/hazard category labels. There are 22 possible product categories and 10 possible hazard categories. There are 1,005 specific products and 116 specific hazards.

We approach this problem by passing two inputs into our neuro-symbolic pipeline: the food recall titles and the labels, independently. These get passed initially into CoCo-Ex, which extracts meaningful concepts from the input and maps them to conjunct concept nodes in ConceptNet. However, many of these nodes are unnecessary/misleading (The phrase "Peanut butter" being processed into

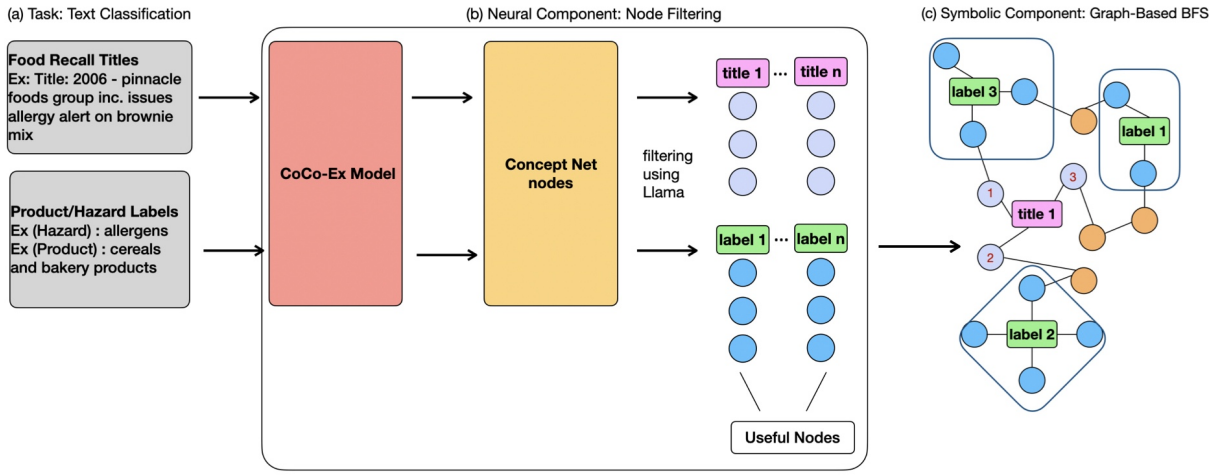

Figure 2: Diagram of our model's architecture, broken up into three parts. Part (a) depicts the inputs, which are both the titles of the food recall notifications as well as the predicted labels. Part (b) details the neural component of our model which shows how we create relevant nodes from the input text. Part (c) shows how we convert the text classification problem into a graph and use BFS to traverse the edges.

"Peanut", "Butter", "Peanut butter") for our specific task, so to further narrow down the nodes to just the relevant one, we use the Llama-3.1-8B-Instruct model with the prompt "For given text and extracted candidate keywords, discard the irrelevant ones and return the relevant ones." This part of our pipeline returns a set of relevant ConceptNet nodes for every label and every title.

From these nodes, we are able to use ConceptNet (v5.7) to establish connections between the label groups and the titles. However, the problem with using ConceptNet is that it includes over 21 million edges and over 8 million nodes. Including too many nodes would bring given any two concepts closer, thereby taking away from the informative nature of the distance metric. Moreover, making a large number of API calls makes traversing the graph extremely slow.

To circumvent these issues, we create a lite version of ConceptNet, essentially by building an offline sub-network of ConceptNet. To do this, we prune ConceptNet by limiting the nodes to the English language, and at the same time, limit the number of relations used. As described above, ConceptNet has 36 relations between nodes. For the given classification task and its domain, we found that using the relations FormOf, Synonym and RelatedTo gave us best results. FormOf relation connects different forms of the same word (meat and meats, oil and oils, etc.). The Synonym relation is useful in ensuring coverage as the same term might be referred to by different names in different concepts (poultry-chicken). We aim at encapsulating all other relations in the network using the RelatedTo relation which allows us to ignore other fine grained relations, as their granularity can add to the complexity while not necessarily improving performance. Curating the edges of ConceptNet gives us control over the size and the complexity of the sub-network that we run queries over.

After the sub-network curation, we proceed to form sub-graphs for each label. We use Breadth First Search (BFS) to keep track of all the nodes present within "n" hops from a given label. Since a given label can amount to multiple concepts ("Coffee and Tea"- "Coffee", "Tea") we use multi source BFS to keep track of neighbors of a given label-cluster. After experimentation, we set "n" (the least number of hops from the source nodes) to the optimal value of 5. As show in Figure 2 part(c), we group each of the nodes associated with each label in a cluster (depicted by the blue squares). For each label cluster, we calculate the distance to each title node, and choose the label that outputs the minimum distance. As can be seen in 2, label 3 is the predicted category, since it has the shortest path (distance 1) to one of the title nodes. This process is iteratively run over the 4 label categories.

An algorithm of this approach is provided in Algorithm 1

6 Evaluation Metrics

We evaluate our approach by measuring its performance on each task by calculating the macro-avg-

Algorithm 1 Food Hazard Classification Pipeline

Require: Titles dataset T, Label categories dataset L, ConceptNet C

Ensure: Classified titles into hazard and product categories

Step 1: Convert ConceptNet to a Graph
Step 2: Extract Keywords
for each title t in T **do**
    Use LLaMA to extract keywords from t
end for
for each label category l in L **do**
    Use LLaMA to extract keywords from l
end for
Step 3: Filter Keywords
Step 4: Perform Multi-Source BFS
Define max distance d
for each start node n in T and L keywords **do**
    Perform BFS up to depth d from n on G
    Record distances of all reachable nodes
end for
Step 5: Generate Subgraphs
for each label category l in L **do**
    Create subgraph S of G containing nodes within d of keywords in l
    Store S
end for
Step 6: Classification
for each title t in T **do**
    Compute distances of t's keywords to hazard and product subgraphs
    Assign t to the category with the minimum distance
end for
Step 7: Evaluation
Compare model predictions against ground truth labels in L
Use evaluation metrics to evaluate performance

F1-score on the predicted labels using the annotated labels as ground truth. Our model's results are compared to the performance of the baseline model — BERT (randlbem, 2024).

7 Results and Discussion

We evaluated the performance of our proposed framework on a subtasks involving four labels, hazard category, product category, hazard, and product. The results were compared against the baseline provided by the SemEval 2025 (randlbem, 2024) task authors. The precision, recall and F1 scores are summarized in Table 1. For the hazard label, our framework shows an improvement over BERT in precision and recall, the slight drop indicating a slight trade-off in the balance of precision and recall.

For the product label, our framework significantly outperformed BERT in both precision and recall.

The results indicate that while our framework introduces improvements over the baseline in specific metrics, the overall performance seems to be limited due to a variety of factors:

1. **Label Space Inflation**: One significant challenge in this study was the labeling scheme of the dataset. Several products, which are either highly similar or essentially identical are treated as distinct labels. For example, spice mix, spices, mixed spices, spice mixture, masala spice mix, masala spice mixture, and spice marinade are all separate product labels. This adds unnecessary complexity since many of these terms are almost semantically identical and hampers the symbolic component's ability for the task of title classification. It's real world application also seems to offer little to no value. This is also evident in one more set of labels involving spinach: spinach, spinach leaves, baby spinach, canned spinach, and frozen spinach.

2. **Inadequacy of links in ConceptNet**: Another challenge we faced in our proposed solution was linked to nodes that are not well represented in ConceptNet, specifically within the English part. We found that for multiple instances ("baking mix", "smoked ham", "staphylococcal enterotoxin", etc.,) there are no links available to other english language concepts. This limits the representation of a label or sometimes completely eliminates a label from being present in our sub graph; causing no input examples to be attributed to that label. While it is a reasonable expectation to have such concepts well connected within english ConceptNet, we hope that such limitations will be addressed in subsequent versions of ConceptNet.

8 Conclusion and Future Work

The Food Hazard Detection Challenge is a classification task aimed at predicting specific prod-

Task	Predicted Label	Precision	Recall	F1-score
BERT	Hazard	0.22	0.19	**0.19**
	Product	0.02	0.04	0.02
Ours	Hazard	**0.25**	**0.22**	0.15
	Product	**0.11**	**0.12**	**0.09**

Table 1: Comparison of performance metrics for the BERT model and our proposed method for the products and hazards.

Task	Predicted Label	Precision	Recall	F1-score
BERT	Product-Category	0.57	0.58	0.57
	Hazard-Category	0.68	0.59	0.61
Ours	Product-Category	0.23	0.17	0.15
	Hazard-Category	0.14	0.19	0.11

Table 2: Comparison of performance metrics for the BERT model and our proposed method for the product and label categories.

ucts/product categories and hazards/hazard categories for food recall titles. We proposed a neuro-symbolic graph-based classification model that is explainable and provides better performance than the baseline on some of the tasks.

The superior performance of our method in the product classification task (1,256 labels) and competitive performance in hazard classification (261 labels) makes a strong case for neuro-symbolic approaches, especially in low-data regimes. We believe that this is because of our method's focus on leveraging the information encoded in label names—as opposed to standard BERT based classification which completely ignore this information.

Our method involved using CoCo-Ex to create nodes from the given inputs and filter out the irrelevant ones using Llama and create a graph using these nodes as vertices and the edges as relations from our ConceptNet sub-graph. We then performed a BFS on the graph to calculate the distance from label clusters to specific titles. Our results showed that our neuro-symbolic approach was able to mostly predict specific products and hazards better than the baseline, while also succeeding in being more explainable in its implementation and working.

Future work includes swapping the 8 billion parameter Llama model for the model Llama-3.1-70B-Instruct, which contains 70 billion parameters. In addition, while we are using this pipeline specifically for food classification, this will work for any text-based classification problem. Generalizing our pipeline for future domains would be a valuable next step. As discussed earlier, current version of ConceptNet is not without limitations. Hurdles like English phrases constantly referring to phrases in other languages come naturally with usage of ConceptNet. The authors believe that, as the rudimentary tools for implementing symbolic concepts in neural models improve, so will the performance and the explainability of neuro-symbolic models.

References

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

Maria Becker, Katharina Korfhage, and Anette Frank. 2021. COCO-EX: A tool for linking concepts from texts to ConceptNet. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jen-

nifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Kornelius Randl, Michail Karvounis, Georgios Marinos, John Pavlopoulos, Tony Lindgren, and Aron Henriksson. 2024. Food recall incidents.

randlbem. 2024. Food hazard detection semeval 2025. `https://github.com/food-hazard-detection-semeval-2025/food-hazard-detection-semeval-2025.github.io`.

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.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. Conceptnet 5.5: An open multilingual graph of general knowledge. *Preprint*, arXiv:1612.03975.

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).