ISSN: 2455-2631 May 2023 IJSDR | Volume 8 Issue 5

Detection of Adverse Drug Reactions Due to Anti-Depressants Using Bio BERT

[1]Samhitha. V.B, [2]Ms.S. Sarjun Beevi, V.B, [3]Dr.V. Balambica

[1]Student, [2]Assistant Professor, [3]Professor&Head
Department of Computer Science and Engineering
Bharath Institute of Higher Education and Research
Chennai, India

Abstract- **An Adverse Drug Reaction is the negative effect experienced by the patient when administered multiple medications. According to research, antidepressants and antipsychotics cause the majority of ADRs in the psychiatry department but unfortunately, it is widely overlooked when compared to other divisions. Although a drug is run through multiple clinical trials before its release in the market, there is always potential for it to cause custom harm in real-life practice. Therefore, post-market safety monitoring of medicines through pharmacovigilance is essential. One needs to track all research papers and health records to identify and prevent potential adverse events related to drugs. However, the avalanche of data available in research archives spotlights the challenge of finding relevant information for pharmacovigilance. NER-Bio BERT (Bidirectional Encoder Representation from Transformers) highlights ways to lower the barrier by automating literature from research repositories to support pharmacovigilance activities, which can further be used as input for our objective of ADE identification in antidepressants.**

Index Terms: **NER, NER-Bio BERT, ADR, Natural Language Processing**

I. INTRODUCTION

Prescription drug usage for mental health continues to be in practice in healthcare strategies to improve the quality of life. Although prescribed for their therapeutic properties, there is always the potential for them to cause unintended, negative adverse effects. Rigorous clinical trials of a drug before production in the market do not warrant desirable effects. A heterogeneous population when catalyzed with various drugs over time might cause benign to life-threatening effects. The morbidity and mortality rates associated with ADRs vary for different drugs in the market.

Mental health is widely misunderstood and overlooked among the Indian population due to lack of awareness and false notions. On the other end of the spectrum, patients who actively seek help to treat mental chemical imbalances are not familiar with the various side effects that come with their medications.

Post-monitoring surveillance in the day-to-day life of individuals is essential in the lifecycle of any drug and the same is carried out by Pharmacovigilance. WHO describes Pharmacovigilance as the field of study and actions concerning the identification, evaluation, comprehension, and avoidance of unfavorable outcomes or other medicine/vaccine-associated concerns.

Evolved from the BERT model, which was explicitly trained on English literature texts, the domain-specific BioBERT is trained on medical corpora for increased accuracy and context-friendliness. BioBERT is fine-tuned and evaluated on three tasks: named entity recognition, relation extraction, and question answers: out of which Named-Entity Recognition (NER) can be used for identifying mentions of drugs and medical concepts in the text as it allows us to determine which text strings correspond to specific medications or medical concepts.

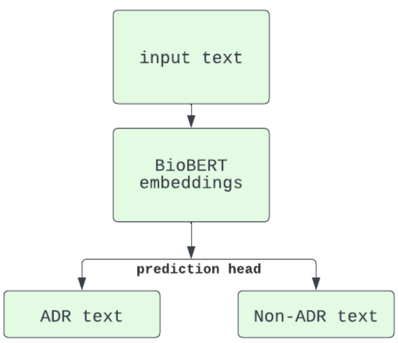

Figure 1. Workflow of the ADR system

To uncover relationships between the responsiveness of antidepressants in different patients, we implement a computational framework built on top of the pre-trained BioBERT model which primarily uses transfer learning, on the crowd-surveyed user reviews from different consumer health forums.

II. LITERATURE SURVEY

[1] **Sajid Hussain et al.** present an end-to-end system for modelling ADR detection from textual input through FARM-BERT(Framework for Adapting Representation Models), which provides support for multitask learning by combining multiple

prediction heads: text classification and ADR sequence; which makes training of the end-to-end systems simpler and computationally

[2] **Zhang Tongxuan et. al** discuss drug safety issues and information extracted through social media channels. He states the difficulty of using both PubMed and social media datasets due to confusing contextual differences. Therefore, aims to improve the generalization of the model and mitigate the noise caused by colloquial expressions in social media. For the problem of imbalanced data between the two different scale datasets, dynamic weights have been imposed to integrate the different datasets' loss functions, which balances the significance of each resource.

[3**] Ed-drissiya El-allaly et. al** recognizes the issue of extraction of mentions of ADR from biomedical texts as many ADR mentions are nested, discontinuous and overlapping. To solve the same, a deep model for Complex Adverse Drug Reaction Mentions Extraction (DeepCADRME) has been proposed. ADR mentions are transferred as an extraction problem in an N-level tagging sequence and fed into an N-level model based on contextual embeddings where the output of the current-level pre-trained model is utilized to create a novel deep contextualized representation for the subsequent level.

[4] **Shuai Chen et. al** have developed a system to extract adverse drug reactions based on bidirectional encoder representations from transformers (BERT). They excel in the automatic classification of adverse effects from social media mining with a score of 0.614 but fail to explicitly distinguish ADR from the reasons mentioned for taking the medications and in the implicit identification of their effects.

[5] **Brandon Fan et. al** draws model comparisons between the latest statistical models, conventional deep learning models, and deep learning models that employ BERT word embeddings and sentence embeddings for the purpose of identifying and extracting ADEs (Adverse Drug Events) from unstructured reviews.

III. EXISTING SYSTEM

Most existing solutions deploy supervised learning methods for extraction of ADR mentions, but they are met with the problem of labelled data scarcity. Lack of labelled data makes it difficult to train a model which requires a huge dataset. The current techniques used for adverse drug reaction (ADR) mention extraction depend on supervised learning approaches, which are hampered by a lack of labeled data.

A few of the conventional machine learning models used for this purpose are Support Vector Machine, Random Forest, and Conditional Random Field. These models depend upon manual feature engineering. Conventional means of identifying adverse drug reactions are dependable but sluggish and only generate a limited amount of information

The prevalent characteristics employed by these models consist of n-grams, negated contexts, semantic types from the Unified Medical Language System (UMLS), Part of Speech (POS) tags, drug names, as well as lexicon-based features and word embeddings. Numerous studies utilize deep learning techniques such as Bidirectional Long Short-Term Memory, Convolutional Neural Networks, and attention-based deep neural networks.

IV. PROPOSED SYSTEM

Consumer Health forums like Askapatient provides comparative drug and healthcare information including unfiltered consumer reviews on medication experiences. With a database of over 4,000 chemically prepared prescription and biological drugs approved by the FDA's Center for Drug Evaluation and Research, one can rely on it for it to be used as a real-time factual dataset. Using such health forums, we can mine user opinions after their experience with their prescription drugs.

We analyse the consumer review of six popularly used medications, where the former three are antidepressants and the latter are primarily used as anxiety disorder pills. The proposed system is a computational framework built on top of the BioBERT model which pertains to transfer learning techniques. The model is further fine-tuned to analyse the transfer learning capacities and efficiencies over cross-domain drugs, on different combinations of the dataset. As a result, we report the adverse drug effects detected by our model; analyze the reason behind the most common complications and their severities and present the top ten ADRS for each drug. The result would then be checked with government-recognised healthcare websites for proof of correctness. A dashboard wrapper would be built around the analysed output for user-friendliness and ease of access.

BioBERT

The significance of text mining in the clinical domain has grown due to the multitude of biomedical documents available, containing valuable data waiting to be decoded and enhanced by natural language processing (NLP) methods. As a state-of-the-art breakthrough in NLP, Google researchers developed a language model known as BERT, which was designed to acquire profound representations by considering both forward and backward contexts of the text simultaneously across all layers of its structure. The encoded representations are particularly beneficial for sequential data, such as text, which heavily depends on context. The development of transfer learning in this domain has facilitated the transfer of encoded knowledge from pre-trained models to downstream tasks. In their pre-training process, the initial English language model utilized two datasets: Wikipedia and BooksCorpus.

BioBERT is a variation of the aforementioned model, which was previously trained using general corpora. In addition to general language, BioBERT is also trained using biomedical corpora from PubMed and PMC, where, PubMed is a database of biomedical citations and abstractions, whereas PMC is an electronic archive of full-text journal articles. This is fine-tuned in order to reap domain-specific results and add more context to the subject matter. This biomedical language representation model can manage three tasks successfully:

Named-entity recognition: Named-entity recognition (NER) is the recognition process of numerous proper nouns that we establish as entity types to be labeled The datasets employed to assess named entity recognition (NER) adhere to the BIO (Beginning, Inside, Outside) format, which is the most widely used tagging scheme for sentence tokens in this task.. Additionally, an important word or letter like can be used to infer a single token. In this manner, the training data is used to obtain the positional prefix and predicted entity type.

Relation Extraction: Being able to automatically extract relationships between entities in free text is useful to analyse data, in our case, clinical data, and therefore extract information. Relation Extraction is the task of forecasting characteristics and connections between entities within a sentence.

Question-Answering: True to its name, this task focuses on extracting relative answers to questions from the subject matter. For the Question Answering System, BioBERT takes two parameters, the input question, and passage as a single packed sequence The input embeddings are computed as the summation of the token embeddings and the segment embeddings. Segment embeddings are used to differentiate the question from the reference text.

Through the utilization of a pre-trained language model that includes both general and biomedical domain datasets, developers and practitioners can now incorporate biomedical terminology that would have been highly challenging for a general model to comprehend.

Architecture

The proposed system is lucid in nature and consists of minimal hardware, hence proving to be cost-efficient. A primary dataset is made from scratch using the user reviews of the chosen medications from popular healthcare websites. This dataset will be used as the training dataset for the system, over which the pretrained BioBERT weights are implemented. Two out of three tasks of BioBERT are made use of to draw a relation between drugs and their adverse events. A pipeline is employed, where first relevant clinical entities are recognized, and next the drug-disease candidate pairs are judged as either ADR or non-ADR. For the first task, domain-trained BIO-NER (Named-Entity Recognition) is used. The NER system recognizes and classifies the biomedical substances from the text. The relation extraction system identifies the relation between drugs and medication-related entities. These two tasks in unison can help segregate drugs from their ADR. Out of many, the top then ADRs are isolated and compared with the standards set by government-recognized healthcare websites for proof of correctness. A dashboard wrapper will be built around this model to display the data analysis results in a more user-friendly and efficient manner.

Methodology

Exploiting the liberal amount of user reviews available on various consumer healthcare welfare websites, one can draw a relation between drugs and the side effects suffered by their consumption. The abstracted version of this process is given below:

A. Data Preparation

Even though the detection of adverse effects of drugs was done prior, there is very less information known about extracting the same for antidepressants. There are no reliable datasets available on the internet for mental health medications. So, it is necessary to build a dataset from scratch.

As the primary data is collected from user reviews from health forums, we employ web scraping techniques to extract information from the internet. Three different health forums are considered along with six different drugs from all those websites. Over 300 reviews for each drug are available on each website. Even with the availability of plentiful data, one can still run into problems due to its quality, as well as hidden biases. The quality of the data needs to be fine-tuned to improve the performance of the system.

B. Data Preprocessing

After diverse data is collected, it is essential to pre-process it into the optimal format. Data cleaning is done to deal with missing values and remove unwanted characters. A customized vocabulary is built by tokenizing the text into a collection of unique words. Along with tokenization, lemmatization and stemming are other steps that take place to extract quality data.

C. BioBERT Embeddings

BioBERT generates contextualized embeddings as it is exclusively trained only on biomedical corpora. Numerous models have been extensively utilized for transforming words into embeddings. However, these models produce word embeddings without considering the surrounding context. These models fail to capture the context-dependent representation, which leads to similar vector representations of a word with different meanings in different contexts. As opposed to the previous models, BioBERT generates contextualized embeddings. Apart from generating token embeddings, BERT also generates sentence embeddings by adding embeddings to each token in the tokenized text, indicating whether the token belongs to the first or the second sentence. In addition, BERT creates position embeddings that indicate the position of each token in the input sequence. Essentially, to represent the input for a specific token, you can combine its corresponding embeddings for the token itself, the sentence it belongs to, and its position within that sentence.

D. Named-Entity Recognition

Bio-NER is the process of identifying and classifying named entities that fall under the biomedical corpora into predefined entity categories. The classifier model should be able to classify two different entities and categorize them accordingly. This is usually a non-ignorable step in information extraction. Bio-NER is employed for mining purposes in freely available user reviews. Annotation is done prior to NER deployment. Only the drug, disease, and the adverse events are considered for annotation purposes, to limit the manual workload and processing time.

E. Relation Extraction:

Most interactions mentioned in the consumer healthcare forums are free text and are not yet contained in structured databases. Relationship extraction, as shown in Fig.2 is the task of extracting semantic relationships from a given text. BIO-relation extraction is exclusively trained on biomedical knowledge, so it is context-friendly for our project. The extracted relationships usually occur between two or more entities of a certain type, in our case, it would be drugs and diseases, and fall into a number of semantic categories, in our context, Adverse effects.

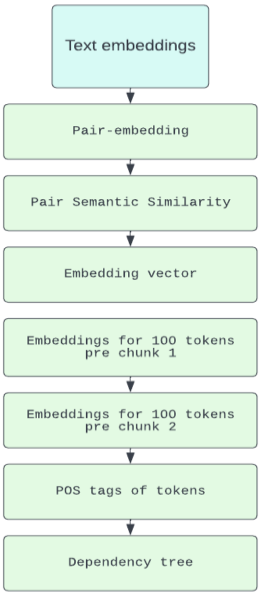

Fig 2. Simple Relation Extraction

V. ARCHITECTURE DIAGRAM

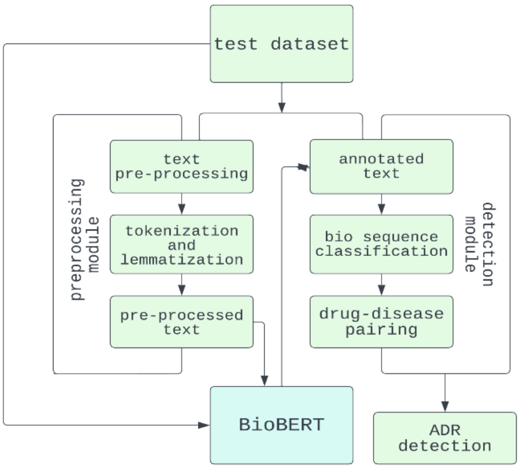

Fig 3. Working of the ADR system

REFERENCES:

1. A. Cocos, A.G. Fiks, A.J. Masino. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. J. Am. Med. Inform. Assoc., 24 (2017), pp. 813-821, 10.1093/jamia/ocw180
2. Brandon Fan, Weiguo Fan, Carly Smith, Harold “Skip” Garner, Adverse drug event detection and extraction from open data: A deep learning approach,Information Processing & Management, Volume 57, Issue 1, 2020, 102131, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2019.102131.(https://www.sciencedirect.com/science/article/pii/S0306457319308623)
3. ChenY. *et al.* Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training(2019) J. Biomed. Inform.
4. El-allaly, Ed-Drissiya & Sarrouti, Mourad & Ennahnahi, Noureddine & Ouatik El Alaoui, Said. (2021). DeepCADRME: A deep neural model for complex adverse drug reaction mentions extraction. Pattern Recognition Letters. 143. 10.1016/j.patrec.2020.12.013.
5. H.J. Dai, M. Touray, J. Jonnagaddala, S. Syedabdul. Feature engineering for recognizing adverse drug reactions from twitter posts. Inform. Int. Interdisciplinary J., 7 (2016), p. 27, 10.3390/info7020027
6. Hussain, Sajid & Afzal, Hammad & Saeed, Ramsha & Iltaf, Naima & Umair, Mir. (2021). Pharmacovigilance with Transformers: A Framework to Detect Adverse Drug Reactions Using BERT Fine-Tuned with FARM. Computational and Mathematical Methods in Medicine. 2021. 1-12. 10.1155/2021/5589829.
7. I.J. Unanue, E.Z. Borzeshi, M. Piccardi. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. J. Biomed. Inform., 76 (2017), pp. 102-109, 10.1016/j.jbi.2017.11.007
8. S.Gupta, S. Pawar, N. Ramrakhiyani, G.K. Palshikar, V. Varma. Semi-supervised recurrent neural network for adverse drug reaction mention extraction. BMC Bioinf., 19 (2018), p. 212, 10.1186/s12859-018-2192-4

9. Wei Li, Yajun Du, Xianyong Li, Xiaoliang Chen, Chunzhi Xie, Hui Li, Xiaolei Li. Named entity recognition in social network combined BERT-BiLSTM-CRF with active learning. Engineering Applications of Artificial Intelligence, Volume 116(2022), ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2022.105460.
10. X. Dong, L. Qian, Y. Guan, L. Huang, Q. Yu, J. Yang. A multiclass classification method based on deep learning for named entity recognition in electronic medical records. Sci. Data Summit (2016), pp. 1-10, 10.1109/NYSDS.2016.7747810
11. Zhang, Tongxuan & Lin, Hongfei & Ren, Yuqi & Yang, Zhihao & Wang, Jian & Duan, Xiaodong & Xu, Bo. (2021). Identifying Adverse Drug Reaction Entities from Social Media with Adversarial Transfer Learning Model. Neurocomputing. 453. 10.1016/j.neucom.2021.05.007.