Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Identification of Causative Drugs for Adverse Events Using BioClinicalBERT and BioGPT on Social Media

Brahami Menaouer

menaouer.brahami@enp-oran.dz

École Nationale Polytechnique d'Oran

Abdeldjouad Fatma Zahra

École Nationale Polytechnique d'Oran

Hadj Benaïchouche Meroua

École Nationale Polytechnique d'Oran

Research Article

Keywords: Adverse Drug Events (ADEs), Large Language Models (LLMs), Natural Language Processing (NLP), BioClinicalBERT, BioGPT, Drug Comments, Text Classification, Pharmacovigilance

Posted Date: July 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10181960/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Identification of Causative Drugs for Adverse Events Using BioClinicalBERT and BioGPT on Social Media

Brahami Menaouer [0000-0003-0045-9797], Abdeldjouad Fatma Zahra[2], and Hadj Benaïchouche Meroua[3]

[1] Computer Systems Engineering Department, National Polytechnic School of Oran, LABAB Laboratory
menaouer.brahami@enp-oran.dz
[2] Computer Systems Engineering Department, National Polytechnic School of Oran, LABAB Laboratory
fatma-zahra.abdeldjouad@enp-oran.dz
[3] Computer Systems Engineering Department, National Polytechnic School of Oran
meroua.hadj-benaichouche@etu.enp-oran.dz

Correspondence should be addressed to Brahami Menaouer at menaouer.brahami@enp-oran.dz
Brahami Menaouer - ORCID: 0000-0003-0045-9797

Abstract. The occurrence of Adverse Drug Events (ADEs) has become a serious social concern for public health and the economy. Early detection of ADEs can lower the risks associated with drug safety and reduce medical expenditures. Various texts, such as drug reviews, clinical texts, and user posts on social media and medical forums, contain substantial information regarding ADEs. However, several transformer-based techniques utilize Natural Language Processing (NLP) to automate ADEs detection from explicit medical text. Similarly, Large Language Models (LLMs) have shown promise in many NLP tasks. In this study, we propose a novel approach that utilizes advanced LLMs, including BioClinicalBERT and BioGPT, for biomedical text classification to automatically identify specific drugs associated with ADEs reported in social media datasets. These classifiers were applied to a combination of drug-related comments extracted from the ADE-Corpus-V2 and PsyTAR "Drug Reviews" datasets. Experimental results from the fusion dataset indicated that the evaluation of individual advanced LLMs achieved an F1 Score of 98% and an accuracy of 98%, thereby demonstrating their effectiveness in identifying drugs associated with adverse events. It is important to use joint LLM techniques and external resources to effectively detect and classify causative drugs for adverse events from clinical narratives, offering a real-time solution for monitoring drug safety and supporting timely clinical interventions.

Keywords: Adverse Drug Events (ADEs), Large Language Models (LLMs), Natural Language Processing (NLP), BioClinicalBERT, BioGPT, Drug Comments, Text Classification, Pharmacovigilance

1 Introduction

Over the last few years, social media has become a prominent platform for discussing health problems, making it a popular source for sharing information about ADR in natural language. Otherwise, drug prescription is a task that doctors face daily with each patient. However, they must remain vigilant about all potential side effects associated with the drugs they prescribe. According to the American Society of Pharmacovigilance (ASP), Adverse Drug Events (ADEs) are common and account for ~250,000 injuries and deaths each year, and drug interactions account for as much as 30% of these ADEs. Adverse drug events (ADEs) pose a significant global public health challenge, impacting both patients and healthcare systems. According to Murphy et al. (2023) [1], ADEs represent a significant medical problem in healthcare, largely due to the increasing prevalence of multimorbidity and the complexity of medical treatment. Many researchers have used other resources to identify ADEs, such as biomedical publications [2; 3], social media [4; 5], and electronic health record (EHR) notes [6; 7].

The use of computers to predict ADEs is critical in pharmacology, as it reduces drug risks and lowers medical costs. Moreover, the use of Artificial Intelligence (AI) techniques, primarily Natural Language Processing (NLP), has made ADEs' predictions more accurate and efficient when handling large textual datasets. In addition, the classification of text in a biomedical context represents a fundamental challenge in the field of NLP, particularly in the pharmacovigilance domain. Similarly, AI enhances the technological aspects of all fields, which affects human life. Some previous research has attempted to extract information on medications, ADEs, and their interrelations utilizing NLP techniques. To improve medication safety, hospitals and clinics must have accurate and

continuous insight into the types of ADEs that occur in their inpatients, including which subpopulations are at a high risk of ADEs. For [8; 9], a wide variety of NLP methods have been previously explored to discover relationships between drugs and symptoms in EHRs, as well as to learn about potential risks from biomedical literature. Likewise, AI technologies, including NLP, offer promising solutions to automate and enhance pharmacovigilance processes.

To address this problem, NLP techniques serve as a potential solution for the precise and automated detection of ADEs from diverse textual sources, which makes it possible to build a sustainable pharmacovigilance system in medical practice. Moreover, NLP has shown promising results for ADE detection in the context of pharmacovigilance. NLP models can significantly improve the performance of ADE-related information extraction. These technological advancements are revolutionizing how we understand drug safety, thereby enhancing the efficiency and proactivity of pharmacovigilance.

In recent years, Large Language Models (LLMs) have shown profound achievements in the field of NLP. These transformer-based models are generally categorized into three main categories. The decoder-only (e.g., GPT) and the encoder-only (e.g., BERT) models. LLMs are sophisticated automated systems that have been trained on extensive corpora of text [10]. These models are designed to comprehend and generate language in a manner that closely resembles human linguistic behavior, thereby enabling them to execute a diverse array of language-related tasks. According to [11], LLM is a sub-branch of NLP that focuses on understanding, generating, and manipulating text for various language-related tasks. Furthermore, integrating LLM output into clinical workflows may help clinicians prevent harm by alerting them to emerging safety signals or prescribing risks.

In contrast, various texts, such as biomedical literature, drug reviews, and user posts on social media and medical forums, contain a wealth of information on ADEs. Recent studies have used word embedding and AI-based natural language processing to automate ADE detection in texts. However, they did not explore the inclusion of explicit medical knowledge about drugs and adverse events or related feature learning. Thus, Generative AI (GenAI) is a field of AI that focuses on creating models able to generate new and original content. Moreover, GenAI empowers healthcare professionals with enhanced diagnostic capabilities, especially personalized treatment plans for ADEs detection and improved patient outcomes. Otherwise, GenAI and LLMs may enable novel approaches that were previously unfeasible with conventional methods. In this study, we propose a novel approach that uses advanced LLMs, including BioClinicalBERT and BioGPT, to classify biomedical text and automatically identify specific drugs associated with ADEs reported in social media datasets. These classifiers were applied to a combination of drug-related comments extracted from the ADE-Corpus-V2 and PsyTAR "Drug Reviews" datasets. Our objective is to conduct experiments on two ADE datasets and answer the following research questions:

- *RQ1*: How do the proposed advanced language models perform in identifying causative drugs for adverse events in texts from various sources compared with other methods?
- *RQ2*: How do BioClinicalBERT and BioGPT with knowledge augmentation enhance the accuracy of identifying causative drugs for adverse events?
- *RQ3*: What is the impact of pretraining on domains and contextualized language representation on the performance of the method in identifying causative drugs for adverse events?
- *RQ4*: How can the proposed advanced language models be compared?

Primarily, our dataset focuses on patient drug reviews and ADE-Corpus-V2 from recent years, analyzing patient comments from around the world to identify the causative drugs for adverse events. This dataset was used to train our advanced language models (BioClinicalBERT and BioGPT) to identify and classify drug-associated adverse events. This study aimed to discover patients' opinions and emotional reactions regarding ADEs and to reveal how they expressed their ADE opinions. The main contributions of the proposed models are as follows:

- Data preprocessing was performed by eliminating punctuation, removing non-alphanumeric characters, removing stop words, and converting text to lowercase. The collected Drug-Corpus-V2 and PsyTAR (Psychiatric Treatment Adverse Reactions) drug review datasets were used. In the context of ADEs identification and classification, preprocessing enhances the overall accuracy of advanced language models by effectively organizing and cleaning the data.
- To enhance the caliber of our dataset, we applied data fusion techniques (Pandas DataFrames) and data augmentation (Text Paraphrasing "T5 Paraphraser"), carefully cleaned, and data split into training, validation, and testing sets.
- We experimented with pre-trained advanced language models to identify causative drugs for adverse events from patient drug reviews and the ADE corpus on social media. We demonstrated better performance than existing artificial intelligence models.
- We aim to make our dataset publicly available for further research, more specifically, in the context of identifying causative drugs for adverse events by improving the current state of knowledge regarding the demographic factors that identify ADEs occurrence.

The rest of the paper is organized as follows: Section 2 introduces the related work review on text classification techniques proposed by various researchers. Section 3 presents the proposed advanced language models used for identifying causative drugs for adverse events from patient drug reviews (PsyTAR) and drug comments on social media, and the experimental results and discussion are covered in Section 4. Finally, Section 5 provides the proposed advanced language models' conclusion, limitations, and planning for future work.

2 Related Work

In recent years, significant attention has been directed towards automating the detection of information pertinent to adverse drug events (ADEs) across diverse data sources. For instance, Khemani et al. [6] proposed a hybrid AI-driven framework that integrates structured and unstructured data to detect ADRs using advanced deep learning and NLP methods. The objective is to outperform traditional signal detection methods and provide interpretable predictions to aid clinicians in real-time. By leveraging advanced Machine Learning (ML) and Deep Learning (DL) techniques, including Random Forests, Gradient Boosting Machines, and Convolutional Neural Networks (CNNs), this model aims to identify potential ADRs across different patient subgroups. Another approach was developed by Ou et al. [12], a predictor that identifies ADRs early in drug discovery, using a DL model designed to fuse multiple drug characteristics. This approach employed four modules to extract one- and two-dimensional sequence structure information of the drug, molecules, drug–protein interaction data, and drug similarity. A fusion model integrated these characteristics to predict the precise probability of ADRs. The study by Geo et al. [5] was conducted to propose a graph-based model with Knowledge Augmented Concept-Aware Graph Embeddings (KnowCAGE). This model utilizes knowledge-augmented contextualized graph embeddings to capture relations for ADE detection. This model leverages medical knowledge from the Unified Medical Language System and proposes a concept-aware attention mechanism that learns distinct features for different types of nodes in the graph. In another study, Dong et al. [13] developed a BERT-based language model tailored to identifying drug adverse events. This model utilized publicly available labeled adverse event data from the ADE-Corpus-V2. Constructing the BERT-based model involved optimizing key hyperparameters, such as the number of training epochs, batch size, and learning rate. Another research study by Sahoo et al. [4] presents a Multi-Modal Adverse Drug Event (MMADE) detection dataset, merging ADE-related textual information with visual aids. The authors introduced a framework that leverages the capabilities of LLMs and VLMs for ADE detection by generating detailed descriptions of medical images depicting ADEs, aiding healthcare professionals in visually identifying adverse events. In the study by [14], a label propagation model was proposed for generating enhanced drug safety signals by utilizing multiple drug features. The authors first constructed multiple drug similarity networks using a range of drug features. They then calculated initial drug safety signals using conventional signal detection algorithms (BCPNN and MGPS). In the study of [1], NLP was applied to detect ADEs in clinical narratives within electronic health records of inpatients. Quantitative and qualitative data items relating to NLP methods were extracted and critically appraised. Out of 1,065 articles screened for eligibility, 29 articles met the inclusion criteria. Moreover, Roosan et al. [15] developed the AI-based NLP algorithm, Tarantula, to automate data extraction from the three patient forums. This context-aware NLP algorithm uses a pre-trained ML algorithm of FastText embeddings to extract contextual information from the database using corpus-based lexicons. Another study [16] presented the design of the ADEC module for classification tasks, where basic machine learning (ML) and deep learning (DL) techniques, such as logistic regression (LR), decision tree (DT), and text-based convolutional neural network (TextCNN), are employed. To extract features from the text data, the TF-IDF and Word2Vec models are employed. The study by [17] combined bidirectional long short-term memory (BiLSTM) and conditional random fields (CRF) neural networks to detect medical entities relevant to ADEs and a combined BiLSTM and attention network to determine relations, including the adverse drug reaction relation between medication and sign or symptom entities. Another study, Wei et al. [18], developed deep learning approaches for extracting medications and their attributes from clinical text and demonstrated their superior performance compared with traditional ML algorithms. Table 1 below presents a comprehensive summary of all literature reviews pertinent to our study.

Table 1. Comparative Analysis of ADEs Detection Models Used in Different Papers

Authors	Data Source	Methods/Techniques	Research Gap
Khemani et al. (2025) [6]	Clinical trial datasets	Advanced ML and DL techniques, including random forests, gradient boosting machines, CNNs, BERT, and GPT, are used to identify potential ADRs.	The models' performance is also highly dependent on the quality and diversity of the training data, and any biases present in the data may be reflected in the model's predictions.

Ou et al. (2025) [12]	SMILES string data, ADRs data, and drug protein interaction data	DL methods: transformer encoder and Graph Attention Network–Graph Convolutional Network Combination (GAT GCN)	The generalization ability by incorporating additional clinical data dimensions, such as genomic information and environmental factors
Geo et al. (2024) [5]	SMM4H, TheMedPub, TwoMad-Twitter, and CADAC	Proposed a model named KnowCAGE (Knowledge-augmented Concept-Aware Graph Embeddings)	Limited in capturing rich contextual information and relational reasoning
Dong et al. (2024) [13]	ADE-Corpus-V2 data	BERT-based language model tailored to identifying DAEs	Ensure the model's utility in clinical settings and real-world applications
Sahoo et al. (2024)[4]	PSB2016 (Twitter), ADE corpus, CADEC, and BioDEX dataset	Multi-Modal Adverse Drug Event (MMADE) detection using the capabilities of LLMs and VLMs	Data privacy concerns, regulatory constraints, and the specialized nature of ADE occurrences.
Padhi et al. (2024) [14]	FAERS DrugBank SIDER	A label propagation framework to generate enhanced drug safety signals for the early detection of ADEs.	Results must be compared with the state-of-the-art to show how the approach is more efficient than other existing techniques and datasets.
Murphy et al. (2023) [1]	MEDLINE EMBASE arXiv	NLP methods for ADE detection in the context of ADE monitoring in hospitals are lacking	Including clinicians as team members in NLP-based ADE detection should become standard practice.
Roosan et al. (2022) [15]	FAERS Social Media UMLS	Tarantula utilized FastText embeddings and an aggregated lexicon to extract contextual data	The number of social media databases used for gathering data.
Rawat et al. 2022 [16]	ADE dataset Pubmed dataset	Drug Adverse Event Detection Using Text-Based Convolutional Neural Networks (TextCNN)	- Entity recognition (MNER) - Multi-model text summarization (MMTS)
Dandala et al. 2019 [17]	1089 de-identified clinical notes of 21 patients with cancer	Detect medical entities relevant to ADEs by a combined BiLSTM and conditional random fields (CRF) neural network	- Incorporating external knowledge in concept and relation extraction. - Use of domain (EHR)-adapted dependency parsers
Wei et al. 2019 [18]	n2c2 corpus from the MIMIC-III (Medical Information Mart for Intensive Care III)	Extraction of medications and associated ADEs from clinical documents	- Integrating existing medical data sets/knowledge - Reduce recognition errors

As aforementioned in these studies, healthcare providers have accordingly attempted to use artificial intelligence techniques designed to detect errors in time to prevent ADEs, which can ultimately reduce the soaring cost of healthcare. Based on these related studies, many researchers have applied advanced language models to identify causative drugs for adverse events. To overcome these problems and limitations, we propose a novel approach that utilizes advanced LLMs, including BioClinicalBERT and BioGPT, for biomedical text classification to automatically identify specific drugs associated with ADEs reported in social media datasets. These classifiers were applied to a combination of drug-related comments extracted from the ADE-Corpus-V2 and PsyTAR "Drug Reviews" datasets.

3 Proposed Advanced Language Models

Based on our objectives and motivations, this study is grounded in several background ideas and research efforts, as shown in Figure 1. In particular, using advanced language models to identify specific drugs associated with ADEs reported in patient drug comments/review datasets and supporting them with text processing is a remarkable idea. A general, widely followed text multi-classification approach, performed using the BioClinicalBERT and BioGPT models, was applied to the drug comment social media and PsyTAR drug reviews dataset (see Fig. 1).

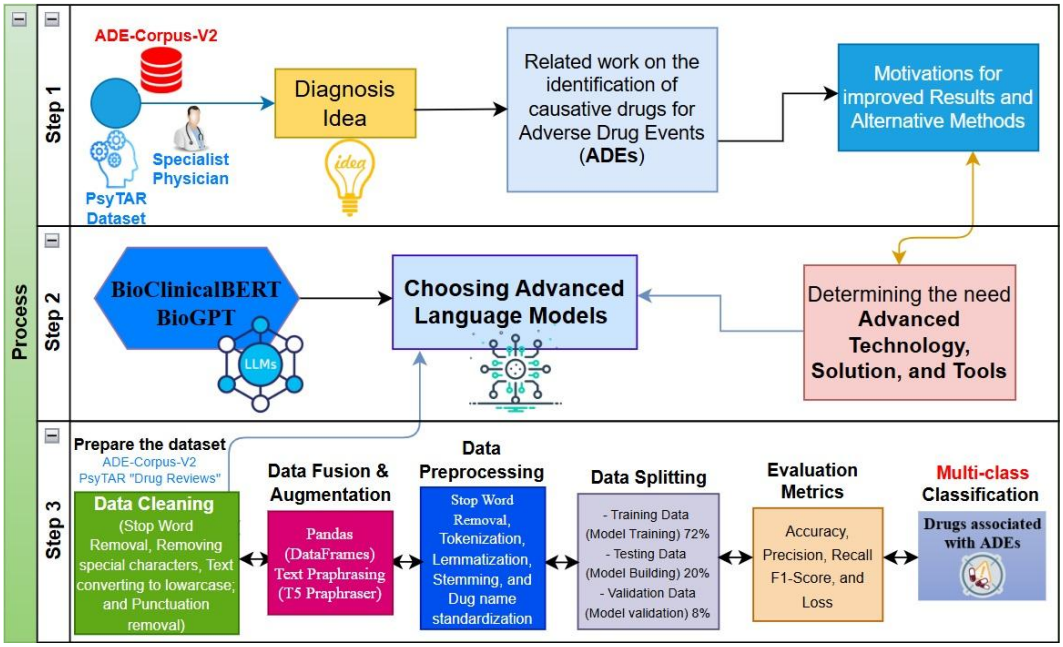


Fig. 1 Ideas and research efforts in the background of this study

In the context mentioned above, this study followed an easy-to-design data preprocessing and LLM models for ADEs text classification by considering the ADE corpus and drug reviews as input data. Initially, the real-time ADE-Corpus-V2 and PsyTAR datasets were collected, and the collected data were pre-processed using a feature engineering technique. Our advanced language models were then used to train and classify the dataset. Finally, various evaluation metrics were used to gauge performance (see Fig. 2).

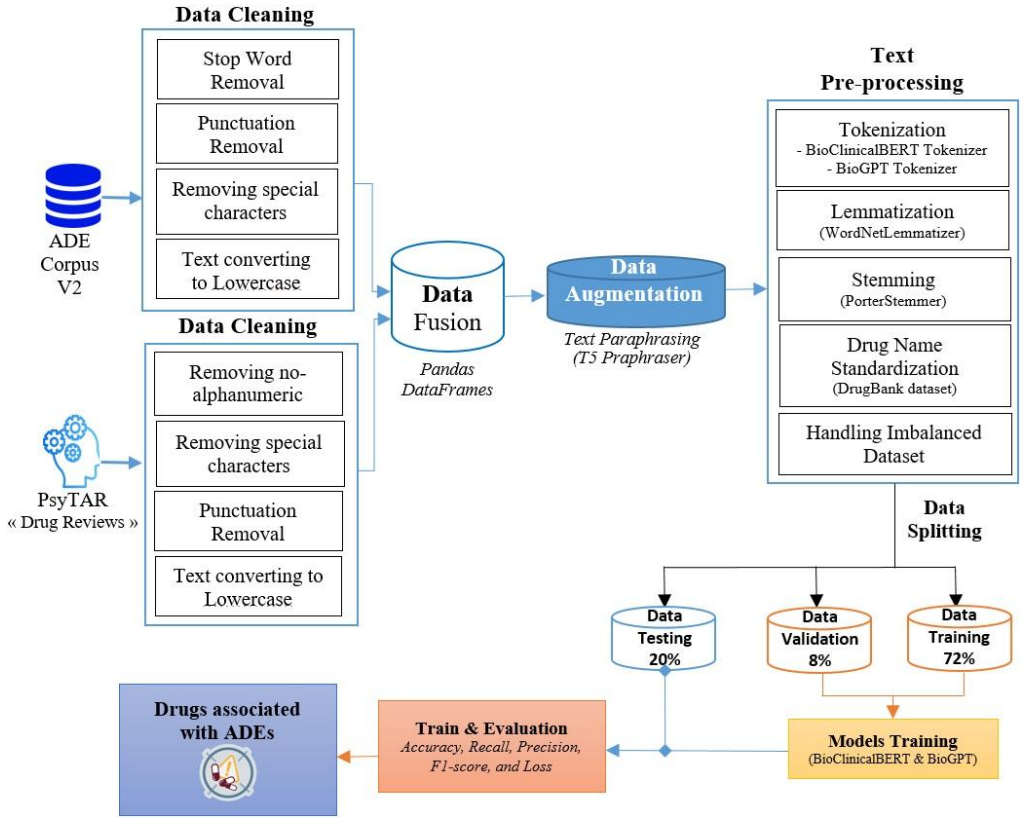


Fig. 2 Main steps of the proposed Advanced Language Models

3.1 Dataset Description

Two datasets were used for the research tasks described in this study. The two datasets were publicly available and were sourced from online medical case reports. We curated a comprehensive dataset from drug reviews and medical forums. The inclusion of various sources ensures a broad representation of the population, enriching the dataset's diversity. In this section, we provide a brief overview of both datasets, including details on the creation of our in-house dataset. Moreover, both datasets needed some preprocessing before being fed to the advanced NLP models for training.

1. ADE_Corpus_V2 Dataset: In the literature, the ADE-Corpus-V2 is a specialized dataset designed for research in the field of pharmacovigilance, with a primary focus on identifying and extracting adverse drug events from text [13; 3]. For [19], this drug corpus serves as a valuable resource for developing and testing advanced NLP models aimed at comprehending drug-related safety issues. The corpus contained 4,271 documents, which included 5,063 drugs and 6,821 adverse events. Among these adverse events, 3,341 unique occurrences were identified, highlighting the breadth and diversity of the dataset in capturing various medical terms and patient experiences. Typically, this dataset includes three columns: (1) sentences describing medical cases, (2) a drug–Medications mentioned within the sentences, and (3) Effect–Adverse reactions or symptoms linked to the drug. Notably, the dataset was carefully annotated to highlight instances of ADEs and the drugs associated with these events. Figure 3 shows detailed information about the sentence annotated with these entities and the relationship between them.

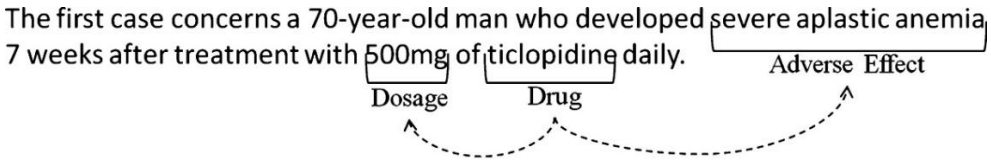


Fig. 3. Example of a sentence annotated with drug, dosage, adverse effect, and the relationships between them [3].

Finally, these annotations are essential for training advanced NLP models (BioClinicalBERT and BioGPT) to identify and classify causative drugs in ADE texts. In this dataset, 50% of the samples were related to our use case involving drugs and adverse events. The other instances were unrelated to the problem we addressed; therefore, we discarded them.

2. PsyTar (Psychiatric Treatment Adverse Reaction) dataset: The second corpus considered in this work is the corpus presented by [20]. According to [21; 20], the PsyTAR dataset contains 891 drug reviews posted by patients on an online healthcare forum (askapatient.com), about the effectiveness and ADEs associated with Zoloft, Lexapro, Cymbalta, and Effexor XR. For each drug review, patient demographic information, treatment period, and level of contentment with the respective drug [22]. According to [23], the reviews were then divided into 6009 separate sentences and labeled for the frequency of ADR, withdrawal symptoms (WDs), signs/symptoms/illness (SSIs), drug indications (DIs), drug effectiveness (EF), drug infectiveness (INF), and others (not applicable). Besides that, the PsyTAR corpus consists of three main components: sentence classification, entity identification, and entity normalization. In general, the patients are instructed to report drug ADRs in the side-effect field and the details of their experience in the comment field. However, patients were noted to report various aspects of their experiences, such as drug effectiveness or perceived distress due to ADRs, in both fields. Likewise, each comment included information about the patient's personal data, length of treatment, and satisfaction with the drugs. Table 2 shows an example of posts for Cymbalta on "askapatient.com".

Table 2. An example of posts for *Cymbalta* in "*askapatient.com*" [20]

Rating	Reason	Comment	Gender	Age	Duration	Date
3	Fibromyalgia/depression	I have only been on **30 mg** of **SSRIs** for 4 days and have **extreme runs**. Upset stomach and no appetite. Pain is minimal, though, and I feel less anxious and depressed.	F	38	4 days	05/10/2009

In this study, we summarize the inclusion criteria involved in some row selection from two PsyTAR sheets (sample and ADR_Identified), including index, drug ID, disorder, rating, comment, gender, age, dosage duration, date, and category ADRs (see Table 3).

Table 3. Representative PsyTar Dataset description

Column Name	Description
index	A unique identifier for each review or comment
comment_index	Another unique identifier for each comment, possibly used to group related comments together
comment_id	A unique identifier for each comment, possibly used to reference the comment in other datasets or systems

drug_id	The name of the medication being reviewed or discussed
rating	The patient's rating of the medication, possibly on a scale of 1-5 or 1-10, indicates their satisfaction with the treatment
disorder	The medical condition or disorder being treated by the medication
side effect	A flag or indicator of whether the patient experienced any side effects from taking the medication
comment	The free-text comment or review written by the patient, describing their experience with the medication
gender	The patient's gender, which may be relevant in understanding how medications affect different populations
age	The patient's age, which can influence how medications are metabolized and their overall effectiveness
dosage_duration	The length of time the patient took the medication can impact the effectiveness and side effects of the treatment
date	The date the review was written, which can help track changes in patient experiences over time
category	A categorization or classification of the review or comment, possibly indicating the type of feedback (e.g., positive, negative, neutral)

3.2 Data Pre-processing

In the domain of ADEs, the efficacy of NLP models is heavily reliant on the quality and variety of data sources used in conjunction with rigorous preprocessing techniques. In this context, text preprocessing is crucial for effective NLP, such as BioClinicalBERT and BioGPT, and plays a pivotal role in ensuring the quality and accuracy of the information extracted by our models. Raw text must be cleaned before feeding it to the model for efficient training and to produce more effective results. To ensure the effective detection of ADEs from corpus data, a comprehensive preprocessing pipeline was developed to address the informal and noisy nature of social media and medical forum text. The main aim of the text preprocessing step was to present the ADE-Corpus and drug reviews datasets in a consistent form and reduce any potential noise. In this study, we used the Pandas, NumPy, and NLTK libraries for data cleaning and processing. In the text preprocessing phase, the following processes were performed after collecting the dataset related to our topic, ADE-Corpus-V2, and PsyTar "drug reviews".

1. **Lowercasing.** The common aspect of preprocessing is converting all characters into the same case. Performance in training and classification is affected by how models handle lowercase words. Because NLP models treat "*Ibuprofen*" and "*ibuprofen*" as distinct tokens, we lowercase all text to ensure both are normalized to "*ibuprofen*", thereby reducing redundant case variants in the vocabulary and improving consistency and generalization in ADEs detection.
2. **Punctuation Removal.** Punctuation is also removed during the cleaning stage. A standard library (e.g., string module in Python) can be used to remove these, as it has a predefined collection of punctuation marks.
3. **URL tag, hashtag, and @ removal.** Within the first stage, preprocessing approaches include URL tag removal to organize information in an arrangement for the classification stage. The URL tag removal (preceded by "http:" and accompanied by "www:"), as well as Twitter handles ('@' symbols). In contrast, we retained hashtag tokens (with the '#' removed) because many encode drug or ADE concepts (e.g., #sideeffects, #ade, #Revlimid, #sideeffects from Humalog). These tokens can explicitly signal adverse events and contribute a useful signal for pharmacovigilance classification.
4. **Stop Words Removal and Special Characters.** In the stop words removal step, NLTK was applied to remove stop words. The NLTK [**24**] library and the regular expression library were used in Python. Stop words do not contribute to the detection of ADEs and can be safely removed to reduce noise. For this, we have removed white spaces and stop words (e.g., you, your, he, she, they, themselves, etc.) because unnecessary spaces and stop words usually will not help the proposed model learn new insights. For example, in the sentence "*The patient was feeling dizzy*" stop words such as "*The*" and "*was*" would be removed, leaving ("*patient*", "*feeling*", "*dizzy*"). Moreover, we have removed special characters (^, $, *, #, ~, /, %, >, <, &, @, etc.). Moreover, we have removed any numerical, character-numerical, and mathematical symbols except punctuation, such as exclamation marks (!) and question marks (?), because these marks help extract texts.
5. **Handling Negations.** Special care was taken to handle negations (e.g., "*not*", "*no*") correctly, as they can significantly alter the meaning of a sentence. Techniques such as appending "*NOT_*" to the following word (e.g., "*not effective*" becomes "*NOT_effective*") were used to preserve the context of negation.
6. **Lemmatization.** Lemmatization and confinement aim to reduce the inflected forms of a word and its derived variants to a basic form that all words share [25]. In this context, we used the *perform_lemmatization* function, which takes a list of tokens as input, applies the WordNetLemmatizer algorithm to each token, and returns a list of lemmatized tokens. For this, the WordNetLemmatizer algorithm uses a lexical

database called WordNet to reduce words to their base forms. For example, if the input tokens are ["slowli", "dosag", "reali"], the lemmatized tokens would be ["slowly", "dosage", "realy"]. The lemmatized tokens were then stored in a new column called "*lemmatized_tokens*" in the original *DataFrame*.

7. **Stemming.** Text stemming modifies words to obtain variant word forms using different linguistic processes, such as affixation (the addition of affixes) [26]. In our study, we used the *perform_stemming* function, which takes a list of tokens as input, applies the PorterStemmer algorithm to each token, and returns a list of the stemmed tokens. The PorterStemmer algorithm uses a set of rules to reduce words to their stems. For example, if the input tokens are ["detoxing", "slowly", "available"], the stemmed tokens would be ["detox", "slowli", "avail"]. The stemmed tokens were then stored in a new column called "*stemmed_tokens*" in the original *DataFrame*.
8. **Tokenization.** It is a crucial preprocessing step in NLP that converts raw text into tokens that can be processed by language models. Modern language models use sophisticated tokenization algorithms to manage the complexity of human language. In this study, we employed tokenizers, such as those provided by Hugging Face's transformers library, which are designed specifically for BERT and GPT models and handle special tokens and subword tokenization. To do this, we utilized domain-specific models, BioClinicalBERT and BioGPT (e.g., *BioClinicalBERT-tokenizer* and *BioGPT-tokenizer*), which include pretrained tokenizers tailored for biomedical text. First, the *BioClinicalBERT tokenizer* processes textual data into subword tokens. It splits words into subwords or characters, which helps handle out-of-vocabulary words and maintain context in biomedical texts. Similarly, the *BioGPT tokenizer* converts the text into tokens that are compatible with the model architecture. The tokenizer handles domain-specific vocabulary and provides embeddings that capture semantic meanings relevant to biomedical data. For example, the sentence "*The patient experienced nausea after taking ibuprofen*" would be tokenized into: ("The", "patient", "experienced", "nausea", "after", "taking", "ibuprofen").

In line with our objectives, we worked with the ADE-Corpus-V2 and PsyTar drug review datasets. To achieve a meaningful prediction, the study required renaming two attributes of the PsyTar dataset: "*comment*" to "*Text*", as well as "*drug_id*" to "*drug*", and the three text files of the ADE-Corpus-V2 dataset: "sentences from medical case reports" to "Text", as well as "drug-AE relations" and "drug-dosage relations" to "drug". Meanwhile, we will combine the two datasets to make use of their respective advantages after the preprocessing step, each separately, to eliminate noise and irrelevant data (See Figure 4). This method of combining the two datasets is consistent and ensures that we achieve an appropriate balance between meaningful detection and data preservation.

index	text	drug
0	Intravenous azithromycin-induced ototoxicity.	Acetaminophen
1	Immobilization, while Paget's bone disease was present, and perhaps enhanced activation of dihydrotachysterol by rifampicin, could have led to increased calcium-release into the circulation.	Ibuprofen
2	Unaccountable severe hypercalcemia in a patient treated for hypoparathyroidism with dihydrotachysterol.	Ibuprofen
3	METHODS: We report two cases of pseudoporphyria caused by naproxen and oxaprozin.	Atorvastatin
4	Naproxen, the most common offender, has been associated with a dimorphic clinical pattern: a PCT-like presentation and one simulating erythropoietic protoporphyria in the pediatric population.	Metformin
5	RESULTS: A 44-year-old man taking naproxen for chronic low back pain and a 20-year-old woman on oxaprozin for rheumatoid arthritis presented with tense bullae and cutaneous fragility on the face and the back of the hands.	Atorvastatin
6	RESULTS: A 44-year-old man taking naproxen for chronic low back pain and a 20-year-old woman on oxaprozin for rheumatoid arthritis presented with tense bullae and cutaneous fragility on the face and the back of the hands.	Aspirin
7	RESULTS: A 44-year-old man taking naproxen for chronic low back pain and a 20-year-old woman on oxaprozin for rheumatoid arthritis presented with tense bullae and cutaneous fragility on the face and the back of the hands.	Atorvastatin
8	RESULTS: A 44-year-old man taking naproxen for chronic low back pain and a 20-year-old woman on oxaprozin for rheumatoid arthritis presented with tense bullae and cutaneous fragility on the face and the back of the hands.	Aspirin
9	BACKGROUND: How to best treat psychotic patients who have had past clozapine-induced agranulocytosis or granulocytopenia remains a problem.	Omeprazole
10	CONCLUSION: The results suggest that olanzapine may be useful in treating patients with clozapine-induced granulocytopenia without the risk of recurrence of hematologic side effects.	Omeprazole

Fig. 4 Dataset global after fusion of ADE-Corpus-V2 and PsyTar drug reviews

3.3 Text Augmentation

In the literature, text augmentation is a crucial technique in NLP for expanding datasets and improving model performance. With the appropriate tools and approaches, text augmentation can lead to substantial improvements in various NLP tasks. Typically, NLP models benefit from large amounts of training data. When the available data are limited, data augmentation techniques can help mitigate this problem by artificially increasing the training set. According to [27], text augmentation techniques like synonym replacement (SR), random insertion (RI), random deletion (RD), and back-translation (BT) are powerful methods to expand and enrich NLP datasets. In this study, we employed a diverse array of text augmentation methods, each chosen for its potential to enhance the model's performance on various NLP tasks. We adopted these text augmentation techniques to enhance the quality, diversity, and size of the datasets. These techniques help models generalize better, especially in domains such as pharmacovigilance, where data can often be imbalanced or scarce. The following key augmentation strategies were applied:

1. **Synonym Replacement (SR).** In synonym replacement, certain words in the dataset are substituted with their synonyms while maintaining the meaning of the original text to introduce lexical variation. This technique introduces variability in the dataset, helping the models encounter different ways of expressing the same ADEs. This enables the model to generalize more effectively across various forms of linguistic expression without altering the meaning. By exposing the model to multiple expressions of the same ADEs, the technique improves the robustness of the model and enhances its ability to handle real-world textual diversity. For example, "*adverse effect*" might be replaced with "*side effect*". For this study, we used WordNet to find synonyms for a given token/word in the original sentence to replace it. Additionally, we used NLP Python packages, such as NLTK and SpaCy, to identify and replace synonyms in the original sentence.
2. **Back translation (BT).** This approach involves translating the text into another language (e.g., English to French) and then translating it back into the original language (e.g., French back to English). In theory, this approach translates a training example X from the source language (SL) to an intermediate language (IL) and then back-translates it to the source language [28]. Back-translation is particularly effective in generating sentence variations without changing their underlying meaning. In the context of this study, it enriches the dataset with paraphrased expressions of ADEs. This technique enhances model generalization by introducing linguistic variability, enabling the model to better comprehend the context, particularly in sentences that are syntactically distinct but semantically equivalent. This enhances the ability to detect ADRs in various textual forms. For example:
    - Original text: "*Ibuprofen caused severe stomach pain and bleeding"*
    - Back-translated variants:
        "*Ibuprofen led to intense abdominal pain and gastrointestinal bleeding*"
        "*The use of ibuprofen resulted in a strong stomachache and GI bleeding*"

3.4 Handling Imbalanced Datasets

In NLP, class imbalance occurs when one or more classes are significantly underrepresented compared to others in a dataset. According to [29], this imbalance can lead to biased models that perform well on the majority classes but poorly on the minority classes, resulting in suboptimal overall performance. In the context of this study, the dataset exhibited a noticeable imbalance among the different drug classes, which could impact the model's ability to accurately predict less frequent classes. To address this challenge, several techniques were employed to ensure a more balanced learning process:

1. **Oversampling.** In oversampling, instances from the minority class (e.g., underrepresented drugs or ADEs) are replicated to balance the dataset. ADE datasets are often imbalanced, with some medicines or ADEs being more prevalent. According to [30], oversampling addresses this issue by artificially increasing the number of samples for underrepresented classes. By balancing the dataset, oversampling helps the model avoid bias toward the majority class. This improves the model's ability to correctly identify rare ADEs, ensuring that the model performs well on both frequent and infrequent classes. For this study, we used SMOTE as an oversampling technique to address the class imbalance in ADE detection. This specifies the problem that SMOTE solves (imbalance between ADE-positive and negative samples in datasets (ADE-Corpus-V2 and PsyTAR). Typically, SMOTE generates synthetic examples by interpolating between existing samples of the minority class, which helps prevent the model from overfitting to the replicated samples and provides more diverse training data. Finally, the synthetic samples generated by SMOTE were incorporated into the model's training pipeline. This integration ensured that the models were trained on a balanced dataset, thereby improving their performance in the minority classes.
2. **Undersampling.** This technique reduces the number of instances in the majority classes to balance the dataset. Although undersampling can help achieve class balance, it may lead to the loss of important data, especially when the original dataset is small [31]. For this study, ADE datasets (ADE-Corpus-V2 and PsyTAR) usually contain many more non-ADE sentences than ADE-positive ones. Training on such skewed datasets can make models biased toward predicting "non-ADE", hurting recall on true ADEs. Therefore, undersampling was used judiciously in combination with oversampling to maintain a balance between the classes without sacrificing too much data size.

By employing these techniques, we aimed to mitigate the effects of class imbalance, enabling the models to learn effectively from all classes and thereby improve their overall predictive performance. This balanced learning approach is crucial in medical text classification, where accurately identifying less frequent but clinically significant drug-related adverse reactions is essential for ensuring patient safety and pharmacovigilance.

3.5 Drug Name Standardization

As per our objective, drug name standardization is a key preprocessing step to ensure consistency and accuracy in our analysis. In pharmacovigilance data, drug names are often mentioned in different forms (e.g., generic names, brand names, and abbreviations). Standardizing these names ensures that all references to the same drug are treated uniformly across the database. In the context mentioned above, we utilized DrugBank version 5.1.4, a comprehensive and widely used database containing 14,315 drugs, to map various drug name variations to their standardized counterparts [32]. DrugBank provides both generic and brand names of drugs, along with their chemical structures and pharmacological information [33; 34]. Therefore, we loaded the DrugBank dataset, which contains standardized drug names and variations (e.g., brand names and abbreviations). We also created a dictionary to map variations in drug names to the standardized version. Finally, we implemented a function to replace non-standard drug names with their corresponding standard names. Table 4 shows an example of this standardization.

Table 4. Example Output of Drug Standardization in the DrugBank dataset

Drug	Drug Standardized
Tylenol	Acetaminophen
Panadol	Acetaminophen
Advil	Ibuprofen
Motrin	Ibuprofen
Lipitor	Atorvastatin
Glucophage	Metformin
Bayer Aspirin	Aspirin
Prilosec	Omeprazole
Amoxil	Amoxicillin
Prinivil	Lisinopril
Ventolin	Albuterol
Zocor	Simvastatin
Cozaar	Losartan
Synthroid	Levothyroxine
Cipro	Ciprofloxacin
….	….
….	….

3.6 Data Transformation

After preprocessing the dataset, we determined whether each text sample was related to ADEs. Therefore, we introduced a new attribute for each dataset to address this issue. We used a dataset of side effects and compared all texts with a side effects dataset. If a side effect was identified, the text was classified as an ADE; otherwise, it was classified as non-ADE. This attribute was used as the target variable for model training.

In the literature, data transformation is the process of converting raw data into a format or structure that is more suitable for model building and data discovery. According to [35], data transformation refers to changing the format, structure, or values of data to improve its quality or make it easier for algorithms to process. The data transformation process is of vital importance in data science and ultimately affects the outcome of the analysis. For this study, the drug column contains string values that must be transformed into numeric values. Figure 5 depicts our dataset after transforming the drug column.

index	text	drug_encoded
0	Intravenous azithromycin-induced ototoxicity.	0
1	Immobilization, while Paget's bone disease was present, and perhaps enhanced activation of dihydrotachysterol by rifampicin, could have led to increased calcium-release into the circulation.	1
2	Unaccountable severe hypercalcemia in a patient treated for hypoparathyroidism with dihydrotachysterol.	1
3	METHODS: We report two cases of pseudoporphyria caused by naproxen and oxaprozin.	2
4	Naproxen, the most common offender, has been associated with a dimorphic clinical pattern: a PCT-like presentation and one simulating erythropoietic protoporphyria in the pediatric population.	3
5	RESULTS: A 44-year-old man taking naproxen for chronic low back pain and a 20-year-old woman on oxaprozin for rheumatoid arthritis presented with tense bullae and cutaneous fragility on the face and the back of the hands.	2
6	RESULTS: A 44-year-old man taking naproxen for chronic low back pain and a 20-year-old woman on oxaprozin for rheumatoid arthritis presented with tense bullae and cutaneous fragility on the face and the back of the hands.	4
7	RESULTS: A 44-year-old man taking naproxen for chronic low back pain and a 20-year-old woman on oxaprozin for rheumatoid arthritis presented with tense bullae and cutaneous fragility on the face and the back of the hands.	2
8	RESULTS: A 44-year-old man taking naproxen for chronic low back pain and a 20-year-old woman on oxaprozin for rheumatoid arthritis presented with tense bullae and cutaneous fragility on the face and the back of the hands.	4
9	BACKGROUND: How to best treat psychotic patients who have had past clozapine-induced agranulocytosis or granulocytopenia remains a problem.	5
10	CONCLUSION: The results suggest that olanzapine may be useful in treating patients with clozapine-induced granulocytopenia without the risk of recurrence of hematologic side effects.	5

Fig. 5 Drug-encoded attribute with these numerical values

To do that, we applied label encoding, one of the most widely used techniques for this type of transformation. The LabelEncoder class in scikit-learn is constructed to convert categorical labels to numerical labels. The "*fit_transform*" method is invoked on the LabelEncoder object, which learns unique labels in column "drug" and returns a numerical representation for each label — stored in a new column "*drug_encoded*" within our dataset. The reason for this encoding is that many NLP algorithms (especially those in scikit-learn) require numeric input; by encoding categorical labels, the data are prepared for use in NLP models (BioClinicalBERT and BioGPT). Figure 6 below illustrates the newly acquired data, highlighting the attributes "*Text*" and "*drug_encoded*".

index	text	drug	drug_encoded
0	Intravenous azithromycin-induced ototoxicity.	Acetaminophen	0
1	Immobilization, while Paget's bone disease was present, and perhaps enhanced activation of dihydrotachysterol by rifampicin, could have led to increased calcium-release into the circulation.	Ibuprofen	1
2	Unaccountable severe hypercalcemia in a patient treated for hypoparathyroidism with dihydrotachysterol.	Ibuprofen	1
3	METHODS: We report two cases of pseudoporphyria caused by naproxen and oxaprozin.	Atorvastatin	2
4	Naproxen, the most common offender, has been associated with a dimorphic clinical pattern: a PCT-like presentation and one simulating erythropoietic protoporphyria in the pediatric population.	Metformin	3
5	RESULTS: A 44-year-old man taking naproxen for chronic low back pain and a 20-year-old woman on oxaprozin for rheumatoid arthritis presented with tense bullae and cutaneous fragility on the face and the back of the hands.	Atorvastatin	2
6	RESULTS: A 44-year-old man taking naproxen for chronic low back pain and a 20-year-old woman on oxaprozin for rheumatoid arthritis presented with tense bullae and cutaneous fragility on the face and the back of the hands.	Aspirin	4
7	RESULTS: A 44-year-old man taking naproxen for chronic low back pain and a 20-year-old woman on oxaprozin for rheumatoid arthritis presented with tense bullae and cutaneous fragility on the face and the back of the hands.	Atorvastatin	2
8	RESULTS: A 44-year-old man taking naproxen for chronic low back pain and a 20-year-old woman on oxaprozin for rheumatoid arthritis presented with tense bullae and cutaneous fragility on the face and the back of the hands.	Aspirin	4
9	BACKGROUND: How to best treat psychotic patients who have had past clozapine-induced agranulocytosis or granulocytopenia remains a problem.	Omeprazole	5
10	CONCLUSION: The results suggest that olanzapine may be useful in treating patients with clozapine-induced granulocytopenia without the risk of recurrence of hematologic side effects.	Omeprazole	5

Fig. 6 Dug and Dug_encoded Attributes with their numerical values

Finally, we saved the preprocessed data in a new file, including only the "*Text*" and "*drug_encoded*" attributes, as these were the only ones required for training. Figure 7 illustrates the distribution of *drug_encoded* within our ADEs Text dataset following the preprocessing phase for frequently used drugs.

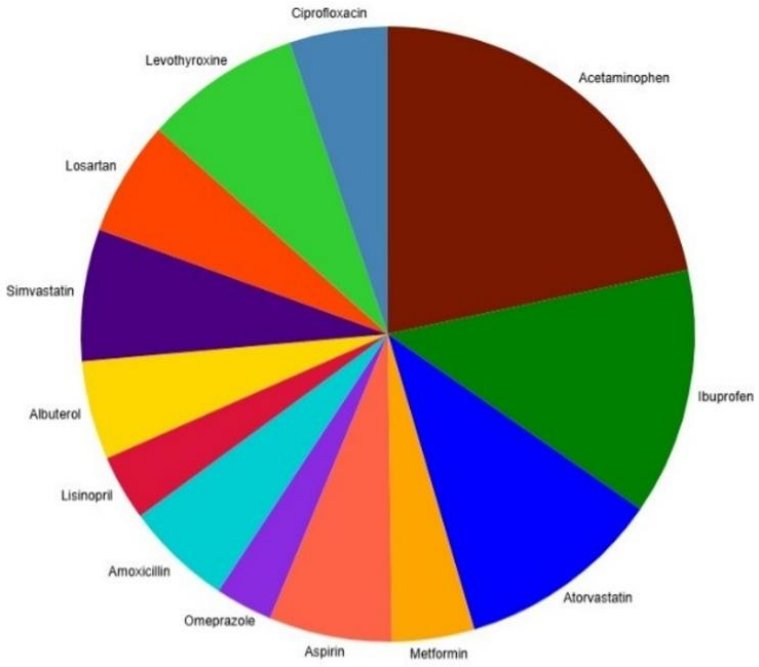


Fig. 7 Distribution of Drugs in our dataset

3.7 Data Splitting

In NLP literature, data splitting is a standard practice to separate data into training, testing, or validation sets. For our study, we divided the data by allocating 72% for training, 8% for validation, and 20% for testing. The data were rearranged before being divided to minimize variation and ensure the generalizability of the models. This is managed in a randomly specified manner through the "*sklearn_model_selection*" module of the "*train_test_split*" directory. Additionally, shuffling the training data enhances the representativeness of the overall data distribution and aids in preventing model overfitting.

4 Advanced Language Models

Pre-trained language models gained increasing attention in the biomedical domain, inspired by their success in the general NLP area. The primary categories of pre-trained language models in the general language domain include BERT (along with its variants) and GPT (and its variants). In this section, we describe how to adapt the two pretrained models, BioClinicalBERT and BioGPT, to relation extraction and text classification tasks.

4.1 BioClinicalBERT Model

In the literature, Bidirectional Encoder Representations from Transformers (BERT) is a Deep Learning (DL) model implemented for NLP that incorporates bidirectional context awareness. Similarly, BERT and its variants, such as BioBERT, RoBERTa, and BioClinicalBERT, have significantly improved tasks such as named entity recognition (NER), information extraction, and relation extraction. BioClinicalBERT is a specialized BERT model that combines the power of BioBERT and is specifically optimized for biomedical data and clinical texts [36]. Moreover, BioClinicalBERT uses the BERT transformer architecture, which allows it to learn the context of words by simultaneously examining both their left and right contexts [37]. BioClinicalBERT learns deep representations of clinical text, which can be used to uncover clinical insights (such as detection of ADEs), identify relationships between drugs and adverse effect outcomes, or create summaries of corpora. Unlike traditional models, BioClinicalBERT reads text bidirectionally, allowing it to better understand the context and relationships between words in clinical text. BioClinicalBERT's ability to understand word meaning in context makes it suitable for the complex sentence structures found in clinical texts [38]. According to Alsentzer et al. [39], its operation is based on Masked Language Modeling (MLM) and Next-Sentence Prediction (NSP) during pre-training. Once fine-tuned on domain-specific data (such as pharmacovigilance data), the model can classify, extract, and predict various biomedical relationships and entities. This specialization makes it highly effective for pharmacovigilance tasks, such as detecting ADEs from clinical text and biomedical data.

For example:

Input: given the sentence "The patient experienced nausea after taking ibuprofen".

Output: Bio-ClinicalBERT can recognize "ibuprofen" as the *drug* and "nausea" as the *side effect*, thus facilitating ADE detection.

4.2 BioGPT Model

In the literature, BioGPT is a generative pretrained transformer model tailored specifically for biomedical text generation and understanding. It builds on the strengths of OpenAI's GPT architecture, which is designed for language generation but is fine-tuned on biomedical datasets such as PubMed and clinical research articles. BioGPT is a language model designed specifically for biomedical texts that assists in drug discovery and clinical decision-making. According to [40], BioGPT is tailored to handle various biomedical NLP tasks, including text generation, relation extraction, document classification, and question-answering. It has demonstrated human parity for text generation, summarization, and question-answering tasks in the biomedical domain; however, evaluating the implications of this new technology for healthcare can prove challenging.

For example:

Input: given the text "The patient reported severe dizziness after taking metformin".

Output: BioGPT can generate a summary or extract "metformin" as the *drug* and "dizziness" as the *side effect*. In general, BioGPT uses a transformer architecture with a decoder-only setup, which means that it is designed for autoregressive text generation [41]**.** In this study, the model predicts the next word in a sequence based on preceding text. When fine-tuned on tasks such as text classification or ADE detection, it can generate relevant outputs, such as classifying drug side effects or generating a summary of clinical findings.

5 Experimental and Results

In the literature, most studies apply various NLP models to detect and predict the drug responsible for an adverse event from a biomedical text dataset to train their models. The suggested framework was implemented using Windows 11 and an Intel® Core™ i5-1035U CPU running at 2.20 GHz to 2.21 GHz with 8 GB of RAM. The CoLab Research Google platform was used for the experiments performed in this study. The programming language used

was Python, as it is well supported in CoLab. Additionally, analyses were performed in Python using important libraries such as the Natural Language Toolkit (NLTK), transformers (Hugging Face), Pandas, NumPy, Matplotlib, Seaborn, TensorFlow, Keras, and scikit-learn. Likewise, this section evaluates the proposed NLP models using the metrics of accuracy, F1-score, precision, and recall, and compares them with the state-of-the-art techniques. Then, we provide the results interpretation and practical and managerial implications.

5.1 Evaluation Criteria

As per mostly authors in the literature, the performance metrics can be used to evaluate a model's ability. It is possible to compare model predictions with known values of dependent features in a dataset using parameters such as accuracy, precision, recall, and the Receiver Operating Characteristic (ROC) curve [42; 43]. The different classification performance properties of the proposed models are evaluated based on the essential measures of True-Positive (TP), True-Negative (TN), False-Positive (FP), and False-Negative (FN). These popular parameters are defined as follows:

$$Sensitivity\ (Recall) = TP/TP+FN \quad (1)$$

The recall metric will tell us how well a model is in finding all of the true positives and is a ratio of true positives over all entities in the testing set.

$$Specificity = TN/TN+FP \quad (2)$$

In general, sensitivity and specificity evaluate the effectiveness of the algorithm on a single class, positive and negative, respectively.

$$Accuracy = TP+TN/TP+TN+FN+FP \quad (3)$$

$$Precision = TP/TP+FP \quad (4)$$

The precision metric will show the ratio of true positives over the total number of detected entities. In other words, this metric will help us understand how well a model is at returning only the true positives and not unrelated entities.

$$F - score = 2 * \frac{precision*recall}{precision+recall} \quad (5)$$

Commonly, accuracy is the most used metric to evaluate classification performance. This metric calculates the percentage of samples that are correctly classified [44]. As well, precision is how "precise" the model is out of those predicted positives, and how many of them are actually positive. A high value of the metric (F1-score) indicates that the model performs better on the positive class. Thus, the F1-score (also known as F-measure) might be a better measure when a balance between Precision and Recall is needed with an uneven class distribution (a large number of Actual Negatives). This metric can be used to show the overall performance of a tool.

5.2 Results and Discussions

The experiments were conducted using the ADE-Corpus-V2 and PsyTAR datasets. The dataset was pre-processed and annotated as described in Section 3.2. This section provides an in-depth analysis and discourse on the results achieved. This experiment uses BERT models, including BioClinicalBERT and BioGPT detect and predict the drug responsible for an adverse event from a biomedical text dataset. To train, validate, and test the BERT models, 72% of the data was used for training, 8% for validation, and 20% for evaluation (testing). As aforementioned, the confusion matrix is an essential metric for assessing a model's performance, as it delineates the quantities of true positives, true negatives, false positives, and false negatives. Accuracy, recall, precision, and F1-score were used to evaluate the performance of the classification models. The confusion matrices presented in Figure 8 offer a comprehensive visual representation of the performance of the four classifiers. Each matrix illustrates the model's ability to classify drugs responsible for adverse events into thirteen categories of standardized drugs: "*Acetaminophen, Ibuprofen, Atorvastatin, Metformin, Aspirin, Omeprazole, Amoxicillin, Lisinopril, Albuterol, Simvastatin, Lisartan, Levothyroxine, and Ciprofloxacin*". A clear diagonal pattern of higher values across all matrices indicates generally good performance in the correct classification. As observed from this matrix, BioClinicalBERT outperformed BioGPT. The diagonal elements indicate the correctly classified instances, whereas the off-diagonal elements represent misclassifications. These insights from the confusion matrices not only highlight each model's strengths and weaknesses in classifying different drug categories responsible for adverse events but also provide valuable directions for potential improvements and model selection in specific ADEs analysis tasks.

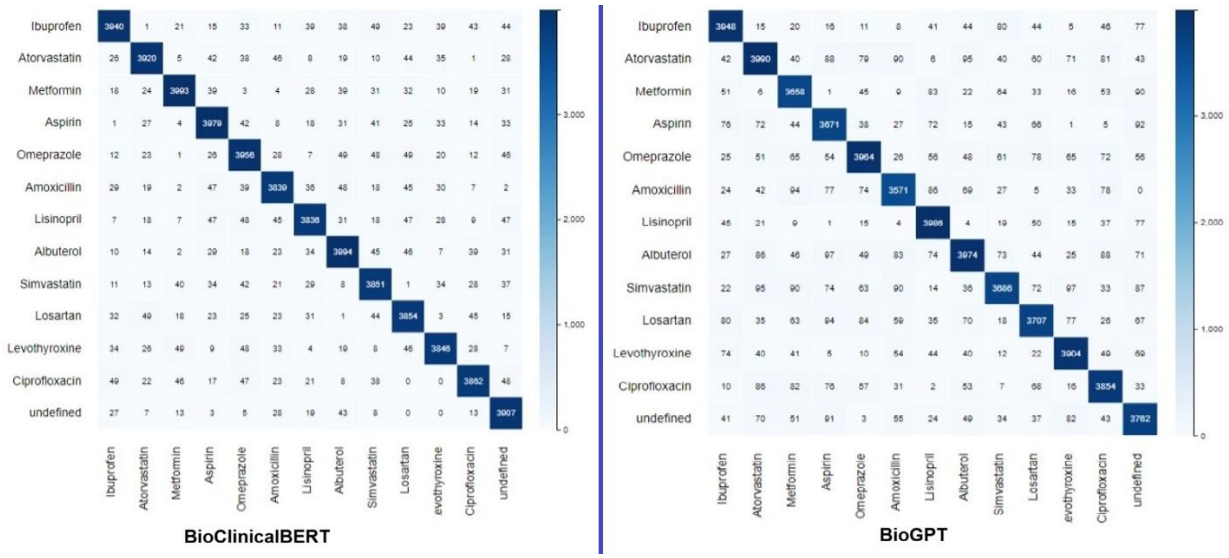


Fig. 8. Confusion matrices of the BioClinicalBERT and BioGPT models, respectively

In our study, the proposed BERT models (BioClinicalBERT and BioGPT) reduced the loss and improved the accuracy of both the testing and training data samples, which revealed the effectiveness and significance of our models for all datasets. The number of times an experiment is repeated depends on factors such as the data variability. Typically, learning curves of accuracy and log loss for the two models are generated to ensure that the results are consistent and reliable, as shown in Figure 9.

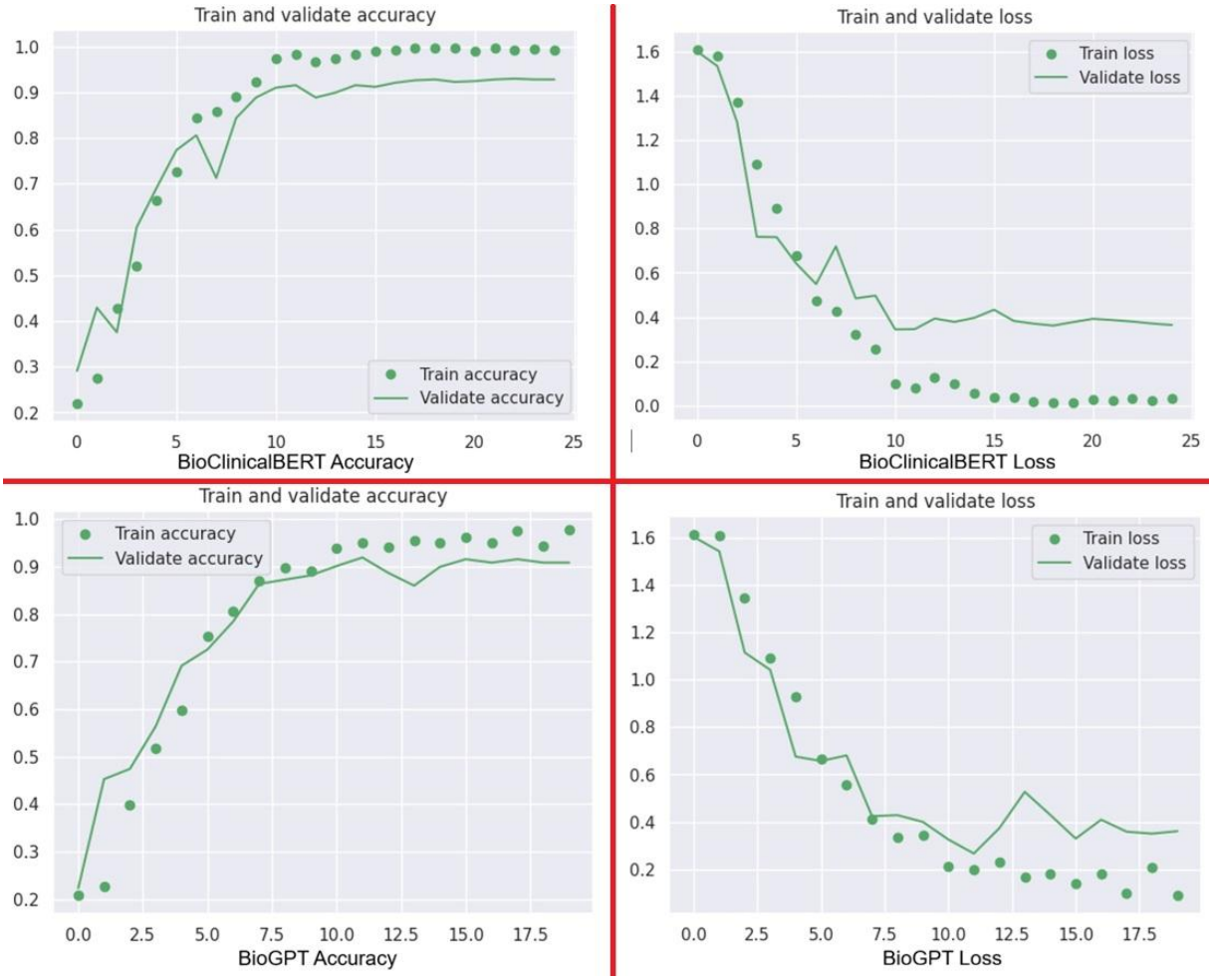


Fig. 9. (a) Training and validation accuracy graph, (b) training and validation loss graph of the BioClinicalBERT and BioGPT models.

Table 5 shows the test, training, and validation accuracies and losses for each of the advanced language models. Interestingly, the findings indicate that our advanced language models demonstrated significant accuracy, with BioClinicalBERT achieving a test accuracy of 0.9792 and a training accuracy of 0.9992, indicating a strong capacity to generalize from the training data while maintaining high performance on the unseen data. The validation accuracy of BioClinicalBERT was 0.9471, demonstrating its robustness in handling data variability. The loss for BioClinicalBERT was relatively low at 0.0500, suggesting minimal deviation from the optimal predictions.
In contrast, BioGPT achieved a test accuracy of 0.9238 and a training accuracy of 0.9971, respectively. Although BioGPT also performed well, its test and validation accuracies (0.9411) were slightly lower than those of BioClinicalBERT. The loss for BioGPT was 0.1000, indicating a slightly higher deviation from the optimal predictions than that of BioClinicalBERT.

Table 5. A Comparative Analysis of Accuracy and Loss Metrics

Model	Test Accuracy	Train Accuracy	Validation Accuracy	Loss
BioClinicalBERT	0.9792	0.9992	0.9471	0.0500
BioGPT	0.9238	0.9971	0.9411	0.1000

In addition, the models were evaluated using several performance metrics, including recall, precision, and F1-score for each drug class. These models were designed to detect and predict the drug responsible for an adverse event from the ADEs text dataset (ADE-Corpus-V2 and PsyTAR drug reviews), and their performance was assessed across each drug class. These metrics offer a comprehensive evaluation of the models' ability to accurately identify causative drugs from ADE texts. The detailed metrics for each model are listed in **Figure 10**.

	precision	recall	f1-score	support
Acetaminophen	0.97	0.99	0.98	90
Ibuprofen	0.94	0.99	0.96	69
Atorvastatin	0.99	0.99	0.99	86
Metformin	0.99	0.98	0.98	86
Aspirin	0.99	0.99	0.99	81
Omeprazole	0.97	0.94	0.96	83
Amoxicillin	0.99	0.97	0.98	86
Lisinopril	1.00	1.00	1.00	77
Albuterol	0.99	0.99	0.99	75
Simvastatin	0.97	0.92	0.95	66
Losartan	0.95	1.00	0.97	72
Levothyroxine	0.98	1.00	0.99	62
Ciprofloxacin	1.00	0.99	0.99	67
accuracy			0.98	1000
macro avg	0.98	0.98	0.98	1000
weighted avg	0.98	0.98	0.98	1000

BioClinicalBERT Model Evaluation

	precision	recall	f1-score	support
Acetaminophen	0.93	0.95	0.94	88
Ibuprofen	0.94	0.94	0.94	80
Atorvastatin	0.90	0.97	0.94	67
Metformin	0.96	0.93	0.94	80
Aspirin	0.92	0.87	0.89	76
Omeprazole	0.88	0.91	0.89	95
Amoxicillin	0.88	0.89	0.89	65
Lisinopril	0.91	0.92	0.92	76
Albuterol	0.93	0.95	0.94	66
Simvastatin	0.98	0.88	0.92	91
Losartan	0.88	0.95	0.92	64
Levothyroxine	0.95	0.93	0.94	81
Ciprofloxacin	0.97	0.96	0.96	71
accuracy			0.93	1000
macro avg	0.92	0.93	0.93	1000
weighted avg	0.93	0.93	0.93	1000

BioGPT Model Evaluation

Fig. 10. Performance Metrics of BioClinicalBERT and BioGPT for each drug class

The results demonstrate that BioClinicalBERT exhibits superior overall performance compared to BioGPT. Specifically, BioClinicalBERT achieved the highest accuracy (98%) and the lowest loss value (0.0500). In addition, BioClinicalBERT reported the highest F1-score for all classes (98%), indicating its ability to maintain consistent performance across diverse categories. The F1-score is particularly useful when the dataset is imbalanced because it penalizes the extreme values of precision or recall, ensuring that both aspects are equally considered. These findings suggest that BioClinicalBERT is the most suitable model for the classification task of drugs responsible for an adverse event, providing a robust balance of accuracy and stability across different data splits and showcasing its superior capacity to generalize from the training data to unseen samples. During our experiments, BioClinicalBERT outperformed BioGPT in this classification task by achieving higher accuracy, lower loss, and superior performance across all evaluated metrics. Its capability to better handle complex medical text data makes it the preferred model for classifying drugs responsible for adverse events in this study.

5.3 Comparison against State-of-the-Art Methods

The results of attribute importance are significant for medical applications. In this study, by integrating advanced language models (BioClinicalBERT and BioGPT), we exemplified the potential of advanced language models to propel the field of identification of causative drugs for adverse event comment analysis in the pharmaceutical

domain. The emphasis is placed on utilizing these advanced language models to conduct a detailed and nuanced analysis of the identification of causative drugs for ADEs based on patient drug reviews and social media comments. This methodological approach aims to pave the way for advancements in understanding social media comments within the context of identifying causative drugs for ADEs, thereby setting a foundation for subsequent evaluations. In this context, we conducted a comparative study of our proposed advanced language models (BioClinicalBERT and BioGPT) with other existing identification and classification models based on the approach used, the many drug comment datasets used in experimentation, the methodology, the features used for identification and classification, and the percentage accuracy achieved. Our results are competitive with previously reported performance on identifying causative drugs for ADE-related tasks, although direct comparison is limited by differing datasets and task definitions (see Table 6).

Table 6. Comparison between our ML models and state-of-the-art models on the ADEs comments dataset.

Methods/Techniques	Data Source	Description & Applications	Performance
Advanced ML and DL techniques, including RF, gradient boosting machines, CNNs, BERT, and GPT	Clinical trial datasets	Detecting Serious Adverse Events (SAEs) in pharmacovigilance to ensure drug safety during and after clinical trials.	BERT: Accuracy 92.7% GPT: Accuracy 94.1%
DL methods: transformer encoder and Graph Attention Network–Graph Convolutional Network Combination (GAT GCN)	SMILES string data, ADRs data, and drug protein interaction data	Identifies ADRs early in drug discovery	F1-score 63,3% ROC--AUC 70% AUPR 66,19%
KnowCAGE (Knowledge-augmented Concept-Aware Graph Embeddings) model	SMM4H, TwiMed-Pub, TwiMad-Twitter, and CADAC	Detection of ADEs from social media content or biomedical literature requires	Precision 87.2% Recall 97% F1-score 91.8%
BERT-based language Model	ADE-Corpus-V2 SMM4H	Accurate DAE Extraction From Social Media	F1-score 98,13% F1-score 97,9%
Multi-Modal Adverse Drug Event (MMADE) detection using the capabilities of LLMs and VLMs	PSB2016 (Twitter), ADE corpus, CADEC, and BioDEX	ADE detection within pharmacovigilance mining	ROUGE 0.211 BLEU 0.197
MGPS and BCPNN algorithms	FAERS DrugBank SIDER	Generate enhanced drug safety signals for the early detection of ADEs	AUC 0.743 (0.743) AUPR 0.226 (0.224) F1-score 0.319 (0.319)
NLP workflow	MEDLINE EMBASE arXiv	Detecting ADE mentions in clinical notes	NER (n = 17; 58.6%) RE (n = 17; 58.6%)
Tarantula utilized Fast Text embeddings and an aggregated lexicon	FAERS Social Media UMLS	Detect existing ADEs from social media	Sensitivity 84.2% Specificity 98%
Text Convolutional Neural Networks (TextCNN), Logistic Regression (LR), and Decision Trees (DT)	- ADE-Corpus-V2 - PubMed dataset	DAEs Detection and Classification	Precision 87% Recall 91% F1-score 89%
BiLSTM and Conditional Random Fields (CRF) neural network	1089 de-identified clinical notes of 21 patients with cancer	Determine relations, including the adverse drug reaction relation between medication and sign or symptom entities	Precision 88,8% Recall 85,5% F1-score 87,2%
Bi-directional LSTM CRF, CNN-RNN	n2c2 corpus from the MIMIC-III (Medical Information Mart for Intensive Care III)	Extraction of medications and associated ADEs from clinical documents	F1-score 93.45%(NER) F1-score 96.30%(RC) F1-score 89.05%(end-to-end evaluation)
Advanced Language Models (BioClinicalBERT and BioGPT)	- **ADE-Corpus-V2** - **PsyTAR drug reviews**	**Identification and Classification of Specific Drugs Associated with ADEs**	**Accuracy 98%** **F1-Score 98%**

The results of the present study were comparatively evaluated using similar parameters, demonstrating that the proposed advanced language models yielded commendable outcomes compared with the findings of other studies. However, much better results were produced in earlier studies, as mentioned above, and better results were obtained with slight differences in the studies using a similar formalization. For better results, since the identification of causative drugs for adverse events is person-specific and can vary from patient to patient, a groundbreaking result can be achieved if the diagnosis information is available for all patients and can be used to determine the

ADEs and their intensity. Moreover, the dataset can be improved in several ways. A patient's blood group, genetic information, chronic disease, smoking, and drinking status can be important factors in determining a specific drug for ADEs or predicting the outcome of a specific drug for ADEs. Furthermore, we plan to investigate other embedding approaches to represent ADEs and help predict the relationships between drugs and new ADEs.

4.4 Practical and Managerial Implications of the Findings

The impact of the identification of causative drugs for adverse events is a major cause of concern with major economic and emotional consequences for different stakeholders, such as pharmaceutical companies, healthcare sectors, patients, and even governments. In this sense, ADE identification and detection from biomedical texts using advanced language models enable proactive pharmacovigilance by automating signal detection from various sources, such as social media, complementing traditional reporting systems. Moreover, advanced language models (BioClinicalBERT and BioGPT) transform ADE/pharmacovigilance detection by automating identification from unstructured biomedical/social media texts, enabling real-time signal detection beyond traditional reporting. Similarly, pharma/healthcare companies and trade actors gain a competitive advantage through faster safety decisions, cost savings, and regulatory compliance (predictive ADEs using BioClinicalBERT and BioGPT). At a broader level, we note that despite the significant improvement in the identification of causative drugs for adverse events achievable according to these results through the application of advanced language models, medical expert knowledge must be employed in both judging the soundness of the medical results and the relevance of a wide range of relevant model metrics. Practical implications include supporting pharmacology physicians, psychologists, and mental healthcare professionals through enhanced screening and assessment tools. However, risks such as privacy violations and misdiagnosis must be addressed. Close collaboration with domain experts is vital to ensure clinical validity and avoid harm to patients.

5 Conclusion

In the literature, the identification of causative drugs from Adverse Drug Event (ADE) texts is a critical aspect of pharmacovigilance, and the application of advanced NLP models has proven to be a powerful tool in this field. This study developed an effective strategy for pinpointing causative drugs for adverse events from biomedical textual data using advanced language models. In this study, we explored various text classification approaches to detect ADEs and identify associated drugs from patient-reported data. We evaluated the performance of two state-of-the-art models, BioClinicalBERT and BioGPT, on a comprehensive dataset that combined clinical notes and social media posts (ADE-Corpus-V2 and PsyTAR drug reviews). Our findings indicate that BioClinicalBERT demonstrated superior performance, achieving the highest accuracy and F1-score, which suggests its strong potential in biomedical text multi-classification tasks. In contrast, BioGPT also showed promising results, particularly in understanding and generating contextual embeddings for biomedical texts. This study contributes to the growing body of knowledge on ADE detection by emphasizing the importance of leveraging medical text data for pharmacovigilance. The integration of patient-reported data from clinical notes and social media platforms offers a robust approach to enhance drug-safety monitoring.

Despite the promising performance achieved by the proposed approach, this study had some limitations. It is essential to have biomedical experts and pharmacovigilance teams validate the findings, considering other critical factors. In addition, significant preprocessing and manual validation by physicians in pharmacology, biomedical doctors, mental healthcare professionals, and pharmacists is an extensive procedure.

Several considerations are worth reporting for future work, such as expanding the dataset to include a wider range of textual sources, exploring additional artificial intelligence methodologies, and integrating these advanced models with existing pharmacovigilance systems. Ultimately, the development of effective tools for ADE detection and causative drug identification will improve patient safety and enable more targeted healthcare intervention. Another future work involves focusing on breaking new ground by integrating text with images, solving a real-world challenge where people affected by ADEs may resort to image sharing for communication when verbally expressing their symptoms is difficult. Additionally, the integration of other clinical datasets (e.g., chatbots) will allow extension of language capabilities in future models. At the same time, the extension includes other pharmaceutical entities (e.g., strength) and additional label data (e.g., medical condition/history).

Conflict/Competing of Interest
The authors have declared that they have no conflicts/Competing of interest.

Funding Information
This research received no external funding.

Ethical and informed consent for the data used
Compliance with Ethical Standards. The data used in this study were exclusively generated by the authors using the Internet to collect the ADE-Corpus-V2 dataset and by accessing the website askapatient.com for PsyTAR drug reviews. No external parties contributed to the data collection process. No field research involving humans or patients in hospitals or psychiatric clinics was conducted.

Data availability and access
The data supporting the findings of this study are available from the corresponding authors upon request.

Acknowledgments
This project is registered in the context of our research team at the LABAB laboratory for generously providing us with computational and storage resources. We want to express our gratitude to the biomedical doctors, physicians in pharmacology (laboratory), psychologists, and mental health care professionals. The dedication and enthusiasm of our team members helped us move forward.

References

[1] Murphy, R. M., Klopotowska, J. E., De Keizer, N. F., Jager, K. J., Leopold, J. H., Dongelmans, D. A., Abu-Hanna, A., & Schut, M. C. (2023). Adverse Drug Event Detection using Natural Language Processing: A Scoping Review of Supervised Learning Methods. PLoS ONE 18(1): e0279842, 1-26. https://doi.org/10.1371/journal.pone.0279842

[2] Wei, C., Peng, Y., Leaman, R., Davis, A. P., Mattingly, C. J., Li, J., Wiegers, T. C., & Lu, Z. (2016). Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database (Oxford) (2016), 1-8. https://doi.org/10.1093/database/baw032

[3] Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., & Toldo, L. (2012). Development of a Benchmark Corpus to Support the Automatic Extraction of Drug-Related Adverse Effects from Medical Case Reports. *Journal of Biomedical Informatics*, 45(5), 885–92. https://doi.org/10.1016/j.jbi.2012.04.008

[4] Sahoo, P., Singh, A. K., Saha, S., Chadha, A., Mondal, S. (2024). Enhancing Adverse Drug Event Detection with Multimodal Dataset: Corpus Creation and Model Development. In Findings of the *Association for Computational Linguistics (ACL'24)*, Bangkok, Thailand, 11214–11226. https://doi.org/10.18653/v1/2024.findings-acl.667

[5] Gao, Y., Ji, S., & Marttinen, P. (2024). Knowledge-augmented Graph Neural Networks with Concept-aware Attention for Adverse Drug Event Detection. In the Proceedings of the *International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING'24)*, Torino, Italia, 9787–9798. https://aclanthology.org/2024.lrec-main.855/

[6] Khemani, B., Malave, S., Shinde, S., Shukla, M., Shikalgar, R., & Talwar, H. (2025). AI-driven pharmacovigilance: Enhancing Adverse Drug Reaction Detection with Deep Learning and NLP. *MethodsX* 15 (2025) 103460, 1-13. https://doi.org/10.1016/j.mex.2025.103460

[7] Christopoulou, F., Tran, T. T., Sahu, S. K., Miwa, M., & Ananiadou, S. (2020). Adverse Drug Events and Medication Relation Extraction in Electronic Health Records with Ensemble Deep Learning Methods. *Journal of the American Medical Informatics Association*, 27(1), 39–46. https://doi.org/10.1093/jamia/ocz101

[8] Luo, Y., Thompson, W. K., Herr, T. M., Zeng, Z., Berendsen, M. A., Jonnalagadda, S. R., Carson, M. B., & Starren, J. (2017). Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review. *Drug Safety*, 40(11), 1075–1089. https://doi.org/10.1007/s40264-017-0558-6

[9] Harpaz, R., Callahan, A., Tamang, S., Low, Y., Odgers, D., Finlayson, S., Jung, K., LePendu, P., & Shah, N. H. (2014). Text Mining for Adverse Drug Events: The Promise, Challenges, and State of the Art. *Drug Safety*, 37(10), 777–790. https://doi.org/10.1007/s40264-014-0218-z

[10] Zitu, M. M., Manne, A., Zhu, Y., Rahat, W. B., & Binkheder, S. (2025). Large Language Models for Drug-Related Adverse Events in Oncology Pharmacy: Detection, Grading, and Actioning. *Pharmacy*, 13(176), 1-21. https://doi.org/10.3390/pharmacy13060176

[11] Kopacheva, E., Lincke, A., Björneld, O., & Hammar, T. (2025). Detecting Adverse Drug Events in Clinical Notes Using Large Language Models. Stud. Intelligent Health Systems, 15(327), 892–893. https://doi.org/10.3233/SHTI250495

[12] Ou, Q., Jiang, X., Guo, Z., Jiang, J., Gan, Z., Han, F., & Cai, Y. A. (2025). Fusion Deep Learning Model for Predicting Adverse Drug Reactions Based on Multiple Drug Characteristics. *Life*, 15(436), 1-22. https://doi.org/10.3390/life15030436

[13] Dong, F., Guo, W., Liu, J., Patterson, T. A., & Hong, H. (2024). BERT-based Language Model for Accurate Drug Adverse Event Extraction from Social Media: Implementation, Evaluation, and Contributions to Pharmacovigilance Practices. *Frontiers Public Health*, 12:1392180, 1-16. https://doi.org/10.3389/fpubh.2024.1392180

[14] Padhi, B., Liu, R., Yang, Y., Peng, X., Li, L., Zhang, P., & Zhang, P. (2024). Using Multiple Drug Similarity Networks to Promote Adverse Drug Event Detection. *Heliyon* 10(22):e39728, 1-13. https://doi.org/10.1016/j.heliyon.2024.e39728

[15] Roosan, D., Law, A. V., Roosan, M. R., & Li, Y. (2022). Artificial Intelligence Context-Aware Machine-Learning Tool to Detect Adverse Drug Events from Social Media Platforms. *Journal of Medical Toxicology*, 18(4), 311–320. https://doi.org/10.1007/s13181-022-00906-2

[16] Rawat, A., Wani, M.A., ElAffendi, M., Imran, A.S., Kastrati, Z., & Daudpota, S.M. (2022). Drug Adverse Event Detection Using Text-Based Convolutional Neural Networks (TextCNN) Technique. *Electronics*, 11(3336), 1-21. https://doi.org/10.3390/electronics11203336

[17] Dandala, B., Joopudi, V., & Devarakonda, M. (2019). Adverse Drug Events Detection in Clinical Notes by Jointly Modeling Entities and Relations Using Neural Networks. *Drug Safety*, 42, 135–146. https://doi.org/10.1007/s40264-018-0764-x

[18] Wei, Q., Ji, Z., Li, Z., Du, J., Wang, J., Xu, J., Xiang, Y., Tiryaki, F., Wu, S., Zhang, Y., Tao, C., & Xu, H. (2019). A Study of Deep Learning Approaches for Medication and Adverse Drug Event Extraction from Clinical Text. *Journal of the American Medical Informatics Association*, 27(1), 13–21. https://doi.org/10.1093/jamia/ocz063

[19] Dietrich, J., & Kazzer, P. (2023). Provision and Characterization of a Corpus for Pharmaceutical, Biomedical Named Entity Recognition for Pharmacovigilance: Evaluation of Language Registers and Training Data Sufficiency. Drug Safety, 46(5), 1-15. https://doi.org/10.1007/s40264-023-01322-3

[20] Zolnoori, M., Fung, K.W., Patrick, T.B., Fontelo, P., Kharrazi, H., Faiola, A., Shah, N. D., Wu, Y. S. S., Eldredge, C. E., Luo, J., Conway, M., Zhu, J., Kyung Park, S., Xu, K., & Moayyed, H. (2019). The PsyTAR dataset: From Social Media Posts to a Corpus of Adverse Drug Events and Effectiveness of Psychiatric Medications. *Data in Brief*, *24*(2019)103838, 1-6. https://doi.org/10.1016/j.jbi.2018.12.005

[21] Sboev, A., Rylkov, G., Selivanov, A., Moloshnikov, I., & Rybka, R. (2021). A Deep Neural Network-Based Data-Driven Model for Evaluating the Recognition of ADR Mentions in the Texts of the PsyTAR Corpus. *19th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2021),* 20–26 September 2021, Rhodes, Greece, Published in *AIP Conference Proceedings*, vol.192, 2023, ISBN: 9780735445895

[22] Yu, D., & Vydiswaran, V. G. V. (2022). An Assessment of Mentions of Adverse Drug Events on Social Media with Natural Language Processing: Model Development and Analysis. *JMIR Medical Informatics*, 10(9):e38140, 1-19. https://doi.org/10.2196/38140

[23] Naderian, S., Rahmani, R., & Samad-Soltani, T. (2024). A Natural Language Processing Framework for Detecting Adverse Drug Reactions in Clinical Structured Drug Reviews. *International Journal of Drug Research in Clinics*, 2: e9, 1-11. https://doi.org/10.34172/ijdrc.2024.e9

[24] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. Publisher(s): O'Reilly Media, Inc. ISBN: 9780596516499

[25] Sabri, M., Brahami, M., & Abid, F. F. Z. (2022). Sentiment Analysis of COVID-19 Tweets Using Adaptive Neuro-Fuzzy Inference System Models. *International Journal of Software Science and Computational Intelligence (IJSSCI)*, 14(1), 1-20. http://doi.org/10.4018/IJSSCI.300361

[26] Brahami, M., Sabri, M., & Matta, N. (2020). Towards a Model to Improve Boolean Knowledge Mapping by Using Text Mining and Its Applications: Case Study in Healthcare. *International Journal of Information Retrieval Research*, *10*(3), 40–58. https://doi.org/10.4018/IJIRR.2020070103

[27] Bonthu, S., Dayal, A., Lakshmi, M.S., & Rama Sree, S. (2022). Effective Text Augmentation Strategy for NLP Models. In: Poonia, R.C., Singh, V., Singh Jat, D., Diván, M.J., Khan, M.S. (eds) Proceedings of *Third International Conference on Sustainable Computing*. Advances in Intelligent Systems and Computing, Vol. 1404. Springer, Singapore, 521-531. https://doi.org/10.1007/978-981-16-4538-9_51

[28] Sennrich, R., Haddow, B., & Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. In Proceedings of the *54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Berlin, Germany. Association for Computational Linguistics, 86–96. https://doi.org/10.18653/v1/P16-1009

[29] Kaur, R., & Gupta, N. (2025). A Novel Boundary Adaptive Oversampling Approach for Intrusion Detection. *Indonesian Journal of Electrical Engineering and Computer Science*, 37(1), 518-529. http://doi.org/10.11591/ijeecs.v37.i1.pp518-529

[30] Yao, Q., & Zhao, X. (2025). An Intrusion Detection Imbalanced Data Classification Algorithm based on CWGAN-GP Oversampling. *Peer-to-Peer Networking and Applications*, 18(21), 1-12. https://doi.org/10.1007/s12083-025-01938-5

[31] Rahma, F., Rachmadi, R. F., Adi Pratomo, B., & Purnomo, M. H. (2023). Assessing the Effectiveness of Oversampling and Undersampling Techniques for Intrusion Detection on an Imbalanced Dataset. In the Proceedings of the *IEEE Industrial Electronics and Applications Conference (IEACon'23)*, 06-07 November, Penang, Malaysia, 1-8. http://doi.org/10.1109/IEACon57683.2023.10370430

[32] Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., & Hassanali, M. (2008). DrugBank: A Knowledgebase for Drugs, Drug Actions, and Drug Targets. *Nucleic Acids Research*, 36(1), 901–906. https://doi.org/10.1093/nar/gkm958

[33] Schreier, T., Tropmann-Frick, M. & Böhm, R. (2024). Integration of FAERS, DrugBank and SIDER Data for Machine Learning-based Detection of Adverse Drug Reactions. *Datenbank Spektrum*, 24, 233–242. https://doi.org/10.1007/s13222-024-00486-1

[34] Padhi, B., Liu, R., Yang, Y., Peng, X., Li, L., Zhang, P., & Zhang, P. (2024). Using Multiple Drug Similarity Networks to Promote Adverse Drug Event Detection. *Heliyon*, 10(22), 1-13. https://doi.org/10.1016/j.heliyon.2024.e39728

[35] Borrohou, S., Fissoune, R., Badir, H. (2025). The Role of Data Transformation in Modern Analytics: A Comprehensive Survey. *Journal of Computer Languages*, 84(101329), 1-16. https://doi.org/10.1016/j.cola.2025.101329

[36] Matondora, L., Mutandavari, M., & Mupini, B. (2024). NLP-Based Prediction of Hospital Readmission Using Clinical-BERT and Clinician Notes. *International Journal of Innovative Science and Research Technology*, 9(7), 2549-2557. https://doi.org/10.38124/ijisrt/IJISRT24JUL1191

[37] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (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

[38] Abdaoui, H., Barki, C., Dergaa, I., Tlili, K., Ceylan, H. İ., Bragazzi, N. L., de Giorgio, A., Ben Salah, R., & Rahmouni, H. B. (2025). Accurate Clinical Entity Recognition and Code Mapping of Anatomopathological Reports Using BioClinicalBERT Enhanced by Retrieval-Augmented Generation: A Hybrid Deep Learning Approach. *Bioengineering*, 13(1), 1-21. https://doi.org/10.3390/bioengineering13010030

[39] Alsentzer, E., Murphy, J., Boag, W., Weng, W., Jindi, D., Naumann, T., & McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. In Proceedings of the *2nd Clinical Natural Language Processing Workshop (NAACL-HLT'2019)*, 07 June, Minneapolis, USA, 72–78. https://aclanthology.org/W19-1909.pdf

[40] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T-Y. (2022). BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *Briefings in Bioinformatics*, 23(6), 1-11. https://doi.org/10.1093/bib/bbac409

[41] Zagirova, D., Pushkov, S., Leung, G. H. D., Liu, B. H. M., Urban, A., Sidorenko, D., Kalashnikov, A., Kozlova, E., Naumov, V., Pun, F. W., Ozerov, I. V., Aliper, A., & Zhavoronkov, A. (2023). Biomedical Generative Pre-trained Based Transformer Language Model for Age-Related Disease Target Discovery. *Aging*, 15(18), 9293-9309. https://doi.org/10.18632/aging.205055

[42] Brahami, M., Abdeldjouad, F. Z., Hadj Benaïchouche, M. (2026). Automated Detection of Adverse Drug Reactions from Twitter and PsyTAR using Supervised Machine Learning Models. Network Modeling Analysis in Health Informatics and Bioinformatics, 15(113), 1-20. https://doi.org/10.1007/s13721-026-00788-2

[43] Cullerne B. W. (2024). Sensitivity and Specificity versus Precision and Recall, and Related Dilemmas. *Journal of Classification*, *41*, 402–426. https://doi.org/10.1007/s00357-024-09478-y

[44] Aydemir, O. (2021). A New Performance Evaluation Metric for Classifiers: Polygon Area Metric. *Journal of Classification*, 38, 16–26. https://doi.org/10.1007/s00357-020-09362-5