Cuet_Data_Wizards at #SMM4H-HeaRD 2026: Multilingual ADE Detection and Influenza Vaccine Effectiveness Estimation from Social Media

Abir Dey, Mohammed Omar Faiaz, Muhammad Ibrahim Khan
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2104005,u2104029}@student.cuet.ac.bd, muhammad_ikhan@cuet.ac.bd

Abstract

We present our systems for Task 1 and Task 3 of the #SMM4H-HeaRD 2026 shared tasks. Task 1 focuses on binary classification of adverse drug event (ADE) mentions across seven languages, including a zero-shot Persian setting without labeled training data. We fine-tune XLM-RoBERTa-large using weighted cross-entropy loss and augment low-resource settings with additional CADEC data and machine-translation-based Persian augmentation. Our system achieves a macro F1 score of 0.582, outperforming the shared task average of 0.547. Task 3 addresses influenza vaccine effectiveness estimation through classification of vaccination status and flu-test results from X posts. We fine-tune `twitter-roberta-large`, achieving micro F1 scores of 0.845 for vaccination status and 0.883 for flu-test classification on the official test set. Post-evaluation experiments with focal loss, test-time augmentation, and head-tail truncation further improve performance. These results highlight the effectiveness of robust transformer adaptation for health-related social media classification.

1 Introduction

Social media provides a valuable real-time source of public health signals for pharmacovigilance, outbreak monitoring, and behavioral surveillance. Platforms such as X contain immediate self-reports of medication effects and vaccination behavior, but extracting reliable information remains challenging due to noisy informal language, class imbalance, and multilingual variation.

We present our systems for two #SMM4H-HeaRD 2026 shared tasks. Task 1 addresses multilingual ADE detection, including a zero-shot Persian setting. Task 3 targets influenza vaccine effectiveness estimation via vaccination status and flu-test classification from X posts.

Our systems emphasize robust modeling under data scarcity and imbalance. For Task 1, we employ multilingual transfer learning with synthetic Persian augmentation and threshold calibration. For Task 3, we use domain-adapted transformers together with head-tail truncation and test-time augmentation to better handle long and noisy posts.

Our main contributions are:

- We develop a cross-lingual transfer pipeline for ADE detection using machine-translation-based augmentation and adaptive threshold optimization, yielding a +0.152 F1 gain on zero-shot Persian while improving multilingual macro F1.
- We combine head-tail truncation with Monte Carlo dropout-based test-time augmentation, preserving informative long-range context and improving prediction robustness on noisy health-related social media text.

2 Related Work

ADE research has progressed from English corpora such as CADEC (Karimi et al., 2015) to multilingual settings via cross-lingual encoders like XLM-RoBERTa (Conneau et al., 2020), with transfer to German (Raithel et al., 2022), Russian (Tutubalina et al., 2021), and zero-shot settings including Persian. Social media has similarly been leveraged for infectious disease surveillance (Paul and Dredze, 2011; Santillana et al., 2015). Task 3 applies the test-negative design (Jackson and Nelson, 2013) to self-reported signals for vaccine effectiveness estimation; Xu et al. (2024) achieved F1 above 0.87 for both vaccination status and flu-test classification using LLM-based chain-of-thought prompting. Domain-specific encoders such as `twitter-roberta` (Barbieri et al., 2020) and BERTweet (Nguyen et al., 2020) consistently outperform general-purpose models on informal text. Techniques such as focal loss (Lin et al., 2017) and Monte Carlo dropout (Gal and Ghahramani, 2016) are widely used to improve robustness under class imbalance and noisy supervision.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 225–229

2.1 Task 1: Multilingual ADE Detection

Task 1 is a binary classification task for detecting adverse drug event (ADE) mentions across seven languages. Training data covers German and French (KEEPHA), Russian (RuDReC), English and Japanese (X), and Mandarin (120ask.com). Persian appears only in the test set, making it a zero-shot transfer setting. Dataset statistics are shown in Table 1.

Lang.	Train	Dev	Test
EN	17,974	902	11,712
JA	14,208	3,045	3,045
RU	10,754	2,670	9,293
ZH	2,248	379	1,144
DE	1,482	634	1,105
FR	977	419	1,104
FA	–	–	–*

Table 1: Dataset statistics for Task 1. *Persian is test-only.

2.2 Task 3: Flu Vaccine Effectiveness Estimation

Task 3 estimates influenza vaccine effectiveness (VE) from X posts using a test-negative design. It contains two subtasks: vaccination status classification and flu test result classification. Both datasets are highly imbalanced, with the "Other" category dominating. Detailed label distributions are shown in Table 2.

Subtask	Label	Train	Dev
Subtask 1: Vaccination	Other	710	97
	Currently-Unvaccinated	499	68
	Currently-Vaccinated	409	56
	Possibly-Vaccinated	228	31
	Previously-Vaccinated	131	18
	Total	**1,977**	**270**
Subtask 2: Test Result	Other	705	96
	Previously-Negative	113	16
	Currently-Negative	82	11
	Currently-Positive	61	8
	Previously-Positive	29	4
	Total	**990**	**135**

Table 2: Dataset statistics for Task 3

3 Methodology

3.1 Task 1: Multilingual ADE Detection

Task 1 is a binary classification task for detecting adverse drug event (ADE) mentions across multiple languages. Our approach focuses on multilingual transfer, imbalance handling, and threshold calibration. Figure 1 summarizes the pipeline.

Data and Preprocessing. We train on the six provided languages (EN, JA, RU, ZH, DE, FR) and merge available German and French CADEC data with the official training split. Inputs are tokenized with the `xlm-roberta-large` tokenizer using a maximum length of 128 tokens.

Model and Training. We fine-tune `XLM-RoBERTa-large`[1] using the Hugging Face Transformers library. The underlying encoder is XLM-RoBERTa-large (Conneau et al., 2020), followed by a linear classification head. To address class imbalance, we use weighted cross-entropy with inverse-frequency class weights. Training uses AdamW (Loshchilov and Hutter, 2019) (2×10^{-5}), cosine scheduling with 200 warmup steps, batch size 32, FP16 training, seed 42, and early stopping (patience=2) based on development macro F1.

Threshold Optimization. Since evaluation uses macro F1, we tune the decision threshold on the development set over $t \in [0.20, 0.69]$ with step size 0.01 and apply the best threshold at test time.

Zero-shot Persian Augmentation. Because Persian is test-only, we translate a subset of English ADE examples into Persian and add them to training data, following prior work showing the effectiveness of synthetic translated data for augmentation (Sennrich et al., 2016).

3.2 Task 3: Flu Vaccine Effectiveness Estimation

Task 3 involves classifying vaccination status and flu test outcomes from noisy X posts with severe label imbalance. Our approach focuses on domain adaptation, preserving informative context, and robust inference. Figure 2 summarizes the post-evaluation pipeline.

Model and Training. Our official evaluation system fine-tuned `twitter-roberta-large-2022`[2] using cross-entropy with label smoothing and cosine learning-rate scheduling. The encoder is based on TweetEval/Twitter-RoBERTa (Barbieri et al., 2020), which is well aligned with social media

[1] https://huggingface.co/FacebookAI/xlm-roberta-large

[2] https://huggingface.co/cardiffnlp/twitter-roberta-large-2022-154m

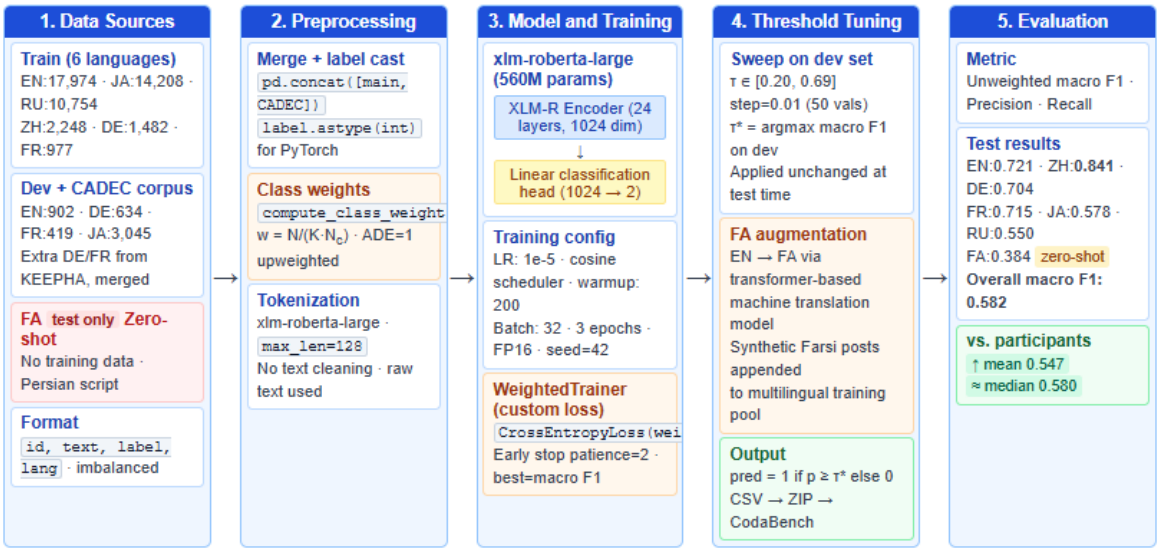

Figure 1: Overall Methodology for Task 1

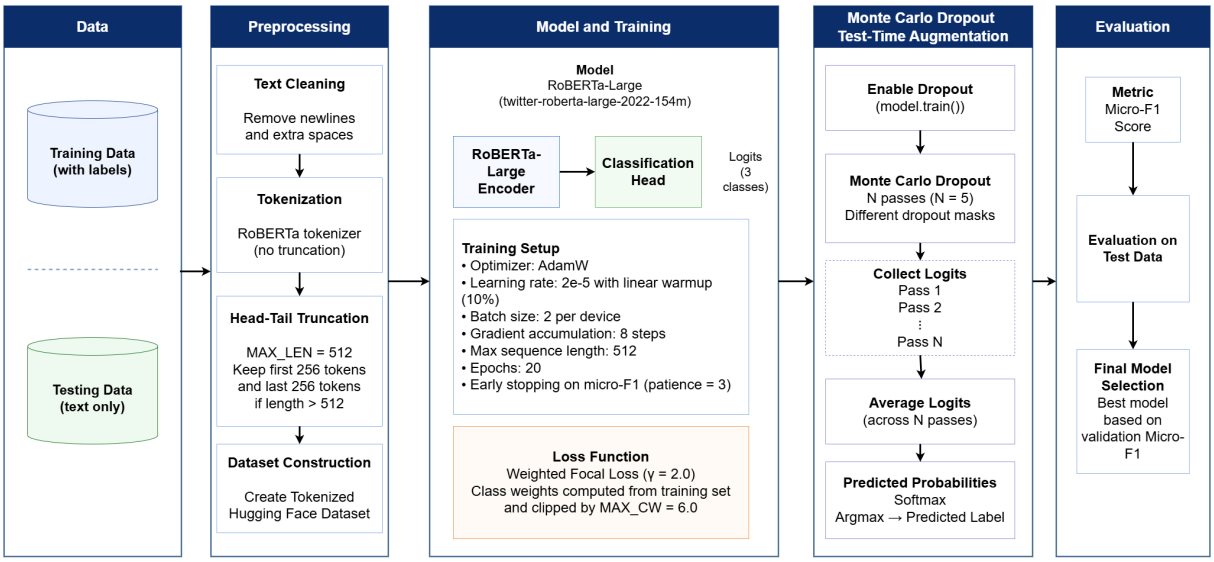

Figure 2: Improved post-evaluation methodology for Task 3

language. In post-evaluation experiments, we replaced this with weighted focal loss ($\gamma = 2.0$) with class weights capped at 6.0. Training uses AdamW ($LR = 2 \times 10^{-5}$), effective batch size 16, and early stopping based on validation micro-F1.

Context Preservation. As part of the post-evaluation system, we apply *head-tail truncation* for inputs longer than 512 tokens, retaining the first and last 256 tokens. This helps preserve diagnostic cues that often appear near the end of long posts.

Inference. In post-evaluation phase, we apply Monte Carlo dropout as test-time augmentation (TTA), keeping dropout active during inference and averaging predictions over five stochastic forward passes. This reduces prediction variance and improves robustness on noisy self-reported posts.

4 Results and Analysis

4.1 Task 1 Results

On the development set, our system achieved macro F1 scores of 0.75 (EN), 0.71 (FR), and 0.70 (RU), with an overall macro F1 of 0.6924. A formatting issue partially affected German predictions, producing an underestimated dev score of 0.50.

Table 3 reports official test set performance for our two configurations. The baseline `XLM-R` system obtained a macro F1 of 0.485, while the **Augmented XLM-R** improved this to **0.582**, exceeding the participant mean (0.547) and median (0.580).

The largest gain came from zero-shot Persian, where synthetic translated data increased F1 from 0.232 to 0.384, demonstrating the value of translation-based augmentation for unseen lan-

System	EN	DE	FR	JA	RU	ZH	FA	DE_c	FR_c	Macro F1
Baseline XLM-R	0.714	0.663	0.707	0.555	0.550	**0.846**	0.232	**0.771**	**0.755**	0.485
Augmented XLM-R	**0.721**	**0.704**	**0.715**	**0.578**	**0.550**	0.841	**0.384**	0.758	**0.755**	**0.582**
Participant Mean	0.685	0.664	0.681	0.534	0.533	0.804	0.367	0.833	0.843	0.547
Participant Median	0.701	0.656	0.696	0.549	0.550	0.821	0.380	0.860	0.883	0.580
BaseLine XLM-R (validation phase)	0.75	0.5	0.7097	*	0.69	*	*	*	*	0.69

Table 3: Task 1 official test set macro F1 scores. DE_c and FR_c denote the German and French CADEC test subsets respectively

Model	Vac F1	Cur-Vac	Cur-Unvac	Test F1	Cur-Pos	Cur-Neg
Twitter-RoBERTa-large + Cosine + LS	0.8452	0.8468	0.8930	0.8830	0.7222	0.7619
Twitter-RoBERTa-base + Cosine + LS	0.8256	0.8219	0.8686	0.8369	0.4865	0.5789
DeBERTa-v3 + Twitter-RoBERTa Ensemble	0.8310	0.8304	0.8736	0.8688	0.5946	0.6977
RoBERTa + BERTweet + Twitter-RoBERTa-Sentiment Ensemble + Voting + Focal†	0.8701	0.8571	0.9199	**0.9113**	**0.7222**	**0.8182**
Twitter-RoBERTa-large + Focal + TTA + Head-Tail (512)†	**0.8790**	**0.8772**	**0.9319**	0.9043	0.7059	0.7660
Participant Mean	0.8499	0.8565	0.8981	0.9071	0.7329	0.7833
Participant Median	0.8523	0.8621	0.9066	0.9024	0.7395	0.7712
RoBERTa-base (validation phase)	0.8037	0.8033	0.8780	0.9111	0.8000	0.8800

Table 4: Task 3 results. Rows marked † denote post-evaluation systems.

guages. Mandarin achieved the highest score (0.841), likely reflecting stronger train–test alignment.

Performance on Japanese (0.578) and Russian (0.550) was more moderate, likely due to noisier social media text and domain mismatch with the Russian training corpus. Despite slightly below-median results on German and French CADEC subsets, the system remained competitive across evaluated languages.

Overall, the results show that multilingual pretrained models provide a strong baseline, while threshold tuning and targeted augmentation are especially effective for zero-shot and low-resource settings.

4.2 Task 3 Results

During validation, our initial `roberta-base` model (Liu et al., 2019) achieved micro F1 scores of 0.8037 for Vaccination Status and 0.9111 for Flu Test Result.

Table 4 shows, in the official evaluation, our primary `twitter-roberta-large` system achieved micro F1 scores of **0.8452** for Vaccination Status and **0.8830** for Flu Test Result. While competitive, performance was limited by severe class imbalance, which reduced minority-class recall, and standard truncation, which could remove diagnostic evidence near the end of longer posts.

Post-evaluation experiments addressed these issues. Replacing cross-entropy with weighted focal loss improved learning on underrepresented labels, while head-tail truncation preserved both initial and distal context in long inputs. MC-dropout TTA further reduced prediction variance during inference.

These changes increased Vaccination Status performance to **0.8790**. For Flu Test Result, our best post-evaluation ensemble configuration reached **0.9113**, exceeding the participant mean. The largest gains appeared in minority classes such as "Currently-Negative" (0.8182) and "Currently-Positive" (0.7222), showing that imbalance-aware training and improved context retention were especially beneficial for rare clinical outcomes.

Overall, the results suggest that domain-specific pretraining provides a strong baseline, while targeted handling of imbalance, long-context posts, and noisy predictions yields further improvements.

5 Conclusion

We presented systems for Task 1 and Task 3 of #SMM4H-HeaRD 2026. For multilingual ADE detection, threshold tuning and synthetic Persian augmentation substantially improved zero-shot transfer. For flu VE estimation, focal loss, ensembling, head-tail truncation, and TTA produced strong post-evaluation performance. Our results highlight the importance of calibration, robust inference, and domain-aware modeling in health-related social media NLP.

Limitations

Task 1 remains challenging for unseen languages such as Persian, where the absence of labeled training data limits cross-lingual transfer performance. Task 3 labels may contain ambiguity and temporal drift, which can affect model generalization over time. In addition, the strongest Task 3 systems were developed after the official evaluation and should be interpreted as follow-up experiments rather than official submissions.

Ethical Considerations

This work uses publicly available social media data for health-related text classification. To reduce privacy risks, we do not attempt to identify users or infer sensitive personal information. Models may reflect dataset biases and show uneven performance across languages, especially in low-resource settings. Therefore, the system should be used only for research and population-level analysis, not for medical diagnosis or individual decision-making.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for social media nlp. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

Alexis Conneau, Kartik Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1050–1059.

Michael L. Jackson and Jennifer C. Nelson. 2013. The test-negative design for estimating influenza vaccine effectiveness. *Vaccine*, 31(17):2165–2168.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81.

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.

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. *arXiv preprint arXiv:1907.11692*. Model checkpoint: https://huggingface.co/FacebookAI/roberta-base.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14. Association for Computational Linguistics.

Michael J. Paul and Mark Dredze. 2011. You are what you tweet: Analyzing health-related topics in twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Lisa Raithel, Sebastian Möller, Joachim Köhler, and Udo Hahn. 2022. Cross-lingual approaches for the detection of adverse drug reactions in german from a patient's perspective. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3637–3649. European Language Resources Association.

Mauricio Santillana, Andre T. Nguyen, Mark Dredze, Michael J. Paul, and John S. Brownstein. 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLOS Computational Biology*, 11(10):e1004513.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Ilseyar Alimova. 2021. The russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics*, 37(2):243–249.

Dongfang Xu, Xinyu Wang, Rui Zhang, and Mike Conway. 2024. Mining social media data for influenza vaccine effectiveness using a large language model and chain-of-thought prompting. In *AMIA Annual Symposium Proceedings*, pages 1404–1413.