TLab at #SMM4H 2024: Retrieval-Augmented Generation for ADE Extraction and Normalization

Jacob Berkowitz[1], **Apoorva Srinivasan**[1],
Jose Miguel Acitores Cortina[1], **Nicholas P Tatonetti**[1]
[1]Department of Computational Biomedicine – Cedars-Sinai
jacob.berkowitz2@cshs.org, nicholas.tatonetti@cshs.org

Abstract

SMM4H 2024 Task 1 is focused on the identification and standardization of Adverse Drug Events (ADEs) in tweets. We introduce a novel Retrieval-Augmented Generation (RAG) method, leveraging the capabilities of Llama 3, GPT-4, and the SFR-embedding-mistral model, along with few-shot prompting techniques, to map colloquial tweet language to MedDRA Preferred Terms (PTs) without relying on extensive training datasets. Our method achieved competitive performance, with an F1 score of 0.359 in the normalization task and 0.392 in the named entity recognition (NER) task. Notably, our model demonstrated robustness in identifying previously unseen MedDRA PTs (F1=0.363) greatly surpassing the median task score of 0.141 for such terms.

1 Introduction

Social media is a potential wealth of information for contemporary public health monitoring, offering real-time insights into the effects of medications as reported by patients (Aichner et al., 2021). X (formerly Twitter), with its continuous flow of user-generated content, is a rich but challenging source for identifying Adverse Drug Events (ADEs), largely due to the informal and diverse language present in tweets (Klein et al., 2023).

The task of extracting standardized ADEs from such unstructured text is complex, as conventional methods typically depend on large training datasets. These not only require heavy computational resources, but also may fail to adapt to the nature of social media language (Yang et al., 2024). Here, we present a Retrieval-Augmented Generation (RAG) methodology to incorporate only relevant examples, side-stepping the need for large-scale data annotation and training processes (Gao et al., 2023).

2 Methods

2.1 ADE Tweet Classification

In the task of classifying tweets for the presence of ADEs, we used Google's BERT-large-uncased model (Devlin et al., 2019). This model is a widely recognized transformer-based neural network pretrained on a large corpus of uncased English text. Our choice of BERT-large-uncased was motivated by its proven capability in text classification tasks and its relative efficiency, making it suitable for processing the high volume of data typically found on social media platforms (Huang et al., 2022; Sakhovskiy et al., 2021).

To adapt BERT to our specific classification task, we fine-tuned the model on the provided training dataset, which included annotated tweets with a binary label showing the presence of an ADE. We fine-tuned the model over 3 epochs, a number selected to balance training time and computational resources. Future work could involve a more systematic exploration of this parameter to potentially enhance model performance.

2.2 Named Entity Recognition (NER)

NER is necessary for identifying and extracting specific text spans from the tweets that have been classified as containing at least one ADE. Our approach involves retrieving similar examples from a vector database to enhance the model's understanding and accuracy. We used the SFR-embedding-mistral model to create embeddings for the tweets in our training dataset. As of the Task 1 submission date, SFR-embedding-mistral leads the Massive Text Embedding Benchmark (MTEB) Leaderboard (Muennighoff et al., 2022). These embeddings capture the semantic similarities between tweets and are stored in a vector database. This storage allows for efficient retrieval of relevant examples.

When processing a new tweet, the same SFR-embedding-mistral model projects the tweet into

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 153–157

the embedding space. The resulting embedding is then compared against the stored embeddings in the tweet vector database using cosine similarity, defined as

$$\text{cosine similarity} = \frac{\sum_{i=1}^{n} t_{ji} * r_{ki}}{\sqrt{\sum_{i=1}^{n} t_{ji}^2}\sqrt{\sum_{i=1}^{n} r_{ki}^2}}$$

for a tweet t_j and reference tweet r_k, with an embedding dimension of n. The top 10 most similar tweets to t_j are retrieved based on these similarity scores. Known as few-shot prompting (Logan-IV et al., 2021), these similar examples provide contextual references and demonstrate the correct response format.

We pass this added information to Meta's Llama 3 7b, and ask it to extract the ADEs from t_j. The prompts fed to the model are shown in Table 1. We selected Llama 3 through consideration of the cost barrier when using GPT-4 with large quantities of tokens.

2.3 Text Normalization

Text normalization is the process of transforming text into a standardized and structured form (Aliero et al., 2023). Here, we map the extracted ADE terms to standardized MedDRA PTs. We used a RAG framework to compare the identified terms to MedDRA PTs and select the most appropriate standardized term.

To normalize an extracted ADE term, we generated its embedding using the SFR-embedding-mistral model and calculated cosine similarity scores (1) between this embedding and the stored embeddings of MedDRA PTs. We selected the top 10 most semantically similar MedDRA PTs based on these similarity scores. These similar terms provide a set of potential matches for the extracted ADE term.

With the retrieved MedDRA PTs, we used GPT-4(version=gpt-4-0125-preview, temperature=0, max new tokens=256), to process the full tweet along with the potential matches. Here, since we are retrieving individual terms rather than phrases, the cost barrier is less drastic than in the NER step. The prompts for this task are shown in Table 2. This strategy allows GPT-4 to consider both the context of the tweet and the extracted term when selecting the most appropriate MedDRA PT. By leveraging the semantic context provided by the full tweet, GPT-4 can map the colloquial language of the tweet to the standardized medical terminology of MedDRA.

3 Results

We evaluated our approach using standard metrics of precision (P), recall (R), and F1 score across three different tasks: named entity recognition (NER), normalization (Norm), and normalization on unseen data (Norm-Unseen). In Table 3, we compare our results with the mean and median scores of the task to demonstrate the effectiveness of our approach. Additionally, we include our results on the provided validation (Val) dataset.

3.1 Performance on Named Entity Recognition (NER) Task

For the NER task, our model was responsible for identifying and extracting specific ADE terms from tweets. The F1-NER score for our model was 0.392, compared to the task mean of 0.327 and median of 0.376. Our precision in the NER task was 0.437 while the recall was 0.355.

3.2 Performance on Normalization Task

In the normalization task, our approach mapped colloquial tweet language to standardized MedDRA terminology. Our method achieved an F1-Norm score of 0.359, which compares to the task mean of 0.283 and median of 0.293. The precision and recall scores of our approach were 0.400 and 0.326, respectively.

3.3 Performance on Normalization Task for Unseen Data

The ability to generalize to unseen data is important for the practical application of any model. In evaluating our approach on MedDRA PTs not seen in the training data, we achieved an F1-Norm-Unseen score of 0.363, compared to the task mean of 0.209 and median of 0.209. This demonstrates our model's robustness and ability to effectively handle the challenge of normalizing terms that were not present in the training data. The precision and recall for unseen data were 0.360 and 0.365, respectively.

4 Discussion

The results of our study highlight the effectiveness of our RAG approach in the context of the SMM4H 2024 Task 1. Our approach, which integrates the capabilities of Llama 3, GPT-4, and the SFR-embedding-mistral model, has demonstrated a particularly strong performance in handling unseen MedDRA Preferred Terms, an important as-

pect of real-world applicability given the evolving nature of language used in social media. This capability can enhance the timeliness and reliability of ADE detection from social media sources, contributing to faster and more informed public health responses.

For future use, there are certain areas of our pipeline that would benefit from experimental validation. For example, we arbitrarily selected the number 10 for both the ADE extraction and text normalization to provide our models with an adequate amount of information without overloading them. In addition to refining our pipeline, future work may aim to explore additional few-shot learning techniques and retrieval methods to enhance the model's ability to adapt to the continuously evolving language on social media. If willing to trade cost for a potential performance boost, it would be worthwhile to assess GPT-4 for the NER task. Or, with a large quantity of training data, the system may perform better with a fine-tuned local LLM (such as Llama 3) for both the NER and normalization steps.

References

T. Aichner, M. Grünfelder, O. Maurer, and D. Jegeni. 2021. Twenty-five years of social media: A review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, Behavior, and Social Networking*, 24(4):215–222.

A. Ahmad Aliero, B. Sulaimon Adebayo, H. Olanrewaju Aliyu, A. Gogo Tafida, B. Umar Kangiwa, and N. Muhammad Dankolo. 2023. Systematic review on text normalization techniques and its approach to non-standard words.

J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Y. Gao et al. 2023. Retrieval-augmented generation for large language models: A survey.

C. Huang et al. 2022. A generic pre-trained bert-based framework for social media health text classification.

A. Z. Klein et al. 2023. Overview of the 8th social media mining for health applications (#smm4h) shared tasks at the amia 2023 annual symposium.

Robert L. Logan-IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2021. Cutting down on prompts and parameters: Simple few-shot learning with language models. *CoRR*, abs/2106.13353.

N. Muennighoff, N. Tazi, L. Magne, and N. Reimers. 2022. Mteb: Massive text embedding benchmark.

A. Sakhovskiy, Z. Miftahutdinov, and E. Tutubalina. 2021. Kfu nlp team at smm4h 2021 tasks: Cross-lingual and cross-modal bert-based models for adverse drug effects.

H. Yang, Y. Zhang, J. Xu, H. Lu, P. A. Heng, and W. Lam. 2024. Unveiling the generalization power of fine-tuned large language models.

A Appendices

A.1 LLM Prompts

A.1.1 Example of Adverse Drug Event Extraction

Role	Content
System	You are a helpful assistant whose job is to, given the user tweet, determine the adverse drug events mentioned. Return a JSON with a key 'uniqueTermCount', as well as a 'terms' key for each adverse drug event including the term name and its minimal text span.
User	@USER_ that sucks. Humira made me so sick. Remicade helped a but, but then seizures put that out. Now waiting, just prednisone now.
Assistant	{"uniqueTermCount": 2, "Terms": [{"Name": "Malaise", "textSpan": "sick"}, {"Name": "Seizure", "textSpan": "seizures"}, {"Name": "Seizure", "textSpan": "seizures"}]}
User	@USER_____ @USER_____ Me too, Remicade for a year, Humira, and had an allergic reaction to Cimzia, almost killed me in November.
Assistant	{"uniqueTermCount": 2, "Terms": [{"Name": "Hypersensitivity", "textSpan": "allergic reaction"}, {"Name": "Apparent death", "textSpan": "almost killed me"}]}
User	@USER_ to it, and now I'm on Humira with just the bloat to battle. Things sometimes get super shitty in every sense. But stay strong xx
Assistant	{"uniqueTermCount": 1, "Terms": [{"Name": "Abdominal distension", "textSpan": "bloat"}]}
User	@USER_____ #humira helped my #crohns but then #respiratory issues ensued & now i can't take anything & have major lung issues
Assistant	{"uniqueTermCount": 2, "Terms": [{"Name": "Respiratory disorder", "textSpan": "Respiratory issues"}, {"Name": "Lung disorder", "textSpan": "lung issues"}]}
User	@USER_ what problems did you have with humira, if you don't mind me asking. I tried infliximab it destroyed my liver.
Assistant	{"uniqueTermCount": 1, "Terms": [{"Name": "Liver injury", "textSpan": "destroyed my liver"}]}
User	@USER_____ @USER_ i feel your pain :(just got 1st #fistula urgh ! humira worked for me for just 3months then got painful reactions
Assistant	{"uniqueTermCount": 2, "Terms": [{"Name": "Fistula", "textSpan": "fistula"}, {"Name": "Pain", "textSpan": "painful reactions"}]}
User	@USER_____ i found the humira to fix all my crohn's issues, but cause other issues. i went off it due to issues w nerves/muscle spasms

Table 1: Example input for Llama 3 in our pipeline, with the last user message being the tweet to evaluate.

A.1.2 Example of Text Normalization

Role	Content
System	Given a tweet, a highlighted span, the medical terminology, and a reference list of MedDRA PTs, identify the most relevant MedDRA PT for the medical terminology. Return only the integer ID of the MedDRA PT.
User	Tweet: @USER_____ i found the humira to fix all my crohn's issues, but cause other issues. i went off it due to issues w nerves/muscle spasms, Span: issues w nerves, Term: Neuralgia, Reference List: ['Neuralgia: 10029223', 'Facial neuralgia: 10061594', 'Morton's neuralgia: 10052288', 'Trigeminal neuralgia: 10044652', 'Trigeminal neuritis: 10074054', 'Post-traumatic neuralgia: 10076781', 'Trigeminal nerve disorder: 10060890', 'Occipital neuralgia: 10068106', 'Glossopharyngeal neuralgia: 10018391', 'Trigeminal nerve injection: 10044651']

Table 2: Example input for GPT-4 in our pipeline, with the retrieved list of MedDRA PT passed as the reference list.

A.2 Performance Metrics

Metric	Named Entity Recognition				Text Normalization			
	Val Results	Task Mean	Task Median	Team Approach	Val Results	Task Mean	Task Median	Team Approach
F1	0.620	0.327	0.376	0.392	0.623	0.283	0.293	0.359
P	0.631	0.356	0.437	0.437	0.634	0.292	0.339	0.400
R	0.609	0.340	0.374	0.355	0.612	0.334	0.326	0.326
	Unseen Text Normalization							
	Val Results	Task Mean	Task Median	Team Approach				
F1	0.200	0.209	0.141	0.363				
P	0.133	0.205	0.144	0.360				
R	0.400	0.287	0.365	0.365				

Table 3: Performance Metrics for Named Entity Recognition, Text Normalization, and Text Normalization of Unseen Terms