Team Paradise at #SMM4H-HeaRD 2026: Multi-Task Approaches for Social Media Health Mining

Dhruv Goyal[1] **Ishita Gupta**[2] **Jatin Bedi**[3]
[1]dgoyal_be23@thapar.edu [2]igupta_be23@thapar.edu [3]jatin.bedi@thapar.edu
Department of Computer Science, Thapar University
Patiala, Punjab, India

Abstract

We present Team Paradise's systems for three tasks in the SMM4H-HeaRD 2026 shared task: multilingual adverse drug event detection (Task 1), influenza vaccine effectiveness estimation via two-subtask classification (Task 3), and opioid impact span extraction (Task 7). For Task 1, threshold-only ablation on XLM-RoBERTa-large achieves macro-F1 0.597, exceeding the field mean (0.547) by +0.050. For Task 3, a three-stage hybrid pipeline combining twitter-RoBERTa-base-2022 with rule-based post-processing achieves Micro-F1 0.8434 (Subtask 1: vaccination status) and 0.8936 (Subtask 2: test results). For Task 7, RoBERTa-large with CRF decoding and sliding-window inference obtains relaxed F1 0.60 despite severe train-test distributional shift. Across tasks, we identify class imbalance, temporal ambiguity, and platform heterogeneity as central challenges.

1 Introduction

The SMM4H-HeaRD 2026 workshop (Lopez-Garcia et al., 2026) presents three complementary problems in social media health mining: detecting adverse drug reactions in multilingual posts (Task 1), estimating flu vaccine effectiveness through two-stage tweet classification (Task 3), and extracting fine-grained impact spans from opioid narratives (Task 7). These tasks share structural challenges—severe class imbalance, platform variation, temporal reasoning—yet demand distinct solutions. Our contributions are methodological rather than architectural: systematic per-language threshold calibration, a three-stage hybrid pipeline exposing failure modes of rule-based temporal reasoning, and CRF-based sliding-window inference with detailed distributional-shift analysis. Figure 1 shows our three system architectures.

2 Task 1: Multilingual ADE Detection

2.1 Task & Approach

Binary classification across six languages (de, fr, ru, en, zh, ja) plus zero-shot Farsi. Training: 47.5k posts with positive rates 2.4–60%. Test: 42.7k documents (Farsi: 15.2k, zero-shot). Challenge: severe imbalance + platform variation (tweets 55 chars vs. forums 450 chars).

We fine-tune xlm-roberta-large (Conneau et al., 2020) (559M params) with: (i) **Focal Loss** (Lin et al., 2017) (γ=2, α=0.25) for 2–60% imbalance; (ii) **language-balanced sampling** ($w_i = N_\ell/(2N_{\ell,y})$) to prevent English/Japanese (17k/14k docs) dominating German/French (1.5k/1k); (iii) **threshold-only ablation**: one trained model, three submission strategies varying only per-language decision thresholds (Figure 1a).

V1 (Manual): Histogram inspection $\rightarrow$ τ_{en}=0.16 (Twitter), τ_{zh}=0.79 (forums). **V2 (Percentile):** Match training prior. **V3 (Feedback):** CodaBench refinement: τ_{de}=0.40, τ_{fr}=0.25.

2.2 Results

Table 1 shows V3 achieves macro-F1 0.597 (field mean: 0.547, median: 0.580). Threshold tuning alone: +0.050 F1 (V1$\rightarrow$V3)—exceeding typical encoder ablations. Zero-shot Farsi: +0.041 vs. mean; Japanese: +0.075. Weakness: German/French forums (–0.04 to –0.05).

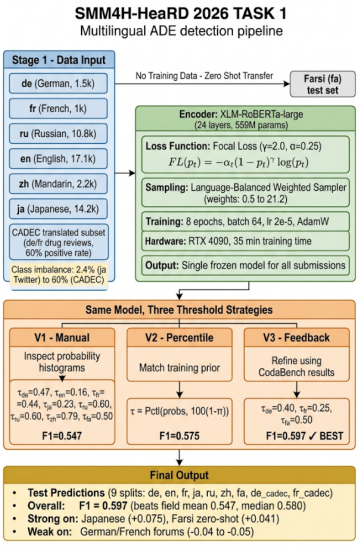
3 Task 3: Flu Vaccine Effectiveness

3.1 Task & Two-Subtask Structure

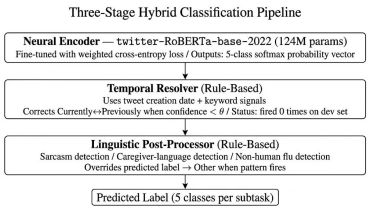
Task 3 estimates influenza vaccine effectiveness (VE) through **two independent but interlinked classification subtasks**:

Subtask 1 (ST1): Classify tweets by vaccination status into 5 classes: *Currently-Vaccinated*, *Currently-Unvaccinated*, *Previously-Vaccinated*, *Possibly-Vaccinated*, *Other*.

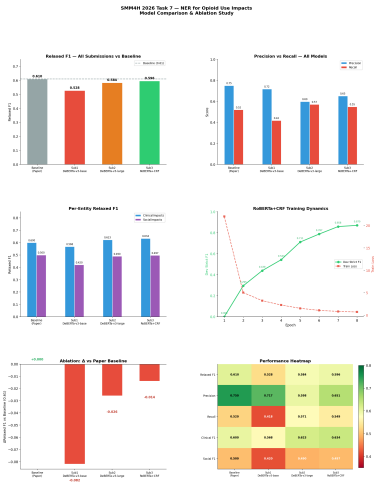
Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 305–308



(a) Task 1: Threshold Ablation



(b) Task 3: 3-Stage Hybrid



(c) Task 7: RoBERTa-large + CRF with sliding-window inference (window=512, stride=128). Overlapping windows are merged via first-window priority; FIXBIO corrects invalid BIO transitions.

Figure 1: System architectures for all three tasks.

Split	V1	V2	V3	Mean	Δ
en	.720	.728	.721	.685	+.036
de	.648	.601	.610	.664	−.054
fr	.628	.604	.634	.681	−.047
ja	.605	.597	.609	.534	+.075
ru	.544	.562	.560	.533	+.028
zh	.808	.826	.823	.804	+.018
fa	.313	.358	.408	.367	+.041
de_c	.891	.860	.857	.833	+.024
fr_c	.904	.855	.887	.843	+.044
Macro	.547	.575	**.597**	.547	**+.050**

Table 1: Task 1 results. Δ = V3 vs. field mean.

Subtask 2 (ST2): Classify tweets by flu test result into 5 classes: *Currently-Positive*, *Currently-Negative*, *Previously-Positive*, *Previously-Negative*, *Other*.

These subtasks feed into odds-ratio estimation: $\text{OR} = \frac{\text{Vac-Pos/Vac-Neg}}{\text{Unvac-Pos/Unvac-Neg}}$ for real-time VE surveillance.

Training: ST1 = 1,977 tweets, ST2 = 990 tweets (2020–2021 flu season, confounded by COVID-19). Test: ST1 = 562, ST2 = 282. Imbalance: ST2 dominated by *Other* (71.2%); *Previously-Positive* only 2.9%.

3.2 Three-Stage Hybrid Pipeline

Figure 1b shows our architecture:

Stage 1 (Neural Encoder): Fine-tune `twitter-RoBERTa-base-2022` (Loureiro et al., 2022) (154M tweets pre-training, Twitter-specific normalization) with two-layer head, class-weighted loss (6.83× for *Previously-Positive*), AdamW (lr = 2×10^{-5}, batch 16, max length 128).

Stage 2 (Temporal Resolver): Rule-based *Currently-*→Previously-** flipper when confidence < 0.65 + timestamp outside 2020–2021 season. **Outcome:** Fired *zero times*—all temporal errors at high confidence (mean 0.955). Negative result: high confidence does not guarantee correctness.

Stage 3 (Linguistic Post-Processor): Regex patterns for sarcasm (*"just joking"*), caregiver language (*"my child tested"*), non-human flu (*"bird flu"*), override to *Other*. Conservative v3 excludes over-broad tokens (*lol*).

3.3 Results

Table 2 shows ST1: 0.8434, ST2: 0.8936 Micro-F1. Post-processing v2 hurt ST1 (–0.0778) via over-broad pattern *get (a|the|your) flu shot* firing on *Currently-Vaccinated*. *Previously-** classes hardest (F1 0.588–0.592): temporal disambiguation needs encoder-level timestamp features, not post-hoc rules.

4 Task 7: Opioid Impact Span Extraction

4.1 Task & Approach

Extract *ClinicalImpacts* (overdose, depression) and *SocialImpacts* (job loss, legal charges) from Reddit

ST1: Vaccination			ST2: Test Result		
Label	**P**	**F1**	**Label**	**P**	**F1**
Other	.96	.92	Other	.96	.95
Curr-Vac	.81	.85	Curr-Pos	.74	.76
Curr-Unvac	.90	.87	Curr-Neg	.75	.70
Poss-Vac	.65	.72	Prev-Neg	.78	.86
Prev-Vac	.62	.59	Prev-Pos	.56	.59
Micro-F1		**.843**	**Micro-F1**		**.894**

Table 2: Task 3 test results for both subtasks.

posts via BIO tagging. RedditImpacts 2.0 (Dey et al., 2025): 842 train / 258 dev *sentences* (mean 20.5 tokens) vs. 578 test *posts* (mean 534 tokens, max 9,009; 159 exceed 512-token limit). Label distribution: 94.1% O tokens; B-SocialImpacts 0.51%.

RoBERTa-large + CRF: 24-layer encoder with Conditional Random Field (Lafferty et al., 2001) output for globally consistent sequences. **Class-weighted loss:** Inverse-frequency weights + multipliers (2.5× *SocialImpacts*, 1.5× *ClinicalImpacts*), capped 15×. **Sliding-window inference:** Window 512, stride 128; first-window priority; FIXBIO post-processing.

Training: 8 epochs, AdamW (lr $= 2\times10^{-5}$ encoder, 2×10^{-4} CRF), batch 16, max length 256. Combined train+dev (1,100 sentences).

4.2 Results

Table 3 shows three submissions. Sub3 (RoBERTa+CRF) achieves test relaxed F1 0.60 (strict 0.48), above task mean (0.55) and median (0.58). DeBERTa-v3-large (Sub2, 0.42) underperformed DeBERTa-v3-base (Sub1, 0.45): overfitting on 842 sentences. CRF + class weights yield largest gains. *SocialImpacts* lags (0.497 vs. 0.634 dev F1): 2× longer spans (6.2 vs. 2.4 tokens), 3× rarer.

Train-test shift: Dev relaxed F1 (0.977) vs. test (0.60) reflects sentence-to-post mismatch. Test includes metadata (submission_title, submission_subreddit) absent from training. GLiNER (Zaratiana et al., 2023) zero-shot: 0.025 F1.

System	Dev	Test	Clin	Soc
Sub1: DeBERTa-base	.528	.45	.568	.420
Sub2: DeBERTa-large	.584	.42	.623	.490
Sub3: RoBERTa+CRF	**.977**	**.60**	**.634**	**.497**

Table 3: Task 7 results. Dev/Test = relaxed F1; Clin/Soc = dev F1.

5 Cross-Task Discussion

Class imbalance: All tasks require mitigation. Task 1 (Focal Loss), Task 7 (class weights) address loss-level; Task 3 (weighted cross-entropy). Threshold calibration (Task 1) and structured decoding (Task 7 CRF) offer post-training interventions.

Temporal reasoning: Task 3's failed temporal resolver (zero dev triggers, high-confidence errors) proves timestamp-dependent distinctions cannot be retrofitted. Encoder-level temporal features necessary.

Platform heterogeneity: Task 1 spans tweets (55 chars) to forums (450 chars); Task 7 trains on sentences, tests on 9,009-token posts. Sliding-window (Task 7) and platform-aware thresholds (Task 1) partially address this, but train-test mismatch remains structural.

6 Conclusion

We presented Team Paradise's systems for SMM4H-HeaRD 2026 Tasks 1, 3, 7. Key findings: (i) Threshold calibration (Task 1, +0.050 F1) rivals architectural changes—underexplored in multilingual NLP. (ii) Rule-based post-processing (Task 3) generalizes poorly; temporal reasoning requires encoder-level features. (iii) CRF decoding (Task 7) benefits token NER, but sentence-to-document transfer demands data augmentation. Future work: timestamp-aware encoders, cross-lingual augmentation, unified sequence-labeling frameworks.

Limitations

Task 1: Single seed, no mBERT comparison. Validation contamination: V3 thresholds were selected using CodaBench leaderboard feedback on the test split, not a held-out validation set, meaning reported V3 gains may partially reflect test-set overfitting. LLM-based approaches (GPT-4, LLaMA-3) were excluded due to GPU memory constraints at 42.7k multilingual document scale. Task 3: Single architecture; post-processing from 13 dev errors. Task 7: Fixed stride; no LLM ensemble; single seed. All: Small dev sets limit generalization. Code: `https://github.com/DhruvGoyal404/SMM4H_TASK1`, `/SMM4H_TASK3`, `/SMM4H_Task7`.

Ethics Statement

All data provided by organizers under signed agreements. Twitter/Reddit data complies with platform

ToS. No PII beyond public content. Misclassification risks exist for automated health classifiers and they should supplement, not replace, clinical surveillance. Models may also embed demographic or language biases in social media datasets and it is important to audit outputs before using them in public health decisions.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Sumon Kanti Dey, Jeanne M Powell, Azra Ismail, Jeanmarie Perrone, and Abeed Sarker. 2025. Inference gap in domain expertise and machine intelligence in named entity recognition: Creation of and insights from a substance use-related dataset. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pages 12–26. World Scientific.

John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z Klein, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, and 8 others. 2026. Overview of the 11th social media mining for health (#smm4h) and health real-world data (heard) shared tasks at ACL 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Twitter-RoBERTa: A robustly optimized BERT pretraining approach for Twitter. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 5771–5782. European Language Resources Association.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. GLiNER: Generalist model for named entity recognition using bidirectional transformer. *arXiv preprint arXiv:2311.08526*.