LHS712_ADENotGood at #SMM4H 2024 Task 1: Deep-LLMADEminer: A deep learning and LLM pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter

Yifan Zheng,[1] Jun Gong,[1] Shushun Ren,[2] Dalton Simancek,[3] V.G.Vinod Vydiswaran[3]
[1]College of Pharmacy, [2]Department of Biostatistics, [3]Department of Learning Health Sciences
University of Michigan, Ann Arbor
{yifzheng, jungong, shushunr, daltonsi, vgvinodv}@umich.edu

Abstract

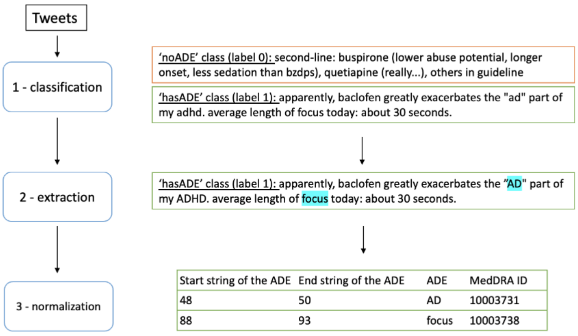
Adverse drug events (ADEs) pose major public health risks, with traditional reporting systems often failing to capture them. Our proposed pipeline, called Deep-LLMADEminer, used natural language processing approaches to tackle this issue for #SMM4H 2024 shared task 1. Using annotated tweets, we built a three part pipeline: RoBERTa for classification, GPT-4-turbo for span extraction, and BioBERT for normalization. Our models achieved F1-scores of 0.838, 0.306, and 0.354, respectively, offering a novel system for Task 1 and similar pharmacovigilance tasks.

1 Introduction

Adverse drug events (ADEs) are significant public health challenges, contributing to substantial morbidity and mortality (Watson et al., 2019). Effective pharmacovigilance, essential for ensuring the safe use of medications, struggles with the underreporting of adverse drug reactions (ADRs), with estimates indicating that over 90% of ADRs go unreported (Hazell and Shakir, 2006). Social media offers a novel avenue for real-time, patient-centered insights into ADRs, supplementing traditional data sources. We developed the **Deep-LLMADEminer** pipeline for the SMM4H-2024 Task 1 (Xu et al., 2024) to extract and normalize ADEs from English-language tweets. This study aims to assess the performance of the three-part pipeline in extracting and normalizing ADEs from tweets.

2 The Deep-LLMADEminer pipeline

In step 1, we train a classifier to detect the presence of ADEs in tweets. In step 2, we train a large language model (LLM) to extract ADE entities and their spans from tweet text. Finally, in step 3, we train a classifier to map the extracted ADE entities to formal IDs in the MedDRA ontology[1], a standardized hierarchical medical terminology (Fig. 1). The #SMM4H 2024 Task 1 dataset (Xu et al., 2024) comprised of 17,974 training tweets annotated with 1,711 ADE mentions and 959 validation tweets annotated with 87 ADE mentions. Additionally, 11,799 test tweets were provided for model evaluation.



'noADE' class (label 0): second-line: buspirone (lower abuse potential, longer onset, less sedation than bzdps), quetiapine (really...), others in guideline

'hasADE' class (label 1): apparently, baclofen greatly exacerbates the "ad" part of my adhd. average length of focus today: about 30 seconds.

'hasADE' class (label 1): apparently, baclofen greatly exacerbates the "AD" part of my ADHD. average length of focus today: about 30 seconds.

Start string of the ADE	End string of the ADE	ADE	MedDRA ID
48	50	AD	10003731
88	93	focus	10003738

Figure 1: Deep-LLMADEminer pipeline

2.1 Step 1: ADE Classification

Using the RoBERTa-base model (Liu et al., 2019), we developed a binary classification system to identify tweets containing adverse drug event (ADE) mentions. The preprocessing involved removing HTML tags, URLs, user mentions, hashtags (converted to plain text), special non-ASCII characters, punctuation, and excess whitespace, and converting all text to lowercase. We fine-tuned RoBERTa on a labeled dataset, where each tweet was tokenized and encoded into input IDs, attention masks, and token-type IDs. Key training parameters were: epochs=8, maximum sequence length=256 tokens, and batch size=16, learning rate=1e-5. The output was processed through a linear layer to classify tweets as containing or not containing ADR mentions. Our workflow for step 1 included data loading, preprocessing, training, validation with

[1] MedDRA® the Medical Dictionary for Regulatory Activities terminology is the international medical terminology developed under the auspices of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 130–132
August 15, 2024 ©2024 Association for Computational Linguistics

cross-entropy loss, and evaluation of accuracy.

2.2 Step 2: ADE Span Extraction

In step 2 of ADE span extraction, we employed the GPT-4 model via the OpenAI API (Achiam et al., 2023) to develop a text span detection method for extracting adverse drug events (ADEs) from tweets. Data preprocessing involved merging datasets that contained tweets with ADE mentions. Various prompts were experimented with to enhance the model's detection capabilities. We provided the model with 50 examples of tweets from training examples using around 10 different prompts. We explored the impacts of linguistic variations in our prompts to optimize the detection of adverse drug events from tweets. Specifically, we conducted experiments with 10 distinct prompts that varied primarily in verb usage, terminology, and the one-shot example. These variations included the use of different verbs such as "identify," "extract," and combinations of both. Additionally, we experimented with terminological changes, alternating between "adverse drug reactions" and "adverse drug effects" to assess any differences in model performance. For our experimental setup, each prompt was tested with a one-shot example tailored to illustrate the specific wording of the prompt. This approach allowed us to evaluate the model's responsiveness to linguistic nuances in a controlled manner. We did not record the results for each prompt but the most effective following prompt format is derived from testing these prompts. The final specific system message was used to guide the model: *"Identify and extract the text of adverse drug effects from the tweets in square brackets."* An example-based few-shot learning approach provided the model with specific examples to cover a wide range of ADE instances. For instance, a prompt used was: *"[I feel like a pile of crap #sick #cold #stomach reacting to some antibiotics. I will never again take #ciprofloxacin #withdrawal gives you chills],"* with the model extracting and formatting the response accordingly. The parameter `temperature` is set to 0 to minimize randomness, fostering deterministic responses from the model. We employ a `top_p` value of 0.95, which allows the model to consider a broader set of possible responses, enhancing the diversity of the output while still focusing on the most probable ones. Both `frequency_penalty` and `presence_penalty` are set to 0, indicating no additional constraints on the frequency or presence of terms in the generated text, thus not artificially influencing the model's natural language processing capabilities. These settings collectively ensure that our model interactions are precisely tailored to maximize accuracy and consistency in identifying and extracting relevant text spans for pharmacovigilance analysis. We also implemented a function to find and return the start and end indices of detected ADE texts.

2.3 Step 3: ADE Normalization

For ADE normalization, we fine-tuned BioBERT (Lee et al., 2020) to map the extracted ADE mentions from tweets to their respective MedDRA Preferred Terms (PTs), making it a multiclass classification task. The preprocessing involved converting all ADE mentions to PT ID levels, utilizing a comprehensive dictionary containing approximately 80,000 entries. This facilitated the accurate alignment of lower-level term (LLT) IDs with PT IDs. The training configurations were: epochs=8, batch size=16, and learning rate=5e-5. Our methodological pipeline comprised data loading, preprocessing to ensure consistent ID levels, and model training on processed data. Subsequent evaluation on validation data assessed the model's performance, ensuring effective normalization of tweets to corresponding PT IDs for enhanced pharmacovigilance. Other attempts to further enhance performance included fine-tuning GPT3.5 Turbo with 1,711 normalization examples. However, time constraints prevented a full evaluation.

3 Results

Models	Accuracy	F1	Precision	Recall
Validation Set (n=932)				
(1) RoBERTa	0.955	0.838	0.817	0.862
Evaluation Test Set				
(2) GPT-4	-	0.306	0.378	0.338
(3) BioBERT	-	0.354	0.395	0.321

Table 1: Performance on the unseen validation set for step 1 and the evaluation test set for steps 2 and 3.

Table 1 includes the performance metrics of RoBERTa on the unseen validation set (n=932) for step 1. We ultimately selected RoBERTa for ADE Classification with an F1-score of 0.838. It also shows the performance metrics on the evaluation dataset for steps 2 and 3. We employed models from GPT-4 for ADE span extraction, achieving an F1-score of 0.306, and utilized BioBERT for ADE normalization, which achieved an F1-score of 0.354. The effectiveness of ADE normalization in step 3 is influenced by the quality of the extracted ADE spans in the preceding task. Therefore, the relatively modest F1 score in step 2 directly impacted the overall performance in step 3.

To facilitate further development and reproducability, we have shared the implementation code for our system participation on GitHub.[2]

4 Conclusion

In our submission to the #SMM4H 2024 Task 1, we evaluated models for ADE classification, span extraction, and normalization steps using RoBERTa, GPT-4, and BioBERT, respectively. Despite some model errors in identification and resource limitations, our methods remained efficient and cost-effective.. Future work will focus on refining these models and addressing data imbalance to improve ADE detection and reporting. Specifically, for step 1, we plan to implement weighting strategies to correct dataset imbalance for more balanced and improved outcomes. For step 2, increasing the number of training examples will potentially boost model accuracy. For step 3, we aim to fine-tune hyperparameters based on validation datasets to enhance model performance.

Limitations

Our approach was not without limitations. In step 1, our methods did not address the potential issue of database imbalance; in step 2, we limited ourselves to one-shot learning to minimize costs; and in step 3, we avoided fine-tuning due to the high costs and lengthy training times associated with such processes. Additionally, we utilized the Unified Medical Language System (UMLS) to obtain synonyms for enhancing our normalization efforts in step 3. However, due to the large dataset size and limited training time, we couldn't fully leverage this approach to improve our performance. Addressing these issues and exploring these additional strategies can potentially lead to improved overall performance. Finally, the notable performance drop in Task 2 bottlenecks performance in Task 3, which challenges the utility of the overall pipeline. Further optimization should prioritize Task 2 performance for the more practical utility of the full end-to-end system.

Ethics Statement

The authors recognize that the tweets provided as part of SMM4H 2024 tasks reference health symptoms and medication use and consider these as private data that is publicly accessible under the ongoing consent provided by Twitter/X's User Agreement Terms and Conditions. The tweets for this study were sourced from the SMM4H 2024 shared task coordinators and were accessed strictly for research purposes. The data were only utilized to participate in the shared task and for no other uses.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lorna Hazell and Saad AW Shakir. 2006. Underreporting of adverse drug reactions. *Drug safety*, 29(5):385–396.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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*.

Sarah Watson, Ola Caster, Paula A Rochon, and Hester den Ruijter. 2019. Reported adverse drug reactions in women and men: aggregated evidence from globally collected individual case reports during half a century. *EClinicalMedicine*, 17:100188. DOI: 10.1016/j.eclinm.2019.10.001.

Dongfang Xu, Guillermo Lopez Garcia, Lisa Raithel, Rolland Roller, Philippe Thomas, Eiji Aramaki, Shuntaro Yada, Pierre Zweigenbaum, Sai Tharuni Samineni, Karen O'Connor, Yao Ge, Sudeshna Das, Abeed Sarker, Ari Klein, Lucia Schmidt, Vishakha Sharma, Raul Rodriguez-Esteban, Juan Banda, Ivan Flores Amaro, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 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.

[2] The software code, written in Python, is available for research use at: https://github.com/NLP4HealthUMich/DeepLLMADEminer