CareLab at #SMM4H-HeaRD 2025: Insomnia Detection and Food Safety Event Extraction with Domain-Aware Transformers

Zihan Liang*, **Ziwen Pan***, **Sumon Kanti Dey, Azra Ismail**

Emory University
Atlanta, Georgia, United States
{zihan.liang, ziwen.pan, sumon.kanti.dey, azra.ismail}@emory.edu

Abstract

This paper presents our system for the SMM4H–HeaRD 2025 shared tasks, specifically Task 4 (Subtasks 1, 2a, and 2b) and Task 5 (Subtasks 1 and 2). Task 4 focused on detecting mentions of insomnia in clinical notes, while Task 5 addressed the extraction of food safety events from news articles. We participated in all subtasks and report key findings across them, with particular emphasis on Task 5 Subtask 1, where our system achieved strong performance—securing first place with an F1 score of 0.958 on the test set. To attain this result, we employed encoder-based models (e.g., RoBERTa), alongside GPT-4 for data augmentation. This paper outlines our approach, including preprocessing, model architecture, and subtask-specific adaptations.

Introduction

The increasing digitization of healthcare and public health communications has led to a proliferation of unstructured text data from clinical records and regulatory announcements. This presents significant opportunities for natural language processing (NLP) to support early risk detection, medical condition monitoring, and health policy response. In this study, we describe our participation in two shared tasks from the SMM4H 2025 initiative, both centered on real-world textual data but addressing distinct domains and challenges. Task 4 involves detecting evidence of insomnia in clinical notes from the MIMIC-III database. Task 5 focuses on identifying food safety-related events—specifically product recalls and disease outbreaks—from news articles and FDA press releases. Together, these tasks highlight the growing need for domain-adapted NLP systems capable of extracting actionable insights from complex health-related text.

Although the two tasks differ in scope, they reflect complementary challenges in applied healthcare NLP. Task 4 requires a nuanced understanding of medical terminology, latent symptom expression, and structured diagnostic guidelines. In contrast, Task 5 emphasizes sentence-level event detection in the public health domain, which benefits from entity-rich expressions and formal news language. Our system achieved top performance in Task 5 Subtask 1, ranking first among all participants with an F1 score of 0.958 in classifying sentences into recall, outbreak, or neither categories. Through careful data augmentation, domain-specific preprocessing, and transformer-based modeling, we demonstrate strong performance in sentence classification.

While our system did not reach leading performance in the span-based subtasks, the comparative results across Task 4 and Task 5 (Subtask 2) highlight how task complexity, label sparsity, and evidence requirements influence NLP model behavior in clinical and public health domains. This study contributes to the growing literature on domain-adaptive NLP and raises further questions about designing robust pipelines for multi-layered textual inference in health-focused tasks.

Datasets

Task 4: MIMIC-III Clinical Database

The dataset used in this task is a subset of the MIMIC-III v1.4 database (Johnson et al. 2016), which contains over two million de-identified clinical notes from ICU admissions. This subset consists of 210 discharge summaries annotated for insomnia using rule-based diagnostic criteria. Each note is labeled with a binary indicator of insomnia presence, five rule-specific diagnostic tags, and corresponding evidence spans. These annotations jointly support three subtasks: overall insomnia classification, multi-label rule-based classification, and extraction of textual evidence, capturing both explicit and implicit expressions of sleep disturbance.

While prior work on sleep disorder detection in MIMIC-III has focused on structured data or general classification (Sivarajkumar and et al. 2022; Li, Liu, and et al. 2021; Irving and et al. 2021), this corpus emphasizes explainability through span-level annotation—a direction still underexplored in clinical NLP.

Task 5: FORCE Food Safety Corpus

Task 5 uses the FORCE (Foodborne disease Outbreak and ReCall Event extraction from openweb) dataset (Jana, Sinha, and Dasgupta 2024), comprising 8,100 English news articles and FDA press releases annotated at sentence and entity levels. Each sentence is labeled as either a food recall, an outbreak, or neither. The entity-level annotations cover

*These authors contributed equally.

six types of entities: **Organization (Org)** refers to the responsible company or agency; **Product (Prdt)** identifies the contaminated or recalled food item; **Cause (Cau)** captures the underlying incident cause (e.g., contamination, mislabeling); **Disease (Dis)** denotes the reported illness (e.g., *Listeria*); **Number of People Affected (#Aff)** indicates how many individuals were impacted; and **Location (Loc)** specifies the geographical site of the event. These labels are used in our Subtask 2 evaluation (see Table 2(b)).

This corpus was annotated by domain expert and underwent quality assurance via inter-annotator agreement to ensure reliability and consistency. Compared to prior datasets (Lee et al. 2023; Goel et al. 2024), FORCE provides uniquely aligned sentence-level and entity-level annotations across real-world regulatory documents.

Methodology

Task 4: Modeling Insomnia Detection from Clinical Notes

We approach the insomnia detection task as a combination of document-level classification and evidence identification, focusing on predicting insomnia status and rule-based indicators from clinical narratives.

To address both Subtask 1 (binary insomnia classification) and Subtask 2A (multi-label rule classification), we implemented and compared three distinct modeling strategies: traditional machine learning baselines, recurrent neural networks, and transformer-based models. Each approach was applied uniformly to both subtasks, with Subtask 2A framed as a multi-output binary classification problem.

First, we trained traditional classifiers—logistic regression, SVM, random forest, and XGBoost—using TF-IDF lexical features. While these models offered the advantage of computational efficiency and provided a solid baseline, they often overfit to surface-level lexical patterns and struggled to capture the deeper contextual meaning of sentences—an essential requirement in the nuanced landscape of healthcare-related text. Their limitations highlighted the need for more context-aware models capable of generalizing beyond keyword cues.

Next, we developed a BiLSTM model with pretrained GloVe embeddings to capture sequential and syntactic dependencies within the text (Pennington, Socher, and Manning 2014; Dheeraj and Ramakrishnudu 2021). Using the same model architecture was applied to both subtasks, we applied sigmoid activation to support multi-label predictions. This approach yielded moderate improvements over traditional baselines, particularly in handling syntactic variants of rule expressions.

Finally, we fine-tuned ClinicalBERT—a domain-specific transformer model pretrained on clinical corpora (Huang, Altosaar, and Ranganath 2019). We used the [CLS] token representation for classification and optimized the model using binary cross-entropy loss across all six labels (one global insomnia label and five rule-specific indicators). ClinicalBERT consistently outperformed other models, benefiting from its contextual embeddings and domain-specific pretraining. The best-performing results are reported in Table 1.

To address class imbalance, we applied class weighting during neural model training and used SMOTE oversampling for the traditional baselines to boost minority class representation during training.

For Subtask 2B, we developed a rule-based evidence extraction system using curated medical keywords and regular expression patterns. This approach was interpretable and easy to implement. However, it lacked the semantic flexibility needed to handle more complex language. Specifically, it struggled with paraphrased expressions, coreference resolution, and negation, which limited its effectiveness in accurately aligning evidence spans with the annotated text. Recent advancements in span-based pre-training techniques—such as Span Selection Pre-training (Glass et al. 2020)—offer promising improvements by teaching models to select spans from relevant passages rather than relying solely on token-level memorization, which could help address paraphrased or indirect expressions.

Task 5: Food Recall and Outbreak Detection in News

We approach Task 5 as a combination of sentence-level classification and entity-level information extraction over food safety–related news articles. Subtask 1 involved classifying each sentence into one of three categories: Food Recall, Foodborne Disease Outbreak, or Neither. Subtask 2 required extracting key entities such as product names, infection types, and affected populations.

For Subtask 1, we fine-tune an encoder-based model, RoBERTa (Liu et al. 2019), to perform 3-way sentence classification. To mitigate strong class imbalance—especially for the `Neither` category—we adopt a two-stage balancing strategy. First, we augmented the training dataset with GPT-4-generated examples (Achiam et al. 2023) that resembled neutral content from FDA documents and improved the classifier's ability to distinguish informative sentences from generic text. An example of the prompt used for this augmentation is shown in Figure 1. Second, we applied class weighting during training to penalize errors on minority classes more heavily. The loss function is defined as:

$$\mathcal{L}_{\text{weighted}} = -\sum_{c=1}^{C} w_c \cdot \mathbf{1}_{\{y=c\}} \cdot \log \hat{y}_c$$

where the weight of the class w_c is inversely proportional to the empirical frequency π_c of the class c. This combination of LLM-based augmentation and class-weighted loss significantly improved recall for underrepresented classes while maintaining high overall accuracy.

To further boost robustness, we used ensemble inference by averaging predictions across five independently trained models using different random seeds. The final system achieved the highest macro-F1 and accuracy scores in Subtask 1 (F1 = 0.958, Acc = 0.959), demonstrating the combined benefit of data-centric augmentation and transformer-based modeling.

For Subtask 2, we adopted a rule-based pipeline that de-

tects entities using spaCy's dependency parser [1] combined with hand-crafted regular expressions. Each entity type is associated with a matching function that extracts candidate text spans from an input sentence. For example, infection-related terms are extracted as:

$$\mathcal{E}_{\text{infection}}(x) = \{s \in x \mid \texttt{regex\_match}(s, \text{r“Listeria—Salmonella—E.coli”})\}$$

Although this approach is interpretable and effective for surface-form patterns, it lacks generalization to semantically implied mentions. As a result, it struggles with indirect expressions and exhibits low recall, particularly in fields such as infection causes and disease names.

Prompt template

You are an FDA press release writer. Your task is to write a formal, FDA-style press release (250–300 words) announcing an update related to food labeling or packaging.

Instructions:

- Do **NOT** discuss food recalls, disease outbreaks, or safety incidents.
- Do **NOT** include contact information, phone numbers, emails, or links.
- Ensure each press release is unique—avoid duplicating previous wording, structure, or content.
- Maintain a professional, public-health tone.
- Vary the focus slightly each time (e.g., nutrition facts, allergen labeling, ingredient transparency, packaging design updates).

Figure 1: Example prompt used for synthetic data generation.

Results

In this section, we present the experimental results of our models on the official shared tasks for Tasks 4 and 5. We report the evaluation metrics defined by the organizers for each subtask, including F1-score for binary and multi-label classification, ROUGE scores for evidence extraction, and F1-scores for entity-level information extraction.

Task 4 Results: Insomnia Detection

Table 1 presents our Task 4 results alongside task-wide mean and median scores. For Subtask 1, ClinicalBERT achieved perfect recall 1 but lower precision 0.648, resulting in an F1-score 0.786—below the task mean 0.877 and median 0.869. This performance reflects a tendency to overpredict positive cases, likely due to class imbalance and a scarcity of explicit negative examples in the training data.

[1] https://spacy.io/api/dependencyparser

Task 4 Subtask 1	**Model**	**F1-score**	**Precision**	**Recall**
CareLab	ClinicalBERT	0.786	0.648	1.000
Mean	–	0.877	0.853	0.913
Median	–	0.869	0.840	0.935
Task 4 Subtask 2A	**Model**	**F1-score**	**Precision**	**Recall**
CareLab	ClinicalBERT	0.769	0.758	0.780
Mean	–	0.718	0.673	0.789
Median	–	0.692	0.650	0.819
Task 4 Subtask 2B	**Model**	**F1-score**	**Precision**	**Recall**
CareLab	Regex Rule-based	0.135	0.097	0.372
Mean	–	0.386	0.423	0.443
Median	–	0.446	0.514	0.487

Table 1: Task 4 evaluation summary showing the performance of our best models (ClinicalBERT and rule-based extractor) across subtasks, compared with task-level benchmarks.

In Subtask 2A, which requires rule-level multi-label classification, ClinicalBERT demonstrated strong results. The model outperformed both the task mean and median in F1-score 0.769, with balanced precision 0.758 and recall 0.780. These results highlight ClinicalBERT's effectiveness in capturing multiple structured diagnostic criteria within clinical narratives. To further mitigate label imbalance, we applied SMOTE-based augmentation, which improved the model's ability to recognize underrepresented rule-specific labels.

In contrast, our approach to Subtask 2B relied on a hand-crafted, regex-based span extractor built on insomnia-related keyword dictionaries. While this method is interpretable and fast, it achieved a low F1-score of 0.135, with weak precision of 0.098 and recall of 0.372. It struggled with paraphrasing, negation, coreference, and nonstandard sentence structures. These limitations emphasize the need for context-aware models, such as span-based transformers with QA-style finetuning or contrastive span supervision, for accurate evidence extraction.

Task 5 Results: Food Recall and Outbreak Detection

(a) Subtask 1: Sentence-level classification

Team	**Model**	**F1-score**	**Precision**	**Recall**	**Accuracy**
CareLab	RoBERTa + Data Augmentation	**0.958**	**0.957**	**0.959**	**0.959**
Best	–	**0.958**	**0.957**	**0.959**	**0.959**

(b) Subtask 2: Entity-level extraction

Team	**Model**	**Avg.**	**Org**	**Prdt**	**Cau**	**Dis**	**#Aff**	**Loc**
CareLab	Rule-based Extractor	0.119	0.120	0.140	0.010	0.000	0.390	0.230
Best	–	**0.576**	0.940	0.620	0.240	0.640	0.700	0.600

Table 2: Performance comparison for Task 5 subtasks. (a) Sentence classification and (b) Entity-level extraction.

As shown in Table 2(a), our system achieved top performance in Task 5 Subtask 1, with a macro-F1 score of **0.958**, precision of **0.957**, and recall of **0.959**—matching the best-performing team. We used a RoBERTa-large model with aggressive data augmentation to strengthen generalization,

particularly for the underrepresented `Neither` class. This strong result reflects three key design choices: (1) augmenting the training set with FDA-derived samples to increase the diversity of "Neither" instances; (2) applying class-weighted focal loss to handle severe class imbalance; and (3) incorporating feature-enhanced modeling and uncertainty-aware post-processing to boost classification robustness.

In contrast, Task 5 Subtask 2 required entity-level information extraction. We adopted a rule-based extraction system, which yielded a much lower average F1-score of 0.119, compared to 0.576 from the top-performing team. As shown in Table 2(b), our model performed poorly across most entity types, particularly for `Cause and Disease`, where recall was nearly zero. Our pipeline relied on regular expressions and handcrafted heuristics to identify six types of entities from food safety news, including product names, infection types, and target organizations.

While the rule-based approach worked reasonably for surface-form mentions—such as explicit brand names or known infections like *Listeria*—it failed to generalize to paraphrased, implicit, or nested entities due to its limited coverage and lack of contextual reasoning, leading to particularly low recall and a large performance gap compared to sequence-labeling methods.

Error Analysis

Task 4

In subtask 1, which involved binary insomnia classification, the ClinicalBERT model demonstrated high recall but relatively lower precision in predicting insomnia presence, suggesting a tendency to over-predict positive cases. This is likely due to the scarcity of explicitly negative examples in the training data and the subtle linguistic cues associated with the absence of insomnia. In many instances, notes contain ambiguous or borderline mentions that were interpreted as positive by the model, especially when no strong negation was present. For example, discharge summaries often list medications like trazodone or melatonin "as needed at bedtime," without context suggesting active insomnia symptoms. In other cases, mentions of depression or blunted affect triggered false positives despite no reference to sleep difficulties. Such cases highlight the limitations of token-level cues in the absence of discourse-level understanding. Although class weighting helped to some extent, future work could explore contrastive pretraining or hard negative mining to reduce false positives.

In Subtask 2A, ClinicalBERT achieved strong overall performance on the rule-level multi-label classification task. However, the complexity of individual labels varied. Rules involving explicit terminology (e.g., mentions of "sleep latency $>$ 30 mins") were easier for the model to learn. In contrast, rules that required temporal reasoning or inference—such as identifying chronic patterns—introduced greater label noise and reduced performance. Additionally, some discharge notes also lacked complete documentation, which led to missing labels and reduced learning signals. Incorporating section-specific encoding and temporal context modeling may help improve per-rule classification consistency.

The rule-based system in Subtask 2B showed low performance (F1 = 0.135), revealing several deeper limitations beyond surface-form mismatch. First, clinical notes often describe insomnia in indirect, nuanced ways—using idiomatic expressions, subjective complaints, or embedded temporal cues—that rule-based systems are ill-equipped to capture. Second, the span annotations are often not contiguous or syntactically isolated, making it difficult for regex patterns to delineate precise evidence boundaries. Third, the system lacked the ability to reason over sentence-level discourse structure, which is crucial when the evidence is scattered across multiple clauses or requires linking antecedents and consequences. These challenges were further compounded by class imbalance and limited training data, which led to an over-reliance on high-precision but low-recall rules, ultimately reducing the system's coverage and generalizability.

Task 5

In subtask 1, the RoBERTa-based classifier achieved high performance in the sentence-level classification task, but some sources of error persisted. Misclassifications were most common between the "Outbreak" and "Recall" classes, especially in cases where overlapping or incomplete information appeared in the same sentence. In some instances, the classifier assigned the correct label only after data augmentation with GPT-4-generated negative examples, highlighting issues related to class imbalance and lexical similarity in the original training data.

In Subtask 2, our rule-based entity extraction system achieved limited performance (F1 = 0.119), falling significantly behind the top-performing system (F1 = 0.576). This performance gap is primarily due to the system's heavy reliance on surface-form pattern matching and its inability to incorporate contextual understanding. A major limitation was poor recall, especially for implicit or semantically diverse expressions. For example, the system frequently failed to identify generic references such as "the item" or "the contamination," which require contextual cues to resolve. It also struggled with paraphrased expressions of key entities—for instance, when the text described the cause of an infection in an indirect way (e.g., "due to improper handling" instead of explicitly stating "bacterial contamination"). Similarly, abstract or implied mentions of affected populations (such as "those impacted" or "the group at risk") were often overlooked, as the system lacked the semantic reasoning needed to associate these phrases with specific entity types.

Moreover, many gold entities were embedded in complex or non-canonical sentence structures, such as passive voice, nested clauses, or ellipses, which the regular expression engine failed to capture. The rule-based approach also struggled with coreference resolution and cross-type disambiguation, leading to frequent confusion when the same phrase could indicate multiple entity types depending on context. These limitations highlight the system's lack of generalization and robustness in handling real-world, linguistically diverse news texts.

To address these challenges, future work will investigate context-aware neural sequence labeling models, such as CRF-augmented transformers trained on the FORCE

corpus. These models can leverage rich contextual cues and capture entity boundaries more flexibly. Joint span-sequence frameworks, such as those proposed by Nguyen et al. (Nguyen et al. 2022), may further improve extraction by modeling both sparse entity spans and their global document context. We also plan to explore weak supervision and distant supervision techniques to expand coverage for low-resource entity types.

Conclusion

This paper presents our approach to SMM4H-HeaRD 2025 shared tasks. In Task 4, ClinicalBERT and BiLSTM models achieved strong results in rule-based classification (Subtask 2A) and competitive performance in binary classification (Subtask 1), while a rule-based span extractor (Subtask 2B) underperformed due to limited flexibility in capturing indirect insomnia mentions.

In Task 5, a RoBERTa-based classifier with LLM-driven data augmentation and class balancing delivered top-ranked results in sentence-level event classification (Subtask 1). However, our rule-based entity extractor in Subtask 2 demonstrated limited generalizability, highlighting the importance of supervised, span-aware architectures in real-world public health applications.

Overall, our findings underscore the value of domain-specific pretraining, data augmentation, and architectural alignment with task structure. In the future, we aim to explore span-aware architectures—like QA-style fine-tuning and contrastive learning (Zheng et al. 2024)—to improve both factual grounding and extraction precision. Our ultimate goal is to develop systems that not only perform well but also better reflect the complexity and richness of health-related language in real-world care. Our final codebase for all subtasks is available at: https://github.com/zihanliang/CARELAB-SMM4H2025.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dheeraj, K.; and Ramakrishnudu, T. 2021. Negative emotions detection on online mental-health related patients texts using the deep learning with MHA-BCNN model. *Expert Systems with Applications*, 182: 115265.

Glass, M.; Gliozzo, A.; Chakravarti, R.; Ferritto, A.; Pan, L.; Bhargav, G. P. S.; Garg, D.; and Sil, A. 2020. Span Selection Pre-training for Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2773–2782.

Goel, M.; Agarwal, A.; Agrawal, S.; Kapuriya, J.; Konam, A. V.; Gupta, R.; Rastogi, S.; Niharika; and Bagler, G. 2024. Deep Learning Based Named Entity Recognition Models for Recipes. *arXiv preprint arXiv:2402.17447*. Available at https://arxiv.org/abs/2402.17447.

Huang, K.; Altosaar, J.; and Ranganath, R. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Irving, J.; and et al. 2021. Using NLP on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk. *Journal of Biomedical Informatics*, 117: 103772.

Jana, S.; Sinha, M.; and Dasgupta, T. 2024. FORCE: A Benchmark Dataset for Foodborne Disease Outbreak and Recall Event Extraction from News. *Proceedings of the 9th SMM4H Workshop*.

Johnson, A. E.; Pollard, T. J.; Shen, L.; et al. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3: 160035.

Lee, N.; Vasanthakumar, U.; Chen, R.; et al. 2023. Predictive Food Safety Risk Monitoring using AI Technologies. In *IEEE International Conference on ICT for Smart Society (ICISS)*.

Li, H.; Liu, J.; and et al. 2021. Development and Validation of a Clinical Prediction Model for Sleep Disorders in the ICU: A Retrospective Cohort Study. *Frontiers in Neuroscience*, 15: 644845.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nguyen, H. S.; Vu, H. M.; Nguyen, T.-A. D.; and Nguyen, M.-T. 2022. Jointly Learning Span Extraction and Sequence Labeling for Information Extraction from Business Documents. *arXiv preprint arXiv:2205.13434*.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Sivarajkumar, H.; and et al. 2022. Extraction of Sleep Information from Clinical Notes of Patients with Alzheimer's Disease Using Natural Language Processing. *arXiv preprint arXiv:2204.09601*.

Zheng, L. A.; Hui, Z.; Liu, Z.; and Hirschberg, J. 2024. Enhancing Pre-Trained Generative Language Models with Question-Attended Span Extraction on Machine Reading Comprehension. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 10046–10063.