Creative Catalysts at #SMM4H-HeaRD 2026: XLM-RoBERTa for Task 1 – Binary Classification of Social Media Posts Containing Adverse Drug Events

Radja Afren[1], Hichem Rahab[1], and Imane Guellil[2]
[1]ICOSI Laboratory, Abbes Laghrour University of Khenchela, Algeria
[2]University of Birmingham, United Kingdom
radja.afren@univ-khenchela.dz
rahab.hichem@univ-khenchela.dz
i.guellil@bham.ac.uk

Abstract

Adverse drug events (ADEs) automatic detection from social media posts has become an important task for healthcare systems with real-world, patient-collected data. The current work deals with ADE on user generated content for Task 1 of the Social Media Mining for Health Research and Applications Workshop (SMM4H 2026), Creative Catalysts. We fine-tuned XLM-RoBERTa, pre-trained model chosen for its robustness in handling multilingual content and linguistic diversity common in social media text. To better handle the class imbalance, we subsequently implemented a class-weighting strategy to increase the model's focus on the underrepresented positive class. This adjusted model improved the validation F1-score to 65%. Our results demonstrate the effectiveness of transformer-based architectures for ADE detection while highlighting the critical need for robust class-balancing techniques and multilingual generalization to handle real-world, imbalanced social media data.

1 Introduction

Social media has become a vital source for monitoring public health, enabling real-world pharmacovigilance through user-generated health discussions (Wang and Leng, 2025; Dong et al., 2025). Detecting Adverse Drug Events (ADEs) from such content is, however, challenging due to its informal, noisy, and highly multilingual nature (Guellil et al., 2026). These challenges are directly addressed by the SMM4H-HeaRD 2026 Task 1 (Lopez-Garcia et al., 2026), which provides a unified benchmark for multilingual ADE classification (Klein et al., 2025). In this work, we present a classification system based on XLM-RoBERTa, assessing how well a single cross-lingual architecture can recognize ADEs across diverse linguistic settings.

2 Related Work

Automatic detection of ADEs from social media has advanced significantly with pre-trained language models (Guellil et al., 2026). While early methods relied on feature-engineered classifiers, transformer-based models like BERT and RoBERTa now dominate due to their contextual understanding (Gokcimen and Das, 2024), with XLM-RoBERTa emerging as the leading choice for multilingual tasks (Prytula, 2024; Saleem et al., 2025).

Recent SMM4H 2024 shared tasks highlight key challenges and solutions: LLM-based augmentation (Li et al., 2024), GPT-4 preprocessing (Mukans and Barzdins, 2024), and retrieval-augmented generation (Berkowitz et al., 2024) improved performance, though encoder-only models remain more reliable for classification due to LLM hallucinations (Zhai et al., 2024). Class imbalance was tackled via augmentation (Fan et al., 2024), two-stage classification (Kadiyala and Rao, 2024), combined loss functions (Hecht et al., 2024), and ensemble methods (Athukoralage et al., 2024; Francis and Moens, 2024). Multilingual ADE detection remains difficult, with the best NER system achieving only F1=0.489 across German, French, and Japanese (Raithel et al., 2024).

Unlike prior work addressing multilingual transfer or imbalance mitigation in isolation, our approach integrates both, fine-tuning XLM-RoBERTa with weighted cross-entropy loss, extended sequence length, and dynamic epoch discovery across six languages under extreme imbalance (93.6% negative vs. 6.4% positive).

3 Dataset

The dataset provided by the #SMM4H-HeaRD 2026 organizers comprises a comprehensive, multilingual corpus of user-generated content (UGC). Table 1 details the partition of the corpus.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 246–251

Table 1: Statistics of the SMM4H 2026 Task 1 Dataset

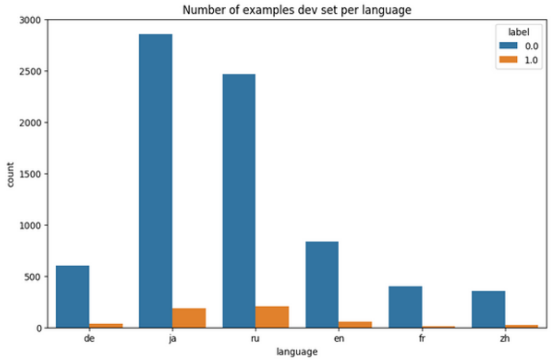
Dataset Split	Samples	Labels
Training Set	46,737	Binary (0 / 1)
Development Set	8,033	Binary (0 / 1)
Test Set	42,736	Unlabeled
Total	**97,506**	–

3.1 Label Distribution and Class Imbalance

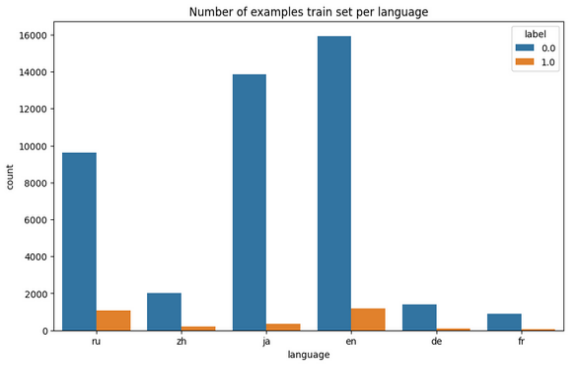
The SMM4H 2026 Task 1 corpus exhibits a stark class imbalance, which is a major technical challenge for automated pharmacovigilance. Analysis of the training data reveals that 93.6% of the samples belong to the negative class (no ADE reported), while only 6.4% are positive instances (explicit mentions of an ADE).

3.2 Multilingual Distribution

Beyond class imbalance, the corpus is extensively multilingual; Figure 1 showcases the distribution of samples across the target languages, including English (en), German (de), French (fr), Russian (ru), Japanese (ja), Chinese (zh).



(a) Number of examples dev set per language



(b) Number of examples Train set per language

Figure 1: Label Distribution by Language

4 Methodology

4.1 Preprocessing

Text was tokenized using the sentencepiece-based tokenizer associated with XLM-RoBERTa. As illustrated in the character length distribution of tweets (Figure2), we set a maximum sequence length of 128 tokens for the submission, which covered the majority of the input text without excessive padding. However, for internal optimization and performance tuning, a sequence length of 256 tokens was utilized to capture additional context in longer posts.

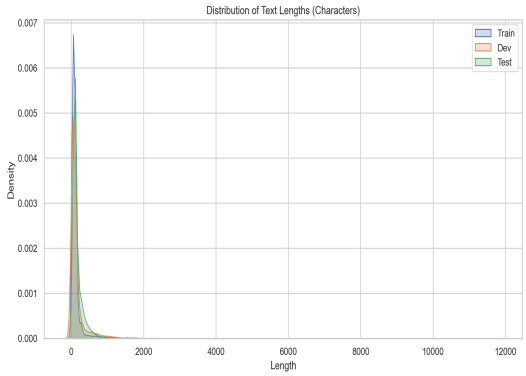


Figure 2: Distribution of Text Length.

4.2 Model Architecture

In this study, we utilized the XLM-RoBERTa-base architecture(Conneau et al., 2020). This model is a transformer-based encoder pre-trained (Gokcimen and Das, 2024) on a massive cross-lingual corpus in over 100 languages (Prytula, 2024). We selected this base variant (125 million parameters) because it provides a highly efficient balance between computational speed and high performance in multilingual sequence classification tasks.

4.3 Training Configuration

Our experimental setup progressed from a baseline configuration utilized for the original submission to an optimized configuration tailored to tackle the specific technical problems of the SMM4H dataset.

4.3.1 Baseline Configuration (Submission file)

The system used in the final shared task submission utilized a standard fine-tuning setup (See Table 2):

4.3.2 Optimized Configuration

Following the submission, we conducted further hyperparameter optimization to improve performance across several dimensions

Parameter	Value
Max Sequence Length	128 tokens
Epochs	5
Learning Rate	2×10^{-5} (with linear weight decay)
Batch Size	16
Loss Function	Standard Cross-Entropy

Table 2: Baseline configuration used for fine-tuning

- **Addressing Class Imbalance:** To mitigate the impact of the 93.6%/6.4% data skew, we implemented a weighted cross-entropy loss. This forced the model to prioritize the rare positive class by penalizing errors on class 1 eight times (8x) more heavily than standard errors.
- **Capturing Long-Form Context:** The maximum sequence length was increased to 256 tokens to capture late-sentence signals in social media threads that were previously truncated, ensuring full linguistic context for the XLM-R tokenizer.
- **Dynamic Training Strategy:** We extended the training window to a maximum of 8 epochs and employed a two-phase "Dynamic Epoch Discovery" strategy. Initially, we used early stopping on a 20% validation split to find the exact peak-performance epoch; subsequently, we retrained the model on 100% of the training data (Train + Val) for that specific epoch count to maximize information density.
- **Computational Efficiency:** Training was executed on Kaggle T4/P100 GPUs using FP16 mixed precision, allowing faster convergence while maintaining high precision for the transformer weights.

5 Results and Discussion

The efficacy of our XLM-RoBERTa system was assessed on the development set under two principal experimental setups: the baseline model utilized for the final submission and the later optimized configuration.

5.1 Official Cross-Lingual Evaluation

To enhance performance validation across the varied language landscape of the assignment, we employed the competition's official scoring script on the development set. Table 3 presents the detailed F1-scores for each target language and the overall global metric. The official evaluation confirms

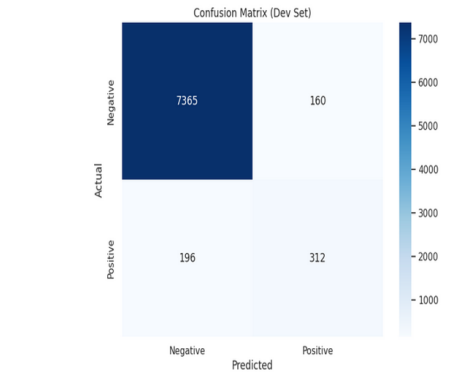
Language	Baseline	Optimized
German (de)	0.3226	0.5714
English (en)	0.8000	0.6957
Japanese (ja)	0.5405	0.4795
Chinese (zh)	0.8205	0.8235
Russian (ru)	0.6420	0.6745
French (fr)	0.5660	0.6486
Global F1	**0.6367**	**0.6546**

Table 3: Performance comparison with gains (green) and losses (red)

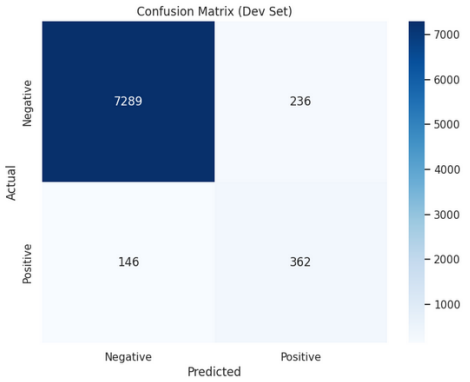
a global F1-score improvement from 0.6367 to 0.6546 in our second model(post-submission). The adjusted setup produced substantial improvements for German (+0.2488) and French (+0.0826), illustrating the efficacy of the weighted loss and context expansion in catching subtleties in non-English social media content. While English and Japanese scores saw a slight drop in this configuration, the overall cross-lingual robustness of the system was substantially improved (Table 3). To further evaluate the stability of the model in the minority ADE class, we analyzed classification reports and confusion matrices for both configurations in the development set (8,033 samples),as illustrated in Figure 3.A closer examination of the confusion matrices highlights the key trade-off introduced by the post-submission optimized configuration. While the submitted baseline model missed 196 positive instances (false negatives), the post-submission optimized model reduced this to 146, a 25.5% recall improvement, driven by the weighted loss shifting the decision boundary toward the minority class. This came at the cost of a modest increase in false positives (160 $\rightarrow$ 236), which is acceptable in pharmacovigilance where missing a genuine ADE report carries a higher practical cost than generating a spurious alert. Overall, true positives increased from 312 to 362, confirming that the post-submission optimized model captures more ADE mentions without substantially degrading specificity on the negative class.

5.2 Test Set Results

Following the blind evaluation period, our system (Team Creative Catalysts) was evaluated on the official SMM4H 2026 test set. Table 4 presents



(a) Baseline Model



(b) Optimized Model

Figure 3: Comparison of Confusion Matrices on the Development Set.

our official F1-scores across all targeted languages and subsets, alongside a comparison with the mean and median performance of all participating teams. In Japanese, our system outperformed the competition mean with a highly competitive result (0.5401), and maintained reasonable performance across the six languages seen during training (en, de, fr, ru, ja, zh). However, the global F1-score of 0.3896, substantially below the mean (0.5465) and median (0.5798), is largely explained by the presence of unseen languages in the test set. Notably, Farsi (fa) and the CADEC domain subsets (de_cadec, fr_cadec) were absent from both the training and validation data, leaving the model without any in-distribution signal for these varieties. This is reflected in our Farsi score of 0.1706, well below the mean of 0.3670, and in the CADEC scores (0.5067 and 0.6279), which fall considerably short of the median. These results highlight a fundamental limitation of supervised fine-tuning under closed-world assumptions: a model trained on a fixed set of languages cannot reliably generalize to languages or domain-specific subsets it has never encountered, regardless of the backbone's multilingual pretraining. Addressing this gap through zero-shot cross-lingual transfer or test-time language adaptation represents a promising direction for future work.

Metric	Mean F1	Median F1	Our Team (F1)
English (en)	0.6845	0.7011	0.5470
German (de)	0.6640	0.6559	0.5455
French (fr)	0.6814	0.6961	0.6022
Japanese (ja)	0.5342	0.5490	0.5401
Russian (ru)	0.5327	0.5504	0.4953
Chinese (zh)	0.8044	0.8210	0.7303
Farsi (fa)	0.3670	0.3797	0.1706
German (de_cadec)	0.8328	0.8598	0.5067
French (fr_cadec)	0.8430	0.8829	0.6279
Global F1	**0.5465**	**0.5798**	**0.3896**

Table 4: Comparison of system performance (F1-score) against mean and median benchmarks across languages

6 Conclusion

In this work, we presented a multilingual classification system for detecting ADEs, developed for the SMM4H 2026 Task 1. Leveraging the XLM-RoBERTa-base model, we addressed the twin challenges of extreme class imbalance (93.6% negative vs. 6.4% positive) and linguistic diversity across six languages. Our post-submission refined model, which included weighted cross-entropy loss, an extended sequence length (256 tokens), and a dynamic epoch discovery technique, increased the validation F1-score from 63.7% to 65.5%, whereas our official submission obtained a global F1-score of 38.96% on the blind test set. Overall, this study emphasizes the potential and ongoing difficulties of using large-scale multilingual models in practical pharmacovigilance tasks.

7 Limitation

Our model failed to generalize to unseen languages (Farsi, CADEC subsets), achieving low F1-scores. Class imbalance remains challenging, increasing false positives. Results may not transfer to other social media platforms or ADE detection tasks. We did not explore ensemble methods or larger models.

References

Dasun Athukoralage, Thushari Atapattu, Menasha Thilakaratne, and Katrina E. Falkner. 2024. LT4SG@SMM4H'24: Tweets classification for digital epidemiology of childhood health outcomes using pre-trained language models. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Jacob S. Berkowitz, Apoorva Srinivasan, Jose Miguel Acitores Cortina, and Nicholas P. Tatonetti. 2024. TLab at #SMM4H 2024: Retrieval-augmented generation for ADE extraction and normalization. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

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*.

F. Dong, W. Guo, J. Liu, T. A. Patterson, and H. Hong. 2025. Pharmacovigilance in the digital age: Gaining insight from social media data. *Experimental Biology and Medicine*, 250:10555.

Yuming Fan, Dongming Yang, and Lina Cao. 2024. CTYUN-AI@SMM4H-2024: Knowledge extension makes expert models. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Sumam Francis and Marie-Francine Moens. 2024. KUL@SMM4H2024: Optimizing text classification with quality-assured augmentation strategies. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

T. Gokcimen and B. Das. 2024. Comparison of pre-trained models for optimized transformer based question answering system. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–5. IEEE.

I. Guellil, Y. Berrachedi, N. E. Chenni, M. N. Abboud, J. Wu, H. Wu, and B. Alex. 2026. Detecting adverse drug events in social media: A brief literature review. *SN Computer Science*, 7(2):199.

Leon Hecht, Victor Martinez Pozos, Helena Gomez Adorno, Gibran Fuentes-Pineda, Gerardo Sierra, and Gemma Bel Enguix. 2024. PCIC at SMM4H 2024: Enhancing reddit post classification on social anxiety using transformer models and advanced loss functions. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Ram Mohan Rao Kadiyala and M.V.P. Chandra Sekhara Rao. 2024. 1024m at SMM4H 2024: Tasks 3, 5 & 6 - self reported health text classification through ensembles. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

A. Z. Klein, T. Dasgupta, I. Flores Amaro, L. Gryboski, S. Jana, S. Khademi, and G. Gonzalez-Hernandez. 2025. Overview of the 10th social media mining for health (SMM4H) and health real-world data (HeaRD) shared tasks at ICWSM 2025. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press.

Hongyu Li, Yuming Zhang, Yongwei Zhang, Shanshan Jiang, and Bin Dong. 2024. SRCB at #SMM4H 2024: Making full use of LLM-based data augmentation in adverse drug event extraction and normalization. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

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, Farnoush Zeidi Kolehparcheh, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Amirali Rezaie Mianroodi, Roland Roller, Judith Rosell, and 10 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.

Eduards Mukans and Guntis Barzdins. 2024. RIGA at SMM4H-2024 task 1: Enhancing ADE discovery with GPT-4. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

M. Prytula. 2024. Fine-tuning bert, distilbert, xlm-roberta and ukr-roberta models for sentiment analysis of ukrainian language reviews. *Machine Learning*, 3(4).

Lisa Raithel, Philippe Thomas, Bhuvanesh Verma, Roland Roller, Hui-Syuan Yeh, Shuntaro Yada, Cyril Grouin, Shoko Wakamiya, Eiji Aramaki, Sebastian Möller, and Pierre Zweigenbaum. 2024. Overview of #SMM4H 2024 – task 2: Cross-lingual few-shot relation extraction for pharmacovigilance in french, german, and japanese. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

H. Saleem, M. Javed, and J. Khan. 2025. Hate speech identification in formal and informal social media text using RoBERTa-base and XLM-RoBERTa-base models. In *BRAIN: Broad Research in Artificial Intelligence & Neuroscience*, volume 16.

X. Wang and X. Leng. 2025. Dialogue pathways and narrative analysis in health communication within the social media environment: An empirical study based on user behavior—a case study of china. *Frontiers in Public Health*, 13:1649120.

Yu Zhai, Xiaoyi Bao, Emmanuele Chersoni, Beatrice Portelli, Sophia Yat Mei Lee, Jinghang Gu, and Chu-Ren Huang. 2024. PolyUCBS at SMM4H 2024: LLM-based medical disorder and adverse drug event detection with low-rank adaptation. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.