

RxBERT: Enhancing drug labeling text mining and analysis with AI language modeling

Leihong Wu¹, Magnus Gray¹, Oanh Dang², Joshua Xu¹, Hong Fang³ and Weida Tong¹

¹Division of Bioinformatics and Biostatistics, FDA National Center for Toxicological Research, Jefferson, AR 72079, USA; ²Office of Surveillance and Epidemiology, FDA Center for Drug Evaluation and Research, Silver Spring, MD 20993, USA; ³Office of Scientific Coordination, FDA National Center for Toxicological Research, Jefferson, AR 72079, USA
Corresponding author: Leihong Wu. Email: Leihong.wu@fda.hhs.gov

Impact Statement

Artificial intelligence has greatly changed the way we analyze text documents. In this study, we presented a new language model, named RxBERT, which is optimized to better understand human prescription drug labeling. We demonstrated that RxBERT showed the state-of-the-art performance in several labeling analysis tasks. This proof-of-concept study also demonstrated a potential pathway to customized large language models (LLMs) tailored to the sensitive regulatory documents for internal application.

Abstract

The US drug labeling document contains essential information on drug efficacy and safety, making it a crucial regulatory resource for Food and Drug Administration (FDA) drug reviewers. Due to its extensive volume and the presence of free-text, conventional text mining analysis have encountered challenges in processing these data. Recent advances in artificial intelligence (AI) for natural language processing (NLP) have provided an unprecedented opportunity to identify key information from drug labeling, thereby enhancing safety reviews and support for regulatory decisions. We developed RxBERT, a Bidirectional Encoder Representations from Transformers (BERT) model pretrained on FDA human prescription drug labeling documents for an enhanced application of drug labeling documents in both research and drug review. RxBERT was derived from BioBERT with further training on human prescription drug labeling documents. RxBERT was demonstrated in several tasks using regulatory datasets, including those involved in the National

Institutes of Technology Text Analysis Challenge Dataset (NIST TAC dataset), the FDA Adverse Drug Event Evaluation Dataset (ADE Eval dataset), and the classification of texts from submission packages into labeling sections (US Drug Labeling dataset). For all these tasks, RxBERT reached 86.5 *F1*-scores in both TAC and ADE Eval classification, respectively, and prediction accuracy of 87% for the US Drug Labeling dataset. Overall, RxBERT was shown to be as competitive or have better performance compared to other NLP approaches such as BERT, BioBERT, etc. In summary, we developed RxBERT, a transformer-based model specific for drug labeling that outperformed the original BERT model. RxBERT has the potential to be used to assist research scientists and FDA reviewers to better process and utilize drug labeling information toward the advancement of drug effectiveness and safety for public health. This proof-of-concept study also demonstrated a potential pathway to customized large language models (LLMs) tailored to the sensitive regulatory documents for internal application.

Keywords: Artificial intelligence, natural language processing, language model, BERT, drug labeling, pharmacovigilance

Experimental Biology and Medicine 2024; 248: 1937–1943. DOI: 10.1177/15353702231220669

Introduction

The US drug labeling documents are an invaluable data resource in the field of regulatory science and research, particularly for Food and Drug Administration (FDA) scientists and drug reviewers.^{1–4} Currently, there are over 140,000 submitted labeling documents containing vital information such as indications, dosage and administrations, adverse events (AEs), and much more. Moreover, these labeling documents are subject to updates over time, incorporating new

evidence from postmarketing pharmacovigilance. However, conventional data analysis methods for these documents often involve manual reading, keywords/pattern matching, or feature-based machine learning. Due to the free-text style of writing in these documents, automating data processing using traditional machine learning approaches proves challenging.⁵ An accurate semantic approach is highly needed for efficient processing and analysis of labeling documents.

In the past 5 years, the transformer architecture has revolutionized the area of natural language processing (NLP).⁶

Bidirectional Encoder Representations from Transformers (BERT) is a typical transformer encoder-only model that was released in 2018 by the Google artificial intelligence (AI) language team⁷ and adopted in many NLP studies shortly thereafter.^{8–10} The original BERT model (or base model) was developed on generic texts. For this reason, the BERT model may not recognize and comprehend the meaning of terms in domain-specific contexts and is thus less effective for analyzing biomedical text or legal documents.¹¹ Consequently, pretraining the BERT base model to accomplish specific task with domain application has become a common place in the BERT-driven NLPs research.

Pretraining a BERT model involves several crucial factors, including the text source, number of training epochs/steps, and size of network, among others, with the text source being of utmost importance. BioBERT¹¹ was pretrained with PubMed abstracts and PMC full-text articles to support mining of publications. ClinicalBERT¹² was pretrained on clinical notes collected from the Medical Information Mart for Intensive Care (MIMIC) III dataset to mine clinical information. Both BioBERT and ClinicalBERT use the BERT base model to set the model's starting weight before pretraining. Other domain-specific BERT models include Legal-BERT,¹³ SciBERT,¹⁴ COVID-Twitter-BERT,¹⁵ etc., all of which were trained BERT with domain-specific text corpus during the pretrain step.

Despite various language models optimized for social or biomedical sciences, only a few models have been designed specifically for regulatory documents. Regulatory documents, such as drug labeling, possess unique data structures and contexts, making them different from generic or scientific texts. Drug labeling, in particular, follows the structured product labeling (SPL) format but includes free-text sections such as "Warnings & Precautions" or "Adverse Reactions," making automatic processing challenging. However, recent efforts like PharmBERT, trained on all drug labeling documents, including over-the-counter (OTC) drugs, have shown the potential of domain-specific labeling BERT models.¹⁶

In this study, we developed a new BERT language model named RxBERT, which is a pretrained BERT model optimized for human prescription drug labeling texts only. Compared to the previous work, our effort was focused on the human prescription drugs and their drug labeling safety-related sections such as Boxed Warnings, Contraindications, Warnings & Precautions, Adverse Reactions, and Drug Interactions, in order to support drug safety reviews and research. RxBERT will be initialized with weights from BioBERT and then continually pre-trained with a text corpus from drug labeling documents. RxBERT's performance will be compared to other NLP models, including other BERT-like models, on named entity recognition (NER) tasks for drug labeling document analysis. We will demonstrate how RxBERT compares to other NLP models in terms of performance, efficiency, and accuracy in extracting important information from drug labeling and predicting AEs associated with known datasets: (1) Text Analysis Challenge Dataset (NIST TAC 2017 dataset); (2) the FDA Adverse Drug Event Evaluation Dataset (ADE Eval dataset); and (3) US Drug Labeling dataset.

Materials and methods

Model pretraining

In this section, we describe the dataset and vocabulary to develop the pretrain RxBERT. The pretrain process in this study is defined as the unsupervised (or self-supervised) training over a large domain-specific corpus (i.e. labeling documents).

Drug labeling dataset. The RxBERT model was trained on a corpus of 44,990 human prescription drug labeling documents, which consists of a total of 1.93 million of labeling sections. It is worth noting that drug manufacturers submit drug labeling to the FDA in accordance with the SPL format, where different sections contain distinct information. For instance, the content of the "Warnings & Precautions" section is independent to the content of the "Adverse Reactions" section, governed by the regulations. Consequently, we separated each section as one sample input to pretrain the BioBERT model. Finally, we only used texts or paragraphs that occurred in the main drug labeling sections. Other drug labeling text sections, such as the highlights, were not used in the pretrain step. As a result, we used over 5.5 million data points, each data point contains one sentence from the labeling section, to develop the pretrained RxBERT.

Drug labeling consists of a highlight section which provides a concise summary of the main labeling sections. We extracted 220k data points from labeling highlights for model validation. This was based on the hypothesis that the model trained from full prescription information should be able to well understand the context in the highlight section.

Vocabulary. Regulatory documents contain some terms or tokens that are rarely used in general texts. Therefore, like many other customized trained BERT models, the vocabulary of RxBERT was first generated using byte pair encoding (BPE) to segment the words. As the result, the vocabulary size of RxBERT is around 28,000.

Model fine-tuning

In this section, the fine-tuning of the pretrained RxBERT model was described to accomplish regulatory-related tasks with a specific focus on two important tasks: (1) NER and (2) sentence classification. The fine-tuning process defined in this study refers to further supervised train the model on a smaller, task-specific dataset. To fine-tune RxBERT for these tasks, labeled training datasets are required. The input of the model was all sentences. For the NER task, we used the annotated words as the output. For the sentence classification task, we used the labeling section name as the output; each input sentence belongs to a labeling section name. All fine-tuning processes were performed on a single GPU node (NVIDIA V100, 32 GB) within our in-house server.

NER

Task description. Given a text input such as document/paragraph/sentence, the NER task is to annotate each word

with appropriate labels, such as drug names, adverse reactions, etc. The goal is to correctly label all words and the performance was evaluated by comparing the NER-derived label with that of the predefined labels.

Method. spaCy, a Python library designed for text mining analysis, was utilized to fine-tune the NER model. The model architecture utilized was “*spacy-transformers.TransformerModel.v3*,” and the RxBERT pretrained model was implemented to initiate the fine-tuning process. The label annotations for the NIST TAC 2017 dataset consisted of “Adverse Reaction,” “Severity,” “Negation,” “Factor,” “Animal,” and “Drug Class.” On the contrary, the FDA ADE Eval dataset had “OSE Labeled AE,” “NonOSE AE,” and “Not AE Candidate” as label annotations (for a comprehensive understanding of these label annotations, please refer to section “Results” of this article).

Datasets

The TAC dataset. Detecting and recognizing AE keywords in regulatory submissions is a critical step in the reviewing process and can help reduce the workload of the reviewers, as well as providing trackable and reproducible results. The NIST TAC 2017 data challenge (<https://tac.nist.gov/2017/>) was a competition hosted by the National Institute of Standards and Technology (NIST) to evaluate the performance of NLP tools on various tasks. In this competition, one of the tasks was NER for drug labeling documents, which involved identifying and extracting specific types of terms from drug labeling texts.³ As part of the competition, the FDA released a set of 200 annotated drug labeling documents (i.e. the TAC dataset), which included six types of terms that were annotated: “Adverse Reaction,” “Severity,” “Negation,” “Factor,” “Animal,” and “Drug Class.” Out of these 200 documents, 101 were released as a training dataset for the competition, with the remaining 99 documents held back for evaluation purposes.

In this study, the 101 training documents and their annotated terms were used to evaluate the performance of the RxBERT model. Specifically, we compared the model’s ability to identify and extract the six types of terms from the drug labeling texts against the “gold standard” annotations provided in the dataset. In order to do that, we normalized the original labeled data file (.xml) into the format accepted by spaCy for model training in order to prepare the labeling annotation. Specifically, we utilized only the “Mention” annotations from the TAC 2017 dataset.

We split the 101 training documents into two random subsets using an 80:20 ratio for training and internal validation. Each document might contain multiple sections, such as “Warnings & Precautions,” “Adverse Reactions,” etc., and as a result, we further divided these documents into sections during data preparation. This resulted in 180 and 59 labeling section texts, for training and validation, respectively. The NER model underwent training on the 180 data sections for 100 epochs with a batch size of 100 per epoch, while the remaining 59 data sections are used for calibration/validation. After the model was trained, it was tested on the

original TAC test set, which had an additional 99 drugs with 237 labeling sections provided by the challenge organizer.

The ADE Eval dataset. The ADE Eval dataset, constructed by the MITRE Corporation and the US FDA, is another valuable resource for evaluating the performance of the RxBERT model.¹⁷ This dataset consists of 100 training drug labeling documents, which were released by the organizers with manually reviewed annotations. Half of these drug labeling documents overlapped with those used in the NIST TAC 2017. However, in contrast, the ADE Eval dataset was annotated for a different use case, using a different criterion, and by different FDA annotators. In addition, this dataset included older drug labeling that were in non-PLR format. The ADE Eval dataset is therefore used as a benchmark dataset for evaluating the performance of various NLP models in identifying ADEs from unstructured text data in drug labeling. We used this dataset to evaluate the performance of our RxBERT model and compare it with other existing models in established studies.

In the ADE Eval dataset, the primary AE terms that are listed in a drug product labeling and associated with exposure to that drug were annotated as “OSE Labeled AE.” In this study, “OSE Labeled AE” is considered positive in the performance evaluation. We used the modified *F*-score measurement defined in the original paper.¹⁷ To elaborate, the main measurement used for this scenario was the “Exact Mention Match—Weighted,” which involved a weighted, microaveraged, corpus-level *P/R/F1* assessment on mentions. The ideal score was assigned to the exact match mention pair, while other components were given weightings to reflect the time and effort required to rectify any inaccuracies. The count of *M* was used for exact match mention pairs, *C* for mention pairs that were detected but differ in span or MedDRA code, *S* for spurious mentions that was incorrectly labeled (typically false positives), and *N* for missed mentions (typically false negatives). Their detailed definition and measurement are as follows; the description in the parenthesis describes the weight determination of each component:

- $M' = M + (0.5 \times C)$ (matches accrue the correct share of the clash).
- $C' = 0.5 \times C$ (errors are weighted 0.5, since correcting a mention is hard but likely not as hard as adding one).
- $S' = 0.25 \times S$ (spurious mentions are weighted 0.25, since deleting a mention is easy).
- $N' = N$ (missing mentions are weighted 1, because adding a mention is hard).

With that, *P/R/F1* measure is then computed as

- $P = M' / (M' + C' + S')$.
- $R = M' / (M' + C' + N')$.
- $F1 = (2 \times P \times R) / (P + R)$.

Note that, the *P/R/F1* defined in ADE Eval dataset is adjusted and is thus not directly comparable to the *P/R/F1* in the TAC2017 results.

In this study, we used 180 of 259 labeling sections for model training and the remaining 79 labeling sections for model validation. Similarly, the NER model was fine-tuned based on the RxBERT architecture.

Sentence classification

Task description. For this task, the goal was to correctly classify sentences from different formats of drug labeling documents into categories based on the physician label rule (PLR) format. The PLR format is considered the current regulation format¹⁸ of prescription drug labeling in the United States. It enhances the safe and effective use of human prescription drugs by providing standardized sections such as drug indications, usages, boxed warnings, etc. and a table of content information for healthcare professionals and patients. We also used the non-PLR format of US Drug labeling documents (i.e. those labeling published prior to 2001) in this study as an additional testing dataset.

Method. HuggingFace Transformers, a Python library designed for sharing and using transformer-based language models, was utilized to fine-tune and evaluate a collection of BERT-based models for the sentence classification task. More specifically, BERT-base, ALBERT-base, DistilBERT-base, RoBERTa-base, and BioBERT-base (v 1.1) were used as a baseline for evaluating the performance of RxBERT. The four sentence classification tasks focused on the following drug labeling sections “Warnings & Precautions,” “Adverse Reactions,” “Indications & Usage,” and “Others” (remaining sections). In particular, the sections “Indications & Usage” and “Others” were included as a non-safety-related comparator to that of the safety-related sections (please refer to section “Results” of this article for more information about the datasets and fine-tuning process used for this sentence classification task).

Datasets

The US Drug Labeling dataset. A total of 45,626 US prescription drug labeling documents were obtained and processed from DailyMed full release of human prescription labeling (retrieved 28 February 2022). Of these documents, 29,709 (65%) were in the PLR format, while 15,917 (35%) were in the non-PLR format. A total of 17,453,802 sentences were extracted using Python and Natural Language Toolkit (NLTK) libraries. These sentences were further mapped with Logical Observation Identifiers, Names, and Codes (LOINC), which are the official codes used to determine the located sections in the labeling documents. Because the PLR format can be considered the “gold standard” of US Drug labeling formats, this was the format used to define the classes within this sentence classification task.¹⁹

For the sentence classification task, a training dataset was developed with PLR-formatted drug labeling documents, randomly pulling 10,000 records for each of the following LOINC: (1) “Warnings & Precautions,” (2) “Adverse Reactions,” (3) “Indications & Usage,” and (4) “Others.” Then, in a similar manner, testing datasets were developed

Table 1. Overview of dataset used in RxBERT pretraining and fine-tuning.

Datasets		
Pretraining	Documents (for training)	Vocabulary
Human prescription labeling	~5.5 million	~28,000
Fine-tuning	Training set	Testing set
TAC-2017	180	59
ADE Eval dataset	180	79
Labeling sentence classification	40,000 (10,000 each category)	40,000 (10,000 each category)

for each of the two formats of drug labeling documents (i.e. PLR and non-PLR). These datasets were used to fine-tune and evaluate several BERT models, including BERT-base, ALBERT-base, DistilBERT-base, RoBERTa-base, BioBERT-base (v 1.1), and the RxBERT model developed in this study. With the HuggingFace Transformers library, the training and testing datasets were tokenized using each model’s respective tokenizer. From here, the tokenized training dataset was split 80% for training and 20% for validation, and each model was fine-tuned using the default parameters. The models were fine-tuned for only 10 epochs, as it was noted that improvements in performance plateaued before or around this stage. Finally, each model was evaluated with the PLR- and non-PLR-formatted testing datasets, each containing 10,000 sentences per endpoint that were new to or unseen by the model.

Results

In this section, we pretrained the RxBERT model and evaluated it on three different regulatory-related NLP tasks. Two involved NER using the NIST TAC 2017 dataset and the FDA ADE Eval dataset. The remaining task was sentence classification applied to the US Drug Labeling dataset. The details of dataset we used for pretraining and fine-tuning are summarized in Table 1.

RxBERT model pretraining

We adopted BioBERT-base (v 1.1) as the pretrained model to initialize the RxBERT training process. RxBERT was trained via a standard masked language modeling (MLM) approach. The pretraining step was performed on an in-house GPU server with seven available GPU nodes (NVIDIA V100, 32GB). The learning rate was set to 0.0001, and the epochs were set to 100. The batch size was set to 64 per device, with the maximum sequence length set to 128. The total number of training steps was 1.23 million. The total training time for 100 epochs (or 1.23 million steps) was 835k seconds (or 230 h or about 8.5 days).

NER tasks of the TAC 2017 dataset

The TAC dataset was released by the FDA as part of the NIST TAC 2017 data challenge, which aimed to assess NLP tools for the applications of identifying AE terms. As described in

Table 2. *F*-score of RxBERT model for the TAC dataset.

Overall performance (all endpoints)	Precision	Recall	<i>F</i> 1
RxBERT	86.6	86.3	86.5
TAC challenge—max	83.8	84.4	82.5
TAC challenge—median	76.8	66.3	70.1
TAC challenge—min	40.5	11.8	18.3
Adverse reaction	Precision	Recall	<i>F</i> 1
RxBERT	89.3	88.5	88.9
TAC challenge—max	86.4	86.9	85.2
TAC challenge—median	78.6	70.8	72.7
TAC challenge—min	42.1	13.4	20.3

Table 3. RxBERT NER model on ADE Eval dataset.

Exact mention match—weighted	Precision	Recall	<i>F</i> 1
<i>RxBERT—ADE Eval</i>	90.8	84.3	87.4
* <i>ADE Eval challenge—max</i>	92	86	89
* <i>ADE Eval challenge—median</i>	87	75	80
* <i>ADE Eval challenge—min</i>	75	19	31

*The result from original ADE Eval challenge only contains two digits.

section “Materials and methods,” there are 180 labeling section texts for model training and 59 sections for validation.

Table 2 summarizes the test results of a comparison study between the RxBERT model and the established TAC challenge results. The comparison was done using the *F*-score metric, which takes into account both precision and recall. Table 1 shows the *F*-score of the RxBERT model for the TAC dataset, as well as the max, median, and min *F*-scores of the TAC challenge participants.³ Overall, the results show that the RxBERT model outperformed the TAC challenge participants in terms of the overall *F*-score for all six endpoints, achieving an *F*-score of 86.5, while the maximum *F*-score achieved by the TAC challenge participants was 82.5. The median and minimum *F*-scores for the TAC challenge participants were 70.1 and 18.3, respectively. In addition, RxBERT also outperformed the TAC challenge participants in recognizing adverse reaction terms, achieving an *F*-score of 88.9, while the maximum *F*-score achieved by the TAC challenge participants was 85.2.³ The median and minimum *F*-scores for the TAC challenge participants were 72.7 and 20.3, respectively.³

In summary, this case study demonstrates the utility of the RxBERT model in recognizing adverse reaction terms from regulatory submission documents and shows that it can achieve high performance in this task.

NER tasks of the ADE Eval dataset

As shown in Table 3, the predictive model fine-tuned with the ADE Eval training data (RxBERT—ADE Eval) obtained 90.8 and 84.2 precision and recall scores, respectively, which yielded an *F*-score of 87.4 for the validation result in the weighted exact mention match. Both RxBERT-based NER models outperformed the median performance from the ADE Eval challenge and showed competitive results compared to the best performing model during the challenge.

Table 4. Overall predictive accuracy of RxBERT-based classification model on labeling classification.

Overall predictive accuracy (multiclass)	PLR	Non-PLR
RxBERT	87%	77%
BERT	84%	74%
ALBERT	84%	72%
DistilBERT	83%	74%
RoBERTa	83%	74%
BioBERT	84%	75%

Sentence classification result

The US Drug Labeling dataset was created using US FDA prescription drug labeling documents obtained from DailyMed. A total of 45,626 documents were processed, with 65% in the PLR format and 35% in the non-PLR format.

The PLR format was considered the “gold standard” and used to define the classes for the sentence classification task. A training dataset was created by randomly selecting 10,000 sentences for each of the following sections: “Indications & Usage,” “Warnings & Precautions,” “Adverse Reactions,” and “Others” remaining sections. Another 10,000 records from each labeling section were used as the testing dataset for each of the two formats of drug labeling documents, respectively.

Table 4 documents the results obtained for each language model, showing the overall predictive accuracy for the two formats of drug labeling. Based on the results, RxBERT outperformed and had lower error rates than other BERT models.

Discussion

We have witnessed a significant increase in the number of US drug labeling documents over time as new drugs are approved and due to continual updates to existing labeling content. Accurate extraction and analysis of crucial insights from this vast corpus of text data play a pivotal role in various domains, including AE detection, drug discovery, and regulatory compliance. However, the traditional methods of manual annotation and information extraction struggle to keep pace with the increasing demand for efficient and comprehensive analysis.

In this study, we developed a domain-specific BERT model, named RxBERT. RxBERT is trained with labeling documents of US human prescription drugs, based on the BERT-base architecture. We then assessed RxBERT on two common NLP tasks: NER and Text Classification, both of which have already been widely applied in the regulatory research of drug labeling.^{20–22} RxBERT showed competitive results compared to previous approaches in both tasks. In particular, for the drug labeling sentence classification task, RxBERT outperforms a collection of BERT-based models, including the biomedical domain-specific model BioBERT, which was pretrained on PubMed abstracts and used to initialize RxBERT. However, it is important to note that RxBERT’s edge in performance may be due to the similarity between its pretraining dataset and the US Drug Labeling dataset, and that other biomedical models may perform better on other NLP tasks. Nonetheless, based on

the results, RxBERT has potential to assist with research regarding drug safety and efficacy, and with the recent release of PharmBERT, there is interest in using such language models in this area.

Over the past decade, the field of NLP has undergone a revolutionary transformation due to the emergence of AI and Deep Learning. Prior to 2012, rule-based systems dominated text mining and NLP studies, with approaches like term frequency-inverse document frequency (TF-IDF) and co-occurrence analysis, which are still widely utilized today. However, since 2013, there has been a paradigm shift in NLP toward embedding-based analysis. In this approach, textual data are first converted into numerical vectors, and various modeling algorithms such as random forest, support vector machine, and most prominently, neural networks are employed.

Word-embedding analysis encompasses popular methods like word2vec,²³ ELMO,²⁴ and transformers.⁶ Notably, BERT, an encoder-only transformer architecture, has emerged as one of the most widely adopted word-embedding approaches in the last 5 years. BERT offers two significant advantages over previous approaches such as word2vec and Recurrent Neural Networks (RNN), namely, context-based learning and transferrable learning weights. These unique characteristics of BERT have not only enhanced model performance, but also improved training efficiency, marking a substantial advancement in NLP.

In addition to BERT, since its introduction in 2020, GPT-3 has swiftly become the “gold standard” for LLMs in the NLP community. Unlike BERT, which has fewer than 1 billion trainable model parameters, larger language models like GPT-3,²⁵ PaLM,²⁶ OPT,²⁷ etc. boast over 100 billion parameters. This significant increase in model size has led to two direct consequences: (1) the training process for LLMs is highly resource-intensive and (2) numerous new capabilities have emerged, such as generalization, few-shot learning, and multimodal capabilities. Interestingly, even though decode-only models like GPT-3 were not originally designed for word embedding, they have acquired the ability to generate embedding vectors, similar to BERT.

However, BERT still possesses unique advantages in the word-embedding process. First, BERT is a smaller-sized model, making it more time and cost-efficient to generate embeddings compared to that of larger language models, particularly when there is a large volume of requests or in real-time environments, such as text queries through web tools. Second, BERT models are easier to migrate across different servers and may be more straightforward to fine-tune in various computing environments. Finally, while larger language models excel in generalization, many real-world regulatory research applications primarily require task-specific applications rather than generalized ones, making BERT a better fit for such purposes.

Although this article is mainly focused on BERT model development and assessment, we are not implying that other LLMs are not fit for regulatory research. LLMs, such as ChatGPT and Bard, have garnered substantial attention not only within the AI community, but also worldwide

due to their rapid development since late 2022. The potential applications of LLMs in the healthcare and regulatory research are evident and have been discussed elsewhere.^{28,29} However, alongside the promising prospects, several concerns have been raised regarding LLMs. One significant concern revolves around data sensitivity. Many widely used LLMs, including ChatGPT, rely on heavily equipped servers, posing challenges and restrictions, particularly for regulatory agencies, when it comes to uploading sensitive data onto these external public servers. Although there have been proposed solutions, such as the development of distilled, closed domain models like LLaMA within agency environments, the creation of such models, which range from 7 to 65 billion parameters, still demands significantly larger computing resources compared to BERT models.

Another major concern associated with LLMs pertains to ethicality, explainability, and trustworthiness. These aspects hold critical importance for regulatory research, as LLMs can become involved in high-stakes decision-making processes. Establishing guidelines for ensuring AI trustworthiness remains an open question within the community, and various discussions on this matter can be found elsewhere.^{30–32} Addressing these concerns is crucial to harness the full potential of LLMs while ensuring their responsible and reliable usage in regulatory research.

Conclusions

To conclude, we propose a customized trained BERT model with human prescription labeling document, named RxBERT. The proposed model was successfully applied on three different NLP regulatory tasks.

Volumes of information associated with FDA-regulated products are continually generated, making it prudent that the agency develops advanced methods to enhance its capabilities for retrieving, organizing, and evaluating large amounts of data efficiently and reliably. AI provides a transformative approach to advance FDA's regulatory mission related to drug safety including pharmacovigilance. In particular, the proposed RxBERT model will enable the analysis of AEs and their patterns in drug labeling. This demonstrates the potential for future understanding and analysis of the vast amount of data associated with FDA-regulated drug products, thereby advancing drug safety.

AUTHORS' CONTRIBUTIONS

LW, JX, and WT conceived the idea and designed the study. LW developed the model. LW, MG, and OD conducted the experiments. LW, MG, and JX analyzed the result. LW, MG, OD, JX, and WT wrote and revised the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

MG thanks the Oak Ridge Institute for Science and Education for their support to the Research Participation Program at the National Center for Toxicological Research, US Food and Drug Administration.

DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

The author(s) received no financial support for the research, authorship, and/or publication of this article.

DISCLAIMER


The views presented in this article do not necessarily reflect those of the US Food and Drug Administration. Any mention of commercial products is for clarification and is not intended as an endorsement.

ORCID IDS

Leihong Wu  <https://orcid.org/0000-0002-4093-3708>

Magnus Gray  <https://orcid.org/0009-0009-6355-7627>

Joshua Xu  <https://orcid.org/0000-0001-5313-5847>

Weida Tong  <https://orcid.org/0000-0003-3488-6148>

REFERENCES

- Fang H, Harris S, Liu Z, Thakkar S, Yang J, Ingle T, Xu J, Lesko L, Rosario L, Tong W. FDALabel for drug repurposing studies and beyond. *Nat Biotechnol* 2020;**38**:1378–9
- Fang H, Harris SC, Liu Z, Zhou G, Zhang G, Xu J, Rosario L, Howard PC, Tong W. FDA drug labeling: rich resources to facilitate precision medicine, drug safety, and regulatory science. *Drug Discov Today* 2016;**21**:1566–70
- Roberts K, Demner-Fushman D, Tonning JM. Overview of the TAC 2017 adverse reaction extraction from drug labels track. https://tac.nist.gov/publications/2017/additional.papers/TAC2017.ADR_overview.proceedings.pdf
- Chen M, Vijay V, Shi Q, Liu Z, Fang H, Tong W. FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discov Today* 2011;**16**:697–703
- Wu L, Ingle T, Liu Z, Zhao-Wong A, Harris S, Thakkar S, Zhou G, Yang J, Xu J, Mehta D. Study of serious adverse drug reactions using FDA-approved drug labeling and MedDRA. *BMC Bioinform* 2019;**20**:129–39
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Polosukhin I. Attention is all you need. *Adv Neur Inform Process Syst* 2017;**30**:3762
- Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. <https://arxiv.org/abs/1810.04805>
- Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput Surv* 2023;**55**:1–35
- Zhou M, Duan N, Liu S, Shum HY. Progress in neural NLP: modeling, learning, and reasoning. *Engineering* 2020;**6**:275–90
- Sun TX, Liu XY, Qiu XP, Huang XJ. Paradigm shift in natural language processing. *Mach Intell Res* 2022;**19**:169–83
- Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;**36**:1234–40
- Huang K, Altsaer J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. <https://arxiv.org/abs/1904.05342>
- Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: the Muppets straight out of law school. <https://arxiv.org/abs/2010.02559>
- Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. <https://arxiv.org/abs/1903.10676>
- Müller M, Salathé M, Kummervold PE. COVID-Twitter-BERT: a natural language processing model to analyse COVID-19 content on Twitter. <https://www.frontiersin.org/articles/10.3389/frai.2023.1023281/full>
- ValizadehAslani T, Shi Y, Ren P, Wang J, Zhang Y, Hu M, Zhao L, Liang H. PharmBERT: a domain-specific BERT model for drug labels. *Brief Bioinform* 2023;**24**:bbad226
- Bayer S, Clark C, Dang O, Aberdeen J, Brajovic S, Swank K, Hirschman L, Ball R. ADE eval: an evaluation of text processing systems for adverse event extraction from drug labels for pharmacovigilance. *Drug Saf* 2021;**44**:83–94
- U.S. Food and Drug Administration. 21 C.F.R. §201.56, 201.57, 201.80, <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=201.56>
- Gray M, Xu J, Tong W, Wu L. Classifying free texts into predefined sections using AI in regulatory documents: a case study with drug labeling documents. *Chem Res Toxicol* 2023;**36**:1290–9
- Wu Y, Liu Z, Wu L, Chen M, Tong W. BERT-based natural language processing of drug labeling documents: a case study for classifying drug-induced liver injury risk. *Front Artif Intell* 2021;**4**:729834
- Wu L, Chen S, Guo L, Shpyleva S, Harris K, Fahmi T, Flanigan T, Tong W, Xu J, Ren Z. Development of benchmark datasets for text mining and sentiment analysis to accelerate regulatory literature review. *Regul Toxicol Pharmacol* 2022;**137**:105287
- Thakkar S, Slikker W Jr, Yiannas F, Silva P, Blais B, Chng KR, Liu Z, Adholeya A, Pappalardo F, Soares MDLC, Beeler PE, Whelan M, Roberts R, Borlak J, Hugas M, Torrecilla-Salinas C, Girard P, Diamond MC, Verloo D, Panda B, Rose MC, Jornet JB, Furuham A, Fang H, Kwegyir-Afful E, Heintz K, Arvidson K, Burgos JG, Horst A, Tong W. Artificial intelligence and real-world data for drug and food safety—a regulatory science perspective. *Regul Toxicol Pharmacol* 2023;**140**:105388
- Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. <https://arxiv.org/abs/1301.3781>
- Sarzynska-Wawer J, Wawer A, Pawlak A, Szymanowska J, Stefaniak I, Jarkiewicz M, Okruszek L. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Res* 2021;**304**:114135
- Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A. Language models are few-shot learners. *Adv Neur Inform Process Syst* 2020;**33**:1877–901
- Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung HW, Sutton C, Gehrmann S. PaLM: scaling language modeling with pathways. <https://arxiv.org/abs/2204.02311>
- Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, Dewa n C, Diab M, Li X, Lin XV. OPT: open pre-trained transformer language models. <https://arxiv.org/abs/2205.01068>
- De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, Rizzo C. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health* 2023;**11**:1166120
- Sonntagbauer M, Haar M, Kluge S. Artificial intelligence: how will ChatGPT and other AI applications change our everyday medical practice? *Med Klin Intensivmed Notfmed* 2023;**118**:366–71
- Vought RT. Guidance for regulation of artificial intelligence applications. <https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf>
- Executive Order 13859. Maintaining American leadership in artificial intelligence. <https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence>
- Executive Order 13960. Promoting the use of trustworthy artificial intelligence in the federal government. <https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government>

(Received August 30, 2023, Accepted November 2, 2023)