RIGA at SMM4H-2024 Task 1: Enhancing ADE discovery with GPT-4

Eduards Mukans
eduards.mukans@lu.lv
University of Latvia, Faculty of Computing

Guntis Barzdins
guntis.barzdins@lumii.lv
University of Latvia, IMCS

Abstract

The following is a description of the RIGA team's submissions for the SMM4H-2024 Task 1: Extraction and normalization of adverse drug events (ADEs) in English tweets. Our approach focuses on utilizing Large Language Models (LLMs) to generate data that enhances the fine-tuning of classification and Named Entity Recognition (NER) models. Our solution significantly outperforms mean and median submissions of other teams. The efficacy of our ADE extraction from tweets is comparable to the current state-of-the-art solution, established as the task baseline. The code for our method is available on GitHub[1].

1 Introduction

The SMM4H-2024 Task 1, as outlined in the overview (Xu et al., 2024), challenged participants to extract and normalize ADEs to MedDRA high-level term identifiers (HLTIs).

Our submission aims to harness the capabilities of large language models (LLMs) to enhance performance. Additionally, we compare the performance of the off-the-shelf submission, which did not involve model training, with a fine-tuned model that combines the original input with the output generated by GPT.

2 Related work

The baseline system (Magge et al., 2021) utilizes a pipeline method for solving the task. The pipeline involves 3 components and are executed sequentially: (1) the ADE classifier for identifying tweets containing ADE mentions, (2) the ADE span extractor or named entity recognition (NER) for extracting ADE mentions, and (3) the ADE normalizer, which maps the extracted ADE mention to MedDRA HLT identifiers. In our submission we utilize the same pipeline components.

Dataset	Train	Dev	Test
Full	18185	965	11799
Contain ADEs	1239	65	N/A

Table 1: Dataset size distribution

The paper concentrates on integrating the GPT model generation with the original text. A comparable methodology was employed in SemEval-2023 (Mukans and Barzdins, 2023), where the task involved token classification with highly specific tags. To streamline the process, the RIGA team utilized GPT as a knowledge database for individuals, entities, food items, and other relevant entities mentioned in the text.

According to the LLM for Generative Information Retrieval Survey (Xu et al., 2023), our method can be classified as a form of data augmentation. Similar approaches have been independently employed in several studies (Amalvy et al., 2023; Chen and Feng, 2023; Li et al., 2023)

3 Data

In contrast to the previous version of the task, the new challenge in the most recent dataset lies in the inclusion of negative samples (falling outside MedDRA categories) in each data split.

As presented in Table 1, the data is highly imbalanced. The amount of tweets containing any ADE is 6.8% for train data split and 6.7% for dev data split.

4 Methodology

In order to address the issue of high data imbalance, our pipeline includes tweet classification as the initial step to filter out the majority of the tweets. Subsequently, for the filtered tweets, we conduct NER to extract the precise spans that contain an ADEs. In the final step, we generate a sentence embedding for the span and identify the nearest

[1] https://github.com/emukans/smm4h2024-riga

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 23–27
August 15, 2024 ©2024 Association for Computational Linguistics

Submission	F1-Norm	P-Norm	R-Norm	F1-NER	P-NER	R-NER	F1-Norm-Unseen
GPT few-shot	**31.8**	**29.5**	**34.6**	40.3	37.7	43.4	21.2
Custom + GPT	10.3	12.1	9	**47.9**	**52.5**	**44.1**	6.5
Baseline	43.9	39.3	49.8	48.1	43.1	54.3	32.3
Mean	28.264	29.244	33.388	32.672	35.625	34.032	20.936
Median	29.3	33.9	32.6	37.6	43.7	37.4	14.1

Table 2: The performance of our submissions

HLTIs using cosine similarity.

We used four Tesla v100 16GB GPUs, provided by our institution, for conducting these experiments.

4.1 Tweet classification

According to Table 1, more than 93% of the tweets do not contain any ADEs. To filter out these tweets, we developed a binary classification model to identify the presence of ADEs in the input tweets. This model is based on a language model fine-tuned from RoBERTa-large (Antypas et al., 2023; Liu et al., 2019).

Before the classification model fine-tuning, all tweets are preprocessed with GPT-4 Turbo (OpenAI et al., 2024) prompt engineering (Brown et al., 2020) to extract mentioned ADEs in the text. The generative model simply needs to mention all ADEs from the provided text in a free-form manner. The prompt used in our submission is detailed in Appendix A.

The GPT output is then concatenated with the original tweet in the following format and used as input to a binary classification model:

`{tweet} <sep> {ADE extracted with GPT}`.

4.2 ADE span extraction

All categorized tweets with ADEs are forwarded to the span extraction stage. We employ a BIO-tagging schema with only three tags: B-ADE, I-ADE, and O.

In this stage, we also incorporate GPT output as additional context for downstream fine-tuning. The prompt utilized in our submission is detailed in Appendix B.

As shown in Table 2, the ADEs generated by `GPT few-shot` demonstrate strong performance in comparison to the mean and median scores. However, a notable limitation of GPT is its verbosity and propensity for hallucinations. Often, the generated spans contain verbs that contribute to coherent sentence structures but are not directly pertinent to ADEs.

Furthermore, the model may generate text that deviates from the original text. For instance, it might produce ADE expressions that do not exactly match the words in the given tweet. This issue goes beyond minor discrepancies, such as differences in American and British spelling, and highlights a broader challenge in utilizing generative models for extracting ADEs from tweets. The foundational model's training datasets, like C4, which predominantly feature texts with American dialects, contribute to this bias (Dodge et al., 2021).

To fine-tune DeBERTaV3 for span extraction (He et al., 2021), we adopt a similar input structure as in the classification step. However, since tweets may contain multiple ADEs, we separate each ADE in the input using the `sep` token.

The output generated by the fine-tuned model `Custom + GPT`, using the following input format is less noisy compared to the original GPT results.

4.3 Span mapping to MedDRA HLTIs

In total, MedDRA contains 23,389 HLTIs, but the training and development data only contain 319 unique identifiers. This indicates that the majority of HLTIs are not present in our dataset.

Training a classifier to map the spans to the HLTIs using the provided data would be futile due to the high variety of HLTIs. Additionally, the test data includes unseen categories that the trained classifier would not be able to identify.

In our submission, we utilized an off-the-shelf solution by leveraging OpenAI's Embedding API. Initially, we computed an embedding representation for all MedDRA HLTIs, followed by doing the same for each ADE span. Subsequently, we identified the closest HLTIs by calculating the cosine similarity between the embeddings.

Unfortunately, we ran out of resources and time to achieve a higher F1-Norm score for the `Custom + GPT` model. Despite using the same approach as the `GPT few-shot` model, the `Custom + GPT` model's performance on F1-Norm suffered.

5 Results

In Table 2, we compare our solutions with the current state-of-the-art solution, which serves as a baseline for the task. The competition evaluates performance using two metrics: F1-Norm and F1-NER. Our primary focus was on the F1-NER metric, where the `Custom + GPT` model demonstrates performance comparable to the baseline and significantly higher than both the mean and median. The `GPT few-shot` submission also achieved results above both the mean and median for both metrics.

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

Arthur Amalvy, Vincent Labatut, and Richard Dufour. 2023. Learning to rank context for named entity recognition using a synthetic dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10372–10382, Singapore. Association for Computational Linguistics.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Feng Chen and Yujian Feng. 2023. Chain-of-thought prompt distillation for multimodal named entity recognition and multimodal relation extraction. *Preprint*, arXiv:2306.14122.

Jesse Dodge, Ana Marasovic, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Conference on Empirical Methods in Natural Language Processing*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Jinyuan Li, Han Li, Zhuo Pan, Di Sun, Jiahao Wang, Wenkun Zhang, and Gang Pan. 2023. Prompting ChatGPT in MNER: Enhanced multimodal named entity recognition with auxiliary refined knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2787–2802, Singapore. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. DeepADEMiner: a deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug event mentions on Twitter. *Journal of the American Medical Informatics Association*, 28(10):2184–2192.

Eduards Mukans and Guntis Barzdins. 2023. RIGA at SemEval-2023 task 2: NER enhanced with GPT-3. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 331–339, Toronto, Canada. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane

Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023. Large language models for generative information extraction: A survey. *Preprint*, arXiv:2312.17617.

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.

A Classification prompt

For tweet classification we used the following prompt:

```
    You will be provided with a tweet.
Summarise it into a brief sentence and
highlight already happened adverse drug
events (ADE) if there are any related to
drugs.
Format:
Summary: {text}
ADE: {text or null}
–
Tweet:
"""
{tweet}
"""
```

The model generates two lines in the output: "Summary" and "ADE." In our submission, we utilize only the "ADE" field. The intention behind the "Summary" field was to classify summarized tweets instead of the original text, potentially simplifying the task by producing summaries in a unified language and style. Unfortunately, this hypothesis did not hold. GPT likely omits important keywords common to many ADE-containing tweets, or the semantics of the generated text do not match the original tweet. It is probable that using a "rewrite" instruction instead of "summarize" would have been more effective.

B ADE extraction prompt

For mining text spans containing ADEs we used the following prompt:

```
    You will be provided with a tweet.
Your task is to identify and highlight
any adverse drug events (ADEs) mentioned
in relation to drug use.  Only the
exact phrases describing the ADEs should
be outputted, without including any
additional context.  Each ADE should
be listed on a new line.  If the same
ADE is mentioned multiple times, each
occurrence should be listed separately.
If multiple different ADEs are identified
within the same tweet, they should be
listed on separate lines. If no ADEs are
found, output "null".
–
Format:
SPAN: {text or null}
–
Samples:
Tweet:
"""
  user
if avelox has hurt your liver, avoid
tylenol always, as it further damages
liver, eat grapefruit unless taking
cardiac drugs
"""
SPAN: hurt your liver
–
Tweet:
"""
losing it. could not remember the word
power strip. wonder which drug is doing
this memory lapse thing.  my guess the
cymbalta. helps
"""
SPAN: not remember
SPAN: memory lapse
Tweet:
"""
is adderall a performance enhancing drug
for mathletes?
"""
SPAN: null
–
Tweet:
"""
{tweet}
"""
```

Since the most of tweets will be filtered out during the classification step, and token classification is more complex task, than sequence classification, the prompt contains more instructions and output samples.