

Uncovering Underreported Adverse Drug Reactions in Epilepsy Communities via Social Media Mining

Ziqi Guo¹, Robert Palermo², Ahmed Abdeen Hamed³, Luis M. Rocha^{*1,4}
rocha@binghamton.edu*

¹School of Systems Science and Industrial Engineering, Binghamton University, Binghamton, NY, USA.

²Department of Mathematics and Statistics, Binghamton University, Binghamton, NY, USA.

³Multidisciplinary Graduate Engineering, Northeastern University, Miami, FL, USA

⁴Universidade Católica Portuguesa, Católica Biomedical Research Centre, Lisbon, Portugal.

Abstract

Post-marketing drug surveillance often suffers from under-reporting and delays in identifying adverse drug reactions (ADRs). While benchmark datasets and shared tasks—particularly from SMM4H—have advanced ADR detection methods, they primarily focus on general populations and known ADRs. This study complements those efforts by focusing on epilepsy and aiming for the discovery of unknown ADRs. We collected user-generated posts from r/Epilepsy and the Epilepsy Foundation of America (EFA) forums, curated a drug-symptom dictionary, and developed a classification pipeline that combines sentiment analysis with relation classification. Sentiment polarity serves as a putative interpretable characterization of patient experience, while relation classification determines whether co-mentioned terms reflect an ADR or a drug indication. Labels from the SIDER database were used for distant supervision, enabling scalable, domain-adaptable automation without manual annotation. Identifying unknown ADRs remains particularly challenging, as they are rarely annotated or included in available databases. Our classifier demonstrates strong generalization to such cases by leveraging patterns in real-world discourse. When evaluated on the 2025 SMM4H Shared Task 1, it achieved a high precision of 0.80—which is desirable since false positives may lead to misleading surveillance hypotheses and costly but unnecessary follow-up efforts. Manual validation on reddit and EFA further demonstrates the ability of our classifier to identify self-reporting of unknown ADRs. Overall, our work demonstrates that community-focused social media mining, informed by sentiment analysis, can enrich pharmacovigilance pipelines and increase interpretability of automated, low-cost drug safety warnings.

Introduction

Adverse drug reactions (ADRs) remain a critical challenge for public health, yet traditional surveillance systems often suffer from underreporting and delayed detection. Social media offers a complementary, real-time source of patient-reported experiences, leading to a surge of computational approaches for ADR detection.

Growing interest in social media mining for ADR detection, has led to the development of benchmark datasets and shared tasks—most notably the Social Media Mining

for Health (SMM4H) initiative. These benchmarks have enabled standardized evaluation of NLP systems for ADR classification, extraction, and normalization (Klein et al. 2023; Xu et al. 2024; Klein et al. 2025). Early methods range from deep learning models using word embeddings (Rezaei et al. 2019) to hybrid approaches that incorporated syntactic, semantic, and sentiment-based features (Zhang, Cui, and Gao 2020). For example, Korkontzelos et al. (2016) demonstrated that incorporating sentiment as an auxiliary feature could improve ADR detection by reducing false positives in token-level classifiers (Korkontzelos et al. 2016). More recently, large language models (LLMs) have driven significant improvements in predictive accuracy and in scalable data annotation through distillation from GPT-based teacher models (Gu et al. 2023), fine-tuning transformer models like BERT for higher recall in ADR mention detection (Dey, Shrivastava, and Kumar 2024), and other techniques.

Despite these advances, most existing methods are optimized for detecting known ADRs within general populations and require substantial manual annotation. These approaches are less suited for discovering previously undocumented ADRs, particularly in underrepresented communities where language is more diverse and context-specific. Identifying such signals is especially challenging because they are rarely labeled.

To approach the unknown ADR problem, we focus on a clinically important population with a high burden of ADRs: epilepsy patients (Bayane et al. 2024). By analyzing user-generated posts from r/Epilepsy and the Epilepsy Foundation of America (EFA) forums, we identify potential ADRs using a domain-adapted NLP pipeline, shown in Figure 1, which combines sentiment analysis with relation classification. Unlike previous approaches, our method treats sentiment polarity as a primary, interpretable signal for ADR detection. We use distant supervision from the SIDER database, without manual annotation, to validate known ADRs signals, and subsequently identify putative unknown ADRs—which we validate with a manual review of ADR self-reporting on Reddit and EFA. The approach is also shown to classify social media posts containing ADRs with high precision, on the SMM4H Shared Task benchmark. Ours is a low-cost and community-aware pipeline for pharmacovigilance, with increased interpretability via sentiment analysis. By shifting the focus from benchmark optimiza-

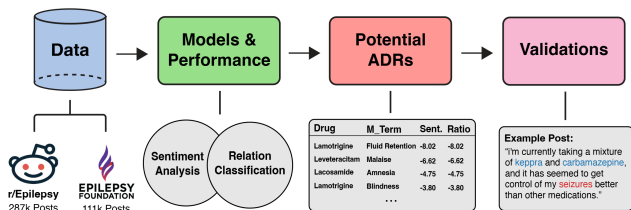


Figure 1: Overview of the proposed pipeline, including data collection, model inference and performance, identification of potential ADRs, and validation through backtracking to the original posts.

tion to community-centered signal discovery, we demonstrate that domain-specific social media mining can uncover novel drug safety concerns and complement existing drug surveillance.

Data and methods

Social media data collection

We collected user-generated text related to epilepsy from the epilepsy-focused Reddit community (r/Epilepsy) and EFA forums. Details on the data selection criteria and cohort definition are provided in a work (Guo et al. 2025), and all analyses involving patient data were conducted under IRB approval (Correia et al. 2025). An overview of the collected data is presented in Table 1. All analyses in this study were conducted at the sentence level, with sentences segmented using the NLTK sentence tokenizer, specifically focusing on target sentences.

Platform	Unique Users	Posts	Median Sentence per Post	Target Sentences*
r/Epilepsy	18,515	287,367	3	111,000
EFA	22,938	111,075	7	120,000

Table 1: Overview of collected social media data. *Target sentences are those containing at least two medical dictionary-matched mentions. The medical dictionary is detailed in the following section

Medical dictionary curation

We built a comprehensive medical dictionary by curating terms from established biomedical ontologies, enabling us to leverage existing medical knowledge. To identify potential adverse drug effects in social media text, we used this dictionary to match key terms of interest — specifically, the *drug terms* and *medical terms* such as conditions and symptoms. The dictionary contains approximately 142K terms, with instances tagged only when an exact match was identified. The drug terms were extracted from Drugbank (v.5.1.0), including the commercial and the chemical names, where all the names will be used for matching but then they will be integrated to a higher level chemical names when we analyze

the relation of the drug-symptom pairs. For instance, the epilepsy drug *Levetiracetam* has the brand-name *Keppra* as a synonym; The medical terms were extracted from MedDra (v15.0) using the Mesh terms, which are in hierarchical structure including the child terms and the parent terms. Similarly to the aggregation with synonymous of drug names, we aggregated the medical terms by mapping synonyms to a parent term, for example, the term *Influenza* has *Flu* and *Flu syndrome* as synonyms. Note that we included the terms related not only to epilepsy but also to other diseases and conditions to facilitate the encoding of relevant associations.

Ground truth annotation

We used the publicly available SIDER database (Campillos et al. 2008) as a distant supervision source to label sentences containing drug-symptom co-mentions as either adverse drug reactions (ADRs), indications, or unreported. SIDER compiles curated information on marketed drugs and their associated adverse effects, providing drug-medical term pairs labeled as either ADRs or indications. Notably, the medical terms in SIDER follow the MedDRA vocabulary and align with the terminology used in our curated dictionary, facilitating seamless integration.

Sentences were annotated based on whether the co-mentioned drug-symptom pair matched an entry in the SIDER database. For example, the sentence shown in Figure 2 is labeled as an ADR because the pair “Dilantin” and “headache” appears in SIDER’s ADR list. To reduce ambiguity from potential drug-drug interactions, we excluded sentences containing multiple drug mentions.

We obtained $\approx 35K$ labeled sentences, which were used for validation in the sentiment analysis and for both training and validation in the relation classification model. The remaining drug-symptom pairs, existing in 48K sentences, are not recorded in the SIDER dataset and are of particular interest, as they represent previously unreported ADRs.

Sentiment analysis

Our methodology is based on the premise that negative sentiments toward a drug-symptom pair likely indicate a side effect, whereas positive sentiments suggest an indication. To accurately capture the context and sentiment associated with each drug-symptom pair, we performed sentiment analysis on sentences where both a drug and a symptom were mentioned. Even though analysis of sentiment at sentence level is noisier than for larger amounts of text, it ensures that sentiment interpretations directly reflect the relevant context, minimizing ambiguity caused by surrounding text. More specifically, we focus on the sentiment associated of direct self-reporting of drug-symptom pairs.

We employed RoBERTa, a robustly optimized BERT approach (Liu 2019), for sentiment analysis of short social media texts at the sentence level. Studies have demonstrated RoBERTa’s promise in analyzing health-related data, accurately identifying emotional contexts in clinical interactions (Gabriel et al. 2024).

Specifically, we applied a pre-trained RoBERTa model to each sentence that contained at least one co-mentioned

Original
 i am now on <unit> of **dilantin**, sometimes i wake up a few hours after going to sleep with raging **headache** or sometimes cant sleep at all.....

Masked
 i am now on <unit> of [DRUG], sometimes i wake up a few hours after going to sleep with raging [MEDICAL_TERM] or sometimes cant sleep at all.....

Figure 2: Example of masking drug and medical term in a sentence. The sentence is preprocessed by replacing the numbers, removing capitalization. We applied exact term matching with the post from the medical dictionary.

drug–symptom pair. The RoBERTa has three dimension outputs of sentiments, which are the probability of being positive, neutral, and negative, and the summation of the probability equals to 1. For ease of calculation purpose, we did the transmission below to reduce into one dimension score P : $P = \frac{p_{\text{positive}} - p_{\text{negative}}}{p_{\text{neutral}}}$. It was found to be the most effective among several alternatives, such as binary 0/1 encoding, and using only p_{positive} as the sentiment factor. The sentiment score P is proportional to the positivity of the sentence—higher values indicate more positive sentiment, and vice versa. We then collected the sentiment scores P from all sentences in which a specific drug–symptom pair was co-mentioned (see Figure 3 for details). For each pair, these scores were aggregated to capture the overall sentiment associated with that pair. Particularly, to prevent the sentiment model from being influenced by the inherent sentiment of specific drug or medical terms (e.g., headache), we masked these terms during sentiment analysis. An example of this masking process is shown in Figure 2.

Model performance To evaluate the effectiveness of sentiment scores in distinguishing ADRs from indications, we analyzed their correlation with the ground truth relationship labels for each drug–symptom pair. We used a box plot to visualize differences in sentiment distributions across the two classes (Figure 4). We also generated precision–recall curves and reported the area under the curve (AUC) as the primary evaluation metric (right plot in Figure 4), which helps mitigate the effects of class imbalance—specifically, an ADR-to-indication ratio of 2:1 (196:99). To reduce noise and randomness, we included only drug–symptom pairs that appeared in more than 20 sentences.

We compared mean and median sentiment scores and found the median to be more effective for distinguishing ADRs from indications. We also evaluated models with and without masking, and since masking performed equally well while reducing bias, we adopted it in our final analysis. Figure 4 shows the results using median sentiment with masking, our chosen setup.

The results show that sentiment is an effective discriminative feature. As illustrated in Figure 4, pairs labeled as side effects exhibit significantly more negative sentiment than those labeled as indications. The model achieved an AUC of 0.91, substantially higher than the naïve baseline of 0.67.

Relation classification

To complement sentiment analysis, we developed a binary relation classifier to determine whether a co-mentioned

drug–symptom pair reflects an adverse drug reaction (ADR) or not. We fine-tuned BioBERT (Lee et al. 2020), a domain-specific language model pre-trained on biomedical literature from PubMed and PMC, which has demonstrated superior performance over the general-purpose BERT model (Devlin 2018) in medical NLP tasks.

We applied a masking strategy during preprocessing to prevent classification based on prior encoding of drug–symptom associations by BioBERT, thereby aiming for generalization to unseen ADRs. Masking (see Section , Figure 2) replaces drug and symptom terms with placeholders, prompting the language model to focus on sentence-level semantics. Although this resulted in slightly lower validation performance, it enhanced generalizability; thus, we adopted the masked setup in our final model (see Model performance Section)

Model performance To ensure robust model evaluation, we performed 5-fold cross-validation on the labeled sentences. There are approximately 28K sentences for training and 7K for validation. The splits were stratified, maintaining a consistent ADR-to-indication ratio of approximately 1:1.17 across all folds. Table 3 presents the results of relation classification using masked and unmasked preprocessing for both BioBERT and BERT models. BioBERT consistently outperformed BERT, even when key terms (drug and symptom names) were masked to prevent reliance on prior knowledge. These results suggest that BioBERT has stronger inference capabilities for biomedical text. To enhance generalizability, we selected the masked BioBERT model for our final setup (highlighted in bold in Table 3).

Model	Preprocessing	Accuracy	MCC
BioBERT	No Mask	0.90	0.84
	Mask	0.83	0.66
BERT	No Mask	0.83	0.61
	Mask	0.67	0.40

Table 2: Relation classification results on the validation set. Accuracy and Matthews Correlation Coefficient (MCC) were used as evaluation metrics.

Results

SMM4H Shared task 1

We applied our method to the Shared Task 1 benchmark on *English*, which focuses on binary classification of social media posts as containing ADRs or not. While our pipeline achieved a high precision of 0.80 (Table 3), its overall F1 score was lower than top-performing shared task submissions. This discrepancy is expected, as our method was designed for a different goal—identifying drug–symptom pairs indicative of potential ADRs, rather than classifying individual posts. Our model aggregates evidence from multiple posts (typically at least 10 co-mentions of a pair) to improve confidence in signal detection. As a result, we prioritizes precision over recall, which is more aligned with our objective of minimizing false positives, under-reported

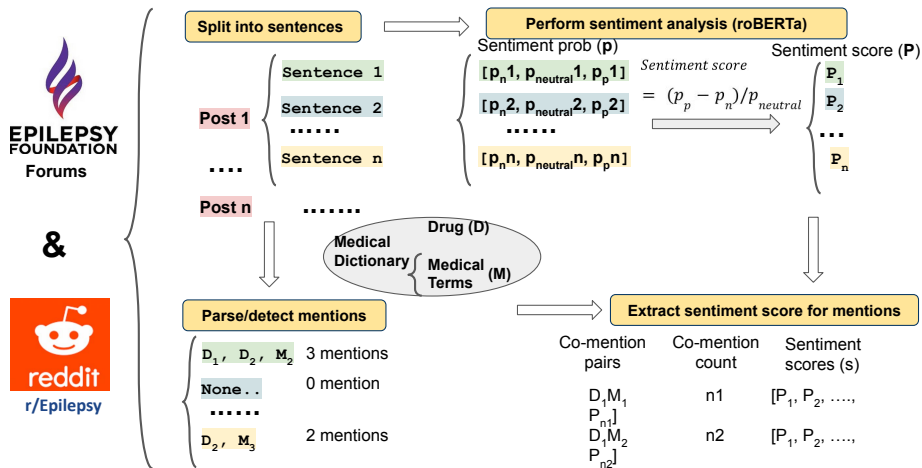


Figure 3: Sentiment analysis scheme.

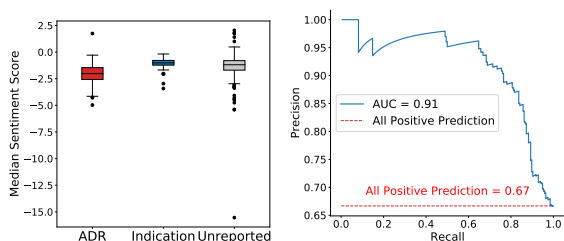


Figure 4: Evaluation of Sentiment Analysis on Drug-Symptom Pairs. Left: Boxplot of median sentiment scores across three ground truth categories. Right: Precision-recall curve for median sentiment scores. The red dashed line represents the baseline performance of a naïve model that predicts all pairs as ADRs.

ADR signals. Our current pipeline is specifically designed for English-language content; for other languages, we developed separate models, which are presented in another work submitted to the shared task track. Despite the task mismatch, our performance on the shared task highlights the generalizability and robustness of our approach.

Prediction on test set		Precision	Recall	F1
Our		0.8036	0.4831	0.6034
All	Mean	0.6374	0.6053	0.5915
	Medium	0.7385	0.6732	0.6924

Table 3: Prediction performance on the test set compared to all submissions.

Predicted ADRs

To uncover previously unreported potential ADRs, we applied our trained sentiment analysis and relation classification models to sentences containing drug-symptom co-mentions that are not labeled in the ground truth dataset

(i.e., not listed in SIDER). We extracted all such pairs from the corpus, aggregated their associated sentiment scores, and recorded the predicted ADR labels from the relation classification model. For each pair, we computed the median sentiment score and the ADR ratio—the proportion of sentences classified as ADRs by the relation classification model among all occurrences of the pair.

To identify high-likelihood ADR candidates, we applied the following filtering criteria to exclude pairs: (1) fewer than 10 co-mentions to reduce noise and sampling bias; (2) implausible medical terms as side effects (e.g., Divorced); (3) medical terms already associated with the drug in SIDER under different synonyms (e.g., Hepatocellular injury vs. liver injury for Levetiracetam) to avoid not novel discoveries; (4) median sentiment score ≥ -1.8 , as such pairs are less likely to indicate negative experiences. After filtering, 75 candidate pairs remained. From these, based on maximizing co-mention count, median sentiment, and ADR ratio, we manually selected the most likely subset of 6 putative ADRs, which are listed in Table 4.

Manual validation of ADR self-reporting

We conducted a manual review of user-generated posts to identify ADRs based on established labeling guidelines. Posts were labeled as ADR-positive only if they described harmful or unintended physical or psychological effects following drug intake, with a clear or reasonably implied causal link. Common indicators of adverse reaction self-reporting include verbs like “gave me”, “made me”, “caused”, “was from”, “triggered”, and temporal cues such as “after starting”, “since taking”, and “when I began”. Notably, all mentions of Lamotrigine and fluid retention were confirmed as true cases of ADR self-reporting.

Discussion

Our study demonstrates the feasibility and power of mining community-specific social media data to uncover unknown or underreported ADRs. Our approach combined sentiment

Drug	Medical Term	Count	Median Sentiment	ADR Ratio
Lamotrigine	Profound mental retardation	10	-8.21	0.80
Lamotrigine	Fluid retention	10	-8.02	1.00
Levetiracetam	Aphasia	17	-6.08	1.00
Lacosamide	Amnesia	21	-4.75	0.73
Clonazepam	Clonic convulsion	16	-4.39	1.00
Levetiracetam	Muscle spasms	37	-3.29	0.90

Table 4: Selected putative unknown ADRs. Count is the number of sentences in which a particular drug and medical term co-occur. Median Sentiment is the sentiment score from the sentiment analysis model, which has been tested as the most effective sentiment metric for this task. ADR Ratio is the proportion of co-mentions that the relation classification model classified as ADRs.

analysis—as an interpretable proxy for patient experience—with relation classification and distant supervision using curated dictionaries and public biomedical resources. It enables low-cost ADR discovery, with increases interpretation afforded by the sentiment analysis. Furthermore, it depends on very minimal human annotation, while it generalizes to uncover putative novel ADRs. The strong precision achieved on the SMM4H benchmark evaluation underscores its reliability in picking credible safety signals with high precision. The manual validation of top putative unknown ADRs based on (epilepsy-focused) social media user self-reporting, further supports our the ability of our approach to detect relevant drug safety warning signs.

More broadly, our work supports a shift toward community-centered, disease-specific, data-driven pharmacovigilance by leveraging real-world narratives and offering a practical approach for identifying emerging ADRs. Given its simplicity, it should be easily extendable for other patient cohorts. As a next step, we aim to validate the putative ADR signals through external evidence, including clinical trial reports, biomedical literature (e.g., PubMed), and drug safety databases, to further assess their credibility and potential clinical relevance. Finally, we encourage readers to consult the supplementary material for additional discussion, including more detailed clarifications of reviewer comments.

References

Bayane, Y. B.; Jifar, W. W.; Berhanu, R. D.; and Rikitu, D. H. 2024. Antiseizure adverse drug reaction and associated factors among epileptic patients at Jimma Medical Center: a prospective observational study. *Scientific Reports*, 14(1): 11592.

Campillos, M.; Kuhn, M.; Gavin, A.-C.; Jensen, L. J.; and Bork, P. 2008. Drug target identification using side-effect similarity. *Science*, 321(5886): 263–266.

Correia, R. B.; Rozum, J. C.; Cross, L.; Felag, J.; Galant, M.; Guo, Z.; Herr, B. W.; Min, A.; Sanchez-Valle, J.; Stungis Rocha, D.; et al. 2025. myAURA: a personalized health library for epilepsy management via knowledge graph sparsification and visualization. *Journal of the American Medical Informatics Association*, ocaf012.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dey, A.; Shrivastava, J. N.; and Kumar, C. 2024. Adverse Drug Reaction Detection from Social Media Review Using BERT Technology. *International Research Journal of Multidisciplinary Scope*, 05(01): 405–416.

Gabriel, R. A.; Litake, O.; Simpson, S.; Burton, B. N.; Watterman, R. S.; and Macias, A. A. 2024. On the development and validation of large language model-based classifiers for identifying social determinants of health. *Proceedings of the National Academy of Sciences*, 121(39): e2320716121.

Gu, Y.; Zhang, S.; Usuyama, N.; Woldesenbet, Y.; Wong, C.; Sanapathi, P.; Wei, M.; Valluri, N.; Strandberg, E.; Naumann, T.; and Poon, H. 2023. Distilling Large Language Models for Biomedical Knowledge Extraction: A Case Study on Adverse Drug Events. *ArXiv:2307.06439 [cs]*.

Guo, Z.; Felag, J.; Rozum, J. C.; Correia, R. B.; Wang, X.; and Rocha, L. M. 2025. Focused digital cohort selection from social media using the metric backbone of biomedical knowledge graphs. *Journal of Biomedical Informatics*. Accepted for publication, in press.

Klein, A. Z.; Banda, J. M.; Guo, Y.; Schmidt, A. L.; Xu, D.; Flores Amaro, I.; Rodriguez-Esteban, R.; Sarker, A.; and Gonzalez-Hernandez, G. 2023. 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.

Klein, A. Z.; Dasgupta, T.; Flores Amaro, I.; Jana, S.; Khademi, S.; Lopez-Garcia, G.; Onishi, T.; Powell, J.; Raitel, L.; Rajwal, S.; Roller, R.; Sarker, A.; Sinha, M.; Thomas, P.; Tutubalina, E.; Xu, D.; Zweigenbaum, P.; and Gonzalez-Hernandez, G. 2025. Overview of the 10th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ICWSM 2025. In *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*. AAAI Press.

Korkontzelos, I.; Nikfarjam, A.; Shardlow, M.; Sarker, A.; Ananiadou, S.; and Gonzalez, G. H. 2016. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of biomedical informatics*, 62: 148–158.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rezaei, Z.; Eslami, B.; Chavoshinejad, R.; and Totonchi, M. 2019. Adverse Drug Reaction Detection in Social Media by Deep Learning Methods. *Cell Journal (Yakhteh)*, 22(3).

Xu, D.; Lopez-Garcia, G.; Raithel, L.; Roller, R.; Thomas, P.; Aramaki, E.; Wakamiya, S.; Yada, S.; Zweigenbaum, P.; O'connor, K.; et al. 2024. Overview of the 9th Social Media Mining for Health Applications (# SMM4H) Shared Tasks at ACL 2024—Large Language Models and Generalizability for Social Media NLP. In *The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, 183–195.

Zhang, Y.; Cui, S.; and Gao, H. 2020. Adverse drug reaction detection on social media with deep linguistic features. *Journal of Biomedical Informatics*, 106: 103437.

Supplementary Material for “Uncovering Underreported Adverse Drug Reactions in Epilepsy Communities via Social Media Mining”

Ziqi Guo¹, Robert Palermo², Ahmed Abdeen Hamed³, Luis M. Rocha^{*1,4}
*rocha@binghamton.edu

¹School of Systems Science and Industrial Engineering, Binghamton University, Binghamton, NY, USA.

²Department of Mathematics and Statistics, Binghamton University, Binghamton, NY, USA.

³Multidisciplinary Graduate Engineering, Northeastern University, Miami, FL, USA

⁴Universidade Católica Portuguesa, Católica Biomedical Research Centre, Lisbon, Portugal.

Our medical dictionary is comprehensive, including nearly every published medical term and drug from established ontologies, reducing the likelihood of unmatched drug and symptom mentions in social media posts. However, the current approach relies on exact matching, which may miss spelling variations, abbreviations, or informal terminology used by patients. To address this, future work will explore fuzzy matching and context-aware term expansion to improve recall without sacrificing precision.

In addition to our current use of BioBERT for relation classification, we plan to evaluate alternative models, such as PubMedBERT and classic machine learning classifiers, for comparative performance analysis. We also intend to incorporate sentiment analysis using ANEW, a dictionary-based approach that offers better interpretability by directly linking sentiment scores to specific words, enhancing the explainability of model outputs.

For the shared task, our approach differed slightly from the main pipeline. We used the full, original posts provided by the shared task dataset, without sentence splitting, to preserve context. Instead of distant supervision from SIDER, we fine-tuned the BioBERT model using the provided training labels. We then applied our medical dictionary to identify drug–symptom co-mentions and incorporated sentiment analysis at the post level, combining the sentiment score with BioBERT predictions for the final classification. This approach ensures that the full context of each post is considered, aligning with the shared task’s evaluation criteria.

Our approach emphasizes high precision to reduce the impact of false positives, which is particularly important for the early stages of ADR signal detection, where filtering out irrelevant signals can reduce the cost of downstream validation. This may come at the cost of recall, potentially missing weaker or less frequently reported signals. Our model is inherently flexible, allowing the trade-off between precision and recall to be adjusted based on the specific goals of the analysis. For example, modifying the decision thresholds, such as the median sentiment score or the relation classification ratio, can shift the model toward higher recall if the priority is comprehensive signal capture. This flexibility makes our approach adaptable to a range of pharmacovigilance contexts, from broad exploratory searches to more

focused, high-precision analyses. Moving forward, a more systematic sensitivity analysis will be necessary to optimize these parameters and better understand their impact on overall model performance.

We also plan to conduct a manual review of original user posts. Specifically, we will sample 5–10 sentences for each of the top 10 drug–symptom pairs with strong negative sentiment, high ADR ratio, and moderate-to-high co-mention counts. Two independent annotators will assess whether each sentence explicitly or implicitly indicates an ADR, using labels such as Yes (ADR indicated), No (not an ADR), or Uncertain. Inter-annotator agreement will be measured to assess the reliability of this process (e.g., Cohen’s Kappa), and the final confirmed ADR proportion will provide a more concrete measure of our pipeline’s effectiveness in surfacing credible, underreported signals.

Our current pipeline relies on the SIDER database as the primary ground truth for known ADRs, providing a standardized and structured annotation source aligned with the MedDRA vocabulary. However, we recognize that some ADRs may be well-documented in broader clinical resources, such as Mayo Clinic guidelines, FDA drug labels, and patient information leaflets, but may not be included in SIDER. In future work, we plan to expand our scope to include these additional sources, providing a more comprehensive definition of known ADRs and further enhancing the precision of our discovery pipeline.

Our current pipeline utilizes both sentiment analysis and relation classification to identify potential ADRs, treating these components as complementary but largely independent signals. There is potential for tighter integration of these models to enhance both interpretability and predictive performance. For instance, the continuous output from the relation classification model could be combined with sentiment scores to generate more nuanced risk assessments for drug–symptom pairs. This could involve calculating performance metrics like AUC for the classification model and comparing them directly with sentiment-based thresholds, or using MCC to assess alignment. Such joint evaluation could provide deeper insights into the interaction between patient sentiment and explicit ADR relationships, potentially improving the pipeline’s overall recall and reducing false positives. This integrated approach will be the focus in future work as we refine our method.

Beyond the technical contributions, this work has broader implications for healthcare systems, patients, and pharmaceutical companies. Early detection of underreported ADRs can significantly reduce healthcare costs by preventing adverse events and hospitalizations, while improving patient safety and treatment adherence. For patients, this approach offers an opportunity to surface potential side effects that might not be captured in clinical trials, empowering individuals to make more informed healthcare decisions. For pharmaceutical companies, these insights can provide early warnings of emerging safety concerns, potentially reducing legal liabilities and improving patient trust. At the same time, maintaining high precision is critical to avoid unnecessary patient anxiety and overburdening healthcare professionals, while high recall is essential to capture subtle, underreported signals. Balancing these objectives will be an important focus as this work continues to evolve.