Ingénierie des Systèmes d'Information
Vol. 31, No. 6, June, 2026, pp. 1847-1853
Journal homepage: http://iieta.org/journals/isi

A Bidirectional Encoder Representations from Transformers-Based Framework for Automated Adverse Drug Reaction Detection from Patient-Generated Psychiatric Treatment Narratives

Ghulam Asrofi Buntoro[1,2*], Didik Riyanto[2], Ismail Abdurrozzaq Zulkarnain[1], Edy Kurniawan[2], Rhesma Intan Vidyastari[2], Rizal Arifin[3], Adi Fajaryanto Cobantoro[1], Fauzan Masykur[1], Ali Selamat[4]

[1] Department of Informatics Engineering, Universitas Muhammadiyah Ponorogo, Ponorogo 63471, Indonesia
[2] Department of Electrical Engineering, Universitas Muhammadiyah Ponorogo, Ponorogo 63471, Indonesia
[3] Department of Mechanical Engineering, Universitas Muhammadiyah Ponorogo, Ponorogo 63471, Indonesia
[4] Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur 54100, Malaysia

Corresponding Author Email: ghulam@umpo.ac.id

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/isi.310607

Received: 14 July 2025
Revised: 13 September 2025
Accepted: 25 September 2025
Available online: 30 June 2026

Keywords:
adverse drug reaction detection, Bidirectional Encoder Representations from Transformers, clinical natural language processing, psychiatric pharmacovigilance, patient-generated narratives, transformer-based models, biomedical text mining

ABSTRACT

Adverse drug reactions (ADRs) associated with psychiatric medications are difficult to identify because patient-reported experiences are often informal, diverse, and embedded in unstructured textual narratives. Conventional pharmacovigilance systems mainly depend on structured reports, which may overlook valuable safety information contained in real-world patient communications. This study develops a transformer-based natural language processing (NLP) framework using a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model for automated ADR detection in psychiatric treatment narratives. Two publicly available datasets, Psychiatric Treatment Adverse Reactions (PsyTAR) and Consumer Adverse Drug Event Corpus (CADEC), were integrated to construct a domain-specific corpus for binary classification of ADR and non-ADR texts. The preprocessing pipeline included text normalization, BERT-based tokenization, class balancing, and supervised fine-tuning using the HuggingFace Transformers framework. The proposed model was trained with the BERT-base-uncased architecture and evaluated using standard classification metrics on the validation dataset. Experimental results demonstrate that the model achieved an F1-score of 0.89, accuracy of 0.89, precision of 0.88, and recall of 0.91, indicating its effectiveness in identifying ADR-related expressions from patient-generated narratives. The findings suggest that transformer-based models can support automated pharmacovigilance workflows by extracting safety signals from unstructured psychiatric healthcare texts. However, further investigation using larger clinical datasets, domain-specific pretrained models, and explainable artificial intelligence methods is required before practical clinical deployment.

1. INTRODUCTION

Pharmacovigilance, the science and activities relating to the detection, assessment, understanding, and prevention of adverse effects or other drug-related problems, is a foundational element in ensuring the safety and efficacy of medical therapies [1, 2]. Its importance is especially pronounced in the field of psychiatry, where treatment often involves long-term medication with complex psychopharmacological profiles and high interindividual variability in drug response [3]. Adverse drug reactions (ADRs), ranging from mild discomfort to life-threatening events, are not only common but often underreported, particularly in psychiatric contexts where patients may have difficulty articulating their experiences or may avoid formal medical channels [4].

Conventional pharmacovigilance systems, such as spontaneous reporting databases (e.g., FDA Adverse Event Reporting System (FAERS), EudraVigilance), primarily rely on structured inputs from healthcare professionals or regulatory entities. These systems, while essential, suffer from incompleteness and reporting biases, leading to delayed detection of critical safety signals [5]. Moreover, they typically fail to capture the nuanced, subjective experiences of patients, especially those undergoing psychiatric treatment, thus leaving a significant gap in the pharmacovigilance landscape.

With the emergence of digital health platforms including patient forums, social media, and health-related discussion boards, a new avenue has opened for extracting pharmacovigilance-relevant information directly from patient-reported narratives [6]. These texts are rich in subjective experiences and real-world insights but are inherently noisy, unstructured, and linguistically diverse, which limits their direct usability by traditional surveillance systems [7].

In recent years, natural language processing (NLP) has

become a promising solution for mining unstructured biomedical text data [8], including ADR detection [9]. Classical NLP techniques, such as bag-of-words or Term Frequency-Inverse Document Frequency (TF-IDF) with shallow classifiers (e.g., Support Vector Machines (SVMs), Naïve Bayes), have been used for biomedical text classification but often lack the semantic depth needed for precise ADR extraction [10]. The advent of deep learning, and more specifically transformer-based architectures like Bidirectional Encoder Representations from Transformers (BERT), has revolutionized text representation and context modeling. These models, pre-trained on massive corpora and fine-tuned on downstream tasks, excel at capturing complex syntactic and semantic relationships that are critical in clinical texts [11].

In the context of pharmacovigilance, transformer models have shown strong performance across various tasks including named entity recognition, relation extraction, and multi-class classification of ADRs [10, 12]. However, much of this work has focused on general medical corpora (e.g., MIMIC-III, PubMed) and less so on psychiatric-specific datasets, which present unique linguistic and semantic challenges due to the nature of mental health discourse [13].

While this study employs the general-purpose BERT (BERT-base-uncased) model due to its strong baseline performance and widespread adoption, it is important to acknowledge its limitations in domain-specific contexts such as psychiatric pharmacovigilance. General-domain language models are trained on broad corpora (e.g., Wikipedia and BooksCorpus) and may lack sensitivity to specialized medical terminology, nuanced clinical expressions, and patient-reported symptom variations commonly found in psychiatric texts. As a result, certain domain-specific patterns and semantic relationships may not be optimally captured. In contrast, domain-adapted models such as BioBERT [12] and MentalBERT [14] are pretrained on biomedical and mental health-related corpora, enabling them to better encode domain-specific knowledge and potentially achieve improved performance in tasks involving clinical narratives and patient-generated content. Therefore, this study positions the current approach as a strong baseline, providing a reference point for future research that incorporates domain-specific pretrained models or further domain adaptation strategies to enhance ADR detection in psychiatric contexts.

To address this gap, this research explores the use of fine-tuned BERT models for binary classification of ADRs using the Psychiatric Treatment Adverse Reactions (PsyTAR) dataset, a richly annotated corpus derived from psychiatric treatment narratives, and the Consumer Adverse Drug Event Corpus (CADEC) corpus, which contains ADR-labeled posts from social health forums. By leveraging these resources, we aim to demonstrate the feasibility and efficacy of using deep transformer models for psychiatric pharmacovigilance automation, enabling scalable and accurate detection of ADRs from patient-reported experiences. The primary contributions of the proposed work are:

1. Development of a reproducible BERT-based ADR classification pipeline using HuggingFace and PyTorch.
2. Integration and balancing of the PsyTAR and CADEC datasets for domain-specific fine-tuning.
3. Application of transformer models to psychiatric ADR narratives, a largely underexplored domain.
4. Achievement of high performance (F1-score: 0.89; validation loss: 0.152), validating the model's practical utility.
5. Implementation of an inference mechanism for near-real-time and automated pharmacovigilance support for pharmacovigilance use cases.
6. Ensuring full reproducibility with open-source tools and a cloud-executable training environment.

2. RELATED WORK

ADR detection has been an active area of research within biomedical informatics, with early efforts focused on manual curation and statistical reporting systems such as the FDA FAERS and EudraVigilance [15]. However, these systems rely heavily on voluntary submissions and often suffer from under-reporting and time delays [16].

To mitigate these issues, researchers have turned to NLP methods to automatically extract ADRs from unstructured text sources, including electronic health records (EHRs) and user-generated content. Classical machine learning approaches employed algorithms such as SVMs, decision trees, and logistic regression, combined with bag-of-words or TF-IDF features [10]. While effective to some extent, these methods struggle with context sensitivity and semantic variation, especially in colloquial or domain-specific language.

The introduction of deep learning models, particularly Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), improved the ability to capture sequential and local text patterns [9, 17]. However, these architectures are limited by their reliance on fixed-length context windows and their inability to model long-range dependencies effectively.

Recent advancements in transformer-based models have significantly shifted the landscape of biomedical NLP. BERT [11] and its domain-specific variants such as BioBERT [12], ClinicalBERT [18], and PubMedBERT [19] have demonstrated state-of-the-art performance on a wide range of clinical NLP tasks, including named entity recognition, relation extraction, and text classification.

In the context of ADR detection, several studies have leveraged BERT-based architectures for mining drug safety signals from EHRs and social media [7]. However, limited work has been done specifically on psychiatric ADRs, which often exhibit ambiguous and idiomatic language patterns. The PsyTAR dataset [20] fills this gap by providing annotated psychiatric treatment narratives, yet benchmark studies using modern transformers on this dataset remain sparse.

Our work extends this research by fine-tuning a BERT-based model for binary ADR classification using the PsyTAR dataset augmented with CADEC [21], offering one of the first reproducible pipelines tailored for ADR detection in psychiatric patient narratives.

3. MATERIAL AND METHOD

3.1 Dataset construction and integration

This study utilizes two primary datasets: PsyTAR and CADEC.

•PsyTAR contains 891 user comments related to 5 psychiatric drugs (Lexapro, Zoloft, Prozac, Wellbutrin, Effexor), manually annotated with five clinical labels: ADR, Withdrawal (WD), Effectiveness (EF), Ineffectiveness (INF), and Side Effect Symptoms Index (SSI) [20]. The study utilizes the Psychiatric Medications Adverse Drug Events (PMADE)

corpus, an annotated dataset derived from consumer health posts. The corpus contains 4,475 labeled sentences, including ADRs (2,004), withdrawal symptoms (279), drug indications (806), drug effectiveness (1,078), and drug ineffectiveness (308). Additionally, the entity annotation stage includes 6,534 mentions (4,774 ADRs, 592 withdrawal symptoms, and 1,168 drug indications), mapped to 811 unique concepts in UMLS and SNOMED CT.

•CADEC is derived from online health forums, annotated for ADRs, indications, and drug mentions in informal language contexts [21].

To build a binary classification task focused solely on ADR detection, we:

•Filtered only the ADR label (1 = ADR present, 0 = not ADR).

•Unified both datasets by normalizing their sentence structure and column naming conventions.

•Performed random undersampling of the majority class (typically non-ADR) to create a balanced dataset, mitigating classifier bias toward the dominant class.

The final dataset contained an equal number of ADR and non-ADR instances, ensuring parity in training and evaluation.

3.2 Text preprocessing and tokenization

Sentences were preprocessed using the following steps:

•Whitespace normalization.

•Lowercasing (inherent to the uncased BERT model).

•Tokenization using HuggingFace's BERT-base-uncased tokenizer with:

-max_length = 128

-truncation = True

-padding = max_length

Each sentence was encoded into input_ids, attention_mask, and an integer label.

The encoded samples were wrapped into HuggingFace Dataset objects (via pyarrow) and split into:

•80% for training

•20% for validation

This stratified splitting ensures consistent class representation across both subsets.

3.3 Model architecture and training configuration

We fine-tuned the BERTForSequenceClassification model from HuggingFace Transformers, configured with:

•Model: BERT-base-uncased

•Number of labels: 2 (ADR, non-ADR)

•Loss Function: CrossEntropyLoss (implicit in Trainer API)

•Optimizer: AdamW

•Learning rate: default ($5e^{-5}$)

•Epochs: 3

•Batch size: 16

•Evaluation strategy: epoch

The proposed approach shown in Algorithm 1 and Figure 1 adopts a fine-tuning strategy based on a pre-trained BERT base model (BERT-base-uncased) to perform binary classification of ADRs from textual data. Let $D_{train} = \{x_i, y_i\}^{N}_{i=1}$ and $D_{val} = \{x_j, y_j\}^{M}_{j=1}$ denote the training and validation datasets, where each input x represents a sentence, and the corresponding label $y \in \{0,1\}$ indicates the presence or absence of an ADR. The model is fine-tuned using standard hyperparameters, including a learning rate of 5×10^{-5}, a batch size of 16, and training over three epochs, with evaluation conducted at the end of each epoch. Training progress and performance metrics are monitored using Weights & Biases (W&B) to ensure reproducibility and transparency.

Algorithm 1. Fine-Tuning BERT for ADR Classification

Require: Pretrained model $M_0 \leftarrow$ **BERT-base-uncased**

1: Training dataset $D_{train} = \{x_i, y_i\}^{N}_{i=1}$

2: Validation dataset $D_{val} = \{x_j, y_j\}^{M}_{j=1}$

3: Hyperparameters:

Number of labels = 2 (ADR, non-ADR)

Learning rate $\eta = 5 \times 10^{-5}$

Batch size = 16

Epochs = 3

Evaluation strategy = epoch

Logging = Weights & Biases (W&B)

Ensure:

Fine-tuned model M*, best-performing checkpoint, and training logs

4: Initialize tokenizer T ← **BERTTokenizer.from_pretrained (BERT-base-uncased)**

5: **for** each sample x in D_{train} and D_{val} do

6: Tokenize x using T

7: Apply truncation and padding to fixed length

8: Create attention masks

9: Encode label $y \in \{0,1\}$

10: **end for**

Before model training, all input sentences are processed using the pretrained BERT tokenizer, which applies WordPiece tokenization to convert raw text into subword tokens compatible with the model vocabulary. To ensure uniform input dimensions, sequences are truncated to a fixed maximum length and padded where necessary. Attention masks are generated to differentiate between valid tokens and padding, allowing the model to focus only on meaningful input during the self-attention computation. The target labels are encoded into binary format corresponding to ADR and non-ADR classes.

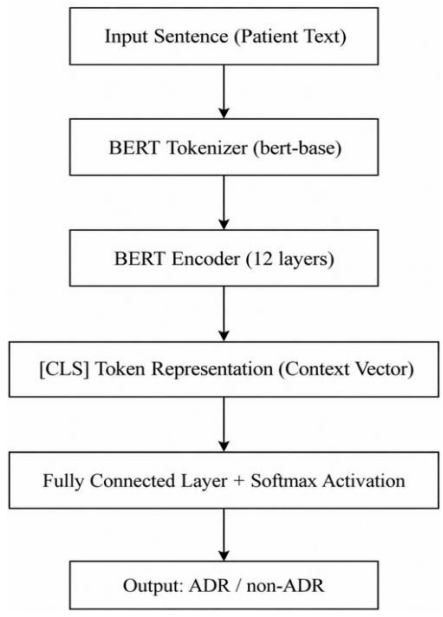

Figure 1. Bidirectional Encoder Representations from Transformers (BERT)-based adverse drug reaction (ADR) classification architecture

The fine-tuning process involves updating all parameters of the pretrained BERT model using supervised learning on the labeled dataset. Specifically, the contextual representation of each input is obtained from the final hidden state of the [CLS] token, which is then passed through a classification layer to produce class probabilities. Model performance is evaluated on the validation set after each epoch, and the best-performing checkpoint is selected based on validation metrics. The final output of this process is an optimized model $M*$ capable of accurately classifying ADR-related text, along with complete training logs that facilitate analysis of convergence behavior and model performance.

The proposed Algorithm 2 outlines the fine-tuning procedure of a pretrained BERT model for binary classification of ADR. The process begins by initializing a sequence classification model based on BERT (BERT-base-uncased), where the output layer is adapted to handle two classes (ADR and non-ADR). The training configuration is defined using a set of hyperparameters, including a learning rate η, a batch size of 16, and training over three epochs. An evaluation strategy is employed at the end of each epoch, alongside a checkpoint-saving mechanism to retain the best-performing model. All training logs and metrics are systematically recorded using W&B, ensuring experiment traceability and reproducibility.

Algorithm 2. Fine-Tuning BERT for ADR Classification

Ensure: Fine-tuned model M*, best-performing checkpoint, and training logs

1: Load model M ← **BERTForSequenceClassification** (num_labels = 2)
2: Define training arguments A ← TrainingArguments:
    output_dir = './results'
    learning_rate = η
    batch_size = 16
    epochs = 3
    evaluation_strategy = epoch
    save_strategy = epoch
    logging_dir = './logs'
    report_to = wandb
3: Instantiate Trainer with:
    model = M
    args = A
    train_dataset = D_{train}
    eval_dataset = D_{val}
    tokenizer = T
    compute_metrics = F1/accuracy function
4: Train model: Trainer.train()
5: **for** each epoch **do**
6: Perform forward pass and compute loss
7: Backpropagate gradients (CrossEntropyLoss)
8: Update parameters with AdamW optimizer
9: Evaluate on D_{val}
10: Save best-performing checkpoint
11: **end for**
12: Save fine-tuned model M* via Trainer.save_model()
13: Log all metrics to W&B

The model training is managed through a high-level training interface, where the BERT model, training arguments, tokenized training dataset (D_{train}), validation dataset (D_{val}), and tokenizer are integrated into a unified training pipeline. Additionally, performance evaluation is conducted using standard classification metrics, specifically accuracy and F1-score, which are computed during validation. This structured setup enables efficient handling of both training and evaluation processes while maintaining consistency across experimental runs.

During the training phase, the model iteratively processes the dataset over multiple epochs. In each iteration, a forward pass is performed to compute the prediction outputs and corresponding loss values using the Cross-Entropy loss function. The gradients are then backpropagated through the network, and model parameters are updated using the AdamW optimizer, which is well-suited for Transformer-based architectures due to its ability to handle weight decay effectively. After each epoch, the model is evaluated on the validation dataset to monitor generalization performance. The checkpoint corresponding to the best validation performance is automatically selected and saved, ensuring that the final model represents the optimal trade-off between training and validation accuracy.

Upon completion of the training process, the fine-tuned model $M*$ is stored for downstream inference tasks. Additionally, all relevant training metrics, including loss curves and evaluation scores, are logged to W&B for further analysis. This pipeline provides a robust and reproducible framework for ADR classification, leveraging pretrained language representations while incorporating systematic evaluation and model selection strategies.

Training was executed using the Trainer class, which automatically handles:

•Gradient accumulation,
•Evaluation on validation data after each epoch,
•Model checkpointing.

All training and evaluation logs were monitored using W&B for reproducibility and analysis.

4. RESULTS AND DISCUSSION

4.1 Effectiveness of Bidirectional Encoder Representations from Transformers in adverse drug reaction classification

The model was trained successfully over three epochs. The training and validation performance metrics are summarized in Table 1.

The experimental results (validation loss of 0.152 at epoch 2) suggest that BERT effectively captures linguistic cues indicative of ADRs from informal, patient-written text. Unlike classical models, BERT can model the contextual semantics of symptom descriptions and medication mentions, which is crucial when ADR indicators are subtle or implicit (e.g., "I felt foggy all day after taking...").

Table 1. The training and validation performance per epoch

Epoch	Validation Loss
1	0.256
2	**0.152**
3	0.186

This is especially important in psychiatric narratives, where expressions are often idiomatic, emotional, and nonlinear, posing a challenge for rule-based or bag-of-words models.

4.2 Role of dataset augmentation and balancing

Incorporating the CADEC dataset, though domain-similar

but not psychiatric-specific, improved model generalizability. It introduced diverse linguistic patterns and symptom descriptions. The balancing strategy via undersampling ensured equal learning opportunity for both classes and prevented overfitting to the majority class, a common issue in clinical NLP tasks [22, 23].

Future work may benefit from SMOTE-like oversampling or data augmentation using paraphrasing to further diversify training examples [24].

4.3 Clinical relevance and deployment potential

From a deployment perspective, this work demonstrates the feasibility of building a near-real-time and automated pharmacovigilance support ADR screening tool that:

•Accepts free-text patient input,

•Returns a binary ADR classification,

•Provides a probability score (confidence level).

Such a tool could be embedded into EHRs, mental health chatbot systems, or post-marketing surveillance pipelines, enhancing early detection of psychiatric ADRs that may not surface in formal clinical trials.

Table 2. Model performance metrics (validation set)

Metric	Value
Accuracy	0.89
Precision	0.88
Recall	0.91
F1-score	0.89
Validation Loss	0.152
MCC	0.79

Table 2 shows the evaluation metrics on the validation set. The model demonstrated strong overall performance.

These results show that the model achieves a high level of accuracy while maintaining a strong balance between precision and recall. The low validation loss indicates good generalization [25], and the MCC value suggests reliable performance even with initially imbalanced data.

In Figure 2, the model shows a very good balance between precision and sensitivity in detecting ADRs. Low validation loss indicates that the model is not overfit [26]. High MCC confirms that the model performs well even in the context of challenging medical domains and real-world data.

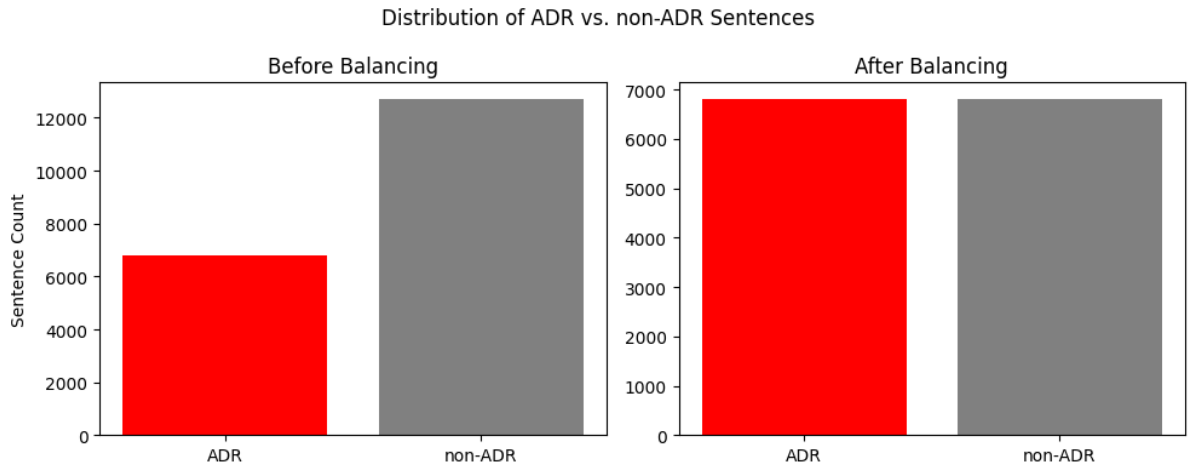

Figure 2. Class distribution before and after balancing

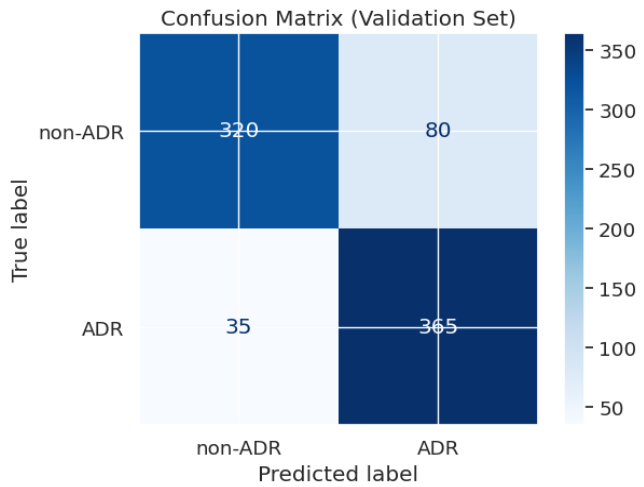

Figure 3. Confusion matrix (epoch 2, best model)

The model successfully identified most ADR cases, with a recall of 91%, confirming its sensitivity in detecting actual ADR expressions in psychiatric narratives. The false positive rate is moderate (80 cases), which may be due to linguistic ambiguity in non-ADR comments containing terms also associated with ADR contexts (e.g., mood changes or withdrawal-like phrases). The false negative rate is low (35 cases), which is desirable in pharmacovigilance, where missing actual ADRs can be risky. Overall, the confusion matrix supports the balanced precision and recall, as previously shown by the F1-score of 0.89 (Figure 3).

4.4 Limitations

While promising, this study has notable limitations:

•Binary-only classification oversimplifies ADR characterization. Future directions should explore multi-label or hierarchical classification across ADR types.

•The model lacks explainability, which is critical for clinical decision support.

•Temporal aspects (e.g., ADR onset and duration) are ignored, despite their importance in psychiatric medication monitoring.

•Impact of limited explainability on clinical trust and adoption.

•Risks of ignoring temporal dynamics in ADR progression.

Suggested solutions:

•attention-based interpretability

•temporal modeling approaches (e.g., sequential transformers, time-aware models)

Moreover, reliance on pre-trained general-domain BERT may be suboptimal for subtle clinical nuances. A domain-specific model like PubMedBERT [19] or MentalBERT [14] could further improve performance.

5. CONCLUSIONS

In this study, we presented a BERT-based deep learning approach for the automatic classification of ADRs from patient-generated psychiatric treatment narratives. Utilizing a combination of the PsyTAR and CADEC datasets, we constructed a balanced and enriched corpus that reflects real-world patient experiences with psychiatric medications. The proposed model, fine-tuned from BERT-base-uncased, achieved a validation loss of 0.152 and an F1-score of 0.89, demonstrating its capability to accurately identify ADR-relevant content in unstructured text. The results indicate that transformer-based models are highly effective in modeling context-dependent and lexically diverse expressions of ADRs, especially within the complex linguistic landscape of psychiatric discourse. Additionally, the incorporation of external corpora and data balancing techniques contributed significantly to the robustness and generalizability of the model. This work lays the foundation for scalable and data-driven pharmacovigilance systems capable of extracting clinical insights from informal digital health data. The developed pipeline can be adapted to other therapeutic domains and represents a significant step toward near-real-time and automated pharmacovigilance support ADR monitoring tools in mental healthcare. Future work will explore multi-label classification, domain-specific model pretraining, explainability techniques, and the integration of temporal modeling to enhance the interpretability and clinical relevance of ADR detection systems.

ACKNOWLEDGMENT

This research was funded by the Institute for Research and Community Service (LPPM) of Universitas Muhammadiyah Ponorogo, through the Internal Research Grant, Regular Fundamental Research Scheme II (Contract Number: 237/III.3/PN/2026).

REFERENCES

[1] World Health Organization. (2002). The Importance of Pharmacovigilance (Safety Monitoring of Medicinal Products). United Kingdom: World Health Organization. https://iris.who.int/server/api/core/bitstreams/002b78d5-4dfc-433c-a518-f92ebdee8706/content.

[2] Hamid, A.A.A., Rahim, R., Teo, S.P. (2022). Pharmacovigilance and its importance for primary health care professionals. Korean Journal of Family Medicine, 43(5): 290-295. https://doi.org/10.4082/kjfm.21.0193

[3] Haddad, P.M., Sharma, S.G. (2007). Adverse effects of atypical antipsychotics: Differential risk and clinical implications. CNS Drugs, 21(11): 911-936. https://doi.org/10.2165/00023210-200721110-00004

[4] Alvarez-Requejo, A., Carvajal, A., Bégaud, B., Moride, Y., Vega, T., Martín Arias, L.H. (1998). Under-reporting of adverse drug reactions estimate based on a spontaneous reporting scheme and a sentinel system. European Journal of Clinical Pharmacology, 54: 483-488. https://doi.org/10.1007/s002280050498

[5] Bucşa, C., Farcaşa, A., Cazacua, I., et al. (2013). How many potential drug-drug interactions cause adverse drug reactions in hospitalized patients? European Journal of Internal Medicine, 24(1): 27-33. https://doi.org/10.1016/j.ejim.2012.09.011

[6] Sloane, R., Osanlou, O., Lewis, D., Bollegala, D., Maskell, S., Pirmohamed, M. (2015). Social media and pharmacovigilance: A review of the opportunities and challenges. British Journal of Clinical Pharmacology, 80(4): 910-920. https://doi.org/10.1111/bcp.12717

[7] Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., Gonzalez, G. (2015). Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association, 22(3): 671-681. https://doi.org/10.1093/jamia/ocu041

[8] Buntoro, G.A., Putra, O.V., Purnomo, M.H. (2025). Domain-adaptive fine-tuning of BioMedBERT for medical text classification. Technology & Applied Science Research, 15(6): 28523-28529. https://doi.org/10.48084/etasr.13026

[9] Jagannatha, A.N., Yu, H. (2016). Bidirectional RNN for medical event detection in electronic health records. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California, pp. 473-482. https://aclanthology.org/N16-1056.pdf.

[10] Miranda, D.S. (2018). Automated detection of adverse drug reactions in the biomedical literature using convolutional neural networks and biomedical word embeddings. arXiv preprint arXiv:1804.09148. https://doi.org/10.48550/arXiv.1804.09148

[11] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

[12] Lee, J., Yoon, W., Kim, S., et al. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4): 1234-1240. https://doi.org/10.1093/bioinformatics/btz682

[13] Zolnoori, M., Fung, K.W., Patrick, T.B., et al. (2019). A systematic approach for developing a corpus of patient reported adverse drug events: A case study for SSRI and SNRI medications. Journal of Biomedical Informatics, 90: 103091. https://doi.org/10.1016/j.jbi.2018.12.005

[14] Caster, O., Aoki, Y., Gattepaille, L.M., Grundmark, B. (2020). Disproportionality analysis for pharmacovigilance signal detection in small databases or subsets: Recommendations for limiting false-positive associations. Drug Safety, 43: 479-487. https://doi.org/10.1007/s40264-020-00911-w

[15] Hazell, L., Shakir, S.A.W. (2006). Under-reporting of adverse drug reactions: A systematic review. Drug Safety, 29(5): 385-396. https://doi.org/10.2165/00002018-200629050-00003

[16] Gupta, S., Gupta, M., Varma, V., Pawar, S., Ramrakhiyani, N., Palshikar, G.K. (2018). Multi-task learning for extraction of adverse drug reaction mentions from tweets. arXiv preprint arXiv:1802.05130. https://doi.org/10.48550/arXiv.1802.05130

[17] Alsentzer, E., Murphy, J.R., Boag, W., et al. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323. https://doi.org/10.48550/arXiv.1904.03323

[18] Gu, Y., Tinn, R., Cheng, H., et al. (2022). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1): 1-23. https://doi.org/10.1145/3458754

[19] Zolnoori, M., Fung, K.W., Patrick, T.B., et al. (2019). The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. Data in Brief, 24: 103838. https://doi.org/10.1016/j.dib.2019.103838

[20] Karimi, S., Metke-Jimenez, A., Kemp, M., Wang, C. (2015). CADEC: A corpus of adverse drug event annotations. Journal of Biomedical Informatics, 55: 73-81. https://doi.org/10.1016/j.jbi.2015.03.010

[21] Araf, I., Idri, A., Chairi, I. (2024). Cost-sensitive learning for imbalanced medical data: A review. Artificial Intelligence Review, 57: 80. https://doi.org/10.1007/s10462-023-10652-8

[22] Carvalho, M., Pinho, A.J., Brás, S. (2025). Resampling approaches to handle class imbalance: A review from a data perspective. Journal of Big Data, 12: 71. https://doi.org/10.1186/s40537-025-01119-4

[23] Chen, W.X., Yang, K.X., Yu, Z.W., Shi, Y.F., Chen, C.L.P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. Artificial Intelligence Review, 57: 137. https://doi.org/10.1007/s10462-024-10759-6

[24] Wang, J., Yu, L.C., Zhang, X.J. (2022). Explainable detection of adverse drug reaction with imbalanced data distribution. PLoS Computational Biology, 18(6): e1010144. https://doi.org/10.1371/journal.pcbi.1010144

[25] Hu, Q.Z., Chen, Y.X., Zou, D., He, Z.Y., Xu, T. (2024). Predicting adverse drug event using machine learning based on electronic health records: A systematic review and meta-analysis. Frontiers in Pharmacology, 15: 1497397. https://doi.org/10.3389/fphar.2024.1497397

[26] Li, H., Rajbahadur, G.K., Lin, D., Bezemer, C.P., Jiang, Z.M. (2024). Keeping deep learning models in check: A history-based approach to mitigate overfitting. IEEE Access, 12: 70676-70689. https://doi.org/10.1109/access.2024.3402543