VerbaNexAI at SemEval-2025 Task 9: Advances and Challenges in the Automatic Detection of Food Hazards

Andrea Menco Tovar and **Edwin Puertas** and **Juan Carlos Martinez-Santos**
Universidad Tecnológica de Bolívar
Cartagena, Colombia
amenco@utb.edu.co, epuerta@utb.edu.co, jcmartinezs@utb.edu.co

Abstract

Ensuring food safety requires effective detection of potential hazards in food products. This paper presents the participation of VerbaNexAI in the SemEval-2025 Task 9 challenge, which focuses on the automatic identification and classification of food hazards from descriptive texts. Our approach employs a machine learning-based strategy, leveraging a Random Forest classifier combined with TF-IDF vectorization and character n-grams (n=2-5) to enhance linguistic pattern recognition. The system a notable performance in hazard and product classification tasks, obtaining notable macro and micro F1 scores. However, we identified challenges such as handling underrepresented categories and improving generalization in different contexts. Our findings highlight the need to refine preprocessing techniques and model architectures to enhance food hazard detection. We made the source code publicly available to encourage reproducibility and collaboration in future research.

1 Introduction

Detecting food hazards is an essential challenge to ensure the safety and quality of food products globally (FAO and WHO, 2007; Nogales et al., 2020). Following this line, the task proposed in SemEval-2025 Task 9: The Food Hazard Detection Challenge focuses on identifying and classifying different types of hazards associated with food products through the analysis of descriptive texts. This task is of great relevance due to the growing need to monitor and ensure food safety and the need for automated systems that can process large volumes of data, which facilitates the early detection of potential hazards in food products globally (Randl et al., 2025). The ability to accurately and efficiently detect these hazards contributes directly to preventing public health incidents and improving food safety standards (WHO, n.d.; USDA, U.S. Department of Agriculture, 2024).

Our system employs a strategy based on machine learning and supervised classification using a Random Forest classifier combined with a TF-IDF vectorization to represent textual features. We implemented an n-gram character analysis approach (n=2-5) to capture relevant linguistic patterns to distinguish between different categories of hazards and products, hazard and product. This method facilitated the extraction of contextual information. It improved the model's ability to generalize from training data, thus optimizing prediction accuracy on new unlabeled datasets. By participating in this task, our system a notable performance compared to other teams, obtaining outstanding micro and macro F1 scores in hazard and product categories and hazard and product detection sub-tasks. However, we identified specific challenges, such as the difficulty in handling underrepresented categories and the need to improve the model's generalization in broader contexts. These findings underscore the importance of refining preprocessing techniques and model architecture to address the inherent complexity of food hazard detection effectively.

Participation in this task revealed promising results, positioning our system competitively against other participating teams. Quantitatively, our model obtained outstanding macro and micro F1 scores in the hazard and product category classification sub-tasks, reflecting high accuracy and robustness. However, significant challenges were identified, such as difficulty in handling linguistic ambiguities, difficulty in handling underrepresented categories, and variability in textual descriptions, suggesting areas of improvement for future developments. These findings underscore the importance of refining preprocessing techniques and model architecture to address the complexity inherent in food hazard detection. We published the source code used for developing and training our model to encourage reproducibility and collaboration in future work related to food hazard detection.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1–6

It is available through the following link [1].

2 Background

SemEval 2025 Task 9: The Food Hazard Detection Challenge focuses on automatically identifying and classifying food hazards from textual descriptions of food products. The input type for this task consists of descriptive texts detailing food safety-related incidents, such as reports of contaminants in food products or warnings about unsafe food handling practices see Table1.

The datasets used in this task include mainly training and validation divisions, covering various textual genres related to food incidents. The genre of the data encompasses food safety incident reports from multiple sources, ensuring a varied representation of contexts and scenarios. In terms of size, the training set contains approximately 5,082 samples. In contrast, the test set includes about 5,984 examples, allowing for robust training and accurate model performance evaluation. The competition has two sub-tasks focused on detecting different levels of granularity in the hazard and product categories. We participated in sub-task ST1, which focused on text classification for food hazard prediction, predicting the type of hazard and product, and sub-task ST2, which focused on food hazard and product vector detection, predicting the exact hazard and product. These approaches contribute to strengthening safety in the food supply chain, reducing the incidence of hazards, and protecting consumer health. In addition, it is possible to design and implement more targeted and effective prevention strategies by having a precise classification and accurate identification of affected products. Also, by clearly identifying the risks and products involved, companies and regulators can allocate resources more efficiently to address the most significant risk areas.

In developing our system, we have employed supervised classification methods that have demonstrated efficacy in similar natural language processing tasks. This approach aligns with previous studies showing the effectiveness of Random Forest-based methods for text classification tasks (He et al., 2024; Onyeaka et al., 2024; Qiu et al., 2025) providing a solid foundation for our methodology. Furthermore, using TF-IDF vectorization techniques combined with Random Forest classifiers is not novel. Research such as (Sabri et al., 2022; Sathishkumar et al., 2023) highlighted the effectiveness of these methods in text classification. However, in our contribution, we have implemented a vectorization strategy based on n-grams of characters in the TF-IDF vectorizer, which improves the capture of complex and contextual linguistic patterns present in food incident descriptions. Furthermore, integrating preprocessing techniques and hyperparameter optimization for the Random Forest classifier represents innovations that enhance the accuracy and robustness of the model in different environments.

[1] https://github.com/VerbaNexAI/SemEval2025

3 System Overview

This section details how we integrated advanced natural language processing techniques and robust machine learning models to transform and analyze the information, allowing the accurate identification of incidents and addressing inherent challenges such as semantic ambiguity. It explains, step by step, the key components from vectorization and the application of the RandomForestClassifier to the modular organization of the pipeline, providing a clear and complete overview of the process see figure1.

3.1 Algorithms

The proposed system combines advanced natural language processing (NLP) techniques with robust machine learning models to address the task of food hazard detection and classification from textual data. The main components of the system are detailed below:

Text Vectorization: TfidfVectorizer with Character N-grams. The representation of textual data is a fundamental step in any PLN system. In this work, the TfidfVectorizer from the scikit-learn library transforms food incident titles into numerical feature vectors. The specific configuration employs n-character frames ranging from 2 to 5, which allows for capturing local and contextual patterns in the texts. In addition, the following parameters strip_accents='unicode' are applied to remove accents from characters to reduce linguistic variability, max_df=0.5 to ignore terms that appear in more than 50% of the documents, which helps to eliminate overly frequent and uninformative words, and min_df=5 which considers only terms that occur in at least five documents, ensuring that features are relevant and representative. This configuration allows a rich and discriminative representation of

Table 1: Expected inputs and outputs.

Input	Output			
Title	**hazard-category**	**product-category**	**hazard**	**product**
Imported frozen duck tongue sample tested positive for COVID-19 virus in Macao	biological	meat, egg and dairy products	virus	duck
P&B (Foods) recalls Ahmed Foods Garlic Pickle in Oil and Mango Pickle in Oil because of undeclared mustard	allergens	herbs and spices	mustard and products thereof	garlic pickle

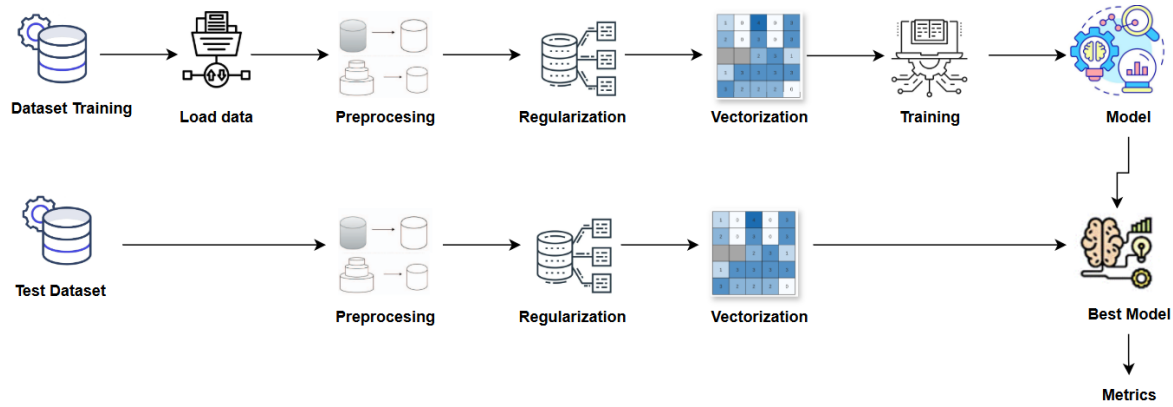

Figure 1: Outline of the proposed model.

the texts, facilitating the subsequent classification task.

Classifier: RandomForestClassifier. For the classification stage, we used RandomForestClassifier from the scikit-learn library. We selected this ensemble model for several reasons. Firstly, it can handle high-dimensional data and provide accurate results without exhaustive hyperparameter tuning. The reduction of overfitting follows this because by training multiple decision trees and averaging their predictions, we decreased the probability of overfitting, improving the model's generalization. Finally, due to the ease of interpretation, it offers some interpretability through feature importance, which can be helpful in understanding which text patterns are most relevant for classification (Nair et al., 2024; Spangenberg et al., 2024). The classifier is set up with 100 estimators and a fixed seed (random_state=42) to ensure the reproducibility of the results.

Machine Learning Pipeline. The system integrates text vectorization and classification into a pipeline that simplifies the training and prediction process. The pipeline is composed of two main stages: **1.** vectorization: application of the TfidfVectorizer to transform texts into feature vectors. **2.** Classification: Training and prediction using the RandomForestClassifier. This modular approach facilitates experimentation and tuning of each system component independently. It ensures that the system takes full advantage of the information available in the data provided without relying on external sources that may introduce biases or inconsistencies.

3.2 Challenges and Solutions

The detection and classification of food hazards from textual descriptions present several inherent challenges. Below are described these issues and the strategies implemented to address them.

3.2.1 Semantic Ambiguity

Challenge: Incident descriptions may be ambiguous or contain terms that have multiple meanings depending on the context.

Solution: Using character n-grams in vectorization allows capturing specific patterns that help dis-

ambiguate terms based on their local context within the text. In addition, the RandomForestClassifier can identify combinations of features that represent specific contexts, improving the model's ability to handle ambiguity.

3.2.2 Linguistic Variability

Challenge: Diversity in linguistic expression, including synonyms, grammatical variations, and typographical errors, can make the classification task difficult.

Solution: Preprocessing includes accent removal and text normalization to reduce variability. In addition, character n-grams allow capturing patterns even in the presence of typographical errors, as they consider more minor character sequences that can be robust to such variations.

3.2.3 Handling Multiple Classification Categories

Challenge: The task involves classifying multiple labels simultaneously, such as hazard categories, product categories, hazards, and specific products.

Solution: Implement a multi-label classification approach by training separate models for each target label within the same pipeline. It allows each model to specialize in a specific task, maintaining consistency and improving the system's overall accuracy.

3.2.4 Data Shortage for Some Categories

Challenge: Some categories may have less data available, affecting the model's ability to learn representative patterns.

Solution: Setting min_df=5 in the TfidfVectorizer helps to focus on terms that appear frequently enough, preventing the model from being affected by extremely rare categories. In addition, using RandomForestClassifier with multiple estimators contributes to better generalization even in unbalanced classes.

4 Experimental Setup

4.1 Division of the Data

We divided the data set provided into three main subsets: training, development, and testing. Initially, we loaded training data from incidents_train.csv and test data from incidents_labelled.csv. Then, we used an additional validation set called incidents.csv to evaluate the final performance of the model. For the division of the training set into training and development, the train_test_split function of scikit-learn is employed with a ratio of 80% for training and 20% for development, using a fixed seed (random_state=2024) to ensure reproducibility of the results.

4.2 Evaluation Measures

We evaluated system performance using two primary metrics: F1macro and F1micro. These metrics are suitable for multi-class and multi-label classification tasks, as they consider both accuracy and recall of predictions. In this sense, F1macro calculates the average F1score for each class, treating all classes equally, regardless of frequency. It is beneficial for evaluating performance in scenarios with unbalanced classes, and F1micro calculates the overall F1score considering the total of true positives, false negatives, and false positives. This metric is more sensitive to frequent classes and provides a global view of model performance. Overall, these metrics allow for a comprehensive evaluation of the system, ensuring that the model is broadly accurate and balanced in its performance across all target classes.

5 Results

The official results of our submission for the ST1 and ST2 sub-tasks are shown in Table 2. We report the macro average F1 score and overall ranking of our system, as well as those of the best-performing team for comparison.

Table 2: Results of tasks ST1 and ST2.

System	F1	Rank
ST1		
Task Best System	0.8223	1/27
VerbaNexAI	0.5165	24/27
ST2		
Task Best System	0.5473	1/26
VerbaNexAI	0.3223	16/26

Our system obtained a macro average F1 score of 0.5165 in Sub-task ST1, ranking 24 out of 27 participating teams. In Sub-task ST2, we achieved a macro F1 score of 0.3223, ranking 16 out of 26 teams. These results compare with the reference system, which led the competition with average macro F1 scores of 0.8223 in ST1 and 0.5473 in ST2, respectively.

We conducted several ablation tests and comparisons of different design decisions to optimize system performance. The ablation tests included varying the vectorizer parameters and using different classifiers. We observed that using longer n-grams improved the capture of specific patterns in the incident titles, albeit at the cost of an increase in the dimensionality of the feature space. Also, the random forest classifier proved robust regarding data variability, providing an appropriate balance between accuracy and recall.

Error analysis revealed that our system presented difficulties in classifying hazard categories and products with semantic similarities or specific technical terms not sufficiently represented in the training set. For example, incidents related to food allergies were frequently confused with bacterial contaminations due to the similarity in terminology used. In addition, we observed that product category predictions showed higher variability, possibly attributed to the diversity and specificity of food products mentioned in the data. To address these errors, we propose future incorporation of more advanced natural language processing techniques, such as contextualized embeddings, which could better capture the semantic subtleties of the terms used in the incidents. Although our system did not achieve a top ranking in the competition, the results provide a solid foundation for future improvements. Quantitative and error analysis has identified key areas where significant improvements can be implemented, such as optimizing text representation and exploring more sophisticated classifiers. These strategies and further enrichment of the training data could boost system performance in future iterations of the SemEval 2025 Task 9 challenge.

6 Limitations of the Approach

The proposed approach is subject to several limitations that could be addressed in future versions to enhance its performance. A primary challenge is the management of under-represented categories, which is prevalent in unbalanced classification problems. Despite adjustments made to parameters such as min_df=5 in the TF-IDF vectoriser to mitigate this issue, under-represented classes still exert an influence on the model's accuracy. Furthermore, the presence of semantic ambiguity in texts, where terms may possess multiple meanings depending on the context, poses an additional challenge. While the utilization of character n-grams assists in capturing contextual patterns, the disambiguation of terms in complex texts remains an area for enhancement. Linguistic variability, arising from synonyms, typos and grammatical variations, also exerts a negative influence on model performance, despite pre-processing endeavors.

7 Conclusion

Our system implemented a strategy based on TF-IDF vectorization and Random Forest classifier to address the food hazard detection and product classification tasks in SemEval-2025 Task 9. The performed tests and error analysis underline the importance of optimizing the text representation and exploring more advanced approaches, such as contextualized embeddings, to improve the classification accuracy of poorly represented categories. In future work, we plan to expand the training dataset, integrate architectures based on deep neural networks, and evaluate new natural language processing methods that effectively address the complexity of textual data. These efforts will lay the foundation for more robust and generalizable systems in the food safety domain.

8 Acknowledgments

The authors express their gratitude to the Call 933 "Training in National Doctorates with a Territorial, Ethnic and Gender Focus in the Framework of the Mission Policy — 2023" of the Ministry of Science, Technology and Innovation (Minciencia). In addition, we thank the team of the Artificial Intelligence Laboratory VerbaNex [2] affiliated with the UTB, for their contributions to this project.

References

FAO and WHO. 2007. WORLD HEALTH ORGANIZATION UNITED NATIONS FOOD AND AGRICULTURE ORGANIZATION. From https://iris.who.int/handle/10665/43718.

Q. He, H. Huang, and Y. Wang. 2024. Detection technologies, and machine learning in food: Recent advances and future trends. *Food Bioscience*, 62:105558.

S. S. Nair, V. N. M. Devi, and S. Bhasi. 2024. Enhanced lung cancer detection: Integrating improved random walker segmentation with artificial neural network and random forest classifier. *Heliyon*, 10(7):e29032.

[2] https://github.com/VerbaNexAI

A. Nogales, R. D. Morón, and Á. J. García-Tejedor. 2020. Food safety risk prediction with deep learning models using categorical embeddings on european union data. *Food Control*, 134.

H. Onyeaka, A. Akinsemolu, T. Miri, N. D. Nnaji, C. Emeka, P. Tamasiga, G. Pang, and Z. Al-sharify. 2024. Advancing food security: The role of machine learning in pathogen detection. *Applied Food Research*, 4(2):100532.

Z. Qiu, X. Chen, D. Xie, Y. Ren, Y. Wang, Z. Yang, M. Guo, Y. Song, J. Guo, Y. Feng, N. Kang, and G. Liu. 2025. Identification and detection of frozen-thawed muscle foods based on spectroscopy and machine learning: A review. *Trends in Food Science Technology*, 155:104797.

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.

T. Sabri, O. El Beggar, and M. Kissi. 2022. Comparative study of arabic text classification using feature vectorization methods. In *Procedia Computer Science*, volume 198, pages 269–275.

R. Sathishkumar, T. Karthikeyan, K. P. Praveen, and S. M. Shamsundar. 2023. Ensemble text classification with tf-idf vectorization for hate speech detection in social media. In *2023 International Conference on System, Computation, Automation and Networking (ICSCAN)*.

G. W. Spangenberg, F. Uddin, K. J. Faber, and G. D. G. Langohr. 2024. Automatic bicipital groove identification in arthritic humeri for preoperative planning: A random forest classifier approach. *Computers in Biology and Medicine*, 178:108653.

USDA, U.S. Department of Agriculture. 2024. Food safety: Prepare for the unexpected. https://www.usda.gov/about-usda/news/blog/food-safety-prepare-unexpected.

WHO. n.d. Food safety. Retrieved January 24, 2025, from https://www.who.int/es/news-room/fact-sheets/detail/food-safety.