

BrightCookies 参加 SemEval-2025 任务 9: 探索数据增强在食品危害分类中的应用

Foteini Papadopoulou^{1,2}, Osman Mutlu¹, Neris Özén¹,
Bas H.M. van der Velden¹, Iris Hendrickx², Ali Hürriyetolu¹

¹Wageningen Food Safety Research, The Netherlands

²Centre for Language Studies, Radboud University, The Netherlands

Correspondence: ali.hurriyetoglu@wur.nl

Abstract

本文介绍了我们为 SemEval-2025 第 9 项任务：食物危害检测挑战开发的系统。该共享任务的目标是评估可解释的分类系统，以对食品召回事件报告中两个粒度级别的危害和产品进行分类。在此工作中，我们提出了一种文本增强技术，以改善在少数类上的表现不佳，并比较它们在各种变换和机器学习模型的每个类别中所产生的影响。我们探讨了三种词级数据增强技术，即同义词替换、随机词交换和上下文词插入。结果显示，变换模型往往具有更好的整体性能。三种增强技术中，没有一种能一致地提高危害和产品分类的整体性能。我们在使用 BERT 模型比较基线与每个增强模型时，观察到了细粒度类别的显著改善 ($P < 0.05$)。与基线相比，上下文词插入增强将少数危害类别预测的准确性提高了 6%。这表明对少数类别进行定向增强可以提高变换模型的性能。

1 介绍

食源性疾病每年影响数百万人。世界卫生组织指出，食物污染导致 200 多种疾病，造成严重的健康并发症，并影响社区和国家的社会经济稳定 (World Health Organization, 2024)。在食品安全相关网站上有大量的公开信息。鉴于早期检测食品危害的重要性，需要及时和准确地分析所有这些公开信息以检测食品危害。

SemEval-2025 任务 9: 食品危害检测挑战赛 (Randl et al., 2025) 的提出旨在促进食品安全相关文件中食品危害的自动分类研究。该任务激发了结合食品安全和自然语言处理 (NLP) 的研究，以实现食品召回事件报告的可解释多类别分类。SemEval-2025 任务 9 包括两个子任务：分类粗略的食品危害和产品类别 (ST1) (hazard-category, product-category)，以及细粒度的危害和产品类别 (ST2) (hazard, product)。

SemEval-2025 任务 9 数据集的一个显著挑战是其显著的类别不平衡。在类别之间存在一种长尾分布，尤其是在精细分类中。这种不平

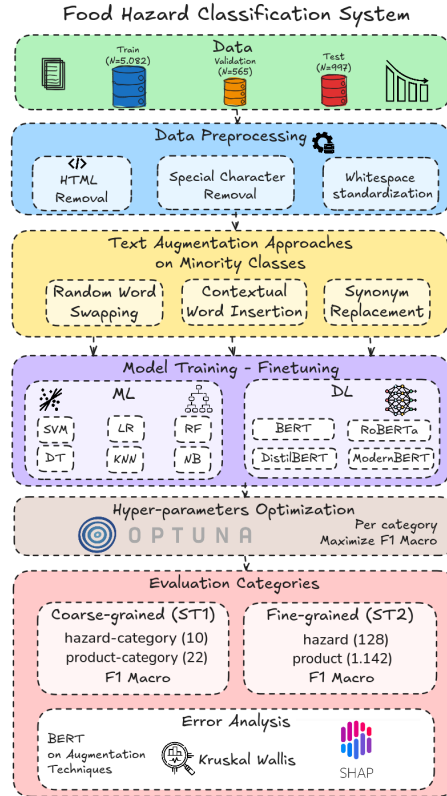


Figure 1: 我们开发的系统架构概述。

衡可能导致分类器表现不佳，特别是对于深度学习 (DL) 模型来说 (Henning et al., 2023)。文本增强技术已被证明在一定程度上能够减轻不平衡数据的影响 (Khan and Venugopal, 2024)。文本增强可以从简单的字符串操作开始，例如在 Easy Data Augmentation (EDA) 中使用的那些 (Wei and Zou, 2019)，到更高级的方法，例如基于 transformer 的文本生成 (Henning et al., 2023)。这有助于提高少数类别的表示，从而可以创造出更平衡的数据集和健壮模型。

我们研究了三种基本的文本增强技术 (同义词替换、上下文单词插入和随机单词交换)，以增强食品召回事件报告多类别分类中代表性不足类别的表现。我们的主要研究问题是：

文本增强技术能否对少数类别提升食品危害多分类器的性能？

我们评估了各种机器学习 (ML) 算法和仅编码器的变压器模型的性能, 包括它们的基线形式和每种增强技术应用后的表现。为了参与该任务, 只允许提交一次。我们在开发集上评估了我们的模型后, 在官方测试集上提交了我们的预测, 并为每个类别选择了表现最佳的模型。在 ST1 中, 我们的系统在 27 个参与者中排名第 15, 距离第一名的宏得分差为 0.0613; 在 ST2 中, 我们在 26 位参与者中排名第 11, 与最高分相差 0.0944 (具体分数见 subsection 2.6)。我们的工作提供了关于文本增强在该领域中有效性的重要见解。¹

2 相关工作

2.1 食品危害分类研究

很少有工作使用文本数据进行细粒度的食品危害分类 (Randl et al., 2024b), 因为大多数现有文献都集中于食品危害的二元分类。最近的一项研究由 Randl et al. (2024b) 介绍了我们在 SemEval-2025 Task 9 中使用的数据集, 并对多种机器学习和深度学习算法进行了基准测试。Randl et al. (2024b) 提出了一个名为符合型上下文学习 (CICLe) 的大型语言模型 (LLM) 参与的框架, 该框架利用符合性预测来优化基分类器预测的上下文长度。通过使用更少、更有针对性的例子, 性能提高了, 同时与常规提示相比, 能耗也减少了。

2.2 少数类别的文本增强

数据增强通过在数据的副本中插入小的变化, 从现有的数据集中创建合成数据 (Shorten et al., 2021)。数据增强缓解了 DL 的类别不平衡问题 (Henning et al., 2023)。根据 Shorten et al. (2021), NLP 中的数据增强方法可以分为两种: 符号和神经。符号技术, 例如基于规则的 EDA (Wei and Zou, 2019), 采用诸如同义词替换和随机插入等简单的词级操作。符号技术在小型数据集中是有效的。神经技术依赖于辅助神经网络, 例如反向翻译或生成增强。一项最近的研究显示, 使用 LLM 进行数据增强, 例如生成新的样本, 可以提高准确性并解决偏斜数据集中的类别不平衡问题 (Gopali et al., 2024)。在我们的研究中, 我们探索符号和简单的神经增强策略, 例如使用 BERT 进行上下文词插入, 以提高分类性能。

此外, 在 2023 年的 SemEval 共享任务中, Al-Azzawi et al. (2023) 探索了数据增强的效果, 特别是对少数类别的回译。他们将其与使用基于变压器模型增强整个数据集进行了比较。他们

观察到, 针对欠代表类别进行增强比广泛的数据集增强更为有效。遵循他们的方法, 我们也将我们的增强策略集中在少数类别而不是整个数据集上。

SemEval-2025 任务 9 中使用的食品召回事件数据集包含 6,644 条用英语撰写的食品召回公告。该数据集分为 5,082 条训练集公告, 565 条开发集公告, 和 997 条测试集公告。数据收集自 24 个不同的网站。样本由描述召回食品产品公告的文本和其他元数据组成。

专家将每个样本手动标记为四个粗略类别的危险 (hazard-category) 和产品 (product-category), 以及细粒度类别 (hazard 和 product)。每个类别的类和类别数量列在 Table 9 中, 示例展示在 Table 10 中。

这四类的类别分布极不平衡, 显示出长尾效应 (Figure 4, Figure 5)。在粗类中, 75% 的类别在 hazard-category 中有 513 个样本, 在 product-category 中有 263 个样本, 而最大的类别包含了 1854 个和 1434 个样本。这种不平衡在细粒度的 hazard 和 product 类别中更加严重, 其中 75% 的类别每 product 最多只有四个样本, 每 hazard 有 24 个样本, 而最大的类别每 product 有 185 个样本, 每 hazard 有 665 个样本。我们使用了 ML 和 DL 并实现了多个数据增强策略。接下来的部分将对此进行更详细的描述。

2.3 机器学习

我们使用术语频率-逆文档频率 (TF-IDF) (Spurck Jones, 1972) 表示文本作为我们机器学习分类器的输入。我们训练了不同的分类器, 并评估它们在每个类别的两个子任务上的性能。使用的分类器包括线性支持向量机 (SVM)、决策树 (DT)、随机森林 (RF)、逻辑回归 (LR)、多项式朴素贝叶斯 (NB) 和 K-最近邻 (KNN)。我们使用了来自 Scikit-learn 库²的实现。

2.4 深度学习

我们使用了基于深度学习的 transformer 语言模型进行序列分类 (Vaswani et al., 2023)。我们选择了仅编码器模型, 它直接生成输入序列的表示, 并将其输入到分类头以进行预测。我们训练了各种 transformers 来处理序列分类任务, 包括 BERT (Devlin et al., 2019)、RoBERTa (Liu et al., 2019)、DistilBERT (Sanh et al., 2020) 和 ModernBERT (Warner et al., 2024) (更多信息可见 ??)。我们利用了 Hugging Face 的 Transformers 库³ (Wolf et al., 2020)。

¹我们的代码可以在 <https://github.com/WFSRDataScience/SemEval2025Task9> 找到

²<https://scikit-learn.org/stable/>

³<https://huggingface.co/docs/transformers>

Operation	Sentence
Original	Certain Stella Artois brand Beer may be unsafe due to possible presence of glass particles
CW	certain notable stella by artois brand beer may be judged unsafe primarily due to his possible presence of glass particles
SR	Certain Frank stella Artois brand Beer may be insecure imputable to potential presence of glass particles
RW	Certain Stella Artois brand Beer may due be unsafe to presence possible of glass particles

Table 1: 应用文本增强技术的示例，通过上下文单词插入 (CW)、同义词替换 (SR) 和随机单词交换 (RW) 对 title 的食品召回进行处理。

除了对上述模型进行基线训练外，我们还探讨了数据增强如何影响每个类别中少数类的表现。

我们采用了三种不同的增强策略，使用了 NLP AUG 库⁴ (Ma, 2019)：随机单词互换 (RW)、同义词替换 (SR) 和上下文单词插入 (CW)。RW 随机互换相邻单词。SR 从英语词汇数据库 (WordNet (Miller, 1995)) 中替换相似的单词。CW 使用来自 BERT 的上下文词嵌入来寻找最相似的单词并插入它们来进行增强。每种技术应用于 title 的示例在 Table 1 中展示。

对于每种策略，我们通过修改标题和文本以保留其内在意义，同时保持标注的类别，为每个类别的少数类在训练数据中生成新的样本。对于粗略的类别 (hazard-category 和 product-category)，我们通过为每个类生成 200 个样本来增强少于 200 个样本的类。对于细粒度的类别 (hazard 和 product)，我们为 hazard 类别中少于 100 个样本的类创建了 100 个样本，为 product 类别创建了 50 个样本。在检查了整个类分布之后，我们选择这些新增样本的数量和低支持类的阈值，因为它们在改善少数类的表示和保持较低计算成本之间提供了一种折中，但不能完全解决不平衡问题。我们首先遍历每个类别的每个代表性不足的类的现有数据样本。然后我们根据使用的增强技术在这些样本之间按比例分配指定的总增强样本数量（调整最后一个以确保新增样本数量与设定目标匹配）来生成新样本。?? 中提供了伪代码描述，?? 中提供了其对类统计的影响。所有方法均在 Python 中实现。

接下来的小节中，我们将进一步描述预处理、超参数微调 and 评估的细节。

2.5 预处理

预处理包括使用来自 title 和 text 的正则表达式去除 HTML 标记和特殊字符（换行符、制

⁴<https://nlpaug.readthedocs.io/>

表符、Unicode 字符符号)，以及文本标准化，如空白标准化。这在消除和过滤不必要格式的同时保留语义内容。

我们使用 Optuna 超参数优化框架中的树结构 Parzen 估计器 (TPE) 采样器，在开发集上微调基线和增强模型的超参数。TPE 是一种基于贝叶斯的优化方法，它使用树结构将超参数与我们的目标函数（最大化每个类别的 F_1 -宏得分）联系起来，以发现最佳超参数。

我们为每个模型和每种增强技术运行了 10 次试验。对于 ST1 的机器学习，由于计算时间较短，我们运行了 50 次试验。我们优化了 TF-IDF 向量化器的参数，比如最低文档频率 (min_df)，以及适用于每个分类器的机器学习超参数，例如 SVM 中的最大迭代次数 (max_iter)，以及适用于深度学习的学习率调度器、批量大小和训练轮数 (subsection 3)。所有涉及 Transformer 模型的实验都在不同的 GPU 集群上进行 (subsection 2)⁵。

我们将我们的结果提交给两个子任务的排行榜，其中最终得分通过将危害 F_1 -宏（计算在所有样本上）与产品 F_1 -宏（仅计算在预测正确的危害样本上）的平均值结合起来，计算粗略 (ST1) 和细粒度类别 (ST2)。例如，如果所有危害都被正确预测，但所有产品都被错误预测，则整体结果将是 0.5 F_1 -宏分数 (??)。

接下来的小节展示了官方测试集中每个模型使用 text 字段（在训练和开发集上训练）进行的定量结果，以及对 BERT 基线模型与其增强训练版本的错误分析。

2.6 定量结果

变压器模型在所有分类中表现优于 ML，如 Table 2 所示，除了 product-category 以外，在所有类别中的基线版本中， $ModernBERT_{base}$ 在变压器模型中领先。

在机器学习中，SVM、LR 和 RF 显示出竞争力的表现： LR_{CW} 在 hazard-category (0.713) 和 product-category (0.682) 中得分最高； RF_{RW} 在 hazard (0.567) 中得分最高， SVM_{CW} 在 product (0.256) 中得分最高。在转换器模型中， $ModernBERT_{SR}$ 在 hazard-category 中得分最高，得分为 0.790，而 $BERT_{CW}$ 在其他类别中得分最高。数据增强提高了性能，但在各个类别之间并不一致。在 ST2 类别中比 ST1 类别更为明显，在 product 类别中， $BERT_{base}$ 和 $BERT_{CW}$ 数据增强之间的最大得分增幅为 0.11。

⁵我们最佳的微调模型可在 <https://huggingface.co/collections/DataScienceWFSR/semeval2025task9-food-hazard-detection-680f43d99cc294f617104be2> 获取。

Model	hazard-category	product-category	hazard	product	ST1	ST2
<i>SVM_{base}</i>	0.701	0.626	0.544	0.234	0.682	0.396
<i>SVM_{CW}</i>	0.655	0.642	0.519	0.256	0.649	0.396
<i>SVM_{SR}</i>	0.707	0.674	0.511	0.234	0.693	0.379
<i>SVM_{RW}</i>	0.687	0.643	0.542	0.246	0.682	0.401
<i>LR_{base}</i>	0.666	0.665	0.511	0.203	0.680	0.368
<i>LR_{CW}</i>	0.713	0.682	0.457	0.209	0.702	0.347
<i>LR_{SR}</i>	0.698	0.677	0.454	0.233	0.691	0.354
<i>LR_{RW}</i>	0.666	0.676	0.522	0.216	0.673	0.380
<i>DT_{base}</i>	0.542	0.445	0.405	0.012	0.484	0.208
<i>DT_{CW}</i>	0.617	0.491	0.427	0.029	0.544	0.230
<i>DT_{SR}</i>	0.576	0.488	0.464	0.037	0.526	0.252
<i>DT_{RW}</i>	0.612	0.475	0.506	0.056	0.542	0.283
<i>RF_{base}</i>	0.691	0.523	0.499	0.129	0.609	0.318
<i>RF_{CW}</i>	0.708	0.597	0.566	0.169	0.642	0.380
<i>RF_{SR}</i>	0.688	0.578	0.455	0.188	0.633	0.331
<i>RF_{RW}</i>	0.698	0.546	0.567	0.202	0.612	0.397
<i>KNN_{base}</i>	0.552	0.497	0.384	0.157	0.527	0.294
<i>KNN_{CW}</i>	0.565	0.490	0.376	0.169	0.534	0.309
<i>KNN_{SR}</i>	0.552	0.507	0.389	0.163	0.537	0.305
<i>KNN_{RW}</i>	0.500	0.491	0.397	0.152	0.515	0.299
<i>NB_{base}</i>	0.553	0.570	0.306	0.064	0.568	0.203
<i>NB_{CW}</i>	0.599	0.586	0.405	0.175	0.603	0.310
<i>NB_{SR}</i>	0.588	0.574	0.444	0.140	0.589	0.314
<i>NB_{RW}</i>	0.603	0.617	0.383	0.167	0.631	0.300
<i>BERT_{base}</i>	0.747	0.757	0.581	0.170	0.753	0.382
<i>BERT_{CW}</i>	0.760	0.761	0.671	0.280	0.762	0.491
<i>BERT_{SR}</i>	0.770	0.754	0.666	0.275	0.764	0.478
<i>BERT_{RW}</i>	0.752	0.757	0.651	0.275	0.756	0.467
<i>DistilBERT_{base}</i>	0.761	0.757	0.593	0.154	0.760	0.378
<i>DistilBERT_{CW}</i>	0.766	0.753	0.635	0.246	0.763	0.449
<i>DistilBERT_{SR}</i>	0.756	0.759	0.644	0.240	0.763	0.448
<i>DistilBERT_{RW}</i>	0.749	0.747	0.647	0.261	0.753	0.462
<i>RoBERTa_{base}</i>	0.760	0.753	0.579	0.123	0.755	0.356
<i>RoBERTa_{CW}</i>	0.773	0.739	0.630	0.000	0.760	0.315
<i>RoBERTa_{SR}</i>	0.777	0.755	0.637	0.000	0.767	0.319
<i>RoBERTa_{RW}</i>	0.757	0.611	0.615	0.000	0.686	0.308
<i>ModernBERT_{base}</i>	0.781	0.745	0.667	0.275	0.769	0.485
<i>ModernBERT_{CW}</i>	0.761	0.712	0.609	0.252	0.741	0.441
<i>ModernBERT_{SR}</i>	0.790	0.728	0.591	0.253	0.761	0.434
<i>ModernBERT_{RW}</i>	0.761	0.751	0.629	0.237	0.759	0.440

Table 2: F_1 -macro 分数表示每个模型在官方测试集中的表现，该测试集由组织者提供，并根据每个类别的 text 字段和子任务得分 (ST1 和 ST2) 计算，结果精确到小数点后三位。我们用加粗字体表示每个类别和子任务的最高得分。

为了理解增强的影响，我们进行了单独的成对克鲁斯卡尔-沃利斯检验，比较了 F_1 -宏分数在 $BERT_{base}$ 模型与增强版本之间的差异，每个版本在每个类别中进行三次训练 (Table 8)。在 product-category 与 RW 的比较中发现了统计显著性 ($P < 0.05$)，在 hazard 中与所有增强技术的比较中，以及在 product 与 CW 和 RW 的比较中 (Table 3)。这表明，BERT 的增强技术在细粒度类别中比在粗粒度类别中更有效地提高了少数类的性能。

我们为每个类别提交了组合的 BERT 和 RoBERTa 模型到排行榜 (subsection A.3)，这导致测试集中 ST1 的 F_1 -宏观分数为 0.761，ST2 为 0.453。这些模型之所以被选择，是因为它们在开发集上显示了最佳的 F_1 -宏观分数。其他模型也在测试集上进行了评估，但未包括在排行榜中。在 ST1 上达到的最好分数是 0.769，而 ST2 是 0.491，这在 Table 2 中以粗体显示。

此外，还进行了仅使用 title 的实验 (其结果可以在 Table 7 中找到)。我们继续使用 text 领域对模型进行误差分析，因为我们观察到了更好的性能。

Category	CW	RW	SR
hazard-category	0.5127	0.2752	0.2752
product-category	0.2752	0.3758	0.0463
hazard	0.0495	0.0495	0.0463
product	0.0463	0.0495	0.5127

Table 3: $BERT_{base}$ 模型与三种增强技术之间的逐项配对 Kruskal-Wallis 测试所得的原始 P 值 (四舍五入至小数点后四位)。

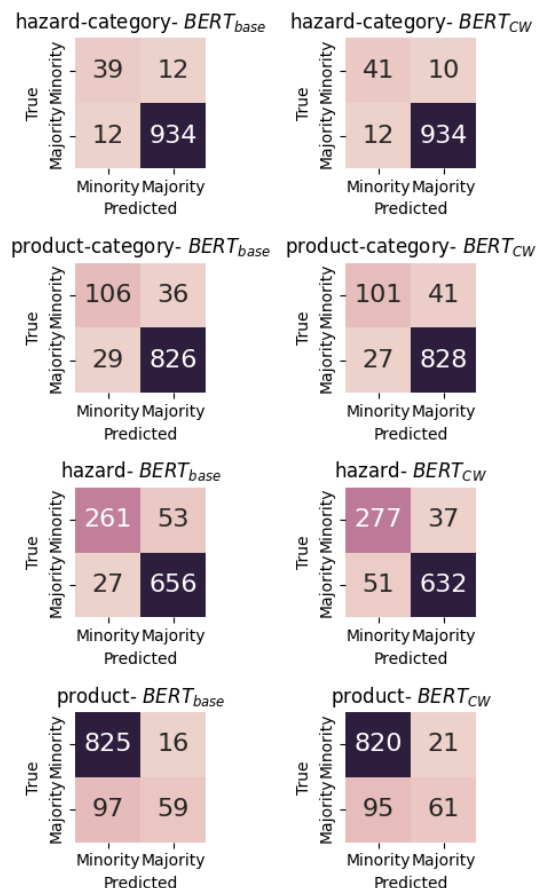


Figure 2: 混淆矩阵比较了 $BERT_{base}$ 和 $BERT_{CW}$ 模型在测试集中的表现，涵盖四个类别，展示了模型在少数类别和多数类别预测上的性能变化。

2.7 错误分析 - 混淆矩阵

我们研究了 BERT 模型的性能和缺陷，该模型在使用 CW 技术后相比基线有了最大的改进。比较增广的多数类和少数类时， $BERT_{CW}$ 模型对少数类的预测略好于 $BERT_{base}$ ，hazard-category 从 39 上升到 41，hazard 从 261 上升到 277 (大约增加 6%) (Figure 2)。然而，模型对多数类的预测略差，从 hazard 的 656 减少到 632，这表明在改善少数类与多数类预测之间存在权衡。另外，尽管增强略微改善了 product-category 的多数类预测，但少数类的预测从 106 个样本减少到 101 个。

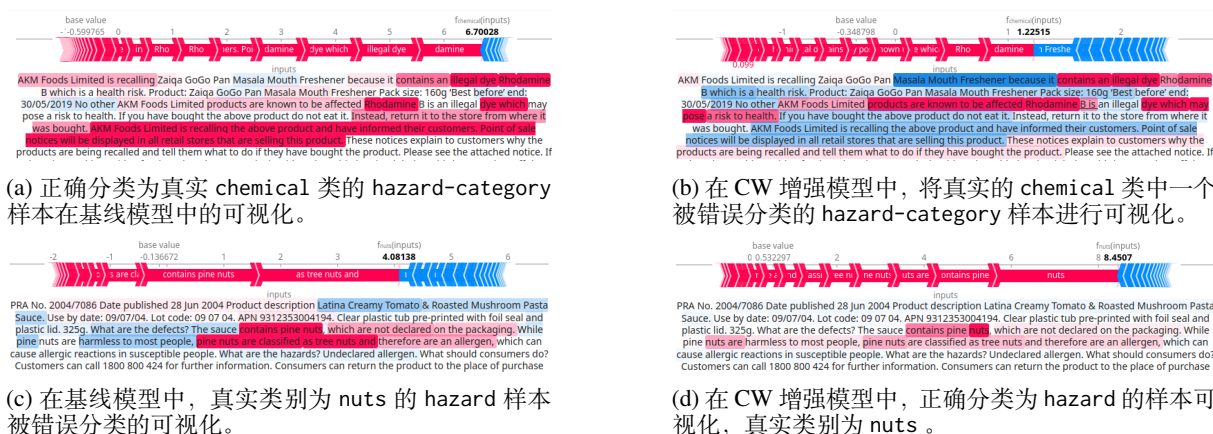


Figure 3: 基于 $BERT_{base}$ 和 $BERT_{CW}$ 模型的预测，针对 hazard-category 和 hazard 这些真实类别样本的 SHAP 值可视化。粉红色文本表示为预测该类别作出正向贡献，而蓝色文本表示负向贡献。在每个样本中不作出贡献的文本已被截断。

2.8 误差分析 - SHAP

我们使用 SHapley Additive exPlanations (SHAP)⁶ 进一步分析 BERT 的预测行为。图 3a 和 3b 展示了一个样本的 SHAP 值，该样本在 hazard-category 中被正确分类为 $BERT_{base}$ ，但错误分类为 $BERT_{CW}$ ，通过可视化展示了导致正确 chemical 类的贡献。图 3c 和 3d 展示了一个样本的 SHAP 值，该样本错误分类为 $BERT_{base}$ ，但在 hazard 中被正确分类为 $BERT_{CW}$ ，通过可视化展示了导致正确 nuts 类的贡献。在 hazard-category 中， $BERT_{base}$ 准确识别了诸如“非法染料”（粉色）的特征，而 CW 增强则有更多负面（蓝色）贡献，使模型的预测偏离正确类别。对于 hazard 类别，尽管两个模型都关注“松仁”等重要术语，基线模型却关注负面的贡献，如“拉丁奶油番茄”，导致误分类，这可能意味着模型错误地将这些特征与不同的危害关联。此误分类模式可以作为未来研究的基础，进一步探索和解释模型的预测，以提高其性能和可靠性。

尽管已进行了多项实验，但在未来的研究中可以解决一些局限性。所使用的数据集仅限于英语，并且所应用的增强技术仅限于词级调整。未来的研究可以探索更复杂的增强方法，例如使用大型语言模型 (LLM) 来生成新样本并验证其质量。引入其他语言的数据集可以提供有关增强技术有效性的见解。进一步的研究还可以集中于优化少数类别的增强样本数量，以提高分类性能，特别是在食品危害分类中，需要可靠的模型以确保安全。最后，为了进一步提升分类器的性能，可以使用更复杂的架构，如集成或层级方法，以比较它们在食品危害分类任务中的增强效果。

⁶<https://shap.readthedocs.io/en/latest/>

3 结论

我们展示了单词级文本增强可以提升少数类的多类分类性能。我们在 SemEval-2025 任务 9 上使用了多种机器学习和变压器模型来评估这些增强的效果。利用 text 领域，我们发现变压器往往优于机器学习。增强技术显示出 F_1 -macro 分数的轻微提升，但这种效果在所有增强方法中并不一致。将 $BERT_{base}$ 与每种增强技术进行比较，对于细粒度类别，发现了统计上显著的改进，这表明对少数类别的增强可以改善变压器在这些类别上的性能。

4

致谢 我们感谢匿名审稿人的反馈和组织者的支持。该研究的资金由欧盟 Horizon Europe 研究和创新计划 EFRA 提供 [赠款编号 101093026]。

References

- Sana Al-Azzawi, György Kovács, Filip Nilsson, Tosin Adewumi, and Marcus Liwicki. 2023. [NLP-LTU at SemEval-2023 task 10: The impact of data augmentation and semi-supervised learning techniques on text classification performance on an imbalanced dataset](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1421–1427, Toronto, Canada. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

- Saroj Gopali, Faranak Abri, Akbar Siami Namin, and Keith S. Jones. 2024. [The applicability of llms in generating textual samples for analysis of imbalanced datasets](#). *IEEE Access*, 12:136451–136465.
- Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. [A survey of methods for addressing class imbalance in deep-learning based natural language processing](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, Dubrovnik, Croatia. Association for Computational Linguistics.
- Reeba Khan and Anoushka Venugopal. 2024. [Exploring deep learning methods for text augmentation to handle imbalanced datasets in natural language processing](#). In *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, pages 1–8.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.
- Edward Ma. 2019. Nlp augmentation. <https://github.com/makcedward/nlpaug>.
- George A. Miller. 1995. [Wordnet: a lexical database for english](#). *Commun. ACM*, 38(11):39–41.
- Korbinian Randl, Manos Karvounis, George Marinos, John Pavlopoulos, Tony Lindgren, and Aron Henriksson. 2024a. [Food recall incidents](#).
- Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2024b. [CICLe: Conformal in-context learning for largescale multi-class food risk classification](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7695–7715, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. [Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter](#). *Preprint*, arXiv:1910.01108.
- Connor Shorten, Taghi M. Khoshgoftaar, and Borko Furht. 2021. [Text data augmentation for deep learning](#). *Journal of Big Data*, 8(1):101.
- Karen Sparck Jones. 1972. [A statistical interpretation of term specificity and its application in retrieval](#). *Journal of Documentation*, 28(1):11–21.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. [Attention is all you need](#). *Preprint*, arXiv:1706.03762.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). *Preprint*, arXiv:2412.13663.
- Jason Wei and Kai Zou. 2019. [EDA: Easy data augmentation techniques for boosting performance on text classification tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- World Health Organization. 2024. Food safety. <https://www.who.int/news-room/fact-sheets/detail/food-safety>. Accessed: February 03, 2025.
- 在本节中，展示了与所提供数据集统计相关的表格和图表。Table 10 展示了数据集中一些示例标题和文本及其标注类别。Table 9 展示了标注类别的数量和名称。Figure 4 和 Figure 5 展示了粗粒度和细粒度类别中危险和产品类别的分布，表明了其遵循的长尾分布，而 Figure 6 展示了每个国家和年份在数据集中出现的分布。Table 4 显示了样本来源的站点域名以及样本数量。
- ### 1 预处理数据集详情
- 使用了 BeautifulSoup⁷ 包来从数据中移除 HTML 内容。用于移除特殊字符的正则表达式如下：
- ```
'[\t\n\r\u200b]|//| '
```
- <sup>7</sup><https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

| Domain                             | Samples |
|------------------------------------|---------|
| www.fda.gov                        | 1740    |
| www.fsis.usda.gov                  | 1112    |
| www.productsafety.gov.au           | 925     |
| www.food.gov.uk                    | 902     |
| www.lebensmittelwarnung.de         | 886     |
| www.inspection.gc.ca               | 864     |
| www.fsai.ie                        | 358     |
| www.foodstandards.gov.au           | 281     |
| inspection.canada.ca               | 124     |
| www.cfs.gov.hk                     | 123     |
| recalls-rappels.canada.ca          | 96      |
| tna.europarchive.org               | 52      |
| wayback.archive-it.org             | 23      |
| healthycanadians.gc.ca             | 18      |
| www.sfa.gov.sg                     | 11      |
| www.collectionscanada.gc.ca        | 8       |
| securite-alimentaire.public.lu     | 6       |
| portal.efet.gr                     | 4       |
| www.foodstandards.gov.scot         | 3       |
| www.ages.at                        | 2       |
| www.accessdata.fda.gov             | 1       |
| webarchive.nationalarchives.gov.uk | 1       |
| www.salute.gov.it                  | 1       |
| www.foedevarestyrelsen.dk          | 1       |

Table 4: 按给定数据集的支持数量排序的公共食品安全机构网站的数据来源。此表格改编自 [Randl et al. \(2024a\)](#)。它还包含非英文数据的来源。

## 2 系统配置详情

实验在不同的机器上运行，使用的是 Python 3.10.16 版本。对于 transformer 模型的微调和训练，使用了 NVIDIA A100 80GB 和 NVIDIA GeForce RTX 3070 Ti。为了确保可重复性，我们在 PyTorch、NumPy 和 Random 包中使用了  $seed = 2025$  作为种子数。为了再运行两次 BERT 模型并计算统计显著性，我们使用了  $seed = 2024$  和  $2026$ 。此外，所使用软件包的版本及其对应的 URL 可以在 Table 5 中找到。

在本节中，我们解释了在实验中使用的仅编码器 transformer 模型的细节和架构。对于 BERT，使用的是包含 110M 参数、12 个编码器层、隐藏状态大小为 768、前馈隐藏状态为 3072 以及 12 个注意力头的模型，作为基础的预训练 transformer 模型。对于 RoBERTa，使用的是大小写敏感的版本，具有 125M 参数，结构为 12 个编码器层、768 维隐藏状态、3072 维前馈网络和 12 个注意力头，并在一个大语料库上训练，利用动态掩码。对于 DistilBERT，使用的是较轻的 BERT 变体，具有 66M 参数和 6 个编码器层，同时保持与 BERT 相似的隐藏状态大小和注意力头。对于 ModernBERT，我们使用了大小写敏感的版本，包含 149M 参数、22 个编码器层、768 的隐藏状态、1152 的中间大小和 12 个注意力头。它在 2 万亿个标记上训练，将标记长度扩展到 8192，并采用其他架构增强以使其比其他 BERT 变体更快、更轻，

| Library        | Version | URL                                                                                                                           |
|----------------|---------|-------------------------------------------------------------------------------------------------------------------------------|
| Transformers   | 4.49.0  | <a href="https://huggingface.co/docs/transformers/index">https://huggingface.co/docs/transformers/index</a>                   |
| PyTorch        | 2.6.0   | <a href="https://pytorch.org/">https://pytorch.org/</a>                                                                       |
| SpaCy          | 3.8.4   | <a href="https://spacy.io/">https://spacy.io/</a>                                                                             |
| Scikit-learn   | 1.6.0   | <a href="https://scikit-learn.org/stable/">https://scikit-learn.org/stable/</a>                                               |
| Pandas         | 2.2.3   | <a href="https://pandas.pydata.org/">https://pandas.pydata.org/</a>                                                           |
| Optuna         | 4.2.1   | <a href="https://optuna.org/">https://optuna.org/</a>                                                                         |
| NumPy          | 2.0.2   | <a href="https://numpy.org/">https://numpy.org/</a>                                                                           |
| NLP AUG        | 1.1.11  | <a href="https://nlpaug.readthedocs.io/en/latest/index.html">https://nlpaug.readthedocs.io/en/latest/index.html</a>           |
| BeautifulSoup4 | 4.12.3  | <a href="https://www.crummy.com/software/BeautifulSoup/bs4/doc/#">https://www.crummy.com/software/BeautifulSoup/bs4/doc/#</a> |

Table 5: 用于本文代码实现的 Python 库及其版本与对应的 URL。

并具有更好的性能。

在算法 ?? 中，展示了使用增强技术创建新样本的函数。从函数的输入开始，它接受以下参数：阈值  $\tau$ ，它是一个类别可以被视为少数类的样本数量；每个少数类要添加的样本数量  $S$ ；包含每个类样本数量的类计数  $C$ ；接受样本和要创建的样本数量的增强函数  $F$ ；原始训练数据集  $D$ ；以及我们希望增强其类的数据集 *category*（例如 hazard）。函数首先通过获取样本数小于给定阈值的类别找到少数类。然后，对每个少数类，收集相应样本，并通过将总样本数除以该特定类的样本数量来计算每个样本需要增强的样本数量，将结果向下取整至最近的整数。对于每个样本，应用增强函数并创建新样本，除了最后一个样本，它是针对剩余所需样本数量进行增强的。新样本被插入到原始训练数据集中，函数返回增强后的数据集。

在 ?? 中，展示了类别统计在每个类别应用增强前后的比较。对于 hazard-category 和 product-category，对于样本数少于 200 的类别，创建的样本数为 200。对于 hazard 和 product，对于样本数少于 100 的类别，分别添加了 100 和 50 个样本。

对于两个子任务，组织者给出的评估标准是预测类和标注类的  $F_1$ -macro 分数。排名基于危险类别，这意味着如果对危险和产品的预测都正确，将获得 1.0 分，而如果危险预测正确但产品预测错误，将得 0.5 分。准确的评分函数可以在算法 ?? 中看到。

### 3 超参数详情

为了调整超参数，采用了 Optuna 优化框架，基于  $F_1$ -macro 分数进行优化。对于机器学习模型，优化了 TF-IDF 向量化器参数，如  $min_{df}$ 、 $max_{df}$  等，以及每个模型的特定参数，如 SVM 和 LR 的  $max_{iter}$ ，以及 NB 的  $alpha$ 。在 ?????????? 中展示了每个模型、类别和领域所使用的超参数。当使用 SpaCy 分词器<sup>8</sup>时，也会从给定文本中移除 SpaCy 的英语停用词。在 SVM、LR、RF 和 DT 模型中使用了平衡的类权重。

对于变压器模型， $batch\_size$ 、 $epochs$  和  $lr\_scheduler$  在所有模型变体中进行了优化，共进行了 10 次试验。对于所有模型，学习率设定为  $5.0e-5$ ，标记器可以生成的最大标记长度设定为 128，因为观察到更高的最大标记长度并没有显著的性能差异。在 ?????????? 中，列出了每个模型、类别和字段所用的超参数。

在调优过程中使用的每个超参数的搜索空间可以在 Table 6 中找到。

| Hyperparameter    | Search Space                                                               |
|-------------------|----------------------------------------------------------------------------|
| $C$               | { 0.1, 1, 5, 10 }                                                          |
| $max\_iter$       | { 100, 1000, 5000 }                                                        |
| $n\_estimators$   | { 100, 200, 300 }                                                          |
| $max\_depth$ (DT) | { 100, 200, 300 }                                                          |
| $max\_depth$ (RF) | { 100, 1000, 5000 }                                                        |
| $max\_features$   | { 1000, 5000, 10000, 50000 }                                               |
| $n\_neighbors$    | { 3, 5, 7, 9, 11 }                                                         |
| $weights$         | { uniform, distance }                                                      |
| $alpha$           | { 0.01, 0.1, 1, 5 }                                                        |
| $analyzer$        | { word, char }                                                             |
| $tokenizer$       | { -, SpaCy }                                                               |
| $min\_df$         | { 1, 2, 5 }                                                                |
| $max\_df$         | { 0.1, 0.3, 0.5 }                                                          |
| $ngram\_range$    | { (1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (2, 3), (2, 4), (2, 5), (3, 5) } |
| $batch\_size$     | { 8, 16, 32 }                                                              |
| $epochs$          | { 3, 5, 10 }                                                               |
| $lr\_scheduler$   | { lin, cos, cosRestarts }                                                  |

Table 6: 用于 Optuna 优化试验的每个超参数的搜索空间，适用于机器学习和变压器模型。对于学习率调度器：cos（余弦退火）、cosRestarts（带重新启动的余弦退火）和 lin（线性）。

## A 更多结果和可解释性分析

### A.1 使用标题的结果

在 Table 7 中，我们展示了使用 title 字段对 ML 和 transformer 模型在测试集上的实验结果。与使用 text 的结果一样，transformer 模型整体上表现优于 ML 模型，但仍低于使用 text 时的结果。每个类别的最佳模型是：BERT<sub>RW</sub> 用于

<sup>8</sup><https://spacy.io/api/tokenizer>

| Model                            | hazard-category | product-category | hazard | product | ST1   | ST2   |
|----------------------------------|-----------------|------------------|--------|---------|-------|-------|
| <i>SVM<sub>base</sub></i>        | 0.644           | 0.692            | 0.436  | 0.250   | 0.670 | 0.363 |
| <i>SVM<sub>CW</sub></i>          | 0.641           | 0.675            | 0.402  | 0.240   | 0.657 | 0.343 |
| <i>SVM<sub>SR</sub></i>          | 0.646           | 0.699            | 0.435  | 0.259   | 0.674 | 0.364 |
| <i>SVM<sub>RW</sub></i>          | 0.646           | 0.690            | 0.432  | 0.253   | 0.670 | 0.372 |
| <i>LR<sub>base</sub></i>         | 0.596           | 0.695            | 0.419  | 0.261   | 0.636 | 0.359 |
| <i>LR<sub>CW</sub></i>           | 0.627           | 0.670            | 0.428  | 0.263   | 0.649 | 0.361 |
| <i>LR<sub>SR</sub></i>           | 0.612           | 0.660            | 0.425  | 0.234   | 0.639 | 0.350 |
| <i>LR<sub>RW</sub></i>           | 0.634           | 0.647            | 0.442  | 0.269   | 0.644 | 0.374 |
| <i>DT<sub>base</sub></i>         | 0.491           | 0.478            | 0.330  | 0.036   | 0.483 | 0.183 |
| <i>DT<sub>CW</sub></i>           | 0.534           | 0.541            | 0.277  | 0.031   | 0.553 | 0.164 |
| <i>DT<sub>SR</sub></i>           | 0.565           | 0.449            | 0.349  | 0.081   | 0.495 | 0.226 |
| <i>DT<sub>RW</sub></i>           | 0.513           | 0.453            | 0.298  | 0.057   | 0.493 | 0.185 |
| <i>RF<sub>base</sub></i>         | 0.611           | 0.633            | 0.420  | 0.287   | 0.616 | 0.369 |
| <i>RF<sub>CW</sub></i>           | 0.592           | 0.640            | 0.446  | 0.232   | 0.615 | 0.367 |
| <i>RF<sub>SR</sub></i>           | 0.638           | 0.527            | 0.422  | 0.207   | 0.590 | 0.329 |
| <i>RF<sub>RW</sub></i>           | 0.629           | 0.635            | 0.372  | 0.244   | 0.638 | 0.328 |
| <i>KNN<sub>base</sub></i>        | 0.519           | 0.598            | 0.349  | 0.187   | 0.566 | 0.299 |
| <i>KNN<sub>CW</sub></i>          | 0.554           | 0.508            | 0.341  | 0.167   | 0.545 | 0.275 |
| <i>KNN<sub>SR</sub></i>          | 0.541           | 0.569            | 0.306  | 0.152   | 0.566 | 0.255 |
| <i>KNN<sub>RW</sub></i>          | 0.536           | 0.551            | 0.335  | 0.174   | 0.558 | 0.278 |
| <i>NB<sub>base</sub></i>         | 0.597           | 0.641            | 0.366  | 0.221   | 0.624 | 0.318 |
| <i>NB<sub>CW</sub></i>           | 0.588           | 0.611            | 0.360  | 0.185   | 0.609 | 0.305 |
| <i>NB<sub>SR</sub></i>           | 0.597           | 0.593            | 0.349  | 0.180   | 0.600 | 0.290 |
| <i>NB<sub>RW</sub></i>           | 0.585           | 0.629            | 0.390  | 0.195   | 0.608 | 0.315 |
| <i>BERT<sub>base</sub></i>       | 0.668           | 0.636            | 0.372  | 0.177   | 0.653 | 0.284 |
| <i>BERT<sub>CW</sub></i>         | 0.654           | 0.714            | 0.502  | 0.249   | 0.693 | 0.392 |
| <i>BERT<sub>SR</sub></i>         | 0.650           | 0.707            | 0.489  | 0.259   | 0.681 | 0.389 |
| <i>BERT<sub>RW</sub></i>         | 0.670           | 0.735            | 0.477  | 0.250   | 0.700 | 0.372 |
| <i>DistilBERT<sub>base</sub></i> | 0.653           | 0.579            | 0.396  | 0.248   | 0.613 | 0.334 |
| <i>DistilBERT<sub>CW</sub></i>   | 0.631           | 0.725            | 0.486  | 0.264   | 0.687 | 0.395 |
| <i>DistilBERT<sub>SR</sub></i>   | 0.640           | 0.695            | 0.503  | 0.262   | 0.667 | 0.400 |
| <i>DistilBERT<sub>RW</sub></i>   | 0.644           | 0.701            | 0.496  | 0.267   | 0.672 | 0.392 |
| <i>RoBERTa<sub>base</sub></i>    | 0.608           | 0.629            | 0.384  | 0.076   | 0.619 | 0.246 |
| <i>RoBERTa<sub>CW</sub></i>      | 0.668           | 0.692            | 0.460  | 0.000   | 0.686 | 0.230 |
| <i>RoBERTa<sub>SR</sub></i>      | 0.639           | 0.718            | 0.471  | 0.000   | 0.673 | 0.236 |
| <i>RoBERTa<sub>RW</sub></i>      | 0.636           | 0.736            | 0.479  | 0.001   | 0.690 | 0.240 |
| <i>ModernBERT<sub>base</sub></i> | 0.586           | 0.671            | 0.393  | 0.275   | 0.627 | 0.353 |
| <i>ModernBERT<sub>CW</sub></i>   | 0.649           | 0.731            | 0.423  | 0.266   | 0.688 | 0.372 |
| <i>ModernBERT<sub>SR</sub></i>   | 0.616           | 0.679            | 0.422  | 0.254   | 0.646 | 0.364 |
| <i>ModernBERT<sub>RW</sub></i>   | 0.641           | 0.697            | 0.385  | 0.263   | 0.668 | 0.351 |

Table 7: 在官方测试集中，组织者利用每个类别的 title 字段和子任务分数 (ST1 和 ST2)，将  $F_1$ -宏成绩四舍五入到小数点后三位。我们用粗体表示每列中的最高分。

hazard-category (0.670), RoBERTa<sub>RW</sub> 用于 product-category (0.736), DistilBERT<sub>SR</sub> 用于 hazard (0.503), 以及 RF<sub>base</sub> 用于 product (0.287)。在 ML 模型中，SVM、LR 和 RF 在各个类别中表现出竞争力，与仅使用 text 字段时的表现相似。虽然基线模型和增强模型之间存在变化，但使用 transformer 模型时，在 product-category 和 hazard 中观察到轻微且一致的增加。

### A.2 统计显著性实验

BERT 模型实验（包括基线和增强版本，每个实验进行三次）的平均  $F_1$ -宏观分数如 Table 8 所示。

### A.3 正式提交的模型

由于评估阶段只允许提交一次，因此提交的模型预测是在开发集上每个类别  $F_1$ -宏观分数最好的模型：RoBERTa<sub>base</sub> 对于 hazard-category，得分为 0.880  $F_1$ -宏观分数；RoBERTa<sub>RW</sub> 对于 product-category，得分为 0.750  $F_1$ -宏观分数；BERT<sub>CW</sub> 对于 hazard，得分为 0.682  $F_1$ -宏观分数；BERT<sub>RW</sub> 对于

| Model         | hazard-<br>category | product-<br>category | hazard | product |
|---------------|---------------------|----------------------|--------|---------|
| $BERT_{base}$ | 0.757               | 0.769                | 0.594  | 0.186   |
| $BERT_{CW}$   | 0.768               | 0.756                | 0.658  | 0.284   |
| $BERT_{RW}$   | 0.751               | 0.752                | 0.662  | 0.256   |
| $BERT_{SR}$   | 0.771               | 0.75                 | 0.652  | 0.189   |

Table 8: 平均每个类别的  $F_1$  -宏分数针对每个  $BERT_{base}$ ，并且数据增强模型使用随机种子编号：2024、2025 和 2026 运行三次。

product，得分为 0.260  $F_1$  -宏观分数（全部在 text 领域训练）。然后，这些模型在训练集和开发集上进行训练，并提供它们在测试集上的预测。当提交这些模型组合时，实现了 ST1 分数 0.761 和 ST2 分数 0.4529，这是我们的官方排行榜分数。

| Category         | Number of Classes | Names of Classes                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|------------------|-------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Hazard Category  | 10                | 'allergens', 'biological', 'foreign bodies', 'fraud', 'chemical', 'other hazard', 'packaging defect', 'organoleptic aspects', 'food additives and flavourings', 'migration'                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| Product Category | 22                | 'meat, egg and dairy products', 'cereals and bakery products', 'fruits and vegetables', 'prepared dishes and snacks', 'seafood', 'soups, broths, sauces and condiments', 'nuts, nut products and seeds', 'ices and desserts', 'cocoa and cocoa preparations', 'coffee and tea', 'confectionery', 'non-alcoholic beverages', 'dietetic foods', 'food supplements', 'fortified foods', 'herbs and spices', 'alcoholic beverages', 'other food product / mixed', 'pet feed', 'fats and oils', 'food additives and flavourings', 'honey and royal jelly', 'food contact materials', 'feed materials', 'sugars and syrups' |
| Hazard           | 128               | 'listeria monocytogenes', 'salmonella', 'milk and products thereof', 'escherichia coli', 'peanuts and products thereof' ... 'dioxins', 'staphylococcal enterotoxin', 'dairy products', 'sulfamethazine unauthorised', 'paralytic shellfish poisoning (psp) toxins'                                                                                                                                                                                                                                                                                                                                                    |
| Product          | 1068              | 'ice cream', 'chicken based products', 'cakes', 'ready to eat - cook meals', 'cookies' ... 'breakfast cereals and products therefor', 'dried lilies', 'chilled pork ribs', 'tortilla chips cheese', 'ramen noodles'                                                                                                                                                                                                                                                                                                                                                                                                   |

Table 9: 四个标注类别的名称和总类数。对于 hazard 和 product，有些类被省略了。对于 product，包括测试数据的总类数为 1,142。

| Title                                                                                     | Text                                                                                                                                                                                                                                                                                                                    | hazard-category | hazard                        | product-category                     | product        |
|-------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|-------------------------------|--------------------------------------|----------------|
| Wismettac Asian Foods Issues Allergy Alert on Undeclared Wheat and Soy in Dashi Soup Base | Wismettac Asian Foods, Inc., Santa Fe Springs, CA is recalling 17.6 oz packages of Marutomo Dashi Soup Base because they may contain undeclared wheat and soy. ... Consumers with questions may contact the company at recall@wismettacusa.com.                                                                         | allergens       | soybeans and products thereof | soups, broths, sauces and condiments | soups          |
| Kader Exports Recalls Frozen Cooked Shrimp Because of Possible Health Risk                | Kader Exports, with an abundance of caution, is recalling certain consignments of various sizes of frozen cooked, peeled and deveined shrimp sold in 1lb, 1.5lb., and 2lb. retail bags. ... Consumers with questions may contact the company at +91-022-62621004/ +91-022-62621009, Mon-Fri 10:00hrs -16:00hrs GMT+5.5. | biological      | salmonella                    | seafood                              | shrimps        |
| Recall Notification: FSIS-024-94                                                          | Case Number: 024-94 Date Opened: 07/01/1994 ... Product: SMOKED CHICKEN SAUSAGE Problem: BACTERIA Description: LISTERIA Total Pounds Recalled: 2,894 Pounds Recovered: 2,894                                                                                                                                            | biological      | listeria monocytogenes        | meat, egg and dairy products         | smoked sausage |

Table 10: 来自食品召回事件数据集的样本，包括标题、文本和注释类别。

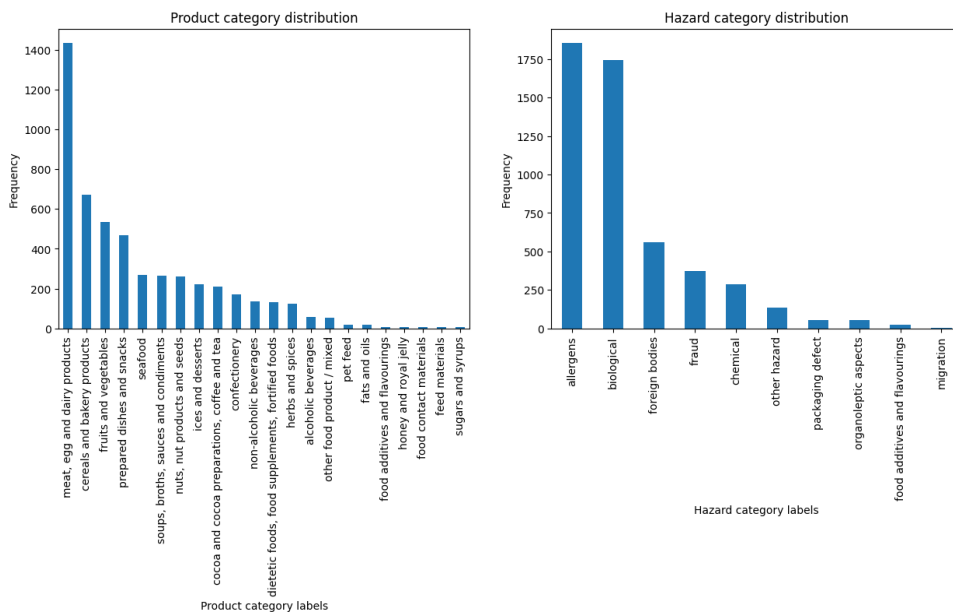


Figure 4: 类出现的 hazard-category 和 product-category 的分布。

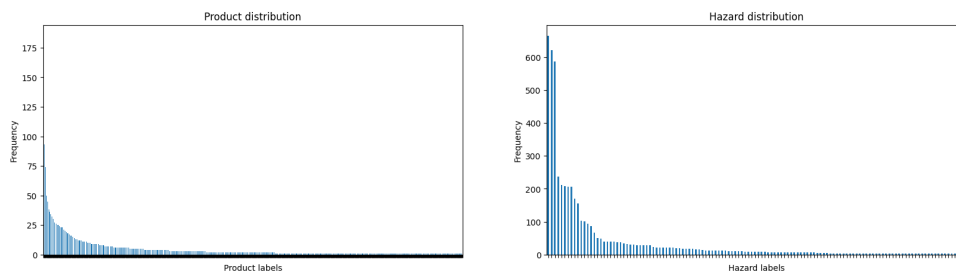


Figure 5: hazard 和 product 在类别出现中的分布。由于类别数量众多且为了图表的清晰性，x 轴上的类别已被省略。

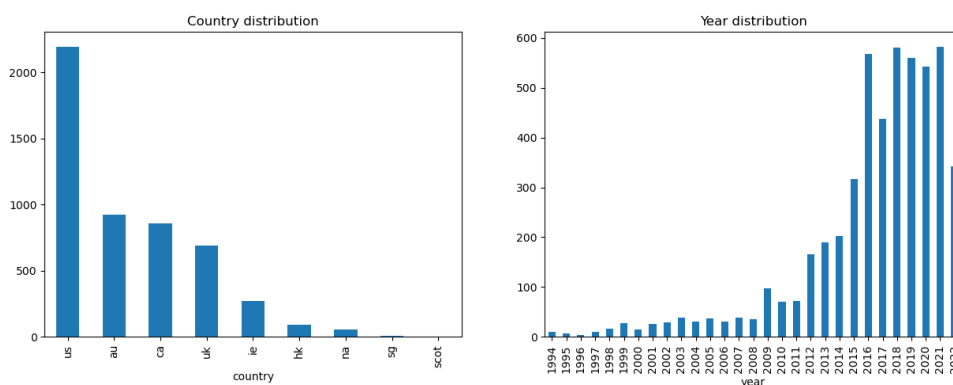


Figure 6: 给定数据集中发布的每个国家（左图）和每年（右图）的发生分布。