Towards the objectives of the United States Food and Drug Administration (FDA) generic drug science and research program, it is of vital importance in developing product-specific guidances (PSGs) with recommendations that can facilitate and guide generic product development. To generate a PSG, the assessor needs to retrieve supportive information about the drug product of interest, including from the drug labeling, which contain comprehensive information about drug products and instructions to physicians on how to use the products for treatment. Currently, although there are many drug labeling data resources, none of them including those developed by the FDA (e.g., Drugs@FDA) can cover all the FDA-approved drug products. Furthermore, these resources, housed in various locations, are often in forms that are not compatible or interoperable with each other. Therefore, there is a great demand for retrieving useful information from a large number of textual documents from different data resources to support an effective PSG development. To meet the needs, we developed a Natural Language Processing (NLP) pipeline by integrating multiple disparate publicly available data resources to extract drug product information with minimal human intervention. We provided a case study for identifying food effect information to illustrate how a machine learning model is employed to achieve accurate paragraph labeling. We showed that the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model is able to outperform the traditional machine learning techniques, setting a new state-of-the-art for labelling food effect paragraphs from drug labeling and approved drug products datasets.

Keywords: information extraction, NLP, FDA drug labels, BERT, product specific guidance
INTRODUCTION

Product-Specific Guidance\(^1\) (PSG) represents the United States Food and Drug Administration (FDA)'s current thinking on the best approaches for demonstrating bioequivalence between a test product and the reference product. The development of product-specific guidelines (PSGs) intends to facilitate generic drug product development, and Abbreviated New Drug Application (ANDA) submission and approval, ultimately promote safe, effective, affordable generic drugs to the public in the United States. During the Generic Drug User Fee Amendments II (GDUFA II), FDA is committed to issuing PSGs for 90 percent of non-complex new chemical entity New Drug Applications that are approved on or after October 1, 2017, at least 2 years prior to the earliest lawful ANDA filing date\(^2\). This commitment, in addition to the demands of developing PSGs for complex drug products, calls for an enhanced PSG developing process. To answer this challenge, one of the solutions is to identify and automate the labor-intensive works during the PSG development. To generate a PSG, FDA staff usually need to take extensive efforts (e.g., 50% efforts for a regular immediate release solid oral dosage form drug product) to collect relevant information from multiple data resources to a PSG review template, e.g., extracting the information related to boxed warning, indication, dosage and administration, clinical pharmacology and pharmacokinetics including absorption, distribution, metabolism, excretion, food effect (ADMEF) from drug labeling as well as the reference listed drug and reference standard (RLD/RS) information from the Orange Book\(^3\). If this part of the PSG development work can be automatically accomplished by a well-designed data analytics tool, the PSG developers can have more time and effort to focus on the human intelligence-required work.

Natural language processing (NLP) has been increasingly used with a specific focus on text mining and information extraction in drug labeling. For example, Fung et al. (2013) extracted information from the indications section of the drug labeling from DailyMed to encode drug-indication pairs, whereas Bisgin et al. (2011) extracted three labeling sections (Boxed Warning, Warnings and Precautions, Adverse Reactions) from DailyMed to group drugs by topics that are associated with the same safety concerns and therapeutic uses. However, most of the previous work focused only on a single data source (for review, see, e.g., Fang et al., 2016). Since none of the data sources can cover all the drug products, retrieving drug labeling information from multiple sources is needed in PSG developing process. The data sources used in our PSG developing process include Orange Book, Drugs@FDA,\(^4\) DailyMed\(^5\) and DrugBank\(^6\), which are described as follows.

Data Sources

Orange Book

Orange Book, formally known as the Approved Drug Products with Therapeutic Equivalence Evaluations, is considered to be the authoritative source of information in the United States on the therapeutic equivalence of FDA approved drug products. It includes currently marketed prescription drug products approved by the FDA through new drug applications (NDAs) and abbreviated new drug applications (ANDAs) with different dosage forms (Center for Drug Evaluation and Research 2020). It also selects the reference standard (RS) which an applicant seeking approval of an ANDA must use in conducting an \textit{in vivo} bioequivalence study required for approval. In this paper, we used Orange Book as the baseline for the FDA application number.

Drugs@FDA

Drugs@FDA is a publicly available resource, which includes the majority of drug labeling, approval letters, reviews, and other information for FDA-approved drug products for human use provided by the FDA. It contains prescription brand-name drug products, over-the-counter brand-name drug products, generic drug products, and therapeutic biological products.

DailyMed

DailyMed is a free drug information resource provided by the United States. National Library of Medicine (NLM) that consists of digitized versions of drug labeling as submitted to the FDA. It is the official provider of the FDA labeling information (package inserts). The documents published use the Health Level Seven (HL7) version 3 Structured Product Labeling (SPL) standard, which specifies various drug label sections (Schadow 2005, 7). It uses Logical Observation Identifiers Names and Codes (LOINC) to link sections and subsections of human prescription drug and biological product labeling.

DrugBank

DrugBank is a richly annotated resource that combines detailed drug data with comprehensive drug target and drug action information provided by the University of Alberta and the Metabolomics Innovation Center (Wishart et al., 2008). It contains FDA-approved small molecule and biologics drugs with extensive food-drug and drug-drug interactions as well as ADMET (absorption, distribution, metabolism, excretion, and toxicity) information (Knox et al., 2011).

As a proof of concept, in this study, we proposed an information extraction pipeline using NLP and machine learning to provide an automatic data-processing workflow to extract, annotate and integrate drug labeling data from multiple publicly available data sources with minimal human intervention for PSG development. However, mining the drug labeling from multiple sources is challenging, since these data sources, including those developed by the FDA, with the various data formats and access methods, are often not easily available to inform the PSG development process.

To address the challenge of interoperation with multiple data sources, our pipeline provided examples to access data via

\(^1\)https://www.accessdata.fda.gov/scripts/cder/psg/index.cfm
\(^2\)https://www.fda.gov/media/101052/download
\(^3\)https://www.accessdata.fda.gov/scripts/cder/ob/index.cfm
\(^4\)https://www.accessdata.fda.gov/scripts/cder/daf
\(^5\)https://dailymed.nlm.nih.gov/dailymed/
\(^6\)https://go.drugbank.com
different methods, parse data in various data formats, and unify identification scheme. At the same time, there are complex scenarios where machine learning models are essential to achieve accurate information extraction. We provided a case study for labelling food effect paragraphs (Food Effect vs. Non-Food Effect) as an example to illustrate how we addressed the scenarios when keyword detection and regular expression were not adequate. In our case study, we leveraged the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model to identify food effect paragraphs from the drug labeling. The BERT (Devlin et al., 2019) is a groundbreaking unsupervised NLP model, which has been trained with huge general language datasets, such as Wikipedia Corpus, and can be fine-tuned on usually small datasets for specific language tasks to achieve the state-of-the-art performance on many NLP tasks. Therefore, in this work, instead of training a new model from scratch, we took advantage of the pre-trained BERT model for labeling food effect paragraphs to obtain better performance. We showed that the BERT-based model is able to outperform the traditional machine learning techniques for identifying food effect-related paragraphs on drug labeling datasets. It is expected that our developed information extraction pipeline will save a great amount of time and effort so that the PSG developers can devote more time to the human-intelligence-required work.

The software package for data collection, preprocessing, and the model training and validation for labeling food effect paragraphs will be provided at https://github.com/Yiwen-Shi/drug-labeling-extraction.

**METHODS**

We collected drug product information from multiple publicly available data sources, including Orange Book, Drugs@FDA, DailyMed, and DrugBank. The raw data varied widely in format, from well-structured XML files to highly irregular free-text PDF files. To create a structured dataset to facilitate the assessment process for PSG development, we attempted to convert various data formats into a structured data frame in four steps. First, we collected raw data from either requesting RESTful (representational state transfer) API (application program interface) or downloading the full database whenever the API is not available. Second, we parsed raw data in various data formats into structured data records. Third, we mapped different identification schemes to the FDA application number, which is a unique six-digit number assigned to a drug product by the FDA. Last, we kept the latest version if multiple versions of drug labeling existed.

We provided a case study for labeling food effect paragraphs, in which a pre-trained BERT model was finetuned on a small set of annotated labels. We showed that the BERT-based model is able to outperform the traditional machine learning techniques for food effect paragraph labeling on drug label datasets.

**Data Processing**

The workflow applied to the aforementioned data sources for processing, i.e., Drugs@FDA, DailyMed and DrugBank, is summarized in Figure 1. It contained four steps in general, which need to be adjusted in a variety of ways to accommodate data source diversity. In this section, we illustrate how the workflow is implemented for each data source.

**Data Collection and Parsing**

Table 1 presented the details of the data format/access method, the download date/version and the number of records extracted from four different sources. In the original data files, Drugs@FDA and DailyMed provided the complete list to access the SPL and PDF files via RESTful API, whereas DrugBank provided the complete dataset in a single XML file.

After pulling the raw data, they were parsed, extracted and annotated based on the 12 sections (For details, see Table 4 below) which are essential for PSG assessment, including boxed warning, indication, dosage and administration, pregnancy, lactation, mechanism of action, pharmacodynamics, and five subsections under pharmacokinetics: absorption, distribution, metabolism, excretion, and food effect.

Drugs@FDA provides the latest FDA-approved drug labeling and previously approved labeling in PDF format. After having converted the PDF file to text, we extracted the 12 sections from free text via regular expression, which is a sequence of characters that define a search pattern. We used regular expression to detect and locate keywords/subtitles, then extract the related sections from the free text converted from pdf files.

DailyMed maintains the labeling in a document markup standard approved by HL7 referred to as SPL, which specifies various drug labeling sections by LOINC codes and names. Table 2 listed a sample of eight sections we collected in this paper. However, since absorption, distribution, metabolism, excretion, and food effect do not have LOINC codes, we extracted their parent section pharmacokinetics via LOINC codes first, then applied keywords detection and regular expression to locate subsections and automatically annotated them.

DrugBank provides the complete dataset in a single XML format. Drugs are represented by <drug> tag. Its children elements contain both pharmacology information and commercial drug products. We extracted their contents from eight elements by tag names, including ADMEF information and the FDA application number whenever it is available. With the well-defined tags, DrugBank is a good complement when the subtitles of ADMEF are not available in Drugs@FDA or DailyMed.

**Unify Identification Scheme**

Different data sources have their own identification schemes. They vary from source to source. For example, the sources developed by the FDA such as Orange Book and Drugs@FDA use the FDA application number, a unique six-digit number to identify the drug, whereas DailyMed uses a unique Set ID to detect the SPL file. DrugBank uses its own DrugBank-ID for drug identification.
In order to provide a unified identifier to retrieve and combine information from multiple data sources, we mapped the Set ID in DailyMed and the DrugBank-ID in DrugBank to FDA application number, which is recognized by both Orange Book and Drugs@FDA. It is possible that one Set ID or DrugBank-ID may have more than one FDA application number. Under such circumstances, all the corresponding FDA application numbers are listed. Figure 2 illustrated a typical example as to how the Set ID was mapped in the parsed result of an SPL file to the FDA application number.
Filtering
Since both Drugs@FDA and DailyMed contain historical drug labeling, we only need to keep the latest version for each FDA drug application number with regard to the effective date of the latest document. Due to the different update schedules, each data source contains an inconsistent list of FDA-approved drugs. As the authoritative source of information in the United States on the therapeutic equivalence of FDA-approved drug products, the Orange Book contains 3,624 unique New Drug Application (NDA) FDA application numbers in the datasets downloaded in Oct. 2020, which is used as a point of reference for the currently valid FDA-approved drugs in this paper.

Paragraph Labeling of Food Effect
As described in the last section, we extracted information from drug labeling including Boxed Warning, Indication, Dosage and Administration, Use in Specific Populations, and Clinical Pharmacology. However, there are often complex scenarios where machine learning models are essential to achieve accurate information extraction. As such, we provide a case study for labeling food effect paragraphs as an example to illustrate how we address the scenarios when keyword detection and regular expression are not applicable.

Food may affect pharmacokinetics by any or all of the following mechanisms: delaying gastric emptying, stimulating bile flow, changing the pH of the gastrointestinal tract, increasing splanchnic blood flow, changing luminal metabolism of a drug, and physically/chemically interacting with a dosage form or drug (Sharma et al., 2013). Food effect paragraphs can appear in different sections in drug labeling. Figures 3A,B, respectively, provide examples of the subsection under pharmacokinetics and the subsection under the absorption section, where each paragraph has been automatically annotated as either “Food Effect” (red border) or “Non-Food Effect” (blue border). There are also the cases where the food effect paragraphs included in the absorption section without any subtitle (Figure 3C), which cannot be automatically extracted and easily annotated by keyword detection and regular expression. In these cases, the content of drug labeling relevant to the food effect is determined by semantic meaning, which motivated us to treat the food effect paragraph labeling as classification task (Food Effect vs. Non-Food Effect). We note that paragraph labeling is only one component (related to food effect) of our drug labeling extraction pipeline which integrates multiple disparate publicly available data resources to extract drug product information for enhancing PSG development.

Dataset
Food effect information is available in both Drugs@FDA and DailyMed. Table 3 shows the statistical summary of food effect sections we extracted from two data sources, where the numbers in the table represent the number of paragraphs. To train the model for identifying the food effect paragraphs from the absorption sections (e.g., Figure 3C), we constructed a dataset in which each paragraph was labeled as either Food Effect or Non-Food Effect. For Figure 3A, we detected the section title with regular expression “(food effect|food effects|effect of food|effects of food)” and annotated the paragraph that followed the title as “Food Effect”. For 3B, we used a slightly different regular expression “(food effect|food effects|effect of food|effects of food)\*\{\}”, to detect if a paragraph belongs to “Food Effect”. We removed the title at the beginning of the paragraph that was detected by regular expression in the dataset. For other paragraphs under the section title detected by regular expression “absorptions\*\{\}”, they were labeled as “Non-Food Effect”. The annotated results were manually checked afterwards to verify the paragraph labeling. Note that there was no change in the annotated label after manual checking. For model training and evaluation, we randomly selected 2,400 records in total (1,200 records from Drugs@FDA, 1,200 records from DailyMed), including both Food Effect and Non-Food Effect. Note that we used the equal number of records from Drugs@FDA and DailyMed (1,200 records each) to keep the data balanced and avoid potential bias during the training.

Rule-Based Methods
We performed two rule-based methods as a baseline to compare the performance of machine learning models in test dataset. For the Rule-Based Method 1, if a paragraph matched the regular expression “(food effect|food effects|effect of food|effects of food)” anywhere, it was labelled as “Food Effect”; otherwise, it was labelled as “Non-Food Effect”. For the Rule-Based Method 2, we used a simple keyword search for “food” for paragraph labeling.

Machine Learning Models
Machine learning algorithms have been widely used for natural language processing, in which the text is represented by word embedding. Among various word embedding techniques, Term Frequency-Inverse Document Frequency (TF-IDF) perhaps is the most used statistical measure that evaluates how relevant a word is to a document (Chowdhury 2010). TF-IDF for a word in a document is calculated by multiplying two different metrics: Term Frequency (TF) and Inverse Document Frequency (IDF). TF-IDF score for the word $t$ in the document $d$ from the document set $D$ is calculated as follows:

$$TF - IDF (t, d, D) = TF(t, d) \cdot IDF(t, D)$$

Where:

$$TF(t, d) = log\left[ 1 + \text{freq}(t, D) \right],$$
FIGURE 2 | Unify identification scheme example (DailyMed SETID: c040bd1d-45b7-49f2-93ea-aed7220b30ac).
12.3 Pharmacokinetics

Absorption

In pediatric patients with ALL, oral absorption of methotrexate appears to be dose dependent; the absorption of doses greater than 40 mg/m² is significantly less than that of lower doses. ...

Effect of Food

The administration of XATMEP with food did not affect the area under the curve (AUC), but decreased the maximal concentrations (Cmax) by 50% and delayed the absorption.

FDA Application Number:
NDA208400

(A) Section in Pharmacokinetics

12.3 Pharmacokinetics

Absorption

Conjugated estrogens are water-soluble and are absorbed from the gastrointestinal tract after release from the drug formulation. ...

Food Effect

The pharmacokinetics of PREMARIN 0.45 mg and 1.25 mg tablets were assessed following a single dose with a high-fat breakfast and with fasting administration. The Cmax and AUC of estrogens were altered approximately 3 - 13%. ...

FDA Application Number:
NDA004782

(B) Subsection in Absorption

12.3 Pharmacokinetics

Absorption

Nintedanib reached maximum plasma concentrations, approximately 2 to 4 hours after oral administration as a soft gelatin capsule under fed conditions. ...

After food intake, nintedanib exposure increased by approximately 20% compared to administration under fasted conditions (90% CI: 95.3% to 152.5%) and absorption was delayed (median tmax: 2.00 hours; fed: 3.98 hours), irrespective of the food type.

FDA Application Number:
NDA205832

(C) No subtitle in Absorption

FIGURE 3 | Three Examples of Food Effect Section. In (A) and (B), the paragraphs in the blue border are annotated as “Non-Food Effect”, and the paragraphs in the red border are annotated as “Food Effect”. In (C), the Food Effect paragraph classification model is able to identify the paragraph in the blue border as “Non-Food Effect”, and the paragraphs in the red border as “Food Effect”. The dashed borders signify the paragraphs labeled by the model, whereas the borders in solid line represent the paragraphs automatically labeled by using regular expressions.

TABLE 3 | Statistical Summary of Food Effect/Non-Food Effect Data Extracted by Keyword Detection and Regular Expression. The numbers represent the number of paragraphs.

|          | Non-food effect | Food effect |
|----------|----------------|-------------|
|          | Food effect in pharmacokinetics section | Food effect in absorption section | Total |
| DailyMed | 3,691           | 822         | 930 |
| Drugs@FDA | 1,166           | 327         | 401 |

\[ IDF(t, D) = \log \left( \frac{N}{\text{count}(d \in D : t \in d)} \right). \]

\( N \) refers to the total number of documents in the corpus and \( \text{count}(d \in D : t \in d) \) corresponds to the number of documents where the term \( t \) appears. The TF-IDF score is then fed to machine learning algorithms such as Logistic Regression and Support Vector Machines, which has been shown improved results over the basic word embedding methods such as Bag-of-Words (BoW). However, the methods based on the concurrence of terms, including the BoW and TF-IDF (that embeds the text by term frequency in the document, which lose the information about the relationship between the words), fail to identify syntactic and semantic relationships between words in the documents.

Language model pretraining has recently advanced the state of the art in many NLP tasks ranging from sentiment analysis, to question answering, natural language inference, named entity recognition, and textual similarity. State-of-the-art pre-trained models include ELMo (Peters et al., 2018), GPT (Radford et al., 2019) and more recently Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019). BERT combines both word and sentence representations in a single very large Transformer (Vaswani et al., 2017); it is pre-trained on vast amounts of text, with an unsupervised objective of masked language modeling and next-sentence prediction and can be fine-tuned with various task-specific objectives. With the pre-trained model, fine-tuning allows BERT to model many downstream tasks, such as text classification. While BERT has dramatically improved outcomes in NLP tasks in the general domain such as optimizing search results, its performance in domain-specific tasks such as drug labeling has not fully been explored.

In this paper, we assessed the predictive performance of three pre-trained BERT-base models: BERT, RoBERTa (Liu et al., 2019),
and DistilBERT (Sanh et al., 2020). We compared these state-of-the-art approaches with three traditional machine learning algorithms: Logistic Regression, Linear Support Vector Classification (SVC), and Random Forest, with the TF-IDF as the word embedding method. We note that the input to the model was a single paragraph and the output was a label for that paragraph to indicate whether it was “Food Effect” or “Non-Food Effect.”

The BERT-base models were implemented by Simple Transformers library7. We used Adam optimizer (Kingma and Jimmy, 2017) with the learning rate of 4e-5. The train batch size was set to 32. To avoid exploding gradients, we clipped the gradients to the maximum norm of one. Other hyperparameters of the models were initialized by default values.

**Evaluation**

We evaluated the classification results from two perspectives: the impact of training and testing dataset and the performance of text classification algorithms.

**RESULTS**

**Information Extraction**

There were total 12 sections (See Table 4 for details) in the drug product information we extracted in this work, only eight sections could be extracted by the LOINC codes (See Table 2 for details; we note that only DailyMed has the LOINC codes). For sections

---

7https://github.com/ThilinaRajapakse/simpletransformers
random experiments. Random Forest and DistilBERT are the best performance evaluation metrics reported are the average of ten for Logistic Regression, Linear SVC, and Random Forest). The traditional machine learning models (0.8889, 0.9004, and 0.9345)

Drugs@FDA and DailyMed. First, the machine learning models

Table 5

| Train                  | Test                  | Random forest | DistilBERT |
|------------------------|-----------------------|---------------|------------|
| Drugs@FDA + DailyMed   | Drugs@FDA + DailyMed  | 0.9345        | 0.9607     |
| DailyMed               | Drugs@FDA + DailyMed  | 0.9204        | 0.9335     |
| Drugs@FDA              | Drugs@FDA + DailyMed  | 0.9067        | 0.9478     |
| DailyMed               | DailyMed              | 0.9388        | 0.9796     |
| Drugs@FDA              | DailyMed              | 0.9149        | 0.9388     |
| DailyMed               | DailyMed              | 0.8866        | 0.9683     |
| Drugs@FDA + DailyMed   | Drugs@FDA             | 0.9313        | 0.9313     |
| DailyMed               | Drugs@FDA             | 0.9242        | 0.9202     |
| Drugs@FDA              | Drugs@FDA             | 0.9219        | 0.9394     |

Classification Performance of Food Effect by Machine Learning Models

Table 4 summarized how much the data was covered by each of the three other sources. DrugBank covered the largest number of unique drugs (91.28%), followed by Drugs@FDA (90.12%). For the drug labeling sections, where the information is available in DrugBank (indication, mechanism of action, pharmacodynamics, absorption, distribution, metabolism, excretion), it also has the largest coverage among the three sources. Drugs@FDA and DailyMed both provided more labeling sections than DrugBank, though Drugs@FDA has higher coverage than DailyMed in general.

To quantify the uniqueness of each source’s contribution, we computed the overlap measure, which is defined as the percentage of the average number of sources that contributed to each drug labeling section to the total number of the data sources. For example, 0% overlap indicates that the data source is the only source provided information for that drug labeling section, which has no overlap with other sources. Similarly, 100% overlap indicates that the information of the drug labeling section that one source provided completely shares with the other two sources. Drugs@FDA had the least overlap for the unique drugs (82.00%).

DISCUSSION AND CONCLUSION

In this paper, we developed an NLP pipeline to perform information extraction from FDA drug labeling for PSG development. We integrated drug product information from multiple data sources (Orange Book, Drugs@FDA, DailyMed, and DrugBank) and have collected information with a variety of data formats via different

**TABLE 6 | Result with different combinations of training and testing on various datasets (Best model performance of each combination is in bold).**

Train                  | Test                  | F1
|------------------------|-----------------------|------------|
| Drugs@FDA + DailyMed   | Drugs@FDA + DailyMed  | 0.9345     |
| DailyMed               | Drugs@FDA + DailyMed  | 0.9204     |
| Drugs@FDA              | Drugs@FDA + DailyMed  | 0.9067     |
| DailyMed               | DailyMed              | 0.9388     |
| Drugs@FDA              | DailyMed              | 0.9149     |
| DailyMed               | DailyMed              | 0.8866     |
| Drugs@FDA + DailyMed   | Drugs@FDA             | 0.9313     |
| DailyMed               | Drugs@FDA             | 0.9242     |
| Drugs@FDA              | Drugs@FDA             | 0.9219     |
access methods and addressed the inconsistent issue of the identification scheme of different data sources, which allows one to request the FDA-approved drug data by a unified identifier. Among these data sources, Drugs@FDA has relatively high coverage and the least overlap with other data sources, making itself ideally the primary source for the PSG assessment. Nonetheless, other data sources can still provide complementary information to Drugs@FDA for a comprehensive assessment of the drug products.

We used food effect paragraph labeling as an example to address the complex scenarios when keyword detection and regular expression are not adequate. For example, some “Non-Food Effect” paragraphs may include a sentence such as “Effect of food on the bioavailability of methadone has not been evaluated,” which matches the regular expression but is not related to food effect study of the drug. Similarly, some “Food Effect” paragraphs may contain the food effect content but without the keyword “food” (e.g., “A high-fat meal increased the extent and rate of naloxegol absorption. The Cmax and AUC were increased by approximately 30 and 45%, respectively. In clinical trials, naloxegol was dosed on an empty stomach approximately 1 h prior to the first meal in the morning.”). We further demonstrated the pre-trained BERT-based model, when fine-tuned on a small set of annotated labels, is able to outperform the traditional machine learning algorithms in text classification for food effect labeling on drug labeling datasets. This result was confirmed by checking the model on unseen test domain. For example, the second paragraph in Figure 3C was correctly identified as “Food Effect” by the BERT, yet completely missed by keywords/regular expression.

Several limitations need to be noted. First, we only apply this classifier to separate food effect paragraphs within the absorption section. As a result, the Non-Food Effect data only contains absorption data from drug labeling. Second, the minimum text unit we extracted and annotated is a paragraph. The food effect is not required to be written in a separate paragraph in the drug labeling and thus has the possibility to be either mentioned in a sentence within other sections or spanned more than one paragraph. In either case, it is possible that a small amount of food effect paragraphs can be mixed with the Non-Food Effect data. Third, when the size of the training dataset is limited, e.g., less than 400 as shown in Figure 4, Random Forest has a better performance than DistilBERT. As the amount of the data increases, the advantage of the BERT-based model becomes clear. Our observation is consistent with what is well known in the deep learning field that a deep learning model requires much more data than a traditional machine learning algorithm.

In summary, we developed a pipeline to integrate four publicly available data sources for drug product information various data formats and access methods. 12 related sections in the drug labeling were extracted from these data sources, which aimed to enhance the assessment process for PSG development. In addition, we demonstrated that the pre-trained BERT model is able to outperform the traditional machine learning techniques, setting a new state-of-the-art on drug labeling datasets to address the classification challenge for labeling food effect paragraphs, which can be adapted to other drug labeling sections or data sources.

**DATA AVAILABILITY STATEMENT**

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

**AUTHOR CONTRIBUTIONS**

All authors wrote the manuscript and performed the research. MH, HL designed the research. YS, HL analyzed the data.

**FUNDING**

This work was partly supported by the FDA Contract #: 75F40119C10106.

**REFERENCES**

Bisgin, H., Liu, Z., Fang, H., Xu, X., and Tong, W. (2011). Mining FDA Drug Labels Using an Unsupervised Learning Technique - Topic Modeling. BMC Bioinformatics. 12 (Suppl. 10), S11. doi:10.1186/1471-2105-12-S10-S11

Center for Drug Evaluation and Research (2020). “Orange Book Preface.” FDA. February. Available at: https://www.fda.gov/drugs/development-approval-process-drugs/orange-book-preface

Chowdhury, G. G. (2010). Introduction to Modern Information Retrieval. London: Facet Publishing.

Devlin, J., Chang, M-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv: 1810.04805 [Cs], May. Available at: http://arxiv.org/abs/1810.04805. doi:10.18653/v1/n19-1423

Fang, H., Harris, S. C., Liu, Z., Zhou, G., Zhang, G., Xu, J., et al. (2016). FDA Drug Labeling: Rich Resources to Facilitate Precision Medicine, Drug Safety, and Regulatory Science. Drug Discov. Today 21 (10), 1566–1570. doi:10.1016/j.drudis.2016.06.006
Fung, K. W., Jao, C. S., and Demner-Fushman, D. (2013). Extracting Drug Indication Information from Structured Product Labels Using Natural Language Processing. *J. Am. Med. Inform. Assoc. JAMIA*. 20 (3), 482–488. doi:10.1136/amiajnl-2012-001291

Kingma, D. P., and Jimmy, B. (2017). Adam: A Method for Stochastic Optimization. *ArXiv:1412.6980 [Cs]*, January. Available at: http://arxiv.org/abs/1412.6980

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., et al. (2011). DrugBank 3.0: a Comprehensive Resource for ‘Omics’ Research on Drugs. *Nucleic Acids Res.* 39, D1035–D1041. doi:10.1093/nar/gkq1126

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv:1907.11692 [Cs]*, July. Available at: http://arxiv.org/abs/1907.11692

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). Deep Contextualized Word Representations. *ArXiv:1802.05365 [Cs]*, March. Available at: http://arxiv.org/abs/1802.05365. doi:10.18653/v1/n18-1202

Radford, A., Jeffrey Wu, R. C., Luan, D., Amodei, D., and Sutskever, I. (2019). “Language Models Are Unsupervised Multitask Learners.” 2019. Available at: paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *ArXiv:1910.01108 [Cs]*, February. Available at: http://arxiv.org/abs/1910.01108

Schadow, G. (2005). HL7 Structured Product Labeling - Electronic Prescribing Information for Provider Order Entry Decision Support. *AMIA Annu. Symp. Proc.* 2005, 1108

Sharma, M. R., Karrison, T. G., Bell, B., Wu, K., Turcich, M., Geary, D., et al. (2013). Evaluation of Food Effect on Pharmacokinetics of Vismodegib in Advanced Solid Tumor Patients. *Clin. Cancer Res.* 19 (11), 3059–3067. doi:10.1158/1078-0432.CCR-12-3839

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*, December. Available at: http://arxiv.org/abs/1706.03762

Wishart, D. S., KnoxGuo, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2008). DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* 36, D901–D906. doi:10.1093/nar/gkm958

Disclaimer: The opinions expressed in this article are the authors’ own and do not reflect the views of the Food and Drug Administration, the Department of Health and Human Services, or the United States government.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Shi, Ren, Zhang, Gong, Hu and Liang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.