Leveraging Large Language Models for Synthetic Data Generation to Enhance Adverse Drug Event Detection in Tweets

Yazdani, Anthony; Rouhizadeh, Hossein; Bornet, Alban; Teodoro, Douglas

How to cite

YAZDANI, Anthony et al. Leveraging Large Language Models for Synthetic Data Generation to Enhance Adverse Drug Event Detection in Tweets. In: Intelligent Health Systems – From Technology to Data and Knowledge. Andrikopoulou, E., Gallos, P., Arvanitis, T.N., Austin, R., Benis, A., Cornet, R., Chatzistergos, P., Dejaco, D., Dusseljee-Peute, L., Mohasseb, A., Natsiavas, P., Nakkas, H. & Scott, P. (Ed.). Glasgow. London : Sage, 2025. p. 778–782. (Studies in Health Technology and Informatics) doi: 10.3233/SHTI250465

This publication URL: https://archive-ouverte.unige.ch/unige:185634
Publication DOI: 10.3233/SHTI250465

Intelligent Health Systems – From Technology to Data and Knowledge
E. Andrikopoulou et al. (Eds.)

doi:10.3233/SHTI250465

Leveraging Large Language Models for Synthetic Data Generation to Enhance Adverse Drug Event Detection in Tweets

Anthony YAZDANI[a,1], Hossein ROUHIZADEH[a], Alban BORNET[a], and Douglas TEODORO[a]
[a] *Department of Radiology and Medical Informatics, Faculty of Medicine, University of Geneva, Geneva, Switzerland*
ORCiD ID: Anthony Yazdani https://orcid.org/0000-0003-3309-6128,
Hossein Rouhizadeh https://orcid.org/0000-0002-6496-6766,
Alban Bornet https://orcid.org/0000-0002-7266-627X,
Douglas Teodoro https://orcid.org/0000-0001-6238-4503

Abstract. Adverse drug event (ADE) detection in social media texts poses significant challenges due to the informal nature of the text and the limited availability of annotations. The scarcity of ADE named entity recognition (NER) datasets for social media hinders the development of robust ADE detection models for this type of corpus. In this paper, we leveraged the generative capabilities of large language models (LLMs) to create synthetic data, addressing this dataset gap. Specifically, we generated 17,000 tweets with ADE annotations and pre-trained NER models on this synthetic data. Our evaluations on an out-of-sample collection of 915 manually annotated tweets revealed that these models outperform state-of-the-art lexico-based and massively pre-trained open NER models. We also show that fine-tuning our synthetically pre-trained models on human-annotated data surpasses the current state-of-the-art in ADE detection on tweets. These findings suggest that synthetic data generated by LLMs can enhance ADE detection performance, offering a promising avenue to explore in response to the scarcity of annotated ADE datasets. The synthetic dataset is available at https://huggingface.co/datasets/anthonyyazdaniml/synthetic-ner-ade-tweets-v1.

Keywords. Adverse Drug Events, Large Language Models, Named Entity Recognition, Synthetic Data Generation, Zero-Shot Learning, Social Media, Tweets, MetaMap

1. Introduction

Detecting adverse drug events (ADEs) from social media platforms like X (formerly known as Twitter) is critical for modern pharmacovigilance but poses significant challenges. The informal and unstructured nature of tweets — characterized by brevity, colloquialisms, misspellings, and abbreviations — renders traditional named entity recognition (NER) methods less effective [1,2]. Additionally, the lack of annotated datasets for ADE detection further complicates model development [3,4]. To address these issues, we leverage the generative capabilities of large language models (LLMs) to

[1] Corresponding Author: Anthony Yazdani; E-mail: Anthony.Yazdani@unige.ch.

create synthetic tweets. Specifically, we trained the Llama-3.1-70B-Instruct model [5] to generate synthetic tweets with ADE annotations using soft prompting [6]. Then, we trained several NER models based on the GLiNER [7] and CONORM [2] architectures, only using our synthetic data. Remarkably, these synthetically trained models outperform the massively pre-trained open NER models and MetaMap [8]. Furthermore, when fine-tuned on a smaller set of human-annotated tweets [9], our best model surpasses the current state-of-the-art in tweet-based ADE detection. These findings highlight the potential of LLM-generated synthetic data to enhance ADE detection, offering new opportunities for advancing pharmacovigilance practices.

2. Methods

2.1. Synthetic data generation

We generated a synthetic dataset of tweets with ADE annotations using the Llama-3.1-70B-Instruct model, chosen for its competitive performance against state-of-the-art models like GPT-4 [5]. To ensure the generation of contextually relevant tweets, i.e., tweets mentioning drugs and ADEs, we used the soft prompting algorithm [6]. This algorithm learns input vectors called "soft tokens" to guide the model's output. Unlike traditional human-readable prompts, soft prompts are continuous embeddings optimized via backpropagation using task-specific data, enabling the model to produce relevant outputs in low-data regimes without requiring full fine-tuning.

Formally, let $x = [x_1, x_2, \dots, x_n]$ represent the sequence of token embeddings corresponding to an input text. Soft prompting introduces a sequence of trainable vectors $p = [p_1, p_2, \dots, p_m]$, where each p_i is a vector expressed in the same embedding space as x. The soft prompt is concatenated to the input to the Llama model, which becomes $x' = [x_1, x_2, \dots, x_n, p_1, p_2, \dots, p_m]$. During training, the soft tokens p are optimized to minimize the cross-entropy loss $\mathcal{L}(G(x'), y)$, where G represents the generative model and y is the desired annotated tweet.

To train the soft token representations, we filtered the SMM4H dataset [9] for instances where drugs could be identified, yielding 2,492 instances in total, half of which also included ADE annotations. Following soft-prompt training, we generated 8,500 positive samples — tweets containing drug names and ADEs — by randomly sampling drug-ADE pairs from the CT-ADE-PT database [10]. This approach allowed us to increase the variety of drugs and ADEs beyond those available in SMM4H. To maintain a balanced dataset, we also generated 8,500 negative samples, which are tweets containing drug names without ADEs. Figure 1 illustrates a soft prompt and its corresponding model-generated output.

To ensure the quality and relevance of the synthetic data, a self-verification step was implemented. After generating each tweet, the model was prompted to verify whether the generated tweet adhered to the provided instructions, responding with either "Yes" or "No." If the response was "Yes," the instance was saved; otherwise, it was discarded, and the generation process was restarted. This verification process further ensures the accuracy and relevance of the generated data.

(a) Input prompt

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are an AI system designed to generate synthetic tweets [...]<|eot_id|><|start_header_id|>user<|end_header_id|>

The generated tweet will mention these drugs: "Atreleuton".

The generated tweet will mention these adverse drug events: "Headache".

<|SOFT_TOKEN_0|> <|SOFT_TOKEN_1|> <|SOFT_TOKEN_2|> [...]<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

(b) Model output

```
@USER_______ i'm on <drug>atreleuton</drug> and it's working great for me. only side effect is
<ade>headaches</ade> when i don't take it at the same time every day.
```

Figure 1. (a) Illustration of the soft prompt to generate synthetic tweets following the official Llama 3 chat template with system, user, and assistant turns. The symbol "[...]" represents omitted parts of the prompt for readability. (b) Response generated after soft-prompt training. The model is able to generate a coherent tweet following the context and format specified by the input text.

2.2. NER model training

To assess the impact of synthetic data on NER for detecting ADEs in tweets, we trained and compared different models using three types of data — synthetic tweets, human-annotated tweets, and combinations of both. Specifically, we used two architectures known for their strong NER performance, GLiNER and CONORM. These architectures were chosen based on their proven effectiveness in prior studies [2,9,11,12]. For training, we used synthetic tweets (as detailed in Section 2.1) and a human-annotated dataset from the SMM4H challenge. Our approach involved building models on pre-trained backbones consistent with those used in the original implementations of GLiNER and CONORM. Specifically, we used DeBERTa-v3 [13] for GLiNER and BERTweet-large [14] for CONORM. The GLiNER models were initially pre-trained on the Pile-mistral-v0.1 dataset[2], and were further trained using either synthetic tweets, SMM4H data, or both. A similar training regimen was applied to the CONORM models.

To thoroughly evaluate the impact of model size and data type, we trained several variants of the GLiNER and CONORM architectures. GLiNER was tested in small (S), medium (M), and large (L) configurations, while CONORM was evaluated in its large (L) configuration. We used the SMM4H official validation set as our internal benchmark to compare the performance of our models, which we refer to as SyNER and SyNORM, against existing GLiNER models, the original CONORM model, and MetaMap.

3. Results

The performance of the NER models was evaluated on the SMM4H official validation set, with results summarized in Table 1. The metrics considered include Precision, Recall, and F1-score, providing a comprehensive view of the models' effectiveness in detecting ADEs in tweets. The results underscore the significant role that synthetic data plays in enhancing model performance, particularly when used in conjunction with other datasets.

[2] https://huggingface.co/datasets/urchade/pile-mistral-v0.1

Table 1. Performance of NER models, reporting Precision, Recall, and F1-score for various model configurations and training datasets. The "Pre-trained" column indicates whether the models were pre-trained on the Pile-mistral-v0.1 NER corpus. The "✓" symbol denotes the datasets used during model training.

Model name	Pre-trained	Synthetic tweets	SMM4H	Precision	Recall	F1
MetaMap				0.0584	0.3103	0.0984
GLiNER-S	✓			0.0625	0.023	0.0336
GLiNER-smm-S	✓		✓	0.6889	0.3563	0.4697
SyNER-S		✓		0.1262	0.4368	0.1959
SyNER-all-S	✓	✓	✓	0.5833	0.5632	**0.5731**
SyNER-smm-S		✓	✓	0.6364	0.1609	0.2569
GLiNER-M	✓			0.0727	0.046	0.0563
GLiNER-smm-M	✓		✓	0.6136	0.3103	0.4122
SyNER-M		✓		0.1525	0.3908	0.2194
SyNER-all-M	✓	✓	✓	0.5918	0.3333	0.4265
SyNER-smm-M		✓	✓	0.6818	0.3448	**0.4580**
GLiNER-L	✓			0.0744	0.1034	0.0865
GLiNER-smm-L	✓		✓	0.8158	0.3563	0.4960
SyNER-L		✓		0.1800	0.4138	0.2509
SyNER-all-L	✓	✓	✓	0.7708	0.4253	0.5481
SyNER-smm-L		✓	✓	0.7368	0.4828	**0.5833**
CONORM-smm-L			✓	0.6582	0.5977	0.6265
SyNORM-L		✓		0.0771	0.3333	0.1253
SyNORM-smm-L		✓	✓	0.6042	0.6667	**0.6339**

The SyNER-S model, trained solely on synthetic tweets, significantly outperformed the pre-trained GLiNER-S (F1-score: 0.1959 vs. 0.0336) and further improved to an F1-score of 0.5731 when combined with massive open NER pre-training and SMM4H fine-tuning (SyNER-all-S). For medium models, SyNER-M surpassed GLiNER-M (F1-score: 0.2194 vs. 0.0563), with SMM4H fine-tuning raising performance further (F1-score: 0.4580). Notably, both medium and large models, including SyNER-smm-M and SyNER-smm-L, achieved their best performance without benefiting from massive pre-training, suggesting that synthetic data may reduce the need for extensive pre-training.

Overall, SyNORM-smm-L, combining synthetic tweets with SMM4H fine-tuning and the CONORM architecture, delivered the highest F1-score (0.6339), setting a new state-of-the-art [2]. In contrast, MetaMap's much lower F1-score (0.0984) underscores the advantage of modern neural models enhanced with synthetic data.

4. Discussion and Conclusions

This study demonstrates the effectiveness of LLMs in generating synthetic data to enhance ADE detection in tweets. Training NER models on synthetic data led to notable performance improvements. Our findings suggest that larger pre-trained backbones, such as DeBERTa-v3-large and BERTweet-large, can effectively leverage synthetic data, potentially bypassing the need for massive open NER pre-training for downstream task performance. Furthermore, this study suggests that LLM-generated data could be a valuable tool in addressing challenges related to data scarcity in ADE detection. Future

research could explore assessing the applicability of this method to clinical notes or enhancing positive transfer from open NER pre-training. These efforts could reinforce the role of synthetic data in improving the performance of various natural language processing systems in pharmacovigilance.

References

[1] Zhou L, Zhang D, Yang CC, Wang Y. Harnessing social media for health information management. Electron Commer Res Appl 2018;27:139–51.
[2] Yazdani A, Rouhizadeh H, Bornet A, Teodoro D. CONORM: Context-Aware Entity Normalization for Adverse Drug Event Detection. medRxiv 2023:2023–09.
[3] Murphy RM, Klopotowska JE, de Keizer NF, Jager KJ, Leopold JH, Dongelmans DA, et al. Adverse drug event detection using natural language processing: A scoping review of supervised learning methods. Plos One 2023;18:e0279842.
[4] Yazdani A, Proios D, Rouhizadeh H, Teodoro D. Efficient Joint Learning for Clinical Named Entity Recognition and Relation Extraction Using Fourier Networks: A Use Case in Adverse Drug Events 2023. https://doi.org/10.48550/ARXIV.2302.04185.
[5] The Llama 3 Herd of Models | Research - AI at Meta n.d. https://ai.meta.com/research/publications/the-llama-3-herd-of-models/ (accessed July 30, 2024).
[6] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. ArXiv Prepr ArXiv210408691 2021.
[7] Zaratiana U, Tomeh N, Holat P, Charnois T. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer 2023.
[8] Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17:229–36.
[9] Klein AZ, Banda JM, Guo Y, Schmidt AL, Xu D, Flores Amaro I, et al. Overview of the 8th Social Media Mining for Health Applications (# SMM4H) shared tasks at the AMIA 2023 Annual Symposium. J Am Med Inform Assoc 2024;31:991–6.
[10] Yazdani A, Bornet A, Zhang B, Khlebnikov P, Amini P, Teodoro D. CT-ADE: An Evaluation Benchmark for Adverse Drug Event Prediction from Clinical Trial Results. ArXiv Prepr ArXiv240412827 2024.
[11] Stepanov I, Shtopko M. GLiNER multi-task: Generalist Lightweight Model for Various Information Extraction Tasks. ArXiv Prepr ArXiv240612925 2024.
[12] Yazdani A, Rouhizadeh H, Alvarez DV, Teodoro D. DS4DH at #SMM4H 2023: Zero-Shot Adverse Drug Events Normalization using Sentence Transformers and Reciprocal-Rank Fusion 2023. https://doi.org/10.48550/arXiv.2308.12877.
[13] He P, Gao J, Chen W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing 2021.
[14] Nguyen DQ, Vu T, Tuan Nguyen A. BERTweet: A pre-trained language model for English Tweets. Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. Syst. Demonstr., Online: Association for Computational Linguistics; 2020, p. 9–14. https://doi.org/10.18653/v1/2020.emnlp-demos.2.