# Overview of the 10th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ICWSM 2025


**Ari Z. Klein[1], Tirthankar Dasgupta[2], Lauren Gryboski[4], Sudeshna Jana[2], Sedigh Khademi[5], Guillermo Lopez-Garcia[3], Diego Mazzotti[4], Takeshi Onishi[3], Jeanne Powell[6], Lisa Raithel[7,8,9], Swati Rajwal[6], Roland Roller[9], Abeed Sarker[6], Manjira Sinha[2], Philippe Thomas[9], Elena Tutubalina[10,11], Dongfang Xu[3], Pierre Zweigenbaum[12], Graciela Gonzalez-Hernandez[3]**

[1]University of Pennsylvania, Philadelphia, PA, USA
[2]TCS Research, West Bengal, India
[3]Cedars-Sinai Medical Center, Los Angeles, CA, USA
[4]University of Kansas Medical Center, Kansas City, KS, USA
[5]Murdoch Children's Research Institute, Melbourne, Australia
[6]Emory University, Atlanta, GA, USA
[7]Technische Universität Berlin, Berlin, Germany
[8]Berlin Institute for the Foundations of Learning and Data, Berlin, Germany
[9]German Research Center for Artificial Intelligence, Berlin, Germany
[10]Artificial Intelligence Research Institute, Moscow, Russia
[11]Kazan Federal University, Kazan, Russia
[12]Université Paris-Saclay, CNRS, LISN, Orsay, France
ariklein@pennmedicine.upenn.edu, Graciela.GonzalezHernandez@csmc.edu



## Abstract

The aim of the Social Media Mining for Health (#SMM4H) shared tasks is to take a community-driven approach for developing and evaluating natural language processing, machine learning, and artificial intelligence methods to utilize publicly available social media data for health research. For the 10th iteration, hosted at the AAAI International Conference on Web and Social Media (ICWSM) 2025, we broadened the scope to include additional web-based sources of "health real-world data" (HeaRD). The 6 tasks represented various data sources (Twitter, Reddit, patient forums, clinical notes, news articles), languages (English, Russian, German, French), health-related topics (adverse drug and vaccine events, nonmedical substance use, dementia family caregiving, insomnia, foodborne disease outbreaks), and methods (binary classification, multi-class classification, named entity recognition). In total, 57 teams registered, representing 17 countries. In this paper, we present an overview of the annotated corpora, participants' systems, and performance results, providing insights into state-of-the-art methods for mining social media and other web-based data sources for health research. To facilitate future work, the datasets remain available by request, and the CodaLab sites remain active for a post-evaluation phase.


## Introduction

With more than 70% of adults in the United States (Auxier and Anderson 2021) and more than 60% of people worldwide (Petrosyan 2025) using social media, the Data Modernization Initiative of the Centers for Disease Control and Prevention (CDC) encourages the use of "non-traditional data sources, including images, audio, social media, and data not specifically collected for public health analysis, such as electronic health records" (Centers for Disease Control and Prevention 2023). The aim of the Social Media Mining for Health (#SMM4H) shared tasks is to take a community-driven approach for developing and evaluating natural language processing, machine learning, and artificial intelligence methods to utilize publicly available social media data for health research. The 10th iteration of the shared tasks was hosted at the AAAI International Conference on Web and Social Media (ICWSM) 2025 and included additional web-based sources of "real-world data" (U.S. Food and Drug Administration 2024). To reflect this broader scope, we extended the name of the shared tasks to *#SMM4H-HeaRD*, where the latter stands for "health real-world data" and is intended to represent the notion of using social media and other web-based data sources as a complementary approach for "listening" to patients.

The #SMM4H-HeaRD 2025 shared tasks consisted of 6 tasks that represented various data sources (Twitter, Reddit, patient forums, clinical notes, news articles), languages (English, Russian, German, French), health-related topics (adverse drug and vaccine events, nonmedical substance use, dementia family caregiving, insomnia, foodborne disease outbreaks), and methods (binary classification, multi-class classification, named entity recognition). Teams were provided with gold standard annotated training and validation sets to develop their systems and, subsequently, a blind test set for the final evaluation. After receiving the test set, teams were given 5 days to submit the predictions of their systems to CodaLab—a platform that facilitates data science competitions—for automatic evaluation, promoting the systematic comparison of performance. Teams could register for a single task or multiple tasks. Among the 57 teams that registered, representing 17 countries, 29 teams participated by submitting system predictions to CodaLab and submitting a short manuscript describing their system: 10 teams for Task 1, 1 team for Task 2, 7 teams for Task 3, 7 teams for Task 4, 3 teams for Task 5, and 10 teams for Task 6. Each system description was peer-reviewed by at least 2 reviewers. In this paper, we present an overview of the annotated corpora, participants' systems, and performance results.

## Tasks

### Task 1: Detection of Adverse Drug Events in Multilingual and Multi-platform Social Media Posts

Adverse drug events (ADEs), defined as "harmful or unpleasant reactions resulting from an intervention related to the use of a medicinal product" (Edwards and Aronson 2000), pose a significant challenge to public health monitoring due to under-reporting and non-sufficient coverage (Hazell and Shakir 2006). In recent years, extracting ADE mentions from user-generated content on social media has emerged as a valuable approach to identifying early signals of drug safety issues. These platforms host large volumes of informal health discussions, expressed in users' own words and languages, offering a rich yet underutilized data source. Despite potential limitations, such as non-representative sampling (Hargittai and Walejko 2008; Wagner et al. 2015), social media enables individuals to share personal experiences anonymously and without fear of stigma or dismissal, factors known to contribute to under-reporting in traditional surveillance systems (Yang et al. 2012; Palleria et al. 2013).

Although the number of datasets related to ADEs has been increasing (Dai et al. 2024), we can still see a lack of diversity with respect to languages and data sources. Therefore, this shared task focused on *multilingual* binary classification of social media posts from *different platforms* to determine the presence or absence of ADE mentions. By developing robust multilingual models, we aim to advance cross-lingual health surveillance and improve the timeliness and reach of ADE signal detection across diverse linguistic and cultural contexts. We provided posts from Twitter and other social media platforms, each labeled according to the presence of an ADE. A post with a positive label contained at least one mention of an ADE, while a post with a negative label did not. The dataset contained 18,876 tweets written in English (Xu et al. 2024), 13,424 tweets (Magge et al. 2021) and drug reviews (Tutubalina et al. 2020) written in Russian, 2,116 posts from patient forums written in German (Raithel et al. 2022), and 1,396 posts from patient forums translated into French (Raithel et al. 2024). The evaluation metric was the $F_1$-score for the "positive" class. The CodaLab site for this task is: https://codalab.lisn.upsaclay.fr/competitions/21886.

### Task 2: Extraction of Clinical and Social Impacts of Nonmedical Substance Use from Reddit

In the United States, nearly 12 million people misuse opioids, contributing to substantial clinical and social harms, including increased mortality and public health burdens (Bolshakova, Bluthenthal, and Sussman 2019). Timely detection of these impacts is essential for guiding interventions and allocating resources (Volkow, Chandler, and Villani 2022). However, people who use drugs often do not disclose illicit use in clinical settings, limiting the utility of traditional data sources (Strike et al. 2020). Researchers have long leveraged social media as a cohort-centered data source to explore the lived experiences of patients, such as those with breast cancer (Rajwal et al. 2024) and migraines (Guo et al. 2023). These studies demonstrate the potential of social media to offer a complementary perspective on patient experiences, which we extend here to the context of opioid misuse. To support the automatic detection of these impacts, we organized a shared task using the Reddit-Impacts dataset, which includes 1,380 Reddit posts manually annotated for clinical and social consequences of nonmedical opioid use (Ge et al. 2024). Participants were challenged to develop named entity recognition (NER) models capable of identifying these nuanced and sparse concepts in highly variable, informal text. The CodaLab site for this task is: https://codalab.lisn.upsaclay.fr/competitions/22203

### Task 3: Detection of Dementia Family Caregivers on Twitter

Internet-based interventions to support family caregivers of people with dementia are valued by caregivers for their easy access (Hopwood et al. 2018) and can have beneficial effects on caregivers' health (Leng et al. 2020). Given that nearly 1 of every 4 adults in the United States already uses Twitter (Gottfried 2024), Twitter may present a novel opportunity to reach caregivers on a large scale. This binary classification task involved automatically distinguishing English-language tweets that reported having a family member with dementia from tweets that merely mentioned dementia, enabling the use of Twitter not only to directly target interventions at family caregivers, but also to inform interventions based on the content of their tweets (Feng et al. 2025). The training, validation, and test sets contained 6,724 tweets, 353 tweets, and 1,769 tweets, respectively: 5,946 (67%) that reported having a family member with dementia and 2,900 (33%) that merely mentioned dementia (Klein et al. 2022). Inter-annotator agreement (Fleiss' kappa), based on 500 tweets that were annotated by all 3 annotators, was 0.82. The evaluation metric was the $F_1$-score for the class of tweets that reported having a family member with dementia. The CodaLab site for this task is: https://codalab.lisn.upsaclay.fr/competitions/22022.

### Task 4: Detection of Insomnia in Clinical Notes

Insomnia is a common sleep disorder with significant health implications, including psychiatric conditions, reduced workplace productivity, and increased risk of accidents. Despite its high prevalence, it remains largely underdiagnosed (Ulmer et al. 2017). Effective methods for detecting insomnia are critical to better understand its prevalence, associated risk factors, progression, and treatment outcomes (Kartoun et al. 2018). We organized the first shared task focused on the automatic identification of patients potentially suffering from insomnia using electronic health records. Structured as a text classification challenge, the task required participants to analyze clinical notes and determine if a patient was likely to have insomnia. We curated an annotated corpus of 164 clinical notes from the MIMIC-III database (Johnson et al. 2016), following a comprehensive set of rules designed to guide the identification of both direct and indirect symptoms of insomnia, as well as the presence of commonly prescribed hypnotic medications. The corpus was divided into training (70 notes), validation (20 notes) and test (74 notes) sets. Each note included a binary label indicating the overall insomnia status ("yes" or "no"), along with rule-level annotations specifying whether each diagnostic rule was satisfied. To promote model transparency and explainability, textual evidence supporting each annotation was also provided.

This shared task was divided into three distinct subtasks:

1. Subtask 1: Binary Text Classification. Participants were required to predict whether a patient described in a clinical note was likely to have insomnia ("yes" or "no"). Evaluation used the $F_1$-score, treating "yes" as the positive class.
2. Subtask 2A: Multi-label Text Classification. Participants were required to evaluate each clinical note against the defined insomnia rules (Definition 1, Definition 2, Rule A, Rule B, and Rule C) and predict "yes" or "no" for each rule item. The primary evaluation metric was the micro-averaged $F_1$-score, considering "yes" as the positive class.
3. Subtask 2B: Evidence-Based Classification. This subtask extended Subtask 2A by also requiring participants to identify and extract text spans from the clinical note that justified each "yes" classification. The evaluation was based on the ROUGE-L metric. This subtask emphasized transparency and explainability by requiring models to provide supporting evidence for their decisions.

A GitHub repository containing resources and evaluation scripts is available at: https://github.com/guilopgar/SMM4H-HeaRD-2025-Task-4-Insomnia. The CodaLab site for this task is: https://codalab.lisn.upsaclay.fr/competitions/22509.

### Task 5: Detection and Extraction of Food Recalls and Foodborne Disease Outbreaks in Online News Articles

The rising incidence of food safety issues remains a critical global concern (Kase, Zhang, and Chen 2017; Boatemaa et al. 2019). Foodborne illnesses continue to pose significant public health challenges, ranking among the leading causes of morbidity and mortality worldwide (Pádua et al. 2019; Lüth et al. 2019). Food and beverage contamination is a multi-factorial issue that can occur throughout various stages of the production and distribution pipeline, including raw material sourcing, transportation, cleaning procedures, thermal processing, packaging, and storage. Outbreaks may emerge before, during, or after these stages (Scallan and Mahon 2012). One of the primary consequences of these incidents is the initiation of food recalls, which can result in considerable economic losses for both industry stakeholders and national economies (Deng, den Bakker, and Hendriksen 2016). These concerns highlight the critical need for identifying root causes and contributing factors in food contamination events (Zhou et al. 2020). Developing a comprehensive understanding of potential contamination pathways is essential for effective outbreak prevention and timely recall interventions (Tao, Yang, and Feng 2020; Zhou, Zhang, and Wang 2021; Jin et al. 2020; Marvin et al. 2017). This shared task focused on the automatic identification of foodborne disease outbreaks and food recall events in online, English-language news articles. The training, development, and test sets contained 3,172 news articles, 357 news articles, and 1,005 news articles, respectively. The task comprised two subtasks: a multi-class classification task to categorize a given article as *Food Recall*, *Foodborne Disease Outbreak*, or *Neither* (Subtask 1) and an NER task to extract *Target Organization*, *Product Name*, *Cause of Incident*, *Disease Caused*, *Number of People Affected*, and *Location* from the the articles (Subtask 2). The CodaLab site for this task is: https://codalab.lisn.upsaclay.fr/competitions/22154.

### Task 6: Detection of Personally Experienced Vaccine Adverse Events on Reddit

The success of vaccination programs relies on comprehensive safety monitoring systems. Although vaccines undergo extensive testing before approval, pre-market trials inherently face challenges in generating comprehensive safety data due to homogeneous participant groups and limited timeframes. Post-licensure surveillance plays a critical role in ensuring vaccine safety (Buttery and Clothier 2022), and social media offers a complementary lens for capturing self-reported adverse events following immunization (AEFIs) (Khademi Habibabadi et al. 2022). To support the detection of personally experienced AEFIs using social media, this task involved binary classification of English-language Reddit posts mentioning shingles (zoster) vaccines. The goal was to distinguish between posts that contained personal experiences of AEFIs and those that did not. The dataset included 3,306 Reddit submissions collected up to April 2024, manually labeled by two experts in healthcare and natural language processing. The evaluation metric was the $F_1$-score for the class of posts that reported personal experiences of AEFIs. The CodaLab site for this task is: https://codalab.lisn.upsaclay.fr/competitions/22159.

## Results

### Task 1: Detection of Adverse Drug Events in Multilingual and Multi-platform Social Media Posts

As a baseline, we fine-tuned an XLM-RoBERTa (Conneau et al. 2020) model, which achieved an $F_1$-score of 0.60 across the multilingual posts in the test set: 11,712 in English, 9,292 in Russian, 1,105 in German, and 1,104 in French. Table 1 presents the performance for the 10 teams that participated in Task 1. PEI achieved the highest $F_1$-score (0.709) by fine-tuning the Mistral-Nemo-Instruct-2407 large

| Team | $F_1$ | P | R | System Summary |
|---|---|---|---|---|
| PEI | **0.709** | 0.700 | 0.717 | Mistral-Nemo-Instruct-2407, fine-tuning, error-driven data augmentation via LLM paraphrasing |
| Y2K | 0.708 | **0.775** | 0.652 | Gemma 3, fine-tuning, PEFT, LoRA, ensemble, translation, data augmentation via Claude 3.7 Sonnet reasoning |
| Deloitte Drocks | 0.677 | 0.627 | 0.736 | 4-bit quantized Phi-4, fine-tuning, PEFT, LoRA, threshold tuning |
| ADETrackers | 0.656 | 0.717 | 0.605 | Llama-3.1-8B-Instruct, fine-tuning, PEFT, LoRA, 4-bit quantization |
| RIGA | 0.656 | 0.624 | 0.691 | RoBERTa-Large, data augmentation via GPT-4o, DrugBank |
| ACSS-PSL | 0.633 | 0.622 | 0.645 | XLM-RoBERTa-Large |
| Baseline | 0.605 | 0.582 | 0.629 | XLM-RoBERTa-Large |
| LLM Pros | 0.592 | 0.468 | **0.803** | Open AI o3-mini, prompting, structured output |
| HSE NLP | 0.567 | 0.462 | 0.733 | GPT-4o, few-shot prompting, EuroBERT, ensemble, UMLS-informed LLM data augmentation |
| BEARCAT | 0.559 | 0.731 | 0.452 | GPT-4o, zero-shot prompting |
| SRMISTAdverseDrug | 0.219 | 0.140 | 0.495 | Logistic Regression, CatBoost, XGBoost, Random Forest, ensemble, translation, TF-IDF, multilingual sentence embeddings, SMOTE |

Table 1: System summaries and $F_1$-score ($F_1$), precision (P), and recall (R) for the detection of adverse drug events in multilingual and multi-platform social media posts (Task 1).

| Team | Relaxed $F_1$ | Strict $F_1$ | Token-level $F_1$ | System Summary |
|---|---|---|---|---|
| Baseline 1 | **0.544** | **0.326** | **0.508** | data augmentation with nearest neighbor (DANN) classifier, few-shot learning |
| LLM Pros | 0.423 | 0.220 | 0.332 | GPT-4o, prompting, structured output |
| Baseline 2 | 0.167 | 0.110 | 0.261 | GPT-3.5, one-shot prompting |

Table 2: System summary and relaxed, strict, and token-level $F_1$-scores ($F_1$) for extraction of clinical and social impacts of nonmedical substance use from Reddit (Task 2).

language model (LLM) and using paraphrasing-based error-driven augmentation. Team Y2K ($F_1$-score=0.708) followed closely with a two-stage self-corrective decoder architecture based on Gemma 3 (Gemma Team 2025) models fine-tuned via PEFT. While Y2K achieved 7.51 percentage points higher in precision, its recall was 6.55 percentage points lower than that of PEI. Team Deloitte Drocks ($F_1$-score=0.677) placed third by using PEFT and threshold tuning on a quantized Phi-4 model. For language-specific performance, PEI achieved the highest $F_1$-scores for English (0.78) and French (0.78), while Y2K achieved the highest $F_1$-scores for Russian (0.65) and German (0.77). Despite a larger amount of data than French and German, Russian achieved a notably lower performance, which may be due to challenges with domain transfer across the multiple Russian data sources. Most high-performing systems used parameter-efficient fine-tuning, such as LoRA (Hu et al. 2021), multilingual models, or combined LLMs with external knowledge. In contrast, prompt-based approaches achieved lower precision and $F_1$-scores, despite relatively strong recall. Overall, results emphasize the difficulty of ADE detection in noisy, multilingual social media text, and the advantage of hybrid strategies combining LLMs with fine-tuning and augmentation. Participants outperformed the baseline with varied approaches, underscoring the value of adaptable, multilingual architectures for pharmacovigilance.

## Task 2: Extraction of Clinical and Social Impacts of Nonmedical Substance Use from Reddit

Table 2 presents the performance for the 1 team that participated in Task 2 and learning-based and prompting-based baselines (Ge et al. 2024). LLM Pros developed a structured and instruction-driven pipeline prompting OpenAI's GPT-4o (OpenAI 2024b) LLM to extract concise impact phrases from Reddit posts. Their approach includes a schema-guided extraction mechanism (implemented via Pydantic) to categorize impacts into clinical, social, or other types, and focuses on minimally sufficient phrases rather than longer contextual spans. While they outperformed the GPT-3.5 prompting baseline for all of the evaluation metrics, they did not outperform the few-shot learning baseline—a data augmentation with nearest neighbor (DANN) classifier—for any of them. Nevertheless, their approach complements the DANN baseline as their model-guided extraction could serve as a high-precision candidate generator, while the baseline could act as a recall-enhancing filter or re-ranker. A hybrid system that first identifies compact phrases via GPT-4o and then expands or validates them using few-shot augmented retrieval could potentially achieve stronger overall balance between precision and recall in future iterations.

## Task 3: Detection of Dementia Family Caregivers on Twitter

In prior work (Klein et al. 2022), a baseline classifier achieved an $F_1$-score of 0.962 (precision=0.946 and recall=0.979) based on fine-tuning a BERTweet-Large (Nguyen, Vu, and Tuan Nguyen 2020) pre-trained model. Thus, participants were encouraged to experiment with approaches based on LLM prompting to compare with the high baseline performance. Table 3 presents the performance for the 7 teams that participated in Task 3. BOUN ($F_1$-score=0.966) marginally outperformed the baseline by

| Team | $F_1$ | P | R | System Summary |
|---|---|---|---|---|
| BOUN | **0.966** | **0.957** | 0.976 | Gemma-2-2B, fine-tuning, LoRA |
| IAI | 0.964 | 0.956 | 0.971 | BERTweet |
| Baseline | 0.962 | 0.946 | 0.979 | BERTweet-Large |
| NoviceTrio | 0.957 | 0.951 | 0.962 | BERT-Base-Uncased, BERTweet, Llama-3.1-8B, few-shot prompting, ensemble |
| Mason NLP-GRP | 0.954 | 0.946 | 0.962 | Llama-3.1-8B, zero-shot prompting |
| LLATMU | 0.948 | 0.911 | 0.987 | BERTweet |
| BingAster | 0.942 | 0.899 | **0.991** | DeepSeek-R1-70B, zero-shot prompting |
| LLM Pros | 0.203 | 0.802 | 0.116 | OpenAI o3-mini, prompting, structured output |

Table 3: System summaries and $F_1$-score ($F_1$), precision (P), and recall (R) for the detection of dementia family caregivers on Twitter (Task 3).

| Team | Subtask 1 | | | Subtask 2A | | | Subtask 2B | | | System Summary |
|---|---|---|---|---|---|---|---|---|---|---|
| | $F_1$ | P | R | $F_1$ | P | R | R-L $F_1$ | R-L P | R-L R | |
| LLM Pros | **0.967** | **0.978** | 0.957 | **0.906** | **0.896** | **0.917** | **0.682** | **0.706** | **0.724** | OpenAI o3-mini, zero-shot prompting, structured output |
| RBG-AI | 0.946 | 0.936 | 0.957 | 0.750 | 0.650 | 0.886 | 0.463 | 0.522 | 0.487 | Gemma-2B, prompting, rule-based extractor |
| HaleLab_NITK | 0.891 | 0.891 | 0.891 | 0.695 | 0.586 | 0.856 | 0.411 | 0.480 | 0.474 | Llama-3-8B, zero-shot prompting, pipeline |
| IAI | 0.875 | 0.840 | 0.913 | 0.689 | 0.682 | 0.697 | - | - | - | MedBERT, Clinical BigBird, SciBERT, ensemble |
| SRMISTNLPInsomnia | 0.842 | 0.816 | 0.870 | 0.657 | 0.654 | 0.659 | - | - | - | AdaBoost, TF-IDF, ClinicalBERT embeddings, SMOTE |
| Bit-UA | 0.809 | 0.837 | 0.783 | 0.607 | 0.739 | 0.515 | 0.446 | 0.514 | 0.450 | BERT-Base, token classification |
| CareLab | 0.786 | 0.648 | **1.000** | 0.769 | 0.757 | 0.780 | 0.135 | 0.097 | 0.372 | ClinicalBERT, rule-based extractor |

Table 4: System summaries and $F_1$-score ($F_1$), Precision (P), Recall (R), ROUGE-L $F_1$-score (R-L $F_1$), ROUGE-L Precision (R-L P), and ROUGE-L Recall (R-L R) for the detection of insomnia in clinical notes (Task 4).

using LoRA (Hu et al. 2021) to fine-tune the Gemma-2-2B LLM with the annotated training data. NoviceTrio ($F_1$-score=0.957) used BERTweet in a majority voting ensemble with BERT-Base and few-shot prompting of the Llama-3.1-8B (Llama Team, AI at Meta 2024) LLM, but did not improve upon the baseline performance. In addition to prompting the LLM for classification, NoviceTrio prompted the LLM to generate explanations, which were concatenated to the corresponding tweets in the training data used to fine-tune the BERTweet and BERT-Base (Devlin et al. 2019) models in the ensemble. Whereas NoviceTrio used Llama-3.1-8B in an ensemble, Mason NLP-GRP ($F_1$-score=0.954) used only zero-shot prompting of Llama-3.1-8B, achieving a performance that was nearly identical to that of NoviceTrio and only marginally lower than the baseline. Similarly, BingAster ($F_1$-score=0.942) achieved comparable performance by using zero-shot prompting of the DeepSeek-R1-70B (DeepSeek-AI 2025) LLM. NoviceTrio and BingAster used prompts with simple questions or instructions that focused on the primary task, whereas LLM Pros ($F_1$-score=0.203) achieved substantially lower performance by prompting the OpenAI o3-mini LLM and using structured output with a complex schema. Nonetheless, the overall results demonstrate that prompting-based approaches—in particular, zero-shot prompting—can be used for this task.

## Task 4: Detection of Insomnia in Clinical Notes

Table 4 presents the performance for the 7 teams that participated in Task 4. Of these, 5 teams participated in all three subtasks, while IAI and SRMISTNLPInsomnia participated only in Subtasks 1 and 2A. LLM Pros achieved the best performance across all three subtasks: Subtask 1 ($F_1$-score=0.967), Subtask 2A ($F_1$-score=0.906), and Subtask 2B (ROUGE-L $F_1$-score=0.682). Their approach relied entirely on a prompt-based pipeline without any task-specific fine-tuning. They utilized the OpenAI o3-mini LLM, employing prompt engineering combined with strictly enforced JSON schemas to guide the model outputs. RBG-AI achieved the second-best performance in Subtask 1 ($F_1$-score=0.946) and Subtask 2B (ROUGE-L $F_1$=0.463), and competitive results in Subtask 2A ($F_1$-score=0.750). Their system combined structured prompting of the open-source Gemma-2B LLM with a regular expression-based medication pattern extractor. HaleLab_NITK (Subtask 1 $F_1$-score=0.891, Subtask 2A $F_1$-score=0.695, and Subtask 2B ROUGE-L $F_1$-score=0.411) leveraged LLMs with a reverse

| Team | Subtask 1 | | | | Subtask 2 | | | | | | | System Summary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | P | R | $F_1$ | Avg | O | P | C | D | A | L | |
| CareLab | 0.96 | 0.96 | 0.96 | **0.96** | 0.12 | 0.12 | 0.14 | 0.01 | 0.00 | 0.39 | 0.23 | Subtask 1: RoBERTa, ensemble, GPT-4 data augmentation<br>Subtask 2: spaCy dependency parser, regular expressions |
| LLM Pros | 0.88 | 0.84 | 0.92 | 0.88 | **0.57** | 0.94 | 0.62 | 0.24 | 0.64 | 0.70 | 0.60 | OpenAI o3-mini, prompting, structured output |
| WITM | 0.93 | 0.92 | 0.93 | 0.92 | 0.48 | 0.89 | 0.43 | 0.22 | 0.60 | 0.56 | 0.42 | Subtask 1: DistilBERT<br>Subtask 2: LLaMA-3.1-8B, few-shot prompting; linguistically informed Bi-LSTM |
| Baseline | 0.87 | 0.83 | 0.89 | 0.86 | 0.53 | 0.88 | 0.53 | 0.25 | 0.50 | 0.66 | 0.61 | BERT-based CNN, BiLSTM, multi-task model |

Table 5: System summaries and accuracy (Acc), $F_1$-score, ($F_1$), precision (P), and recall (R) for the detection (Subtask 1) and extraction (Subtask 2) of food recalls and foodborne disease outbreaks in online news articles (Task 5). The evaluation metric for Subtask 2 is the average (Avg) of the accuracy for each entity type: Organization (O), Product (P), Cause (C), Disease (D), Number of People Affected (A), and Location (L).

reasoning pipeline that prioritized evidence extraction before classification. They used Llama-3-8B (Llama Team, AI at Meta 2024) to first extract relevant text spans for each insomnia rule (Subtask 2B), which were then used to predict rule satisfaction (Subtask 2A) and the insomnia status (Subtask 1). The remaining teams focused on BERT-based text classification approaches. IAI (Subtask 1 $F_1$-score=0.875 and Subtask 2A $F_1$-score=0.689) and CareLab (Subtask 1 $F_1$-score=0.786, Subtask 2A $F_1$-score=0.769, and Subtask 2B ROUGE-L $F_1$-score=0.135) fine-tuned BERT-based models—MedBERT (Vasantharajan et al. 2022), Clinical BigBird (Li et al. 2022), SciBERT (Beltagy, Lo, and Cohan 2019), and ClinicalBERT (Huang, Altosaar, and Ranganath 2020)—for Subtasks 1 and 2A, framing Subtask 1 as a binary classification task and Subtask 2A as a set of independent binary classification problems. Notably, CareLab achieved the second-best performance in Subtask 2A. SR-MISTNLPInsomnia (Subtask 1 $F_1$-score=0.842 and Subtask 2A $F_1$-score=0.657) combined ClinicalBERT embeddings with TF-IDF features and applied class balancing strategies such as SMOTE (Chawla et al. 2002). Finally, Bit-UA (Subtask 1 $F_1$-score=0.809, Subtask 2A $F_1$-score=0.607, and Subtask 2B ROUGE-L $F_1$-score=0.446) implemented a hybrid pipeline using Finite Context Models for classification and a BERT-based token classification model for evidence extraction in Subtask 2B. Overall, prompt-based approaches leveraging LLMs achieved the highest performance across all subtasks, highlighting their potential for addressing complex clinical information extraction tasks using limited annotated real-world health data.

## Task 5: Detection and Extraction of Food Recalls and Foodborne Disease Outbreaks in Online News Articles

In prior work (Jana, Sinha, and Dasgupta 2024), a baseline multi-task model combining a BERT-based CNN and BiLSTM achieved an $F_1$-score of 0.86 for classifying news articles as *food recall*, *foodborne disease outbreak*, or *neither* (Subtask 1), and prompting a Llama-2 (GenAI, Meta 2023) LLM achieved an average accuracy of 0.53 for extracting entities (Subtask 2). Table 5 presents the performance for the 3 teams that participated in Task 5. CareLab achieved the highest $F_1$-score (0.96) for Subtask 1, addressing the class imbalance in the "neither" class by augmenting the training set with examples generated by prompting GPT-4 (OpenAI 2024a), followed by fine-tuning a RoBERTa (Liu et al. 2019) model for classification. WITM achieved an $F_1$-score of 0.92 by fine-tuning a DistilBERT (Sanh et al. 2020) model on the original training set. While LLM Pros achieved an $F_1$-score of 0.88 for Subtask 1 by prompting OpenAI's o3-mini LLM and using a complex output schema, they achieved the highest average accuracy (0.57) for Subtask 2 by using this same approach. WITM achieved an average accuracy of 0.48 by using structured few-shot prompting of the Llama-3.1-8B (Llama Team, AI at Meta 2024) LLM to extract *Organization*, *Product*, *Disease*, *Number of Affected People*, and *Location*, and a linguistically informed BiLSTM model to extract *Cause*. CareLab achieved an average accuracy of 0.12 by using a rule-based pipeline that combined spaCy's dependency parser with hand-crafted regular expressions. Overall, participants adopted a diverse mix of traditional and LLM-based approaches, demonstrating that, while legacy transformer models, such as RoBERTa and DistilBERT, remain competitive for straightforward classification tasks, LLM-based methods offer enhanced robustness and adaptability for more complex extraction tasks. Nevertheless, issues such as hallucination and inconsistency persist in LLM-generated outputs, underscoring the need for more sophisticated, multi-stage prompting frameworks in future work.

## Task 6: Detection of Personally Experienced Vaccine Adverse Events on Reddit

In prior work (Khademi et al. 2024), a baseline classifier achieved an $F_1$-score of 0.95 (precision=0.93 and recall=0.96) based on fine-tuning a Twitter-RoBERTa pretrained model, outperforming few-shot and chain-of-thought prompting of GPT-4 (OpenAI 2024a), which achieved an $F_1$-score of 0.90 (precision=0.86 and recall=0.93). Table 6 presents the performance for the 10 teams that participated in Task 6. BioNLP1 achieved the highest $F_1$-score (0.959) by using an SVM classifier with TF-IDF features in an en-

| Team | $F_1$ | P | R | System Summary |
|---|---|---|---|---|
| BioNLP1 | **0.959** | 0.946 | 0.973 | SVM, TF-IDF, RoBERTa-Large sentence embeddings, ensemble |
| GooSeek | 0.957 | 0.949 | 0.966 | Llama 3.1-8B, chain-of-thought prompting |
| PEI | 0.954 | 0.934 | 0.976 | Mistral-Nemo-Instruct-2407, fine-tuning, error-driven data augmentation via LLM paraphrasing |
| ACSS-PSL | 0.951 | 0.940 | 0.962 | DeBERTa-V3-Base, class weights, threshold tuning |
| BrynMawrNLP | 0.950 | 0.928 | 0.973 | RoBERTa |
| Baseline | 0.946 | 0.930 | 0.962 | Twitter-RoBERTa, data augmentation via GPT-4-Turbo chain-of-thought prompting |
| UoT | 0.945 | 0.916 | 0.976 | Twitter-RoBERTa-Large-2022-154M, class weights |
| beatAVE | 0.943 | **0.951** | 0.935 | Twitter-RoBERTa-Large-2022-154M |
| NU Health Miners | 0.943 | 0.900 | **0.990** | RoBERTa-Large ensemble |
| HpiVaxVigil | 0.919 | 0.883 | 0.959 | BERTweet-Large, data augmentation via GPT-4o chain-of-thought prompting |
| LLM Pros | 0.882 | 0.843 | 0.924 | OpenAI o3-mini, prompting, structured output |

Table 6: System summaries and $F_1$-score ($F_1$), precision (P), and recall (R) for detection of personally experienced vaccine adverse events on Reddit (Task 6).

semble with mean-pooled sentence embeddings weighted by attention masks, based on fine-tuning RoBERTa-Large (Liu et al. 2019). GooSeek achieved a nearly identical $F_1$-score (0.957) by using chain-of-thought prompting of the Llama 3.1-8B-Instruct (Llama Team, AI at Meta 2024) LLM, and PEI achieved a similar $F_1$-score (0.954) by fine-tuning the Mistral-Nemo-Instruct-2407 LLM and also using the LLM to paraphrase misclassifications for error-driven data augmentation. Most of the other teams used fine-tuned BERT-based models. The task was challenging due to confusion between vaccine side effects and symptoms of shingles or other illnesses, and because positive labels included adverse events from vaccines other than shingles. The results demonstrate that LLMs can outperform fine-tuned models when they are guided by a deep understanding of the data and well-crafted prompts that capture such nuances.

## Conclusion

This paper presented an overview of the #SMM4H-HeaRD 2025 shared tasks, providing insights into state-of-the-art methods for mining social media and other web-based data sources for health research. In general, 18 of the 29 participating teams used LLMs in their approaches, including the top-performing teams for Task 1, Task 3, Task 4 (all 3 subtasks), and Task 5 (both subtasks), outperforming teams that used BERT-based models. In particular, 5 teams fine-tuned LLMs, 10 teams prompted LLMs for classification or extraction, and 6 teams used LLMs for data augmentation. To facilitate future work, the datasets remain available by request, and the CodaLab sites remain active to automatically evaluate new systems against the blind test sets, promoting the ongoing systematic comparison of performance.

## Acknowledgments

The work for Task 1 was supported by the trilateral ANR-DFG-JST AI Research project KEEPHA (ANR-20-IADJ-0005-01 and DFG-442445488), the German Federal Ministry of Education and Research (BIFOLD25B), the National Library of Medicine (R01LM011176), and the Russian Science Foundation (23-11-00358). The work for Task 2 was supported by the National Institute on Drug Abuse (R01DA057599). The work for Task 3 was supported by the National Library of Medicine (R01LM011176). The authors thank those who contributed to annotating the data, the program committee of the #SMM4H-HeaRD 2025 Workshop, and additional peer reviewers of the system description papers.

## References

Auxier, B.; and Anderson, M. 2021. Social Media Use in 2021. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/. Accessed: 2025-04-02.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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)*, 3615–3620. Association for Computational Linguistics.

Boatemaa, S.; Barney, M.; Drimie, S.; Harper, J.; Korsten, L.; and Pereira, L. 2019. Awakening from the Listeriosis Crisis: Food Safety Challenges, Practices and Governance in the Food Retail Sector in South Africa. *Food Control*, 104: 333–342.

Bolshakova, M.; Bluthenthal, R.; and Sussman, S. 2019. Opioid Use and Misuse: Health Impact, Prevalence, Correlates and Interventions. *Psychol Health*, 34(9): 1105–1139.

Buttery, J. P.; and Clothier, H. 2022. Information Systems for Vaccine Safety Surveillance. *Hum Vaccin Immunother*, 18(6): 2100173.

Centers for Disease Control and Prevention. 2023. Artificial Intelligence and Machine Learning: Applying Advanced Tools for Public Health. https://www.cdc.gov/surveillance/data-modernization/technologies/ai-ml.html. Accessed: 2025-04-03.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16(1): 321–357.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-Lingual

Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Association for Computational Linguistics.

Dai, X.; Karimi, S.; Sarker, A.; Hachey, B.; and Paris, C. 2024. MultiADE: A Multi-domain Benchmark for Adverse Drug Event Extraction. *J Biomed Inform*, 160: 104744.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Deng, X.; den Bakker, H. C.; and Hendriksen, R. S. 2016. Genomic Epidemiology: Whole-Genome-Sequencing–Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. *Annu Rev Food Sci Technol*, 7: 353–374.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Association for Computational Linguistics.

Edwards, I. R.; and Aronson, J. K. 2000. Adverse Drug Reactions: Definitions, Diagnosis, and Management. *Lancet*, 356(9237): 1255–1259.

Feng, Y.; Hou, B.; Klein, A.; O'Connor, K.; Chen, J.; Mondragón, A.; Yang, S.; Gonzalez-Hernandez, G.; and Shen, L. 2025. Exploring Semantic Topics in Dementia Caregiver Tweets. *Alzheimers Dement*, 20(Suppl 4): e093035.

Ge, Y.; Das, S.; O'Connor, K.; Al-Garadi, M. A.; Gonzalez-Hernandez, G.; and Sarker, A. 2024. Reddit-Impacts: A Named Entity Recognition Dataset for Analyzing Clinical and Social Effects of Substance Use Derived from Social Media. arXiv:2405.06145.

Gemma Team. 2025. Gemma 3 Technical Report. arXiv:2503.19786.

GenAI, Meta. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Gottfried, J. 2024. Americans' Social Media Use. https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/. Accessed: 2025-04-02.

Guo, Y.; Rajwal, S.; Lakamana, S.; Chiang, C.-C.; Menell, P. C.; Shahid, A. H.; Chen, Y.-C.; Chhabra, N.; Chao, W.-J.; Chao, C.-J.; Schwedt, T. J.; Banerjee, I.; and Sarker, A. 2023. Generalizable Natural Language Processing Framework for Migraine Reporting from Social Media. *AMIA Jt Summits Transl Sci Proc*, 2023: 261.

Hargittai, E.; and Walejko, G. 2008. The Participation Divide: Content Creation and Sharing in the Digital Age. *Information, Communication & Society*, 11(2): 239–256.

Hazell, L.; and Shakir, S. A. W. 2006. Under-Reporting of Adverse Drug Reactions : A Systematic Review. *Drug Saf*, 29(5): 385–396.

Hopwood, J.; Walker, N.; McDonagh, L.; Rait, G.; Walters, K.; Iliffe, S.; Ross, J.; and Davies, N. 2018. Internet-Based Interventions Aimed at Supporting Family Caregivers of People with Dementia: Systematic Review. *J Med Internet Res*, 20(6): e216.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Huang, K.; Altosaar, J.; and Ranganath, R. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv:1904.05342.

Jana, S.; Sinha, M.; and Dasgupta, T. 2024. FORCE: A Benchmark Dataset for Foodborne Disease Outbreak and Recall Event Extraction from News. In *Proceedings of The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, 163–169. Association for Computational Linguistics.

Jin, C.; Bouzembrak, Y.; Zhou, J.; Liang, Q.; Van Den Bulk, L. M.; Gavai, A.; Liu, N.; van den Heuvel, L. J.; Hoenderdaal, W.; and Marvin, H. J. P. 2020. Big Data in Food Safety - A Review. *Current Opinion in Food Science*, 36: 24–32.

Johnson, A. E. W.; Pollard, T. J.; Shen, L.; Lehman, L.-W. H.; Feng, M.; Ghassemi, M. M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a Freely Accessible Critical Care Database. *Sci Data*, 3(160035).

Kartoun, U.; Aggarwal, R.; Beam, A.; Pai, J.; Chatterjee, A. K.; Fitzgerald, T. P.; Kohane, I. S.; and Shaw, S. Y. 2018. Development of an Algorithm to Identify Patients with Physician-Documented Insomnia. *Sci Rep*, 8(7862).

Kase, J. A.; Zhang, G.; and Chen, Y. 2017. Recent Foodborne Outbreaks in the United States Linked to Atypical Vehicles — Lessons Learned. *Current Opinion in Food Science*, 18: 56–63.

Khademi, S.; Palmer, C.; Dimaguila, G. L.; Javed, M.; and Buttery, J. 2024. Exploring Large Language Models for Detecting Online Vaccine Reactions. *Stud Health Technol Inform*, 318: 30–35.

Khademi Habibabadi, S.; Delir Haghighi, P.; Burstein, F.; and Buttery, J. 2022. Vaccine Adverse Event Mining of Twitter Conversations: 2-Phase Classification Study. *JMIR Med Inform*, 10(6): e34305.

Klein, A. Z.; Magge, A.; O'Connor, K.; and Gonzalez-Hernandez, G. 2022. Automatically Identifying Twitter Users for Interventions to Support Dementia Family Caregivers: Annotated Data Set and Benchmark Classification Models. *JMIR Aging*, 5(3): e39547.

Leng, M.; Zhao, Y.; Xiao, H.; Li, C.; and Wang, Z. 2020. Internet-Based Supportive Interventions for Family Caregivers of People with Dementia: Systematic Review and Meta-analysis. *J Med Internet Res*, 22(9): e19468.

Li, Y.; Wehbe, R. M.; Ahmad, F. S.; Wang, H.; and Luo, Y. 2022. Clinical-Longformer and Clinical-BigBird: Transformers for Long Clinical Sequences. arXiv:2201.11838.

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:1907.11692.

Llama Team, AI at Meta. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Lüth, S.; Boone, I.; Kleta, S.; and Al Dahouk, S. 2019. Analysis of RASFF Notifications on Food Products Contaminated with Listeria Monocytogenes Reveals Options for Improvement in the Rapid Alert System for Food and Feed. *Food Control*, 96: 479–487.

Magge, A.; Klein, A.; Miranda-Escalada, A.; Ali Al-Garadi, M.; Alimova, I.; Miftahutdinov, Z.; Farre, E.; Lima López, S.; Flores, I.; O'Connor, K.; Weissenbacher, D.; Tutubalina, E.; Sarker, A.; Banda, J.; Krallinger, M.; and Gonzalez-Hernandez, G. 2021. Overview of the Sixth Social Media Mining for Health Applications (#SMM4H) Shared Tasks at NAACL 2021. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, 21–32. Association for Computational Linguistics.

Marvin, H. J. P.; Janssen, E. M.; Bouzembrak, Y.; Hendriksen, P. J. M.; and Staats, M. 2017. Big Data in Food Safety: An Overview. *Crit Rev Food Sci Nutr*, 57(11): 2286–2295.

Nguyen, D. Q.; Vu, T.; and Tuan Nguyen, A. 2020. BERTweet: A Pre-trained Language Model for English Tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9–14. Association for Computational Linguistics.

OpenAI. 2024a. GPT-4 Technical Report. arXiv:2303.08774.

OpenAI. 2024b. GPT-4o System Card. arXiv:2410.21276.

Pádua, I.; Moreira, A.; Moreira, P.; de Vasconcelos, F. M.; and Barros, R. 2019. Impact of the Regulation (EU) 1169/2011: Allergen-Related Recalls in the Rapid Alert System for Food and Feed (RASFF) Portal. *Food Control*, 98: 389–398.

Palleria, C.; Leporini, C.; Chimirri, S.; Marrazzo, G.; Sacchetta, S.; Bruno, L.; Lista, R. M.; Staltari, O.; Scuteri, A.; Scicchitano, F.; and Russo, E. 2013. Limitations and Obstacles of the Spontaneous Adverse Drugs Reactions Reporting: Two "Challenging" Case Reports. *J Pharmacol Pharmacother*, 4(Suppl1): S66–S72.

Petrosyan, A. 2025. Worldwide Digital Population 2025. https://www.statista.com/statistics/617136/digital-population-worldwide/. Accessed: 2025-04-02.

Raithel, L.; Thomas, P.; Roller, R.; Sapina, O.; Möller, S.; and Zweigenbaum, P. 2022. Cross-Lingual Approaches for the Detection of Adverse Drug Reactions in German from a Patient's Perspective. In *Proceedings of the Language Resources and Evaluation Conference*, 3637–3649. European Language Resources Association.

Raithel, L.; Yeh, H.-S.; Yada, S.; Grouin, C.; Lavergne, T.; Névéol, A.; Paroubek, P.; Thomas, P.; Nishiyama, T.; Möller, S.; Aramaki, E.; Matsumoto, Y.; Roller, R.; and Zweigenbaum, P. 2024. A Dataset for Pharmacovigilance in German, French, and Japanese: Annotating Adverse Drug Reactions across Languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 395–414. ELRA and ICCL.

Rajwal, S.; Pandey, A. K.; Han, Z.; and Sarker, A. 2024. Unveiling Voices: Identification of Concerns in a Social Media Breast Cancer Cohort via Natural Language Processing. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, 264–270. ELRA and ICCL.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2020. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv:1910.01108.

Scallan, E.; and Mahon, B. E. 2012. Foodborne Diseases Active Surveillance Network (FoodNet) in 2012: A Foundation for Food Safety in the United States. *Clin Infect Dis*, 54(Suppl 5): S381–S384.

Strike, C.; Robinson, S.; Guta, A.; Tan, D. H.; O'Leary, B.; Cooper, C.; Upshur, R.; and Chan Carusone, S. 2020. Illicit Drug Use while Admitted to Hospital: Patient and Health Care Provider Perspectives. *PLoS One*, 15(3): e0229713.

Tao, D.; Yang, P.; and Feng, H. 2020. Utilization of Text Mining as a Big Data Analysis Tool for Food Science and Nutrition. *Compr Rev Food Sci Food Saf*, 19(2): 875–894.

Tutubalina, E.; Alimova, I.; Miftahutdinov, Z.; Sakhovskiy, A.; Malykh, V.; and Nikolenko, S. 2020. The Russian Drug Reaction Corpus and Neural Models for Drug Reactions and Effectiveness Detection in User Reviews. *Bioinformatics*, 37(2): 243–249.

Ulmer, C.; Bosworth, H.; Beckham, J.; Germain, A.; Jeffreys, A.; Edelman, D.; Macy, S.; Kirby, A.; and Voils, C. 2017. Veterans Affairs Primary Care Provider Perceptions of Insomnia Treatment. *J Clin Sleep Med*, 13(8): 991–999.

U.S. Food and Drug Administration. 2024. Real-World Evidence. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Accessed: 2025-04-03.

Vasantharajan, C.; Tun, K. Z.; Thi-Nga, H.; Jain, S.; Rong, T.; and Siong, C. E. 2022. MedBERT: A Pre-trained Language Model for Biomedical Named Entity Recognition. In *Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 1482–1488.

Volkow, N. D.; Chandler, R. K.; and Villani, J. 2022. Need for Comprehensive and Timely Data to Address the Opioid Overdose Epidemic without a Blindfold. *Addiction*, 117(8): 2132–2134.

Wagner, C.; Garcia, D.; Jadidi, M.; and Strohmaier, M. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, 454–463. Association for the Advancement of Artificial Intelligence.

Xu, D.; Garcia, G.; Raithel, L.; Thomas, P.; Roller, R.; Aramaki, E.; Wakamiya, S.; Yada, S.; Zweigenbaum, P.; O'Connor, K.; Samineni, S.; Hernandez, S.; Ge, Y.; Rajwal, S.; Das, S.; Sarker, A.; Klein, A.; Schmidt, A.; Sharma, V.; Rodriguez-Esteban, R.; Banda, J.; Amaro, I.; Weissenbacher, D.; and Gonzalez-Hernandez, G. 2024. Overview of the 9th Social Media Mining for Health Applications (#SMM4H) Shared Tasks at ACL 2024 – Large Language Models and Generalizability for Social Media NLP. In *Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and*

*Shared Tasks*, 183–195. Association for Computational Linguistics.

Yang, C. C.; Yang, H.; Jiang, L.; and Zhang, M. 2012. Social Media Mining for Drug Safety Signal Detection. In *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, 33–40. Association for Computing Machinery.

Zhou, Q.; Zhang, H.; and Wang, S. 2021. Artificial Intelligence, Big Data, and Blockchain in Food Safety. *International Journal of Food Engineering*, 18(1): 1–14.

Zhou, Z.; Alikhan, N.-F.; Mohamed, K.; Fan, Y.; Achtman, M.; Brown, D.; Chattaway, M.; Dallman, T.; Delahay, R.; Kornschober, C.; et al. 2020. The EnteroBase User's Guide, with Case Studies on Salmonella Transmissions, Yersinia Pestis Phylogeny, and Escherichia Core Genomic Diversity. *Genome Res*, 30(1): 138–152.