

# RIGA at SMM4H-HeaRD 2025: Context-enriched classification pipeline

Eduards Mukans, Guntis Barzdins

University of Latvia  
eduards.mukans@lu.lv, guntis.barzdins@lumii.lv

## Abstract

The following is a description of the RIGA team’s submissions for the SMM4H-HeaRD 2025 Task 1: Detection of adverse drug events (ADEs) in multilingual and multi-platform social media posts. Our approach leverages Large Language Models (LLMs) and knowledge databases to design a set of informative features that enhance the fine-tuning of a sequence classification model. Experimental results demonstrate that this method improves the F<sub>1</sub> score and leads to more balanced predictions.

**Code** — <https://github.com/emukans/smm4h2025-riga>

## Introduction

The SMM4H-HeaRD 2025 Task 1 (Klein et al. 2025) challenged participants to detect adverse drug events in multilingual and multi-platform social media posts.

Our submission focuses on leveraging the capabilities of LLMs and knowledge databases to enhance classification performance. We propose a pipeline-based approach that incorporates various external sources to construct features, which are combined with the textual input to fine-tune a deep neural network.

Additionally, we compare the performance of an off-the-shelf LLM model, which does not involve any fine-tuning.

## Related Work

This work focuses on integrating GPT-generated outputs with the original text input. A similar methodology was employed in SemEval-2023 (Mukans and Barzdins 2023) and SMM4H-2024 (Mukans and Barzdins 2024), where the tasks involved token and sequence classification. In the first study, the team used GPT as a knowledge base to retrieve information about individuals, food items, and other relevant entities mentioned in the text. In the second, GPT-generated predictions were combined with the original input.

In our submission, we utilized predictions generated by OpenAI’s GPT-4o (OpenAI et al. 2024) and information from the DrugBank database version 6 (Knox et al. 2023).

According to a recent survey (Xu et al. 2023), our approach can be categorized as a form of data augmentation.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Dataset	Train	Dev	Test
Full	31187	4625	23214
Stratification by category			
Contain ADEs	2451	398	N/A
English	17974	902	11712
Russian	10754	2670	9293
German	1482	634	1105
French	977	419	1104
Tweet	24521	2538	20805
Sentence	3810	935	0
Forum post	2459	1053	2209
Review	397	99	200

Table 1: Dataset size distribution

Source type	Letters	Words
Tweet	105	17
Sentence	77	12
Forum post	577	96
Review	753	115

Table 2: Mean text length by source.

Similar strategies have been independently explored in several studies (Amalvy, Labatut, and Dufour 2023; Chen and Feng 2023; Li et al. 2023).

## Data

This year’s dataset is multilingual and compiled from multiple sources (Magge et al. 2021; Tutubalina et al. 2020; Xu et al. 2024; Raithel et al. 2022, 2024).

As shown in Table 1, the dataset is highly imbalanced, with only around 8% of texts containing any ADEs. Most of the texts are in English and Russian, while German and French are significantly underrepresented. The primary sources for German and French texts are “forum posts” and “reviews,” which, as outlined in Table 2, are considerably longer than “tweets” and “sentences.” Consequently, handling German and French inputs requires additional preprocessing to align them with the predominant input format in the dataset.

## Methodology

To address high data variability and class imbalance, we fine-tuned six separate models on custom datasets, each constructed using different feature sets. Table 3 summarizes the results and the features used in each case.

For all submissions, we fine-tuned a monolingual RoBERTa-large model (Antypas et al. 2023; Liu et al. 2019). Although we also evaluated multilingual models such as EuroBERT (Boizard et al. 2025) and XLM-RoBERTa (Conneau et al. 2019), these performed worse on the development subset.

All experiments were conducted using four Tesla V100 16GB GPUs provided by our institution.

For generating predictions we used GPT-4o and did not evaluate models from other providers.

### 1. Text Preprocessing

The first step in every dataset creation pipeline is common across all runs. At this stage, we perform a series of unification procedures to align textual characteristics across different languages and sources. This includes special tag normalization, whitespace stripping, anonymization, emoji decoding, and similar preprocessing operations.

### 2. Translation

This step is applied only to Russian texts, as these inputs are typically short and can be fully included in the model input. The prompt used for translation is provided in Appendix A.

### 3. Translation and Summarization

This step is specific to German and French texts, which are generally much longer than most texts in the dataset. The prompt used for combined translation and summarization is provided in Appendix B.

### 4. Drug Mining and Normalization

This is a central step in our pipeline, as most features depend on the drugs identified in the input text. Initially, we extract all drug names mentioned in the original text using the "drug mining prompt" provided in Appendix C.

The output from the LLM is then validated and normalized against the DrugBank database. We employ the Jaro-Winkler distance (Jaro 1989; Winkler 1990) with a 95% similarity threshold to match GPT outputs with actual drug names or their synonyms/product names. The resulting list of normalized drug names is subsequently used for further feature extraction or as standalone features to enrich the textual input.

### 5. Feature Extraction from DrugBank

At this step, we collect relevant information from the DrugBank database for all drugs identified in the previous step. In our submission, we utilize the following fields: description, classification, names of drugs with known interactions (along with interaction descriptions), known food interactions, mechanism of action, indication, and toxicity.

GPT	Drugs	D.I. <sup>1</sup>	Food	Prec.	Rec.	F <sub>1</sub>
X	X	X	X	0.689	0.557	0.616
X	X	X	-	0.689	0.620	<b>0.653</b>
X	X	-	-	0.618	0.680	0.648
X	-	-	-	0.624	0.691	<b>0.656</b>
-	-	-	-	0.703	0.565	0.627
X*	-	-	-	0.352	0.405	0.37
Comparison				Prec.	Rec.	F <sub>1</sub>
Median				0.617	0.631	0.627
Mean				0.544	0.566	0.5394

Table 3: Submission results by features.

<sup>1</sup> Drug interactions

\* This submission relies solely on GPT output, labeling a sample as positive if the LLM returns a non-null prediction.

### 6. ADE Extraction with GPT

For this step, we use the information about identified drugs to enrich the GPT prompt context. The prompt and context are provided in Appendix D. The result of this extraction is a list of symptoms associated with the use of the identified drugs.

### 7. Has Known Drug Interaction

This feature is derived from DrugBank, based on the drugs found in the original text. If multiple drugs are mentioned, we check for known interactions between them as recorded in the database.

### 8. Has Known Food Interactions

For some drugs in DrugBank, there are documented interactions with specific foods. If any such interaction is identified, it is included as a feature in the model input during fine-tuning.

### 9. Sample Input

To illustrate the structure of our model input during fine-tuning, we provide the following sample:

```
[drug] moxifloxacin [drug]
acetaminophen [sep] has known drug
interaction [sep] has known food
interaction [gpt] Liver damage [sep]
[user] if #avelox has hurt your liver,
avoid tylenol always, as it further
damages liver, eat grapefruit unless
taking cardiac drugs
```

## Results

Incorporating GPT-generated output into the model input during fine-tuning led to a 3-point improvement in the F<sub>1</sub> score. Adding additional features alongside the GPT output the same F<sub>1</sub> score, but led to a more balanced performance in terms of precision and recall.

We hypothesize that the additional features were ineffective because a larger model is required to capture the relationships in the provided context.

## Acknowledgments

This work has been supported by the EU Recovery and Resilience Facility projects Language Technology Initiative (No 2.3.1.1.i.0/1/22/I/CFLA/002) and Latvian Quantum Initiative (No. 2.3.1.1.i.0/1/22/I/CFLA/001).

## References

- Amalvy, A.; Labatut, V.; and Dufour, R. 2023. Learning to Rank Context for Named Entity Recognition Using a Synthetic Dataset. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10372–10382. Singapore: Association for Computational Linguistics.
- Antypas, D.; Ushio, A.; Barbieri, F.; Neves, L.; Rezaee, K.; Espinosa-Anke, L.; Pei, J.; and Camacho-Collados, J. 2023. SuperTweetEval: A Challenging, Unified and Heterogeneous Benchmark for Social Media NLP Research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- Boizard, N.; Gisserot-Boukhlef, H.; Alves, D. M.; Martins, A.; Hammal, A.; Corro, C.; Hudelot, C.; Malherbe, E.; Malaboef, 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.
- Chen, F.; and Feng, Y. 2023. Chain-of-Thought Prompt Distillation for Multimodal Named Entity Recognition and Multimodal Relation Extraction. arXiv:2306.14122.
- Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *CoRR*, abs/1911.02116.
- Jaro, M. A. 1989. Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. In *Journal of the American Statistical Association*, 414–20.
- 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.
- Knox, C.; Wilson, M.; Klinger, C. M.; Franklin, M.; Oler, E.; Wilson, A.; Pon, A.; Cox, J.; Chin, N. E. L.; Strawbridge, S. A.; Garcia-Patino, M.; Kruger, R.; Sivakumaran, A.; Sanford, S.; Doshi, R.; Khetarpal, N.; Fatokun, O.; Doucet, D.; Zubkowski, A.; Rayat, D. Y.; Jackson, H.; Harford, K.; Anjum, A.; Zakir, M.; Wang, F.; Tian, S.; Lee, B.; Liigand, J.; Peters, H.; Wang, R. Q. R.; Nguyen, T.; So, D.; Sharp, M.; da Silva, R.; Gabriel, C.; Scantlebury, J.; Jasinski, M.; Ackerman, D.; Jewison, T.; Sajed, T.; Gautam, V.; and Wishart, D. S. 2023. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Research*, 52(D1): D1265–D1275.
- Li, J.; Li, H.; Pan, Z.; Sun, D.; Wang, J.; Zhang, W.; and Pan, G. 2023. Prompting ChatGPT in MNER: Enhanced Multimodal Named Entity Recognition with Auxiliary Refined Knowledge. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2787–2802. Singapore: Association for Computational Linguistics.
- 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. *CoRR*, abs/1907.11692.
- 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.
- Mukans, E.; and Barzdins, G. 2023. RIGA at SemEval-2023 Task 2: NER Enhanced with GPT-3. In Ojha, A. K.; Doğruöz, A. S.; Da San Martino, G.; Tayyar Madabushi, H.; Kumar, R.; and Sartori, E., eds., *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 331–339. Toronto, Canada: Association for Computational Linguistics.
- Mukans, E.; and Barzdins, G. 2024. RIGA at SMM4H-2024 Task 1: Enhancing ADE discovery with GPT-4. In Xu, D.; and Gonzalez-Hernandez, G., eds., *Proceedings of the 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, 23–27. Bangkok, Thailand: Association for Computational Linguistics.
- OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; and et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.
- 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 Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 3637–3649. Marseille, France: 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 Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceed-*

ings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 395–414. Torino, Italia: ELRA and ICCL.

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.

Winkler, W. E. 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proceedings of the Section on Survey Research Methods. American Statistical Association*, 354–359.

Xu, D.; Chen, W.; Peng, W.; Zhang, C.; Xu, T.; Zhao, X.; Wu, X.; Zheng, Y.; and Chen, E. 2023. Large Language Models for Generative Information Extraction: A Survey. arXiv:2312.17617.

Xu, D.; Lopez Garcia, G.; Raithel, L.; Roller, R.; Thomas, P.; Aramaki, E.; Yada, S.; Zweigenbaum, P.; Tharuni Samineni, S.; O’Connor, K.; Ge, Y.; Das, S.; Sarker, A.; Klein, A.; Schmidt, L.; Sharma, V.; Rodriguez-Esteban, R.; Banda, J.; Flores 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. In *Proceedings of The 9th Social Media Mining for Health Research and Applications Workshop and Shared Tasks*. Bangkok, Thailand: Association for Computational Linguistics.

### A. Translation prompt

Translate to English. Output just the result of the translation without any supplementary text. Keep the original semantic, orthography and punctuation.

Text for translation:  
{text}

### B. Translation and summarization prompt

Summarize and translate to English. Output just the result of the translation without any supplementary text. Keep the original semantic, orthography and punctuation. The summarization should focus on detection of adverse drug events. Irrelevant and formal information, such as greetings, closing, etc, could be omitted. Keep the named and nominal and named entities related to drugs, symptoms and drug effects. The output should be up to 5 sentences.

Text to process:  
{text}

### C. Drug mining prompt

Extract a list of drugs that a mentioned in the text. If there are multiple options for a valid drug name, then provide them in brackets as comma-separated. If no drug in the text, then output null. The output should be provided as a list where each drug is on a new line. Every line starts with a bullet point \*

Text for translation:  
{text}

### D. ADE extraction prompt

The context is different for every entry, depending on drugs that we found in the texts and the present information in the DrugBank database. The context could add the following prompt chunks:

Drug: {drug\_name}  
Description: {drug\_description}  
Classification: {drug\_classification}  
Indication: {drug\_indication}  
Mechanism of action:  
{drug\_mechanism\_of\_action}  
Toxicity: {drug\_toxicity}  
Interactions: {drug\_interaction\_description}

If the input text contains multiple drugs, then we provide a context for every drug dividing them with ---

The prompt template is the following:

You are provided with a text. List all already happened adverse drug effects caused by taking medications. Ignore symptoms that are cured by the mentioned drugs. You are also provided with additional context of drug names mentioned in the text, corresponding drug descriptions, and adverse drug effects that could cause the drug. Output just the list of adverse effects without any supplementary. The output text should be in English. Each line could contain only one adverse effect. If no adverse effects, then output null.

Context:  
{context}

Text for translation:  
{text}