ADR Identification using BERT for Classification and NER

Divij Mishra
dmishra45@gatech.edu

Aravind Rajeev Nair
anair353@gatech.edu

Naveen Sethuraman
nsethuraman3@gatech.edu

Abstract

Adverse Drug Reactions (ADRs) are a significant concern in medical practice, necessitating effective methods for their identification and management. This study explores the application of advanced natural language processing (NLP) techniques to classify sentences in medical reports indicative of ADRs and extract relevant information about drugs and their adverse effects. We show that BERT-based models demonstrated high accuracy on these tasks, with BioBERT achieving the best performance, surpassing standard BERT and BioClinicalBERT. These findings underscore the importance of pre-training on domain-specific data, as BioBERT, trained on biomedical text, outperformed its counterparts. Overall, this study highlights the efficacy of Transformer-based NLP models, particularly BioBERT, in automating the detection and extraction of ADR-related information from medical text. Future research could focus on refining model architectures and exploring additional sources of domain-specific data to further enhance performance in ADR detection and extraction tasks.

1 Introduction

Adverse Drug Reactions are an ever-pertinent issue in medical fields. The increasing collection of patient and diagnostic data has allowed for natural language analysis to identify instances of ADRs in medical reports for further data collection and analysis.

This study seeks to achieve two goals: to classify whether a given sentence is indicative of an adverse drug reaction or not, and to extract the drug and its adverse effect by name from the sentence. We tested BERT, BioBERT, and BioClinicalBERT for both tasks.

2 Previous Work

One study (Murphy et al., 2023) conducts a scoping review to comprehensively assess the use of NLP methods for ADR detection in hospital settings. LSTMs and CRF methods were found to be commonly employed, although performance evaluation reveals challenges in predicting ADR entities and relations. Recommendations include exploring semi-automated annotation methods and examining NLP implementation in clinical practice. Another group (Gurulingappa et al., 2012) has created a benchmark dataset of ADRs to evaluate NLP methods. A third study (Fei et al., 2021) proposes an enriched contextualized language model (BioKGLM) that integrates extensive biomedical knowledge graphs, This enhanced capability proves to be useful for detecting ADRs accurately and efficiently in medical reports and electronic health records, where subtle mentions of adverse effects and their relationships with drugs require sophisticated information extraction techniques. Advancements in BioIE may significantly contribute to the automated detection, monitoring, and understanding of ADRs, ultimately enhancing patient safety and pharmacovigilance practices.

3 Dataset

We used the Adverse Drug Event benchmark dataset (Gurulingappa et al., 2012) to train and test our models. In particular, we use the version available on HuggingFace (`ade_corpus_v2`). The dataset contains two necessary dataframes for our analysis. The first dataframe (23.5k entries) has labelled pairs of sentences and a binary value indicating whether the sentence describes an adverse drug reaction. The second dataframe (6.8k entries) contains all sentences in the dataset that do indicate an adverse drug reaction, each labelled with the start and endpoints of the substrings referring to the drug and adverse effect in question.

4 Classification

Our first objective was to train a model to accurately classify sentences as referring to an ADR or not. We chose to accomplish this task using the BERT, BioBERT, and BioClinicalBERT models. The input data contains sentences sourced from medical reports. Due to the jargon used and the specialized nature of the subject, classifying these manually is not a trivial task, especially for a non-physician. Training a model to identify ADRs in medical records would enable this difficult task to be accomplished at scale.

4.1 BERT

BERT (Bidirectional Encoder Representation from Transformers) is a Transformer-based pre-trained language model that can be fine-tuned to perform various NLP tasks (Devlin et al., 2019). It has emerged as a groundbreaking model due to its ability to understand contextual information bidirectionally within text. By considering the entire context of a word, BERT generates deeply contextualized word embeddings, enabling it to capture nuanced semantic relationships. This contextual understanding is particularly advantageous for tasks like ADR classification, where accurately identifying whether a sentence is related to ADRs relies heavily on understanding the surrounding context. One of BERT's strengths lies in its pre-trained representations which encode a broad understanding of language patterns and structures and enable efficient adaptation to downstream tasks with fine-tuning. For ADR classification, this means BERT would be able to leverage its pre-trained knowledge to grasp the intricacies of ADR-related language patterns, even with the relatively small amounts of task-specific labeled data at our disposal.

We took advantage of the pre-trained BERT language model and fine-tuned it with our classification training data (80% of the classification data was used as training data, with 10% each reserved for validation and testing). The model was fine-tuned and was able to achieve 84.5% accuracy on the validation data.

4.2 BioBERT

As medical reports often make use of medical jargon and have linguistic features distinct from those of normal English text data, it may prove useful to use a BERT model pre-trained on text data of this nature. The BioBERT model (Lee et al., 2019) uses the BERT architecture but is pre-trained on a large biomedical corpus. We fine-tuned a BioBERT model on the same training data to examine any possible improvements it could provide. The BioBERT model achieved a significantly improved 94.7% accuracy on the validation data.

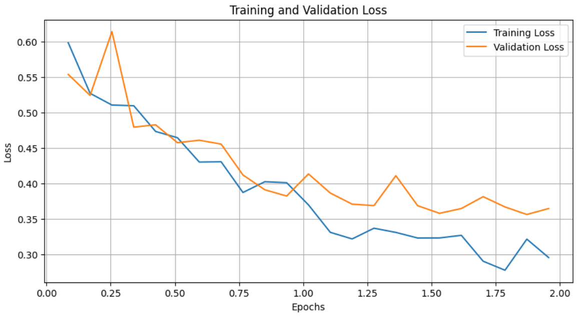


Figure 1: Training and Validation loss of BERT

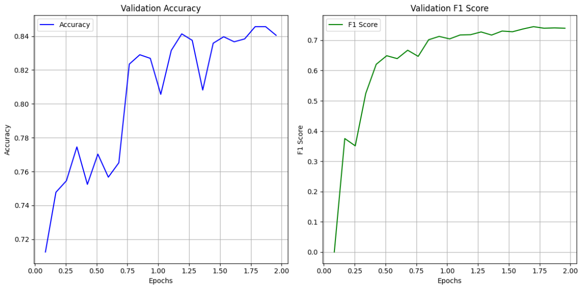


Figure 2: BERT Validation Accuracy and F1-Score as training progresses

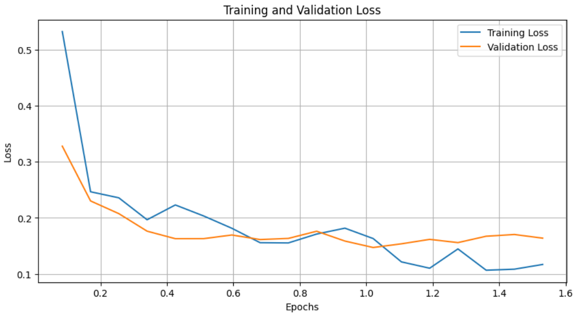


Figure 3: Training and Validation loss of BioBERT

4.3 BioClinicalBERT

Our objective centred specifically on medical data. The BioClinicalBERT model (Alsentzer et al., 2019) is a BioBERT with additional fine-tuning already applied using clinical training data. As such a model may be better suited to our task and data set, we ran the same fine-tuning with a BioClinicalBERT model. The model achieved a similar 93.8% accuracy on the validation data.

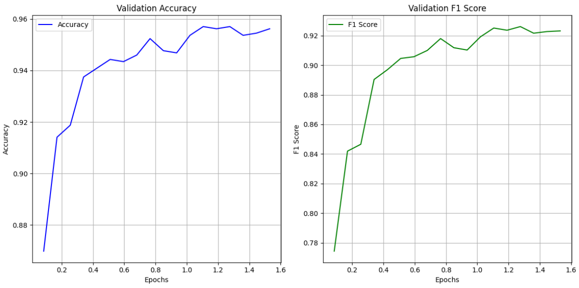


Figure 4: BioBERT Validation Accuracy and F1-Score as training progresses

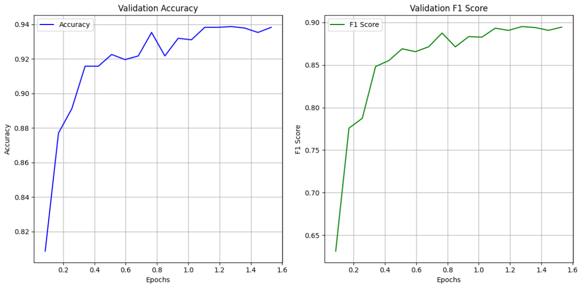


Figure 6: BioClinicalBERT Validation Accuracy and F1-Score as training progresses

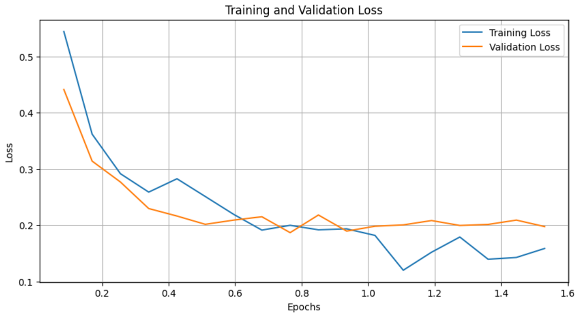
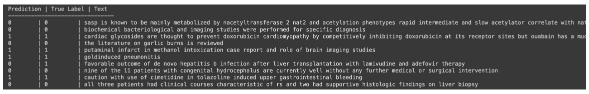


Figure 5: Training and Validation loss of BioClinical-BERT



```
Prediction | True Label | Text
--------------------------------
0        | 0          | sasp is known to be mainly metabolized by nacetyltransferase 2 nat2 and acetylation phenotypes rapid intermediate and slow acetylator correlate with nat
0        | 0          | biochemical bacteriological and imaging studies were performed for specific diagnosis
1        | 1          | cardiac glycosides are thought to prevent doxorubicin cardiomyopathy by competitively inhibiting doxorubicin at its receptor sites but ouabain has a muc
0        | 0          | the literature on garlic burns is reviewed
1        | 1          | putaminal infarct in methanol intoxication case report and role of brain imaging studies
1        | 1          | goldinduced pneumonitis
0        | 1          | favorable outcome of de novo hepatitis b infection after liver transplantation with lamivudine and adefovir therapy
0        | 0          | nine of the 11 patients with congenital hydrocephalus are currently well without any further medical or surgical intervention
1        | 1          | caution with use of cimetidine in tolazoline induced upper gastrointestinal bleeding
0        | 0          | all three patients had clinical courses characteristic of rs and two had supportive histologic findings on liver biopsy
```

Figure 7: Sample predicted values from BioClinical-BERT

4.4 Comparison

All three models performed fairly well on the task. The BioBERT and BioClinicalBERT had significantly metrics 1. The BioBERTs outperforming the base BERT stands to show the advantageous effect of the specialized data used to train BioBERT. The even more specialized BioClinicalBERT fared slightly poorer than BioBERT. This suggests that the hyper-specialized nature of BioClinicalBERT's training data did not provide much advantage towards accomplishing the objective of identifying ADR-related sentences over the already specialized BioBERT.

Model	F1-score	Precision	Recall
BERT	0.7605	0.6939	0.8413
BioBERT	0.911	0.8947	0.927
BioClinBERT	0.8939	0.8868	0.9010

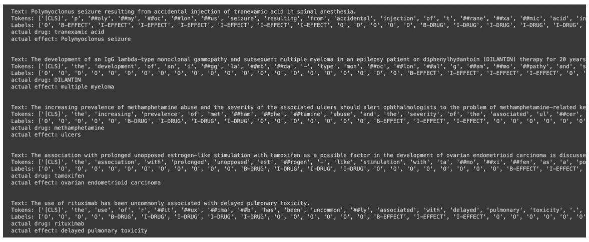
Table 1: Model Performance Metrics

5 NER Tagging

Our second objective in this project was to take sentences referring to an ADR and identify the named entities corresponding to the drug and the adverse reaction it caused. For this task, we chose to continue with the same set of three BERT models (BERT, BioBERT, and BioClinicalBERT). As the BERT architecture is flexible and can be fine-tuned to perform for a variety of tasks, we have opted to used it here for NER as well.

5.1 Preprocessing for NER

The dataset contained only the sentence and the start and end points of the two relevant substrings. In order to use it to train BERT, the sentences needed to be tokenized and the substring indices needed to be used to label each token. Firstly, wt tokenize the input texts, ensuring uniform length through padding or truncation up to a maximum length of 512 tokens. Subsequently, we initialize the labels for each token as 'O'. The code then iterates through each text and identifies word boundaries. It then tokenizes the drug and effect entities and assigns labels to the tokens based on their positions within the text. The labeling scheme follows the NER convention, where entities are denoted by labels starting with 'B-' (indicating the beginning of an entity) and followed by 'I-' (indicating inside an entity) for subsequent tokens of the same entity.



```
Text: Polymyoclonus seizure resulting from accidental injection of tranexamic acid in spinal anesthesia.
Tokens: ['[CLS]', 'p', '##oly', '##my', '##oc', '##lon', '##us', 'seizure', 'resulting', 'from', 'accidental', 'injection', 'of', 't', '##rane', '##xa', '##mic', 'acid', 'in
Labels: ['O', 'B-EFFECT', 'I-EFFECT', 'I-EFFECT', 'I-EFFECT', 'I-EFFECT', 'I-EFFECT', 'I-EFFECT', 'O', 'O', 'O', 'O', 'O', 'B-DRUG', 'I-DRUG', 'I-DRUG', 'I-DRUG', 'I-DRUG',
actual drug: tranexamic acid
actual effect: Polymyoclonus seizure


Text: The development of an IgG lambda-type monoclonal gammopathy and subsequent multiple myeloma in an epilepsy patient on diphenylhydantoin (DILANTIN) therapy for 20 years
Tokens: ['[CLS]', 'the', 'development', 'of', 'an', 'i', '##gg', 'la', '##mb', '##da', '-', 'type', 'mon', '##oc', '##lon', '##al', 'g', '##am', '##mo', '##pathy', 'and', 's
Labels: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-EFFECT', 'I-EFFECT', 'I-EFFECT', 'I-EFFECT', 'O', '
actual drug: DILANTIN
actual effect: multiple myeloma


Text: The increasing prevalence of methamphetamine abuse and the severity of the associated ulcers should alert ophthalmologists to the problem of methamphetamine-related ke
Tokens: ['[CLS]', 'the', 'increasing', 'prevalence', 'of', 'met', '##ham', '##phe', '##tamine', 'abuse', 'and', 'the', 'severity', 'of', 'the', 'associated', 'ul', '##cer',
Labels: ['O', 'O', 'O', 'O', 'O', 'B-DRUG', 'I-DRUG', 'I-DRUG', 'I-DRUG', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-EFFECT', 'I-EFFECT', 'I-EFFECT', 'O', 'O', 'O', 'O', 'O', 'O'
actual drug: methamphetamine
actual effect: ulcers


Text: The association with prolonged unopposed estrogen-like stimulation with tamoxifen as a possible factor in the development of ovarian endometrioid carcinoma is discusse
Tokens: ['[CLS]', 'the', 'association', 'with', 'prolonged', 'unopposed', 'est', '##rogen', '-', 'like', 'stimulation', 'with', 'ta', '##mo', '##xi', '##fen', 'as', 'a', 'po
Labels: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-DRUG', 'I-DRUG', 'I-DRUG', 'I-DRUG', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-EFFECT', 'I-EFFECT',
actual drug: tamoxifen
actual effect: ovarian endometrioid carcinoma


Text: The use of rituximab has been uncommonly associated with delayed pulmonary toxicity.
Tokens: ['[CLS]', 'the', 'use', 'of', 'r', '##it', '##ux', '##ima', '##b', 'has', 'been', 'uncommon', '##ly', 'associated', 'with', 'delayed', 'pulmonary', 'toxicity', '.',
Labels: ['O', 'O', 'O', 'O', 'B-DRUG', 'I-DRUG', 'I-DRUG', 'I-DRUG', 'I-DRUG', 'O', 'O', 'O', 'O', 'O', 'O', 'B-EFFECT', 'I-EFFECT', 'I-EFFECT', 'O', 'O', 'O', 'O', 'O', 'O'
actual drug: rituximab
actual effect: delayed pulmonary toxicity
```

Figure 8: A sample of the preprocessed data

5.2 Comparison

All three models performed fairly similarly on NER tagging. BioBERT had a slightly higher F1 score as before, though BioClinicalBERT had a slightly higher recall.

Model	F1-score	Precision	Recall
BERT	0.6332	0.5789	0.6987
BioBERT	0.6552	0.602	0.7187
BioClinBERT	0.6437	0.5804	0.7225

Table 2: Model Performance Metrics

Conclusions and Recommendations

The successful application of Transformer-based NLP models in classifying ADR-related sentences and extracting drug-adverse effect pairs signifies a promising avenue for streamlining ADR detection processes in healthcare. The superior performance of BioBERT over standard BERT underscores the significance of pre-training on domain-specific biomedical text. Future efforts should prioritize the acquisition and curation of large-scale biomedical corpora to facilitate the development of more robust NLP models tailored to medical applications. Although all models performed adequately in NER tagging, there is room for improvement.

In conclusion, while Transformer-based NLP models show great promise in automating ADR detection and extraction tasks, further efforts are needed to address existing challenges and optimize model performance for real-world healthcare applications. By leveraging the strengths of these models and continually refining their capabilities, we can enhance patient care and contribute to the advancement of pharmacovigilance practices in the medical field.

Limitations

The data available at our disposal was limited in size and nature. There were only 16k observations. Though this may be sufficient to get decent results with a BERT model, there can always be more improvements made using larger sources of data. Secondly, the data consisted solely of medical reports. This limits the scope of the study to the analysis and extraction of ADRs from medical documents, i.e. data collected post-diagnosis. This will not aid in identifying ADRs from patient symptom descriptions, as such examples are not contained within our training data.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. 2021. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3):bbaa110.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 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 – 892. Text Mining and Natural Language Processing in Pharmacogenomics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

R. M. Murphy, J. E. Klopotowska, N. F. de Keizer, K. J. Jager, J. H. Leopold, D. A. Dongelmans, A. Abu-Hanna, and M. C. Schut. 2023. Adverse drug event detection using natural language processing: A scoping review of supervised learning methods. *PloS One*, 18(1):e0279842.