ADE Oracle at #SMM4H 2024: A Two-Stage NLP System for Extracting and Normalizing Adverse Drug Events from Tweets

Andrew S. Davis Billy Dickson Sandra Kübler
Department of Linguistics
Indiana University
{ad7,dicksonb,skuebler}@iu.edu

Abstract

This study describes the approach of Team ADE Oracle for Task 1 of the Social Media Mining for Health Applications (#SMM4H) 2024 shared task. Task 1 challenges participants to detect adverse drug events (ADEs) within English tweets and normalize these mentions against the Medical Dictionary for Regulatory Activities standards. Our approach utilized a two-stage NLP pipeline consisting of a named entity recognition model, retrained to recognize ADEs, followed by vector similarity assessment with a RoBERTa-based model. Despite achieving a relatively high recall of 37.4% in the extraction of ADEs, indicative of effective identification of potential ADEs, our model encountered challenges with precision. We found marked discrepancies between recall and precision between the test set and our validation set, which underscores the need for further efforts to prevent overfitting and enhance the model's generalization capabilities for practical applications.

1 Introduction

This paper outlines Team ADE Oracle's participation in the 9th Social Media Mining for Health Research and Applications (#SMM4H) 2024 workshop's Task 1 (Xu et al., 2024), which involved extracting and normalizing adverse drug events (ADEs) from tweets into Medical Dictionary for Regulatory Activities (MedDRA) terms[1]. The task complexity increased in 2024 by combining ADE detection with normalization, a challenge heightened by the informal and diverse language used on social media (Xu et al., 2024). Addressing ADEs through social media enhances pharmacovigilance, providing critical data for public health interventions (Huynh et al., 2016; Alimova and Tutubalina, 2019; Vydiswaran et al., 2019; Magge et al., 2021; Liu et al., 2022; Lee et al., 2023). Our approach employed a spaCy-based NLP pipeline, retraining a Named Entity Recognition (NER) module to extract ADEs, and a RoBERTa model for aligning text with MedDRA standards (Weissenbacher et al., 2022), navigating the trade-offs between recall and precision. While our system effectively identified many ADEs, the prevalence of false positives points to a need for further refinement to enhance the accuracy and utility of our methods for public health surveillance.

2 Dataset

This study employed the #SMM4H 2024 Task 1 dataset, comprising 30,949 tweets distributed across 18,185 training, 965 validation, and 11,799 test tweets (Klein et al., 2024; Xu et al., 2024).

3 System Description

Our methodology for the #SMM4H 2024 Task 1 involves a two-stage process: We use an NER package to extract ADEs, followed by the normalization of these entities against the MedDRA using vector similarity techniques (Yazdani et al., 2023a,b).

3.1 Preprocessing

For preprocessing the dataset, we implemented a two-step approach to optimize data for training. Initially, all labeled entities representing ADEs were converted to lowercase to ensure consistency and address case discrepancies between labels and their occurrences in the tweet text. Subsequently, we employed the tokenizer[2] from the blank, spaCy "en" model[3] to tokenize the text (Dai et al., 2017).

3.2 NER for ADE Extraction

We chose to use the blank spaCy model "en" for training a customized NER model tailored

[1] https://www.meddra.org

[2] https://spacy.io/api/tokenizer

[3] https://spacy.io/usage/models

Proceedings of The 9th Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Tasks, pages 117–120
August 15, 2024 ©2024 Association for Computational Linguistics

to the extraction of ADE entities due to its robust handling of English syntax and adaptability to the specialized domain of pharmacovigilance (Dai et al., 2017; Jiang et al., 2022). Specifically, we trained the model's span categorizer[4] component to identify and label ADE spans within tweets effectively. The span categorizer comprises two main components: a suggester function and a labeler model. The suggester function, employing the spacy.ngram_suggester.v1, was selected to propose candidate spans with designated lengths—specifically one to five tokens. These candidates, which may overlap, are presented in a ragged array format comprising two columns that denote the start and end positions of each span. Subsequently, the labeler model evaluates each candidate span, assigning the ADE label as appropriate based on the predictive outcomes.

This model was trained on the 18,185 labeled tweets of the official training set. Optimization was achieved over 49 epochs with a batch size of 8 and a dropout rate of 0.5, selecting the best-performing iteration for our analyses.

3.3 Vectorization and Normalization

For the normalization stage, we employed the base model RoBERTa to vectorize the ADE entities and MedDRA entries (Liu et al., 2019; Gencoglu, 2020; Weissenbacher et al., 2022). We did not further finetune the base RoBERTa model, as our focus was solely on utilizing its semantic representation capabilities. We extracted and vectorized ADE entities from the validation set using our span categorizer, and vectorized the textual descriptions of MedDRA adverse event terms. The MedDRA vectors were stored in a vector database from Facebook's Faiss library, which is designed for efficient similarity searching of dense vectors at scale (Johnson et al., 2019; Douze et al., 2024). We then iterated through our extracted entities and used Euclidean distance (L2 distance) to identify the closest match between each ADE entity vector and the MedDRA term vectors in the database.

3.4 Evaluation Metrics

The performance of our NER model in identifying ADEs, along with the pipeline's effectiveness in matching ADEs to MedDRA terms, was evaluated using the official metrics of #SMM4H 2024 Task 1, specifically F1, precision, and recall.

[4] https://spacy.io/api/spancategorizer

Task & Metric	F1	P	R
ADE Extraction	16.6	15.1	18.4
ADE Normalization	8.4	7.5	9.4

Table 1: Validation Set Scores for ADE Tasks

Task & Metric	F1	P	R
ADE Extr. official	13.2	8.0	37.4
ADE Norm. official	8.2	5.0	23.7
ADE Norm. unseen IDs	1.4	0.7	10.0

Table 2: Comprehensive Test Set Scores for ADE Tasks

4 Results

Our system consisting of the NER model for ADE extraction and RoBERTa for the normalization task is evaluated in Tables 1 and 2 on the validation set and the official test set, respectively.

4.1 ADE Extraction

Table 2 shows that the ADE extraction model achieved an F1 of 13.2 on the test set, with precision and recall scores of 8.0% and 37.4%, respectively. These results highlight the model's higher success in recall, indicating its effectiveness in identifying ADE mentions. However, the model's low precision of 8.0% highlights a significant challenge in specificity. The model's unexpectedly high recall on the test set compared to the validation set, where recall and precision were more balanced, indicates differences in the distribution and complexity of the validation and test set.

4.2 ADE Normalization

For ADE normalization, the official scores in Table 2 show a precision of 5.0%, a recall of 23.7%, and an F1 of 8.2. Additionally, when evaluating the model's performance on previously unseen MedDRA IDs, it returned significantly lower metrics (precision: 0.7%, recall: 10.0%, F1: 1.4). This considerable drop suggests challenges in generalizing to new, unseen ADE terms, reflecting potential limitations in the model's generalizing capability. The results on the validation set were somewhat consistent, with an F1-score of 8.4 and slightly lower precision and recall of 7.5% and 9.4%, respectively.

4.3 Discussion

Our results point to the challenges inherent in biomedical NLP tasks, especially in balancing pre-

cision and recall and generalizing to new data. The low precision observed shows issues with generalizability beyond training data in NER. The results may also reflect the complexities of social media language, complicating ADE detection and normalization.

Moreover, the differences in results between validation and official test data underline the importance of robust cross-validation strategies to mimic real-world performance and prevent overfitting. Further efforts need to focus on integrating domain-specific knowledge bases to heighten normalization accuracy and better manage new ADE identifiers.

5 Conclusion

Our contribution to #SMM4H 2024 Task 1 consists of an NER model retrained to identify ADEs and a similarity-based RoBERTa model to normalize them. The findings from our system underline the challenges and opportunities presented by the use of NLP in detecting and normalizing ADEs from social media. Despite achieving high recall, our model's low precision highlights a significant challenge in accurately identifying relevant ADEs amid the informal language prevalent on platforms like Twitter. Furthermore, the task has demonstrated that while our current methodology is capable of initial identification, it falls short in scenarios involving generalizing to data different from the training data, which is crucial for practical applications.

For future work, we will investigate enhancing model precision through advanced linguistic analysis, employing models pre-tuned on ADE datasets, fine-tuning RoBERTa for vectorization of ADE entities and MedDRA entries, and incorporating additional ADE data.

References

Ilseyar Alimova and Elena Tutubalina. 2019. Detecting adverse drug reactions from biomedical texts with neural networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 415–421, Florence, Italy. Association for Computational Linguistics.

Xiang Dai, Sarvnaz Karimi, and Cecile Paris. 2017. Medication and adverse event extraction from noisy text. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 79–87, Brisbane, Australia.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. https://arxiv.org/abs/2401.08281.

Oguzhan Gencoglu. 2020. Sentence transformers and Bayesian optimization for adverse drug effect detection from Twitter. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 161–164, Barcelona, Spain (Online). Association for Computational Linguistics.

Trung Huynh, Yulan He, Alistair Willis, and Stefan Rueger. 2016. Adverse drug reaction classification with deep neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 877–887, Osaka, Japan. The COLING 2016 Organizing Committee.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the Tweebank corpus on named entity recognition and building NLP models for social media analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7199–7208, Marseille, France. European Language Resources Association.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Ari Z Klein, Juan M Banda, Yuting Guo, Ana Lucia Schmidt, Dongfang Xu, Ivan Flores Amaro, Raul Rodriguez-Esteban, Abeed Sarker, and Graciela Gonzalez-Hernandez. 2024. Overview of the 8th social media mining for health applications (# smm4h) shared tasks at the amia 2023 annual symposium. *Journal of the American Medical Informatics Association*, 31(4):991–996.

Seunghee Lee, Hyekyung Woo, Chung Chun Lee, Gyeongmin Kim, Jong-Yeup Kim, and Suehyun Lee. 2023. Drug_snsminer: standard pharmacovigilance pipeline for detection of adverse drug reaction using sns data. *Scientific Reports*, 13(1):3779.

Xi Liu, Han Zhou, and Chang Su. 2022. PingAnTech at SMM4H task1: Multiple pre-trained model approaches for adverse drug reactions. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 4–6, Gyeongju, Republic of Korea. 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. *arXiv preprint arXiv: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.

V.G.Vinod Vydiswaran, Grace Ganzel, Bryan Romas, Deahan Yu, Amy Austin, Neha Bhomia, Socheatha Chan, Stephanie Hall, Van Le, Aaron Miller, Olawunmi Oduyebo, Aulia Song, Radhika Sondhi, Danny Teng, Hao Tseng, Kim Vuong, and Stephanie Zimmerman. 2019. Towards text processing pipelines to identify adverse drug events-related tweets: University of Michigan @ SMM4H 2019 task 1. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 107–109, Florence, Italy. Association for Computational Linguistics.

Davy Weissenbacher, Juan Banda, Vera Davydova, Darryl Estrada Zavala, Luis Gasco Sánchez, Yao Ge, Yuting Guo, Ari Klein, Martin Krallinger, Mathias Leddin, Arjun Magge, Raul Rodriguez-Esteban, Abeed Sarker, Lucia Schmidt, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the seventh social media mining for health applications (#SMM4H) shared tasks at COLING 2022. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 221–241, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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.

Anthony Yazdani, Hossein Rouhizadeh, David Vicente Alvarez, and Douglas Teodoro. 2023a. Ds4dh at# smm4h 2023: zero-shot adverse drug events normalization using sentence transformers and reciprocal-rank fusion. *arXiv preprint arXiv:2308.12877*.

Anthony Yazdani, Hossein Rouhizadeh, Alban Bornet, and Douglas Teodoro. 2023b. Conorm: Context-aware entity normalization for adverse drug event detection. *medRxiv*, pages 2023–09.