Named Entity Recognition (NER) and Relation Classification on Drug-Related Adverse Effects

Diana Nguyen and Nan Li
UC Berkeley School of Information

Abstract
Drug-related adverse events (ADE) are a public health concern. The Centers for Disease Control and Prevention (CDC) states that there are over a million emergency department visits each year caused by adverse drug events. This problem will get worse as people age and take more medication for treatment. With that in mind, the objective is to use named entity recognition (NER) and relation classification to improve safety, clinical decision-making and knowledge management making it easier to search, access, and analyze. Models (SVM, Gradient Boosting, BERT, and BioBERT) will be compared and incorporated to categorize relevant information such as drug names and the adverse effects, and to extract relevant relationships between them.

Introduction
Drug-related adverse events (ADE) are negative reactions followed by the use of medication. These negative reactions are a public health concern, and it can range from being mild to severe to life threatening. Drug-related adverse events can happen to anyone, and can occur for many reasons including: drug side effects, multiple drug interactions, allergic reactions, existing medical conditions, incorrect drug dosage, etc. There are over a million emergency department visits each year due to negative drug events. For this reason, it is important to establish pharmacovigilance, coined by the World Health Organization (WHO), to detect, assess, understand, and prevent pharmacological adverse effects.

To support pharmacovigilance, there is a need for reviewing large volumes of medical texts to identify the relationship between drugs and effects. This can be time consuming and error prone. Natural language processing (NLP) techniques, such as named entity recognition (NER) and relation extraction (RE), can be useful to identify and extract relevant information and relationship from unstructured textual data of medical documents. NER can learn to recognize and classify specific entities such as drug names and its corresponding drug effects, and RE can extract and determine the relevance of the causal relationship between drugs and effects for given texts.

Many have approached this problem using a combination of models such as Bidirectional Long Short-Term Memory (BiLSTM) along with conditional random fields (CRF), convolutional neural network (CNN), or other classical machine learning models.
For this study, Bidirectional Encoder Representations from Transformers (BERT),

and a biomedical language representation model (BioBERT) will be compared for NER. Furthermore, Support Vector Machines (SVMs), Gradient Boosting(GB), and BioBERT will be compared for RE. Classification metrics including precision, recall, F1, and accuracy will be used as the evaluation metric. It is expected for BioBERT to do better than BERT, SVM, and GB due to the fact that it is pretrained on domain specific biomedical texts. Lee et al. had pre-trained BioBERT on biomedical domain corpora after starting with weights from BERT which was pre-trained on general domain corpora [10].

Background and Related work

Recent advances in natural language processing (NLP) have led to the development and application of various techniques for detecting adverse drug events (ADE) from unstructured medical texts.

A number of studies have focused on named entity recognition (NER) models to identify drugs and effects in clinical text. For example, Christopoulou et al. [1] and Li et al. [3] used a combination of bidirectional long short-term memory (BiLSTM) networks and conditional random fields (CRF) models for NER, while Haq et al [8] and Yang et al. [6] introduced new NER models based on BiLSTM and CNN that achieved state-of-the-art model metrics. In addition to predominant neural networks, the pretrained BERT model was also applied in this field by fine-tuning with a highly modular Framework for Adapting Representation Models (FARM) [9].

In addition to NER models, several studies have also explored relation extraction (RE) models to identify the relationship between drugs and effects/symptoms in clinical text. Christopoulou et al. [1] uses a convolutional neural network (CNN) for RE, while Miller et al.[4] employs both a deep learning-based approach that combines BiLSTM and attention mechanisms and a support vector machine (SVM) classifier. Yang et al. [6] applied traditional machine learning algorithms, such as support vector machines, random forests and gradient boosting, for relation classification tasks. Haq et al [8] also introduces two new RE models, one based on BioBERT and the other utilizing crafted features over a Fully Connected Neural Network (FCNN), both of which performed on par with existing state-of-the-art models.

Several studies developed the end-to-end relation extraction, which performs a joint entity and relation extraction from clinical text. A deep neural model called SpERT.PL based on SciBERT/BioBERT [5] and a pipeline system [1] based on BiLSTM was proposed. An integrated system which employs a conditional random field (CRF) model for named entity recognition (NER) and a random forest model for relation extraction (RE) also achieved state-of-the-art performance [7].

To evaluate the performance of these models, various metrics have been used such as F1 score, precision, recall, and accuracy. For instance, Chapman et al. [7] reports an F1 score of 80.9% for NER, 88.1% for RE, while Miller et al. [4] reported an F1 score

of 0.91 and 0.90 for NER using SVM and LSTM based models, respectively, and 0.91 and 0.93 for relation classification using SVM and LSTM based models, respectively.

The studies reviewed above indicate that great effort has been put into the development of advanced NER and RE models to handle the variability and complexity of clinical text. Deep neural networks such as BiLSTM and CNN are the most commonly used models for these tasks. While few studies have utilized variants of the BERT model, those that have done so have shown promising results for NER and RE tasks [5]. Interestingly, traditional machine learning models also demonstrated promising results in some cases and can be comparable to neural networks. In this study, we aim to employ pre-trained transformer (BERT and BioBERT) models and traditional machine learning models for conducting NER and relation classification tasks, with the goal of achieving high accuracy and computational efficiency.

Dataset

The Adverse Drug Events (ADE) Corpus [2] from the Journal of Biomedical Informatics (JBI) will be utilized for this study. Development of the ADE corpus was generated from detailed medical case reports, which were collected from PubMed. The ADE corpus was narrowed down from 30,000 drug therapy and adverse effect related medical case documents to 3,000 randomly selected documents for annotation. The ADE corpus contains three files including relations between drugs and adverse effects with 6,821 rows of texts and 8 columns, drugs and dosages with 279 rows of texts and 8 columns, and all sentences that do not contain any drug-related adverse effects with 16,695 rows of texts and 1 column.

For the NER task, the relations between drugs and adverse effects dataset was loaded from Hugging Face [11]. The dataset contains sentences, drug and effect annotations and their offset. From the exploratory data analysis, out of 6,821 rows in the dataset, 4,271 rows have unique sentences with 1,319 unique drugs, and 3,341 unique effects. After removing duplicate sentences, an extra column was added to the dataset for inside-outside-beginning (IOB) tagging. This was done by mapping the offsets. The dataset was then split: 80% trained, 10% validation, and 10% test.

For the RE task, the classification dataset was loaded from Hugging Face [11], which includes two files, both relations between drugs and adverse effects, and sentences that do not contain any drug-related adverse effects. This dataset contains a total of 23,516 sentences, and a column for positive (1) and negative (0) relationship classification. The dataset was split into 80% train and 20% test.

Methodology

NER

Named entity recognition is performed to identify and extract drug and effect entities. Each token of each sentence is tagged with

0, 1, 2, 3, or 4, or “O”, “B-Drug”, “I-Drug”, “B-Effect”, and “I-Effect” respectively (IOB tagging) for a total of 5 classes. The numerical tags will be used as the labels for training, validation, and test. Additionally, each sentence has inconsistent lengths, thereby padding is required for dimensional reasons. Note that due to class imbalance, with more “O” than the other classes, accuracy metric is expected to be high. The texts are then encoded for input ids, token type ids, and attention masks, which are used as training, validation, and test input. Since this is a multiclass classification problem, the loss is specified as sparse categorical cross entropy. Five epochs were run with a batch size of 16, and learning rate as 1E-5. This is done for both BERT as the baseline model and BioBERT. We also experimented with all lowercase and removing stop words, but the changes in results were insignificant; therefore, it was left out of the analysis.

Relation Classification

We performed relation classification to identify the relationships between drug names and the adverse symptoms they caused in the free-text data. The annotation of drug and symptom entities as relevant or irrelevant allowed us to approach the development of the relation classification module as a supervised classification problem. To tackle this, we utilized two classical machine learning models - support vector machines and gradient boosting - in conjunction with pre-trained word embedding models developed by the spaCy library. Class weights were applied to address the class imbalance issue in the dataset. In addition, we implemented a third system that leveraged the BioBERT model for the binary text classification task. Five epochs were run with a batch size of 16 and the learning rate as 2E-5.

Results

NER

The classification metrics for NER are listed in Table 1. The results show that BioBERT did better than BERT for both drug and effect entities, as well as overall with an F1 score of 0.82 for BERT and 0.88 for BioBERT. The results suggest that training on domain specific corpora is recommended in order to get the best outcome.

Relation Classification

Table 2 summarizes the performance of relation classification for all models on the test set. The BioBERT model achieved the best performance with an F1 score of 0.94, outperforming the SVM and GB models. The results indicate that the application of advanced transformer models significantly improve the accuracy of the relation classification task.

Table 1. Classification metrics for named entity recognition on the test set.

	Class	Precision	Recall	F1	Accuracy
BERT (baseline)	Drug	0.89	0.94	0.91	
	Effect	0.65	0.84	0.73	
	Overall	0.76	0.89	0.82	0.98
BioBERT	Drug	0.92	0.96	0.94	
	Effect	0.78	0.86	0.82	
	Overall	0.85	0.92	0.88	0.99

Table 2. Classification metrics of relation classification on the test set. (Precision, recall and F1 scores are macro values.)

	Precision	Recall	F1	Accuracy
SVM (baseline)	0.75	0.79	0.76	0.78
GB	0.84	0.83	0.83	0.86
BioBERT	0.93	0.95	0.94	0.95

Conclusion

For the named entity recognition task, with room for improvement, the BioBERT model did better compared to BERT with an F1 score of 0.88. For the relation classification task, the BioBERT model stood out with its exceptional performance, achieving a state-of-the-art F1 score of 0.94 and accuracy of 0.95. The findings suggest that utilizing models trained on domain-specific datasets is advisable to achieve optimal results.

Acknowledgement

The NER code framework was adopted with major changes from

https://github.com/jsylee/personal-projects/blob/783f84980f5209be9c39edf03f931ff778ffa3ee/Hugging%20Face%20ADR%20Fine-Tuning/SciBERT%20ADR%20Fine-Tuning.ipynb

References

1. 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: JAMIA, 27(1), 39–46. https://doi.org/10.1093/jamia/ocz101

2. 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), 885–892. https://doi.org/10.1016/j.jbi.2012.04.008.

3. Li F., Liu W., Yu H. (2018), Extraction of Information Related to Adverse Drug Events from Electronic Health Record Notes: Design of an End-to-End Model Based on Deep Learning, *JMIR Med Inform* , 6(4):e12159. doi: 10.2196/12159. https://medinform.jmir.org/2018/4/e12159/

4. Timothy Miller, Alon Geva, and Dmitriy Dligach. (2019). Extracting Adverse Drug Event Information with Minimal Engineering. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 22–27, Minneapolis, Minnesota, USA. Association for Computational Linguistics. https://aclanthology.org/W19-1903/

5. T.Y.S.S, S., Chakraborty, P., Dutta, S., Sanyal, D.K., & Das, P.P. (2021). Joint Entity and Relation Extraction from Scientific Documents: Role of Linguistic Information and Entity Types. *EEKE@JCDL*. https://ceur-ws.org/Vol-3004/paper2.pdf

6. Yang, X., Bian, J., Fang, R., Bjarnadottir, R. I., Hogan, W. R., & Wu, Y. (2020). Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. *Journal of the American Medical Informatics Association : JAMIA*, *27*(1), 65–72. https://doi.org/10.1093/jamia/ocz144

7. Chapman, A.B., Peterson, K.S., Alba, P.R. et al. Detecting Adverse Drug Events with Rapidly Trained Classification Models. Drug Saf 42, 147–156 (2019). https://doi.org/10.1007/s40264-018-0763-y

8. Haq, H.U., Kocaman, V., Talby, D. (2023). Mining Adverse Drug Reactions from Unstructured

Mediums at Scale. In: Shaban-Nejad, A., Michalowski, M., Bianco, S. (eds) Multimodal AI in Healthcare. Studies in Computational Intelligence, vol 1060. Springer, Cham. https://doi.org/10.1007/978-3-031-14771-5_26

9. Hussain, S., Afzal, H., Saeed, R., Iltaf, N., & Umair, M. Y. (2021). Pharmacovigilance with Transformers: A Framework to Detect Adverse Drug Reactions Using BERT Fine-Tuned with FARM. Computational and mathematical methods in medicine, 2021, 5589829. https://doi.org/10.1155/2021/5589829

10. Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics, Volume 36, Issue 4, February 2020, Pages 1234–1240, https://doi.org/10.1093/bioinformatics/btz6

11. Hugging Face. (2021). Dataset: ade_corpus_v2. Retrieved March 5, 2023, from https://huggingface.co/datasets/ade_corpus_v2