

# HSE NLP Team at #SMM4H-HeaRD 2025: Hybrid LLM and Multilingual BERT Ensemble for Adverse Drug Event Detection

Airat Valiev\*

HSE University  
Moscow, Russia  
aa.valiev@hse.ru

## Abstract

We present a hybrid system for multilingual adverse drug event (ADE) detection in social media, developed for the #SMM4H-HeaRD 2025 shared task. Our approach combines large language models (LLMs) with domain-adapted BERT ensembles, addressing the challenges of extreme class imbalance and linguistic diversity across German, French, Russian, and English user-generated texts. To improve recall in low-resource languages, we generated synthetic ADE-positive samples using LLM-based data augmentation informed by biomedical NER and UMLS knowledge. Our pipeline dynamically integrates medication-specific few-shot prompts, language-specific BERT checkpoints, and ensemble decision strategies, including the use of BERT expert predictions as hints for LLMs. On the official test set, our best submission—GPT-4o few-shot with EuroBERT ensemble hints—achieved an F1-score of 0.57 for the positive class, outperforming most baseline and ensemble configurations. These results demonstrate that fusing LLM reasoning, biomedical entity linking, and targeted augmentation can substantially improve ADE detection in multilingual, imbalanced social media datasets.

## Introduction

The detection of Adverse Drug Events (ADEs) in user-generated content has become a vital task for digital pharmacovigilance, enabling early identification of drug safety signals from real-world discussions. Social media platforms and online patient forums provide a rich, yet challenging, source of such data, characterized by informal language, diverse linguistic expressions, and a high degree of ambiguity in symptom and medication mentions. Recent work has demonstrated the effectiveness of cross-lingual and multimodal BERT-based models, as well as the integration of drug structure embeddings, for ADE detection in social media across English, Russian, and French, achieving state-of-the-art results in recent SMM4H shared tasks (Miftahutdinov, Sakhovskiy, and Tutubalina 2020; Sakhovskiy, Miftahutdinov, and Tutubalina 2021; Sakhovskiy and Tutubalina 2022). The Social Media Mining for Health/Health Real-World Data (#SMM4H-HeaRD) 2025 shared task

(Task 1) specifically targets the problem of ADE detection across four major languages—German, French, Russian, and English—reflecting the growing need for robust, multilingual natural language processing (NLP) solutions in global health monitoring (Klein et al. 2025).

## Task and Evaluation

Task 1 is formulated as a binary classification problem: given a social media post, systems must predict whether it contains at least one mention of an ADE (label 1) or not (label 0) (Klein et al. 2025). The task is particularly challenging due to the extreme class imbalance, with only about 1% of posts being ADE-positive in most language splits. This setup closely mirrors real-world pharmacovigilance, where relevant signals are rare and often expressed in ambiguous, non-standard language (Klein et al. 2025). Despite progress, ADE detection remains challenging due to phenomena such as negation, speculation, and the frequent conflation of intended drug effects with adverse reactions in user narratives (Klein et al. 2025).

## Data

The multilingual dataset for the task includes user-generated social media texts in four languages: German, French, Russian, and English. The German (Thomas et al. 2022, 2024) and French (Klein et al. 2025) data are sourced from the "lifeline.de" forum, and the French texts are translations of the German posts, distinct from the original German subset. For Russian and English, Twitter posts from the past SMM4H editions (Magge et al. 2021; Klein et al. 2025) are provided. Additionally, the Russian dataset also includes user reviews from the RuDReC corpus (Tutubalina et al. 2020). The training set comprises 17,974 English, 10,754 Russian, 1,482 German, and 977 French documents, while the development set contains 2,670 Russian, 902 English, 634 German, and 419 French documents. In both splits, the data is highly imbalanced, with the positive class (ADE present) constituting less than 10% of all samples.

## System and Approaches

This section details the architectures, strategies, and experimental results of our submissions for Task 1. Our methodology was shaped by the challenges of extreme class imbalance, the scarcity of positive ADE examples in German and

\*Corresponding author.

Table 1: Positive class precision (P), recall (R), and F1-score (F1) for German, French, Russian, and English subsets of the SMM4H 2025 Task 1. We report the results of few-shot GPT-4o prompting *FS* and GPT-4o+EuroBERT ensembling (*ens*). For the GPT-4o+EuroBERT set-up, we experiment with OR/majority ensemble aggregation.

Approach	German			French			Russian			English			Average		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT-4o + EuroBERT <i>ens</i> .	0.50	0.81	0.62	0.40	0.62	0.48	0.56	0.90	0.69	0.56	0.85	0.68	0.46	0.73	<b>0.57</b>
GPT-4o FS + NER (3)	0.43	0.89	0.58	0.39	0.73	0.51	0.55	0.86	0.67	0.52	0.93	0.67	0.42	<b>0.82</b>	0.56
GPT-4o FS + EuroBERT hint	0.47	0.85	0.61	0.29	0.73	0.42	0.54	0.92	0.68	0.40	0.93	0.56	0.38	0.81	0.51
OR/majority Meta- <i>ens</i> .	0.82	0.22	0.35	0.41	0.38	0.40	0.80	0.19	0.31	0.67	0.50	0.58	<b>0.52</b>	0.31	0.39
Baseline EuroBERT (1)	0.74	0.19	0.31	0.35	0.36	0.35	0.57	0.15	0.24	0.55	0.34	0.42	0.44	0.27	0.34

French, and the heterogeneity of language and platform. We explored several approaches, ranging from baseline multilingual models to complex hybrid systems integrating LLMs and BERT ensembles.

**Data Augmentation** To overcome a high class imbalance, we generate synthetic positive ADE samples for French and German by augmenting the original texts with LLM (see Appendix ). Prompts were constructed using medication and side effect information identified by NER in the English part.

**BERT-based Approaches** As an initial baseline, we adopt EuroBERT (Boizard et al. 2025), a recent multilingual encoder, optimized for multiple natural language tasks across 15 European languages, including English, German, French and Russian, and equipped with modern architectural features: wide context window, grouped query attention and rotary position embeddings<sup>1</sup>. We explore the following fine-tuning set-ups: (i) training a single multilingual data on all four languages; (ii) training a separate model for each language on augmented data only; (iii) an ensemble of three EuroBERT checkpoints trained on either the original data or the union of the original and augmented data. The ensembling is aimed at softening the high variability of BERT predictions. The aggregation is performed with majority voting.

**Few-Shot LLM Inference** We explored few-shot inference with state-of-the-art LLMs such as DeepSeek (DeepSeek-AI 2024) and GPT-4o (OpenAI 2024), providing biomedical entities extracted with the BERN2 (Sung et al. 2022) NER tool as additional context. As few-shot examples, we selected in-language examples, prioritizing samples mentioning the same medication, and prompted the LLM with the post and extracted context.

**LLM+BERT Ensembling** Our top-performing system provided GPT-4o with few-shot examples and predictions of the language-specific EuroBERT ensemble formulated as an "expert opinion" hint. In addition to majority voting, we experimented with AND/OR aggregation.

## Results

We evaluated our approaches on the official SMM4H-HearD 2025 Task 1 test set, focusing on Precision, Recall,

and macro F1-score for the positive (ADE present) class, micro-averaged across all four languages.

Table 1 summarizes the overall performance of our main submissions, ordered by F1-score. The highest F1-score of 0.57 is achieved by a hybrid system combining GPT-4o few-shot inference with hints from a language-specific EuroBERT ensemble. The few-shot LLM pipeline with NER context (Approach 3) also performed strongly. Both significantly outperformed the baseline EuroBERT model (Approach 1).

The top-performing system achieved relatively high F1-scores in most challenging German (0.62) and French (0.48), demonstrating the effectiveness of the hybrid approach, potentially aided by synthetic data augmentation for the underlying ensemble. Performance in English was also strong (F1: 0.68). The Russian subset proved less challenging, with high recall (0.90) but low precision (0.56), resulting in a worthy F1-score (0.69).

## Discussion

Our best system achieved an overall F1-score of 0.57. Compared to the reported median F1 of all teams in a similar prior task (e.g., 0.63, though direct comparison requires caution), our system performed competitively, particularly given the dataset challenges. The largest gains across our experiments were observed in recall, which is critical for pharmacovigilance applications aiming to capture as many potential ADE signals as possible. However, precision remained a challenge, especially in Russian and low-resource corpora, where the diversity of ADE expressions and potential domain mismatch between the RuDReC training reviews and the mixed test set likely contributed to false positives.

Integrating BERT ensemble predictions as hints within LLM prompts leveraged complementary strengths, improving recall and robustness compared to individual models. LLM-generated synthetic data, guided by biomedical knowledge, proved effective for boosting performance in low-resource languages (German, French), particularly for the underlying BERT ensembles. Selecting few-shot examples based on language and medication likely improved the LLM’s contextual understanding and ability to generalize.

<sup>1</sup><https://huggingface.co/EuroBERT>

## Conclusion

Our participation in the #SMM4H-HearD 2025 Task 1 demonstrated that a hybrid architecture—combining the reasoning capabilities of LLMs (GPT-4o) with the specialized knowledge captured by language-specific BERT ensembles (EuroBERT), enhanced by biomedical entity linking (NER, UMLS) and targeted data augmentation—can effectively address the challenges of multilingual ADE detection in highly imbalanced, under-resourced social media corpora. The strategy of using ensemble predictions as hints for the LLM proved particularly effective, yielding our best results.

## References

- Boizard, N.; Gisserot-Boukhlef, H.; Alves, D. M.; Martins, A.; Hammal, A.; Corro, C.; Hudelot, C.; Malherbe, E.; Malaboeuf, E.; Jourdan, F.; Hautreux, G.; Alves, J.; El-Haddad, K.; Faysse, M.; Peyrard, M.; Guerreiro, N. M.; Fernandes, P.; Rei, R.; and Colombo, P. 2025. EuroBERT: Scaling Multilingual Encoders for European Languages. arXiv:2503.05500.
- DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. <https://huggingface.co/deepseek-ai/DeepSeek-V3-0324>. Accessed: 2025-04-20, arXiv:2412.19437.
- Klein, A. Z.; Dasgupta, T.; Flores Amaro, I.; Jana, S.; Khademi, S.; Lopez-Garcia, G.; Onishi, T.; Powell, J.; Raithel, L.; Rajwal, S.; Roller, R.; Sarker, A.; Sinha, M.; Thomas, P.; Tutubalina, E.; Xu, D.; Zweigenbaum, P.; and Gonzalez-Hernandez, G. 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.
- 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 Magge, A.; Klein, A.; Miranda-Escalada, A.; Al-garadi, M. A.; Alimova, I.; Miftahutdinov, Z.; Farre-Maduell, E.; Lopez, S. L.; Flores, I.; O'Connor, K.; Weissenbacher, D.; Tutubalina, E.; Sarker, A.; Banda, J. M.; Krallinger, M.; and Gonzalez-Hernandez, G., eds., *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, 21–32. Mexico City, Mexico: Association for Computational Linguistics.
- Miftahutdinov, Z.; Sakhovskiy, A.; and Tutubalina, E. 2020. KFU NLP Team at SMM4H 2020 Tasks: Cross-lingual Transfer Learning with Pretrained Language Models for Drug Reactions. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, 51–56. Barcelona, Spain (Online): Association for Computational Linguistics.
- OpenAI. 2024. GPT-4o System Card. <https://openai.com/index/hello-gpt-4o/>. Accessed: 2025-04-20, arXiv:2410.21276.
- Sakhovskiy, A.; Miftahutdinov, Z.; and Tutubalina, E. 2021. KFU NLP Team at SMM4H 2021 Tasks: Cross-lingual and Cross-modal BERT-based Models for Adverse Drug Effects. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, 39–43. Mexico City, Mexico: Association for Computational Linguistics.
- Sakhovskiy, A.; and Tutubalina, E. 2022. Multimodal model with text and drug embeddings for adverse drug reaction classification. *Journal of Biomedical Informatics*, 135: 104182.
- Sung, M.; Jeong, M.; Choi, Y.; Kim, D.; Lee, J.; and Kang, J. 2022. BERN2: An Advanced Neural Biomedical Named Entity Recognition and Normalization Tool. *Bioinformatics*, 38(20): 4837–4839.
- Thomas, P.; Raithel, L.; Roller, R.; Tutubalina, E.; Zweigenbaum, P.; and Klein, A. Z. 2022. The Lifeline Corpus: Annotated German Patient Forum Data for Adverse Drug Event Detection. In *Proceedings of the 2022 Conference on Language Resources and Evaluation (LREC)*.
- Thomas, P.; Raithel, L.; Roller, R.; Tutubalina, E.; Zweigenbaum, P.; and Klein, A. Z. 2024. The Lifeline Corpus: Annotated German Patient Forum Data for Adverse Drug Event Detection. In *Proceedings of the 2024 Conference on Language Resources and Evaluation (LREC)*.
- 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.

## Data Augmentation Strategy

Given the extreme class imbalance in the SMM4H-HearD 2025 dataset—with the positive (ADE) class constituting less than 10% of the samples in each language—we prioritized data augmentation for the most underrepresented languages and classes. Specifically, German and French had both the smallest training sets and the lowest absolute number of ADE-positive examples, making them the focus of our augmentation efforts.

## Selection of Languages and Classes for Augmentation

To maximize the impact of synthetic data, we first analyzed the class distributions across all languages. English and Russian had sufficiently large training sets and a relatively higher number of positive examples, while German (1,482 train docs) and French (977 train docs) had both low total and positive counts. We therefore targeted the ADE-positive (`label=1`) and ADE-negative (`label=0`) classes in German and French for augmentation, aiming to improve recall and model robustness for these low-resource settings.

## LLM-Based Synthetic Data Generation

We employed a large language model (GPT-4o) to generate realistic, language-specific ADE-positive examples. For each synthetic instance, the LLM was prompted with:

- **Task description:** A brief instruction to generate a social media post describing an adverse drug event.
- **Drug name:** Selected from the list of medications observed in the training data for the target language.
- **Possible side effects:** Extracted from UMLS or MeSH for the given drug, ensuring medical plausibility.
- **Target language:** The prompt explicitly requested output in German or French, as appropriate.

An example prompt is as follows:

*Generate a forum post in German where a patient describes experiencing one or more of the following side effects after taking [DrugName]: [SideEffect1], [SideEffect2], ... Make the post informal and realistic.*

### **Volume and Integration of Synthetic Data**

Using this approach, we generated over 10,000 synthetic ADE-positive and negative examples for German and 10,000 for French, approximately doubling the number of samples in each language. The synthetic posts were reviewed for fluency and medical plausibility, then combined with the original training data. This augmented dataset was used to fine-tune both monolingual and multilingual BERT models, as well as to provide more diverse few-shot examples for LLM prompting.