Dec 11th, 12:00 AM

# DrugExBERT for Pharmacovigilance – A Novel Approach for Detecting Drug Experiences from User-Generated Content

Eva Bohnen
*Ulm University*, eva.bohnen@uni-ulm.de

Stefanie Erlebach
*Ulm University*, stefanie.erlebach@uni-ulm.de

Steffen Zimmermann
*Ulm University*, steffen.zimmermann@uni-ulm.de

# DrugExBERT for Pharmacovigilance – A Novel Approach for Detecting Drug Experiences from User-Generated Content

*Completed Research Paper*

**Eva Bohnen**
Ulm University
Helmholtzstraße 22
89081 Ulm, Germany
eva.bohnen@uni-ulm.de

**Stefanie Erlebach**
Ulm University
Helmholtzstraße 22
89081 Ulm, Germany
stefanie.erlebach@uni-ulm.de

**Steffen Zimmermann**
Ulm University
Helmholtzstraße 22
89081 Ulm, Germany
steffen.zimmermann@uni-ulm.de

## Abstract

*Pharmaceutical companies have to maintain drug safety through pharmacovigilance systems by monitoring various sources of information about adverse drug experiences. Recently, user-generated content (UGC) has emerged as a valuable source of real-world drug experiences, posing new challenges due to its high volume and variety. We present DrugExBERT, a novel approach to extract adverse drug experiences (adverse reaction, lack of effect) and supportive drug experiences (effectiveness, intervention, indication, and off-label use) from UGC. To be able to verify the extracted drug experiences, DrugExBERT additionally provides explications in the form of UGC phrases that were critical for the extraction. In our evaluation, we demonstrate that DrugExBERT outperforms state-of-the-art pharmacovigilance approaches as well as ChatGPT on several performance measures and that DrugExBERT is data- and drug-agnostic. Thus, our novel approach can help pharmaceutical companies meet their legal obligations and ethical responsibility while ensuring patient safety and monitoring drug effectiveness.*

**Keywords:** Pharmacovigilance System, Adverse Drug Reactions, Natural Language Processing, User-Generated Content, Design Science Research

## Introduction

Each year, approximately 2 million patients in the United States are hospitalized due to severe adverse drug reactions, resulting in roughly 100,000 deaths. This makes adverse drug reactions the fourth to sixth leading cause of death in the United States (Lazarou et al. 1998; FDA 2018). As a major cause of morbidity and mortality, the economic burden associated with adverse drug reactions is also substantial, with annual costs reaching $136.8 billion in the United States (Johnson and Bootman 1997; FDA 2018).

To mitigate these enormous consequences, pharmaceutical companies have both an ethical responsibility and a legal obligation to maintain the ongoing safety and effectiveness of drugs. For post-marketing surveillance, pharmaceutical companies have to operate a pharmacovigilance system to monitor the usage

of drugs for cases of adverse drug experiences (FDA 2001; FDA 2023). Adverse drug experiences are defined as adverse events associated with the use of a drug and include both harmful and unintended drug reactions (*adverse reactions*) and failure of expected pharmacological action (*lack of effect*) (FDA 2023). When adverse drug experiences are identified, pharmaceutical companies are required to submit reports to regulatory authorities such as the Food & Drug Administration (FDA) in the United States. These reports consist of general information and a holistic view of the adverse drug experience. To this end, and if accessible, pharmaceutical companies are prompted to provide further information on drug experiences in order to support the assessment of individual cases. These supportive drug experiences include in particular the dosage, frequency, and route of administration for usage (*intervention*), and the diagnosis for usage (*indication*) (FDA 2023). To better assess individual cases, pharmaceutical companies should further evaluate the adverse drug experiences with regard to whether a drug was taken for an unapproved diagnosis for usage (*off-label use*) (FDA 2001). They also have an ethical responsibility to carefully consider the ongoing use of their drugs. Therefore, it is common practice to evaluate supportive drug experiences in terms of whether the drug achieved the expected pharmacological effect (*effectiveness*) for various *interventions*, *indications*, and *off-label uses*. Table 1 provides an overview of the drug experiences that are a critical component of pharmacovigilance systems.[1] The description of each drug experience is closely aligned with the FDA (2001) and FDA (2023) and supplemented with examples derived from our data.

| Category | Class | Description | Example |
|---|---|---|---|
| **Adverse drug experiences** | Adverse reaction | Harmful and unintended drug reaction | Headache occurs after intake of allergy pills |
| | Lack of effect | Failure of expected pharmacological effect | Symptoms do not disappear after intake of allergy pills |
| **Supportive drug experiences** | Effectiveness | Expected pharmacological effect | Symptoms disappear after intake of allergy pills |
| | Intervention | Dosage, frequency, and route of administration | Single intake of allergy pills after an allergic reaction |
| | Indication | Diagnosis for usage | Intake of allergy pills against allergic reaction |
| | Off-label use | Unapproved diagnosis for usage | Intake of allergy pills to promote sleeping |

**Table 1. Drug Experiences**

In recent years, user-generated content (UGC) has emerged as a valuable source of information to complement pharmacovigilance systems (Borchert et al. 2019). UGC generally refers to any type of content that is created and shared by users on a digital platform or website and includes valuable first-hand experiences of users with a product (Goh et al. 2013). Thus, unlike clinical trials that follow strict protocols, UGC provides real-world drug experiences (Gosal 2015), where users report their experiences explicitly, implicitly, or even unconsciously. However, for pharmaceutical companies, detecting drug experiences from UGC is often a difficult and lengthy process (Nikfarjam and Gonzalez 2011). This is further complicated by the ever-increasing volume of UGC (Schouten and Frasincar 2016). While this large volume of UGC provides valuable insights to maintain ongoing drug safety and efficacy in the post-marketing setting, it makes its manual evaluation almost impossible and necessitates approaches to automatically detect drug experiences from UGC. As a result, researchers and practitioners have made serious efforts to process this high volume of UGC by developing data-driven approaches to detect drug experiences (see, e.g., Pilipiec et al. 2022 for literature a review). For such an approach to be useful to pharmaceutical companies, it needs to have the ability to accurately extract adverse and supportive drug experiences (see Table 1) from UGC. Especially for adverse drug experiences, it is crucial to achieve high performance while minimizing type 2 errors to ensure that as few adverse drug experiences as possible are missed. Second,

[1] *Note that while we derive the crucial drug experiences for pharmacovigilance from ethical responsibilities and regulatory obligations in the United States (FDA 2023; FDA 2001), they are very similar for pharmaceutical companies in other regions, such as, e.g., the EU (European Medicines Agency 2017). This is confirmed by a leading global pharmaceutical company that also operates in Europe and is collaborating with us to evaluate our approach.*

and to the same end, pharmaceutical companies are prompted to provide a description of the adverse drug experience and include concise medical narratives (FDA 2023). In practice, this is typically achieved by including the relevant UGC phrases from which the adverse drug experiences were deduced in reports (*explication*). Providing this explication not only increases the transparency of the approach, but also enables pharmaceutical companies and regulatory authorities to verify that the extracted adverse drug experiences are reasonable. Moreover, UGC on different digital platforms can be very diverse, ranging from unstructured text data such as social media posts, medical forum posts, or online customer reviews (OCR), to semi-structured text data such as OCR templates (Gosal 2015). However, approaches that are limited to a specific type of UGC may miss valuable information or may simply not apply to various data structures. Thus, for effective drug safety surveillance, it is crucial that an approach is data-agnostic in order to work with different types of UGC and across platforms (i.e., data-agnosticism). The agnostic nature of the approach must also be ensured with respect to the drugs of interest. Approaches, therefore, need to be developed or trained independently of drug-specific experiences in order to be transferable to new drugs (i.e., drug-agnosticism).

Thus, we aim to build and evaluate a novel approach for pharmacovigilance that (1) is able to detect the drug experiences shown in Table 1 from UGC, (2) provides an explication for the detected drug experiences, and (3) is data- and drug-agnostic. We follow a design science approach (Hevner et al. 2004) and use these requirements as design principles to build our design artifact, DrugExBERT, representing a method for detecting drug experiences from UGC. Our approach uses state-of-the-art natural language processing (NLP) techniques such as transformers (Devlin et al. 2018) in combination with powerful medical systems (Bodenreider 2004) to detect adverse and supportive drug experiences (see Table 1). We evaluate DrugExBERT in multiple dimensions. We first evaluate its performance against state-of-the-art approaches from the literature. Second, we evaluate DrugExBERT against generative AI tools on the market, including ChatGPT. In both evaluation steps, we outperform existing approaches in terms of accuracy and even more in terms of recall. We further evaluate its data- and drug-agnosticism by applying DrugExBERT to real data and drugs from a leading pharmaceutical company on which DrugExBERT was not trained. In this setting, our approach also outperforms ChatGPT. Overall, according to Gregor and Hevner (2013), our approach represents an "improvement" and a promising solution to the challenge of detecting pharmacovigilance-relevant drug experiences from UGC. By improving the efficiency and accuracy of this process, our approach can help pharmaceutical companies to meet their legal obligations while ensuring patient safety and monitoring the effectiveness of their drugs.

# Related Literature

Our proposed approach relates to two distinct bodies of literature in the context of UGC: research on NLP of UGC and research on drug experience detection.

## *Natural Language Processing of User-Generated Content*

NLP aims to process and analyze natural language text (Locke et al. 2021). Its application in IS research is guided by real-world problems (Liu et al. 2017). NLP encompasses a wide range of tasks, including but not limited to aspect extraction, sentiment analysis (Nazir et al. 2022), text classification, text generation, or word sense disambiguation (Liu et al. 2017). For our approach, aspect extraction and text classification are the most relevant NLP tasks, as it aims to extract aspects (i.e., *adverse reactions*) from UGC and to assign a meaning to the text units of UGC by classifying them into predefined categories (i.e., drug experiences).

### Aspect Extraction from User-Generated Content

Aspect extraction aims to extract specific expressions (i.e., aspects) of a product, service, or domain from unstructured text data (Nazir et al. 2022). In the context of UGC, aspect extraction is a particularly challenging task as UGC contains both explicit and implicit aspects. Explicit aspects refer to aspects that are directly mentioned in the UGC (e.g., "dizziness"), whereas implicit aspects are solely implicitly annotated and have to be inferred from the UGC (e.g., "everything was spinning") (Schouten and Frasincar 2016). There are many models for aspect extraction (see, e.g., Nazir et al. 2022, Schouten and Frasincar 2016 for a literature review). The majority of existing models are designed to extract explicit aspects (e.g., Li and Lam 2017; Gu et al. 2017), while only a few are available to extract implicit aspects (e.g., Tubishat et

al. 2018) or both explicit and implicit aspects (e.g., Ma et al. 2018). Extracting implicit aspects can still be particularly difficult and may sometimes even require additional context or domain knowledge (Schouten and Frasincar 2016). Domain applications pose a huge challenge to aspect extraction, as different domains often have unique terminologies, concepts, and language conventions. For example, medical texts contain specific jargon and medical terms that are not commonly used in other domains. Thus, to improve the accuracy of aspect extraction in specific domains, many researchers have explored the use of domain-specific knowledge and resources (e.g., O'Connor et al. 2014; Cavalcanti and Prudêncio 2017; Imani and Noferesti 2022). In the medical domain, MetaMap (Aronson 2001) has emerged as one of the most widely used resources for aspect extraction (Imani and Noferesti 2022). MetaMap performs syntactic and semantic analysis of texts and includes a rule-based mapping module that links mentions of biomedical concepts to the Unified Medical Language System (UMLS) (Bodenreider 2004). MetaMap is able to detect acronyms and abbreviations, search the Metathesaurus for concepts that are only remotely related to the input text, detect negations, and perform word sense disambiguation (Aronson and Lang 2010). This multitude of powerful features makes it particularly effective for extracting explicit and implicit aspects from UGC in the medical domain, and thus for our research.

**Classification of User-Generated Content**

Text classification is the process of assigning predefined categories to units of text data, such as documents, paragraphs, sentences, or phrases. Traditionally, there are two main approaches for text classification: rule-based approaches, which require deep domain knowledge and predefined rules, and machine learning-based approaches, which learn to classify text based on observational data (Minaee et al. 2022). In recent years, deep-learning approaches for text classification tasks have emerged and surpassed traditional machine learning approaches (see, e.g., Minaee et al. 2022 for a literature review). These approaches range from recurrent neural networks (e.g., Cheng et al. 2016) to convolutional neural networks (e.g., Kalchbrenner et al. 2014) to transformer-based models (e.g., Devlin et al. 2018). Transformer-based models, introduced by Vaswani et al. (2017), are particularly noteworthy for their ability to capture long-range dependencies and complex linguistic patterns. The Generative Pre-Trained Transformer (GPT) series, such as ChatGPT (Brown et al. 2020) includes large-scale models. These models have been pre-trained on extensive text data and have shown remarkable performance on a variety of NLP tasks, such as text classification or text generation. Recent studies reveal that other transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT), show exceptional performance on various NLP tasks, outperforming traditional text classification models (Devlin et al. 2018). In recent years, various extensions of BERT have been introduced to address the specific challenges and nuances of particular domains. Examples include ClinicalBERT for clinical text (Huang et al. 2019), BioBERT for biomedical research (Lee et al. 2019), and RoBERTa for UGC (Liu et al. 2019). Consequently, BERT models are particularly suitable for our research due to their superior performance in text classification tasks and their available domain-specific extensions.

## *Drug Experiences Detection in User-Generated Content*

Detecting adverse reactions in UGC such as OCR or social media data, is a well-known and discussed problem with numerous data-driven models available in the existing literature (see, e.g., Sarker et al. 2015, Dreisbach et al. 2019, Pilipiec et al. 2022, Kaas-Hansen et al. 2022 for a literature review). Models to extract *adverse reactions* range from lexicon-based approaches (e.g., Leaman et al. 2010) to approaches that focus on machine learning (e.g., Yang et al. 2013), sentiment analysis (e.g., Sharif et al. 2014) or deep learning (e.g., Xia et al. 2017). Although these approaches achieve reasonable performance, they do not correspond to good pharmacovigilance practice, as they do not extract *lack of effect* and supportive drug experiences (i.e., *effectiveness*, *intervention*, *indication*, *off-label use*) from UGC.

To meet regulatory obligations, Adams et al. (2017) apply sentiment analysis to detect safety concerns in terms of *adverse reactions* and *lack of effect* from OCR on amazon.com. Other approaches have been proposed to detect not only a*dverse reactions* and *lack of effect*, but also *effectiveness* from UGC. Based on semi-structured OCR on medical forums, these approaches use sentiment analysis (e.g., Gräßer et al. 2018; Ajibade et al. 2022), or generic foundation models (e.g., Unnikrishnan et al. 2023) to detect drug experiences. However, they still fail to provide a sufficient level of detail with respect to the detection of supportive drug experiences. To this end, other approaches provide a more holistic consideration of drug

experiences by additionally detecting *interventions* and *indications* from UGC (e.g., Na and Kyaing 2015; Cavalcanti and Prudêncio 2017; Imani and Noferesti 2022). Na and Kyaing (2015) develop an approach based on sentiment analysis. More recently, Cavalcanti and Prudêncio (2017) and Imani and Noferesti (2022) propose an approach that combines medical systems with generic foundation models. While these approaches provide a more detailed detection of supportive drug experiences, they still fail to distinguish between *lack of effect* and *effectiveness* of drugs (e.g., Na and Kyaing 2015; Cavalcanti and Prudêncio 2017; Imani and Noferesti 2022), even though this is crucial for reporting adverse drug experiences to regulatory authorities. To make matters worse, to the best of our knowledge, none of the approaches meeting regulatory requirements can detect *off-label use* from UGC or provide *explications* to verify the detected adverse drug experiences. Moreover, only a few of those approaches evaluate data-agnosticism (e.g., Gräßer et al. 2018; Unnikrishnan et al. 2023) and drug-agnosticism (e.g., Gräßer et al. 2018; Ajibade et al. 2022; Unnikrishnan et al. 2023). Those approaches that are able to provide a more holistic view of drug experiences (e.g., Na and Kyaing 2015; Cavalcanti and Prudêncio 2017; Imani and Noferesti 2022) only provide an evaluation based on the same drugs and data, which were used to train their approaches.

| | | Adams et al. (2017) | Gräßer et al. (2018) | Ajibade et al. (2022) | Unnikrishnan et al. (2023) | Na and Kyaing (2015) | Cavalcanti and Prudêncio (2017) | Imani and Noferesti (2022) | Our approach (DrugExBERT) |
|---|---|---|---|---|---|---|---|---|---|
| **Adverse drug experiences** | Adverse reaction | x | x | x | x | x | x | x | x |
| | Lack of effect | x | x | x | x | x | x | x | x |
| **Supportive drug experiences** | Effectiveness | | x | x | x | | | | x |
| | Intervention | | | | | x | x | x | x |
| | Indication | | | | | x | x | x | x |
| | Off-label use | | | | | | | | x |
| **Explication** | | | | | | | | | x |
| **Agnosticism** | Data-agnosticism | | x | | x | | | | x |
| | Drug-agnosticism | | x | x | x | | | | x |

**Table 2. Related Literature and Research Gap**

In summary, and as shown in Table 2, existing approaches still have some limitations that make them unsuitable for pharmaceutical companies. They either do not provide a sufficient level of detail to meet crucial regulatory obligations (e.g., Na and Kyaing 2015; Cavalcanti and Prudêncio 2017; Imani and Noferesti 2022), or they fail to detect or isolate relevant supportive drug experiences (e.g., Gräßer et al. 2018; Ajibade et al. 2022; Unnikrishnan et al. 2023; Cavalcanti and Prudêncio 2017; Imani and Noferesti 2022; Na and Kyaing 2015). Furthermore, none of these approaches is able to detect *off-label use* from UGC or to provide *explications* for adverse drug experiences.

## DrugExBERT – A Novel Approach for Detecting Drug Experiences

In this section, we present our novel approach DrugExBERT to detect adverse drug experiences along with supportive drug experiences from UGC. As shown in Figure 1, the approach can be broadly divided into three main phases: drug experience extraction (Phase 1), drug experience classification (Phase 2), and explication (Phase 3). Thereby, it is important to distinguish between the training path (dotted lines) and the application of the pre-trained approach to new, previously unseen UGC (solid lines).

In Phase 1, the UGC is separated into phrases. In parallel, UGC aspects that address a drug experience are extracted and mapped to corresponding medical terms and tags of their semantic type. Phase 2 classifies these phrases into either one of the classes *adverse reaction*, *lack of effect*, *effectiveness, intervention*, *indication*, or *other* (Classifier). The outputs of Phase 2 are classified phrases and indirectly classified aspects. Finally, Phase 3 takes the classified phrases as input and provides an aggregated classification of the entire UGC, along with an *explication* of this classification. Together, these phases form DrugExBERT, a comprehensive and robust approach to detecting drug experiences in UGC. In the remainder of this section, the phases and individual steps are described in detail.

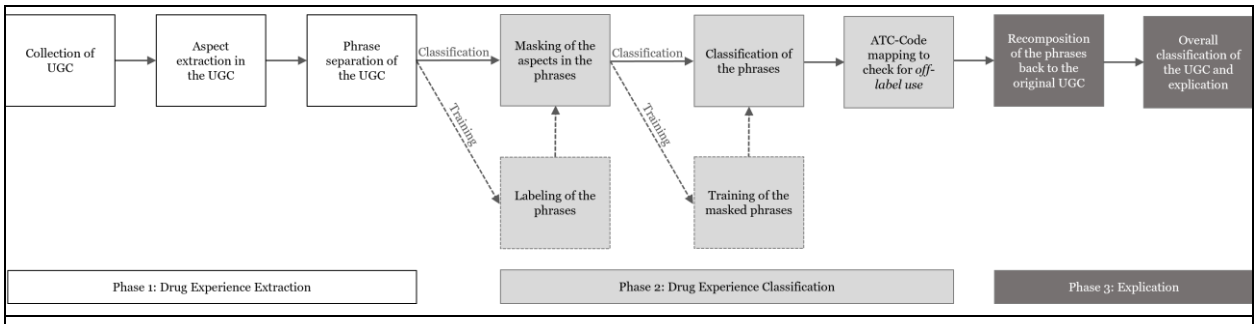


**Figure 1. DrugExBERT**

## *Drug Experience Extraction*

After the collection of UGC, the relevant aspects are extracted for each UGC, i.e., the concrete expressions that mention a drug experience. This aspect extraction is performed at the UGC level and not at the phrase level to ensure that the context of the entire UGC is preserved. To perform the aspect and medical term extraction, we use MetaMap (Aronson 2001), which is an essential component of the UMLS (Bodenreider 2004). Using MetaMap allows us to conduct word sense disambiguation and extract the aspects that mention a drug experience in the UGC. Typically, a UGC consists of colloquial language, so the aspects tend to paraphrase the experience. For this reason, the extracted aspects are also assigned to a corresponding medical term. UMLS has collected over one million different biomedical concepts and categorized them into 130 groups known as semantic types (Aronson and Lang 2010). Each UGC may or may not consist of one or more aspects. Each aspect can have multiple semantic types associated with it. After a detailed examination of the 130 different types, we filtered out the three most relevant types for pharmacovigilance. These types describe concrete medical experiences, as they include symptoms, diseases, and dysfunctions. Table 3 shows these three semantic types as well as their abbreviations. We will refer to the abbreviations as "tags" and the assignments of the tags as "tagging". This tagging process is critical to understanding the extracted aspects and their role in the overall reported drug experience.

| Tag | Semantic Type |
|---|---|
| sosy | Sign or Symptom |
| dsyn | Disease or Syndrome |
| mobd | Mental or Behavioral Dysfunction |

**Table 3. Relevant Semantic Types**

Further, a UGC can be broken down into different sections of meaning, allowing it to be split into individual phrases after pre-processing. This is done in Phase 1 of DrugExBERT by the phrase separation using Python's SpaCy library (Honnibal et al. 2020). The separator identifies the main verb of each phrase and its conjuncts. For each head identified (i.e., the main verb and its conjuncts), the function collects its subtrees (words directly or indirectly related to the head). These subtrees are stored as chunks, later sorted, and returned as phrases.

In summary, the output of Phase 1 of our approach consists of a UGC divided into different phrases. In addition to the phrases, the corresponding aspects that mention drug experiences are provided, mapped to a medical term, and tagged with their semantic type. For example, in Figure 2, the phrase "and after that I

felt sick" consists of one aspect describing a drug experience. This aspect is "felt sick", which is tagged as a sign or symptom and mapped to the medical term "nausea" in the ULMS MetaMap.

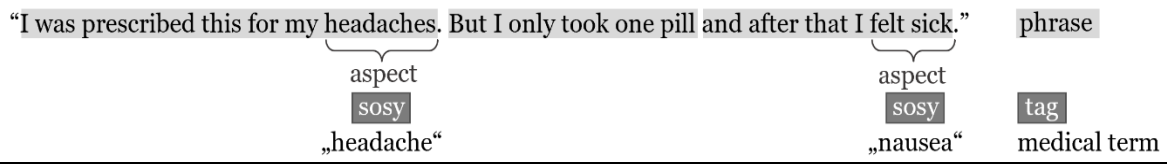


**Figure 2. Example of Aspect Extraction with MetaMap Tagging**

## *Drug Experience Classification*

Phase 2 represents the classification part of DrugExBERT, which is crucial for determining the meaning of the extracted drug experience. This phase consists of the four parts masking, labeling, training, and classification. First, we take the output of Phase 1 and perform what is known as masking. That is, for each phrase we replace the extracted aspect with the corresponding tag. An example of this is shown in Figure 3. By masking, we ensure that the classifier solely learns the syntax of the sentence and not the specific aspect during training. For example, a symptom can be an *adverse reaction*, but it can also be an *indication* (e.g., "I took the pill because I had a headache" vs. "I took the pill and now I have a headache."). Since a phrase can consist of multiple aspects, it is possible that a phrase will eventually be masked in multiple places as well. By masking the symptoms, the classifier is trained on the structure of the phrase and does not learn the symptom itself. This approach has several advantages. It allows us to better generalize to unseen data, making it more robust and adaptable to variations in language and expression. It also prevents overfitting to the training data by avoiding the memorization of specific aspect expressions that may not be universally applicable. In the training pipeline, each phrase is given a corresponding label, classifying it as an *adverse reaction*, *lack of effect*, *effectiveness*, *intervention*, *indication*, or *other*[2]. It is important to distinguish between the labels and the tags. The tags are only used for masking and the labels are the labels that the classifier is trained on. As shown by the dotted lines in Figure 1, the labeled and masked phrases are then used as a database to train a classifier which in turn can be used to classify new, unseen UGC.

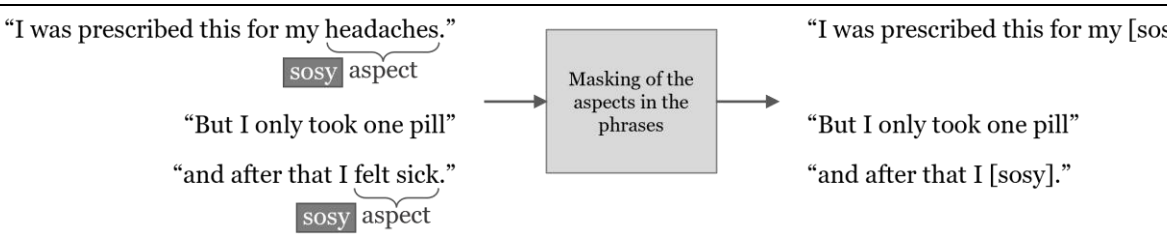


**Figure 3. Example of Masking**

For training purposes, we use state-of-the-art transformer models such as BERT (Devlin et al. 2018). Transformer models are particularly suitable in this context due to their superiority in capturing long-range dependencies and complex linguistic patterns in text. In the context of pharmacovigilance, these capabilities are critical for accurately classifying the phrases and extracted aspects and detecting relevant drug experiences. By using transformer models, we can ensure that the classification task is both accurate and robust. We have tested various pre-trained BERT models for their suitability in this context, including ClinicalBERT (Huang et al. 2019) and BioBERT (Lee et al. 2019). However, and in line with Unnikrishnan et al. (2023), we selected RoBERTa (Liu et al. 2019) based on its superior performance. RoBERTa was developed by Meta, which means that it is based on social media data, making it particularly suitable for UGC. In addition, it is pre-trained on a larger dataset than BERT and performs longer training with additional optimization. This pre-trained RoBERTa is then fine-tuned with the training dataset consisting of the labeled and masked phrases. The result is a customized RoBERTa-based multi-class classifier capable of predicting the labels of masked phrases and assigning them to predefined categories. It classifies whether

[2] *The label "other" describes all content that is not covered by the defined classes for drug experiences. This includes, for example, other product-, purchase-, seller-, or customer-related content that is typically addressed in UGC such as, e.g., OCR (Zhu et al. 2017; Züllig et al. 2023).*

the phrase mentions an *adverse reaction*, a *lack of effect,* the *effectiveness* of the drug*,* the *intervention*, the *indication*, or anything *other* than that. A further step is necessary to classify *off-label use* as well, as this classification is drug-specific and cannot be generalized. Since *off-label use* describes an unapproved indication, the phrases that are labeled as *off-label use* are a subset of those predicted to be an *indication*. Therefore, we take all phrases that are predicted to be *indications* and compare them to the originally intended indication of the drug to determine if it is an *off-label use* or an approved indication. We derive this information from an Anatomical Therapeutic Chemical (ATC) code mapping. The WHO ATC classification system is a method of categorizing drugs based on their active ingredients. This system considers the organ or system targeted by the drug as well as its therapeutic, pharmacological, and chemical properties to provide a comprehensive and organized classification (Nahler 2009). Since we know which drug the UGC is about, we also know which active ingredient the drug is based on. This allows us to assign the ATC code to the drug and thus know its originally intended indication.

To summarize, Phase 2 of our approach takes the phrases of a UGC along with the extracted aspects, tagged with a semantic type as input. The generated output consists of labeled phrases that are classified into different classes of drug experiences.

## *Explication*

Phase 3 identifies the critical part of the UGC that has led to an adverse drug experience classification. To do this, we first combine all the classifications for each phrase for each UGC to obtain an overall classification. Figure 4 shows this process for the overall classification of the UGC. We reassemble the phrases into the original UGC and consider all the labels for each phrase together. If the UGC contains an adverse drug experience (i.e., *adverse reaction*, *lack of effect*, question 1 in Figure 4), the UGC needs to be reported to regulatory authorities. Further effort is then required to gather supportive drug experiences from the other labels. Therefore, the first step is to check whether the set of labels for all phrases contains the label *adverse reaction* (question 2 in Figure 4). Then, the entire UGC is classified as an *adverse reaction*. If this is not the case, the same is checked for *lack of effect*, otherwise, the UGC is classified as *other*. If the UGC is classified as one of the adverse drug experiences, the UGC is searched for the supportive drug experiences (questions 3 to 6 in Figure 4). This process is staged, meaning that the UGC is searched first for the *indication* (question 3 in Figure 4), then for *off-label use* (question 4 in Figure 4), then for *effectiveness* (question 5 in Figure 4), and then for *intervention* (question 6 in Figure 4). To provide the explication, we extract the crucial phrases in the UGC that need further investigation by pharmaceutical companies (i.e., the explicit phrases classified as an *adverse reaction* or *lack of effect*). If an *adverse reaction* or a *lack of effect* is due to *off-label use*, we highlight the discrepancy between the originally intended indication for each drug and the *indication* reported by the patient as additional explication.

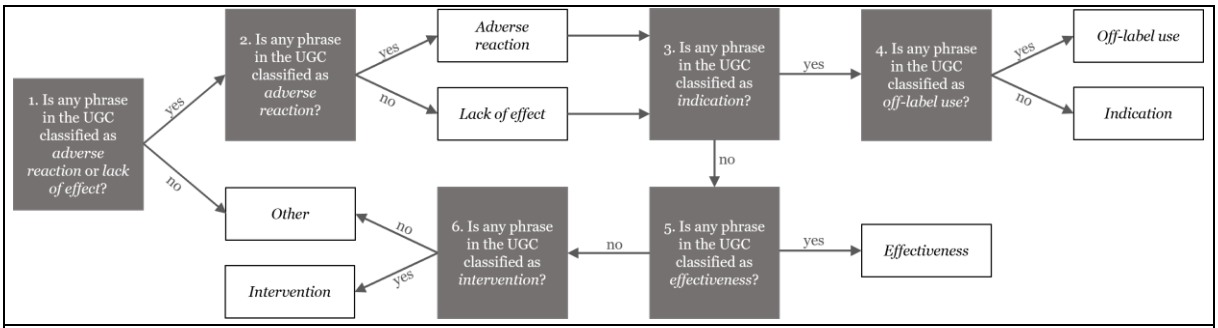


**Figure 4. Process of the Overall Classification**

Supporting this explication of the classification together with the medical term that was extracted in the first phase, not only enhances the transparency of our approach. It also provides pharmaceutical companies and regulatory authorities with a comprehensive understanding of the adverse drug experience. By pinpointing the exact phrases that led to a particular classification, they can more accurately assess the validity and relevance of the extracted information. This ultimately leads to the final output of our proposed approach, contributing to improved drug safety and more efficient monitoring of potential adverse reactions or off-label uses. An example of this output is shown in Figure 5.


**Figure 5. Example Output**

## Demonstration and Evaluation

To demonstrate the applicability of DrugExBERT, we first instantiate the approach using a real-world dataset from amazon.com consisting of drug OCR. To evaluate DrugExBERT, we first compare our results with competing approaches from the literature (E1). Second, we compare its performance with the performance of ChatGPT on the amazon.com dataset (E2). Finally, we investigate the performance of DrugExBERT on a new, unseen dataset. This dataset is a real-world pharmacovigilance dataset from the pharmacovigilance department of a leading pharmaceutical company. It contains not only UGC on drugs other than those on which DrugExBERT has been trained but also other types of UGC. This allows us to demonstrate the data- and drug-agnosticism of our approach and its transferability to the wide variety of existing drugs and UGC.

### *Demonstration*

To demonstrate DrugExBERT, we instantiate our approach on a real-world dataset consisting of OCR from amazon.com. We first crawled OCR for various drugs across a wide range of therapeutic areas. This results in a collection of 1,695 OCR. Details of the dataset and the therapeutic areas are shown in Table 4. We pre-processed the data by checking for spelling mistakes, converting to lower case and removing special characters such as emoticons, as these can make it difficult to accurately extract and classify aspects (Laboreiro et al. 2010). Stop words were not removed as this could lead to incorrect syntactic sentence analysis. As described in the previous section, we used MetaMap for aspect extraction and tagging and performed phrase separation on the OCR. As described above, DrugExBERT is based on the already pre-trained RoBERTa. Accordingly, the OCR of our dataset are only used for fine-tuning and context specification of DrugExBERT.

| **Therapeutic Area** | **# OCR** | **# Phrases** |
|---|---|---|
| Allergies | 188 | 614 |
| Anxiety and Stress | 353 | 1382 |
| Flu-like Symptoms | 199 | 617 |
| Dietary Supplements | 17 | 98 |
| Heartburn | 152 | 458 |
| Insomnia | 110 | 419 |
| Emergency Contraception | 97 | 313 |
| Nasal Spray | 283 | 926 |
| Nausea | 100 | 406 |
| Pain Management | 196 | 518 |
| **Sum** | **1,695** | **5,751** |

**Table 4. Content and Size of the amazon.com Dataset**

To label these phrases, we used Amazon SageMaker Ground Truth with MTurk masters to ensure a high degree of reliability in the labeling process (Lovett et al. 2018). The human coders were tasked with

assigning one of the six possible labels to each phrase. The class *off-label use* is not explicitly labeled by the human coders as it requires further knowledge about the drug and is derived from the phrases labeled as an *indication*. To assist the human coders in this process, they were given access to both the label definitions (see Table 1) and relevant examples. This reference material remained accessible throughout the labeling process to provide guidance when needed. For each labeling task, the human coders were provided with three key pieces of information to facilitate accurate labeling: the drug category (e.g., "antihistamine, allergy relief"), the entire OCR for context, and the specific phrase that needs to be labeled. It was important for the human coders to understand that they should focus only on the given phrase and not on the entire OCR. This means, that each phrase of one OCR gets its own label from the given fixed set of possible labels. To ensure the quality of the labeling process, we measured the quality of the annotations by comparing them to "gold standard" (i.e., expert-labeled) labels on the same data (Snow et al. 2008). We achieved this gold standard by having a subset of the phrase set labeled by pharmacovigilance experts. These experts work in the pharmacovigilance department of a leading global pharmaceutical company, which is working with us to evaluate our approach.

To ensure intercoder reliability, maintain quality control, and manage discrepancies, each phrase was labeled by two human coders. In cases of disagreement between the two coders, a third human coder was consulted. The majority vote of the three independent labels was then used to determine the final label assigned to the phrase. This process helped to minimize the impact of subjective judgment and to maintain a higher level of consistency in the labeling process. The quality and objectivity of the labeled data were assessed using percent agreement, a measure of intercoder reliability that quantifies the degree of agreement between coders (Lombard et al. 2002). Our data achieved a percent agreement of 83.08%, indicating an almost perfect level of agreement between our human coders (Landis and Koch 1977). After the phrases are labeled by the coders, they are masked using the tags obtained from MetaMap. This dataset was then split into two distinct parts: a 70% training set and a 30% test set. This split resulted in 4,106 phrases (corresponding to 1,077 OCR) being allocated for training purposes. The training set was then used to fine-tune a pre-trained RoBERTa model to create a custom classifier specifically designed for pharmacovigilance to extract adverse drug experiences. To identify the most effective parameter settings for RoBERTa, we performed a sklearn grid search (Pedregosa et al. 2011). This technique systematically searches through a range of potential parameter configurations to find the one that gives the best performance in terms of classification accuracy. Once the optimal parameters were determined, the phrase-level classifier was applied to the ATC-code mapping process. As previously described, this classifier was applied to individual phrases and the results were aggregated at the OCR level. Thus, we obtained an overall classification of each OCR, including an explication of this classification.

## *Evaluation*

In the following, we evaluate the performance of DrugExBERT with respect to the classification of OCR using widely accepted performance metrics, including accuracy, precision, recall, and F1-measure. For each classification (i.e., *adverse reaction*, *lack of effect*, *effectiveness*, *intervention*, *indication*, and *other)*, we define true positives (TP) as the number of cases in which a label is correctly classified. The number of cases in which a label is incorrectly classified as the given label is defined as false positives (FP). The same applies to true negatives (TN) and false negatives (FN), respectively. Accuracy (A) measures the proportion of correctly classified labels, while precision (P) is the proportion of correct labels out of all instances classified as a given label. Recall (R) is the proportion of labels that are correctly identified as belonging to a given label and the F1-measure (F) represents the harmonic mean of precision and recall. Specifically for pharmacovigilance, the recall of adverse drug experiences is of particular interest because it is important to be accurate in detecting adverse drug experiences. It is better that a few cases are labeled false positives as adverse drug experiences and are investigated further by pharmacovigilance than that these important cases are missed. In other words, it is important to keep the number of false negatives as low as possible.

### E1: Comparison with Competing Approaches

We compare our approach with Na and Kyaing (2015), Cavalcanti and Prudêncio (2017), and Imani and Noferesti (2022). These authors have also proposed a pharmacovigilance approach capable of identifying *adverse reactions*, terms related to *effectiveness* (without distinguishing between *effectiveness* and *lack of effect*), *interventions,* and *indications* from OCR. Na and Kyaing (2015) used WebMD OCR of drugs for

diabetes, depression, ADHD, and slimming and sleeping pills. Cavalcanti and Prudêncio (2017) and Imani and Noferesti (2022) used Druglib.com and Drugs.com OCR of drugs for ADHD, AIDS, and anxiety. The results are presented in Table 5.[3] Based on these results, our approach significantly outperforms all the other approaches in terms of classifying *adverse reactions* and *effectiveness* together with *lack of effect*. The relatively lower performance in *intervention* and *indication* classifications can be attributed to the small size of the data subset, which precludes a comprehensive comparison in these specific classes. Despite these limitations, the results show that DrugExBERT outperforms competing approaches in the classes most relevant to the extraction and classification of drug experiences.

| | | P | R | F | A |
|---|---|---|---|---|---|
| **Adverse reaction** | DrugExBERT | 84.10% | 82.41% | 83.25% | 89.32% |
| | Na and Kyaing (2015) | 44.67% | 68.69% | 69.64% | 84.26% |
| | Cavalcanti and Prudêncio (2017) | 73.18% | 82.41% | 83.25% | 89.32% |
| | Imani and Noferesti (2022) | 70.61% | 78.04% | 75.53% | 87.88% |
| **Effectiveness and lack of effect** | DrugExBERT | 89.64% | 84.40% | 86.94% | 85.28% |
| | Na and Kyaing (2015) | 51.67% | 60.67% | 53.33% | 75.00% |
| | Cavalcanti and Prudêncio (2017) | 86.31% | 74.64% | 80.05% | 81.72% |
| | Imani and Noferesti (2022) | 76.57% | 78.06% | 77.31% | 80.90% |
| **Intervention** | DrugExBERT | 0.00% | 0.00% | | 99.03% |
| | Na and Kyaing (2015) | 66.33% | 73.00% | 69.00% | 82.00% |
| | Cavalcanti and Prudêncio (2017) | 87.56% | 93.00% | 90.20% | 98.15% |
| | Imani and Noferesti (2022) | 77.63% | 81.72% | 79.62% | 96.14% |
| **Indication** | DrugExBERT | 40.00% | 100.00% | 57.14% | 99.51% |
| | Na and Kyaing (2015) | 58.33% | 69.67% | 60.67% | 68.00% |
| | Cavalcanti and Prudêncio (2017) | 59.09% | 74.05% | 65.73% | 86.31% |
| | Imani and Noferesti (2022) | 70.85% | 69.04% | 69.94% | 86.47% |

**Table 5. Comparison with Competing Approaches**

**E2: Comparison with ChatGPT**

In the second step, we measure and compare the performance of DrugExBERT to powerful language models such as the GPT series, including ChatGPT[4] (Brown et al. 2020) on the test set crawled from amazon.com. ChatGPT has demonstrated its potential in various domains such as reasoning, text generation, human-machine interaction, and scientific research (Liu et al. 2023). Although ChatGPT was not originally developed for pharmacovigilance, its versatility makes it both relevant and appropriate to compare the performance of DrugExBERT with that of ChatGPT to evaluate its performance in the context of drug experience extraction and classification. The results are shown in Table 6 and indicate that DrugExBERT consistently outperforms ChatGPT across all classes and performance measures. Both approaches face challenges in classifying *interventions* due to the small size of this subset. However, DrugExBERT still achieves a higher accuracy than ChatGPT, demonstrating its resilience in dealing with limited data. Similarly, for *indication* classification, despite the small subset size, DrugExBERT manages to achieve a

*[3] Note, that we were limited to evaluating the performance of their approaches on their test set and ours on our test set. As we do not have access to both the full code and the original dataset, we cannot make a detailed comparison between the approaches. For this reason, we consider the respective performance measures calculated from the confusion matrices, presented in their studies.*

*[4] For the evaluation we used the openai API. We tested several engines, of which GPT-3.5 was the best. In this paper, we will refer to it as "ChatGPT".*

remarkable recall and F1-measure, further solidifying its effectiveness in extracting valuable insights. For *off-label use,* only one case occurred in the whole dataset. Both DrugExBERT and ChatGPT were able to detect this *off-label use*, with ChatGPT having one additional OCR labeled as *off-label use*. In summary, DrugExBERT consistently outperforms ChatGPT in terms of precision, recall, F1-measure, and accuracy. This highlights its superiority in the area of drug experience extraction and classification.

| | | P | R | F | A | TP | FP | TN | FN |
|---|---|---|---|---|---|---|---|---|---|
| **Adverse reaction** | DrugExBERT | 84.10% | 82.41% | 83.25% | 89.32% | 164 | 31 | 388 | 35 |
| | ChatGPT | 81.12% | 79.90% | 80.51% | 87.54% | 159 | 37 | 382 | 40 |
| **Lack of effect** | DrugExBERT | 83.42% | 79.59% | 81.46% | 88.51% | 156 | 31 | 391 | 40 |
| | ChatGPT | 82.05% | 65.31% | 72.73% | 84.47% | 128 | 28 | 394 | 68 |
| **Effectiveness** | DrugExBERT | 84.77% | 78.53% | 81.53% | 90.61% | 128 | 23 | 432 | 35 |
| | ChatGPT | 74.26% | 61.96% | 67.56% | 84.30% | 101 | 35 | 420 | 62 |
| **Intervention** | DrugExBERT | 0.00% | 0.00% | | 99.03% | 0 | 4 | 612 | 2 |
| | ChatGPT | 0.00% | 0.00% | | 97.90% | 0 | 11 | 605 | 2 |
| **Indication** | DrugExBERT | 40.00% | 100.00% | 57.14% | 99.51% | 2 | 3 | 613 | 0 |
| | ChatGPT | 0.00% | 0.00% | | 98.87% | 0 | 5 | 611 | 2 |
| **Off-label use** | DrugExBERT | 100.00% | 100.00% | 100.00% | 100.00% | 1 | 0 | 617 | 0 |
| | ChatGPT | 50.00% | 100.00% | 66.67% | 99.84% | 1 | 1 | 616 | 0 |
| **Other** | DrugExBERT | 56.58% | 76.79% | 65.15% | 92.56% | 43 | 33 | 529 | 13 |
| | ChatGPT | 37.72% | 76.79% | 50.59% | 86.41% | 43 | 71 | 491 | 13 |

**Table 6. Comparison with ChatGPT on the amazon.com Dataset**

### E3: Transferability to New Datasets and Drugs

As often noted in the literature and experienced in various contexts (Christen 2007), high-quality labeled training data is often scarce. Therefore, it is crucial for an approach like ours to be transferable between different application settings, especially for different drugs and other UGC. This data-agnosticism and drug-agnosticism allow DrugExBERT, once trained on one dataset, to identify adverse drug experiences in other datasets without the need for additional labeling. For evaluation, we use a real-world pharmacovigilance dataset from a leading pharmaceutical company. Using this dataset ensures that the true labels are reliable and verifies DrugExBERT's ability to perform pharmacovigilance tasks effectively. This dataset includes not only OCR of various drugs but also excerpts from customer Q&A on amazon.com. It also covers therapeutic areas different from our training set, such as drugs for dementia syndromes and herpes creams. The results of the comparison of the performance of DrugExBERT and ChatGPT on this dataset are shown in Table 7. DrugExBERT outperforms ChatGPT on this previously unseen dataset for *adverse reactions*. In the *lack of effect* class, ChatGPT shows better precision, F1-measure, and accuracy.

| | | P | R | F | A | TP | FP | TN | FN |
|---|---|---|---|---|---|---|---|---|---|
| **Adverse reaction** | DrugExBERT | 50.00% | 64.29% | 56.25% | 96.28% | 9 | 9 | 353 | 5 |
| | ChatGPT | 34,78% | 57.14% | 43.24% | 94.41% | 8 | 15 | 347 | 6 |
| **Lack of effect** | DrugExBERT | 39.13% | 81.82% | 52.94% | 95.74% | 9 | 14 | 351 | 2 |
| | ChatGPT | 100.00% | 63.64% | 77.78% | 98.94% | 7 | 0 | 365 | 4 |
| **Other** | DrugExBERT | 98.51% | 94.02% | 96.21% | 93.09% | 330 | 5 | 20 | 21 |
| | ChatGPT | 97,09% | 95.16% | 96.12% | 92.82% | 334 | 10 | 15 | 17 |

**Table 7. Transferability to Pharmacovigilance Dataset**

However, DrugExBERT achieves a higher number of true positives and a lower number of false negatives, resulting in a significantly better recall, which is the critical performance metric in pharmacovigilance. For the *other* class, the focus shifts to the false positives, since adverse drug experiences, in this case, the negatives, are more important for pharmacovigilance. In this category, DrugExBERT also shows superior performance, with a higher number of true negatives and a lower number of false positives. In conclusion, this evaluation demonstrates the successful transferability of DrugExBERT to a previously unseen dataset, while achieving commendable performance on a real pharmacovigilance dataset. This validates the effectiveness of DrugExBERT in addressing practical pharmacovigilance tasks and supports its data-agnosticism and drug-agnosticism.

# Conclusion and Discussion

In this paper, we present DrugExBERT, a novel approach for extracting drug experiences from UGC. UGC serves as a valuable source of patient-generated information about drug experiences, providing a large amount of data for pharmacovigilance. DrugExBERT combines state-of-the-art methods for NLP with leading concepts of the UMLS Metathesaurus from biomedical texts, MetaMap. As a result, it successfully meets the needs of pharmaceutical companies by not only identifying adverse drug experiences (i.e., *adverse reactions*, *lack of effect*). It also extracts additional supportive drug experiences such as *intervention*, *indication*, and *off-label use*. In addition, DrugExBERT considers the *effectiveness* of drugs and provides an *explication* in the form of relevant UGC phrases describing the adverse drug experience. We have demonstrated that DrugExBERT performs well on our datasets from amazon.com and have shown its drug-agnosticism and data-agnosticism by transferring it to other datasets, including other types of UGC and drugs. Further, we performed an evaluation on a real-world dataset from a leading pharmaceutical company, labeled by pharmacovigilance experts, and achieved excellent results. In our comparisons, DrugExBERT demonstrates its superior performance and consistently outperforms ChatGPT and other state-of-the-art competing approaches. A key strength of DrugExBERT is its tendency to identify and classify a higher number of UGC as adverse drug experiences while minimizing type 2 errors. This is particularly important for pharmacovigilance, as it ensures that potentially adverse drug experiences are not overlooked. By using DrugExBERT, researchers, and practitioners can harness the wealth of information in UGC to improve drug safety monitoring and contribute to better patient outcomes.

## *Novelty of DrugExBERT*

In this section, we discuss how DrugExBERT contributes to the field of pharmacovigilance and NLP in the context of UGC. Taking advantage of the best of both approaches, DrugExBERT combines RoBERTa's suitability for UGC and MetaMap's domain-specific knowledge. The use of RoBERTa demonstrates DrugExBERT's adaptability and effectiveness in a wide range of applications, including those with specialized contexts. Unlike other approaches, such as Imani and Noferesti (2022), DrugExBERT uses MetaMap directly for drug experience extraction rather than drug experience classification, demonstrating its usefulness in identifying relevant concepts from UGC. By incorporating MetaMap into our approach, we are able to improve the precision and recall of our approach, leading to a better aspect extraction in the context of drug experiences. Masking is a key component of our approach, allowing it to learn sentence syntax without relying on specific expressions. This is particularly important in the medical domain, as it allows us to distinguish between different contexts in which the same term might be used (e.g., "headache" as an *adverse reaction* or *indication*). Masking increases the ability of the approach to generalize and adapt to different situations, ultimately improving its performance. The masking technique used in DrugExBERT also contributes to its transferability to other drugs and datasets. By focusing on sentence structures rather than specific expressions, our approach can effectively process and classify information from different sources, making it a valuable tool in the field of pharmacovigilance. Consistent with Gregor and Hevner's (2013) design science research knowledge contribution framework, we consider our work to be an "improvement". With DrugExBERT, we have developed a new solution to a known problem, while simultaneously improving and extending competing solutions to a broader problem space. In this sense, DrugExBERT not only extends the problem space of existing approaches that either do not provide a sufficient level of detail to meet regulatory obligations (e.g., Na and Kyaing 2015; Cavalcanti and Prudêncio 2017; Imani and Noferesti 2022) or fail to detect or isolate relevant supportive drug experiences and to provide explications (e.g., Gräßer et al. 2018; Ajibade et al. 2022; Unnikrishnan et al. 2023; Cavalcanti and Prudêncio 2017; Imani and Noferesti 2022; Na and Kyaing 2015). It also outperforms competing

approaches (e.g., Na and Kyaing 2015; Cavalcanti and Prudêncio 2017; Imani and Noferesti 2022) and ChatGPT with its unique combination of powerful medical systems with NLP techniques and masking tailored for its application to UGC. In summary, DrugExBERT highlights the benefits of using RoBERTa, MetaMap, and masking techniques in the context of drug experience extraction and classification. These components contribute to its superior performance, transferability, and adaptability.

### *Contribution to Practice*

DrugExBERT contributes to pharmaceutical companies by improving their pharmacovigilance systems. First, it automates the extraction and classification of relevant information from UGC, significantly reducing the manual labor required for pharmacovigilance tasks. This boosts efficiency and productivity as pharmacovigilance experts can focus on more critical components of their work. In terms of performance, ChatGPT is the closest to our approach, but it is ethically questionable to use in sensitive contexts such as adverse drug experiences. For this reason, it is important to use DrugExBERT as a reliable alternative for this purpose. Second, DrugExBERT can be employed by pharmaceutical companies, healthcare organizations, and regulatory authorities to monitor and analyze UGC related to drugs in the post-marketing setting. This capability facilitates the detection of potential adverse drug experiences, ultimately improving drug safety and enabling swift intervention when necessary. Through this, also previously unknown adverse reactions can be detected. Third, DrugExBERT enhances "reconciliation", the process of cross-checking data to be reported. In practice, however, only a fraction of this data undergoes this check. Thus, DrugExBERT can enable a more comprehensive reconciliation, further enhancing drug safety. Finally, DrugExBERT offers insights to better understand patient experiences, preferences, and concerns related to specific drugs. This can improve the development of personalized medical strategies and support pharmaceutical companies to provide better-informed treatments for their patients. By incorporating patients' real-world experiences, DrugExBERT bridges the gap between clinical data and individual patient needs, ultimately contributing to improved patient outcomes. In summary, DrugExBERT goes beyond its immediate improvement of pharmacovigilance systems by automating extraction and classification of drug-related information from UGC. Its potential to enhance drug safety monitoring and personalized medical strategies makes it a valuable asset for advancing healthcare practice and patient relationship management.

### *Limitations and Future Research*

DrugExBERT has limitations that need to be further investigated in future research. First, DrugExBERT was designed for its application on UGC. While we are confident that it can be applied to other types of content that include drug experiences (e.g., patient interviews, physician interviews, blogs, and literature), further research is needed to confirm its transferability to this content. In addition, it has been trained and tested primarily on over-the-counter products (i.e., products that don't require a prescription). Given its demonstrated transferability, we are confident that DrugExBERT can be applied to prescription drugs as well. However, further research is needed to validate its performance in this context. Furthermore, DrugExBERT has only been evaluated against competing approaches using their own reported results. Further evaluation, including a comparison with competing approaches on the same dataset, should be performed. In addition, the validation robustness of DrugExBERT could be further improved by increasing the dataset size. Although DrugExBERT is designed to detect adverse reactions, its current scope does not include distinguishing between serious and non-serious adverse reactions. The inclusion of such a severity classification in future research could be beneficial as it would allow pharmaceutical companies to prioritize investigations based on urgency (FDA 2001). DrugExBERT's ability to learn about adverse reactions could also be extended by future research to distinguish whether an adverse reaction is expected or unexpected. Identifying unexpected adverse reactions would increase the usefulness of the approach in detecting potential safety concerns and contribute to a more comprehensive understanding of drug safety profiles. In summary, while DrugExBERT represents a promising approach to extracting drug experiences from UGC, some limitations warrant further investigation to enhance its applicability and utility in pharmacovigilance.

## References

Adams, D. Z., Gruss, R., and Abrahams, A. S. 2017. "Automated discovery of safety and efficacy concerns for joint & muscle pain relief treatments from online reviews," *International Journal of Medical Informatics* (100), pp. 108-120.

Ajibade, S.-S. M., Zaidi, A., Tapales, C. P., Ngo-Hoang, D. L., Ayaz, M., Dayupay, J. P., Dodo. Y. A., Chaudhury, S., and Adediran, A. O. 2022. "Data Mining Analysis of Online Drug Reviews," in *2022 IEEE 10th Conference on Systems, Process & Control*, Malacca, Malaysia.

Aronson, A. R. 2001. "Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program," *Proceedings of the AMIA Symposium*, pp. 17-21.

Aronson, A. R., and Lang, F.-M. 2010. "An overview of MetaMap: historical perspective and recent advances," *Journal of the American Medical Informatics Association* (17:3), pp. 229-236.

Bodenreider, O. 2004. "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Research* (32), pp. D267-D270.

Borchert, J. S., Wang, B., Ramzanali, M., Stein, A. B., Malaiyandi, L. M., and Dineley, K. E. 2019. "Adverse Events Due to Insomnia Drugs Reported in a Regulatory Database and Online Patient Reviews: Comparative Study," *Journal of Medical Internet Research* (21:11), e13371.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. 2020. "Language Models are Few-Shot Learners," in *34th Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 523-531.

Cavalcanti, D., and Prudêncio, R. 2017. "Aspect-Based Opinion Mining in Drug Reviews," in *18th EPIA Conference on Artificial Intelligence*, Porto, Portugal, pp. 815-827.

Cheng, J., Dong, L., and Lapata, M. L. 2016. "Long Short-Term Memory-Networks for Machine Reading," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language*, Austin, TX, pp. 551-561.

Christen, P. 2007. "A two-step classification approach to unsupervised record linkage," in *Proceedings of the sixth Australasian conference on Data mining and analytics-Volume*, Gold Coast, Australia, pp. 111-119.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, pp. 4171-4186.

Dreisbach, C., Koleck, T. A., Bourne, P. E., and Bakken, S. 2019. "A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data," *International Journal of Medical Informatics* (125), pp. 37-46.

European Medicines Agency. 2017. *Guideline on good pharmacovigilance practices (GVP): Module VI - Collection, management and submission of reports of suspected adverse reactions to medicinal products (Rev 2)*. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-good-pharmacovigilance-practices-gvp-module-vi-collection-management-submission-reports_en.pdf. Accessed April 24, 2023.

FDA. 2001. *Guidance for Industry: Postmarketing Safety Reporting for Human Drug and Biological Products Including Vaccines*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/postmarketing-safety-reporting-human-drug-and-biological-products-including-vaccines. Accessed April 24, 2023.

FDA. 2018. Preventable Adverse Drug Reactions: A Focus on Drug Interactions. https://www.fda.gov/drugs/drug-interactions-labeling/preventable-adverse-drug-reactions-focus-drug-interactions. Accessed April 10, 2023.

FDA. 2023. *Title 21 of the Code of Federal Regulations (CFR): PART 314-APPLICATIONS FOR FDA APPROVAL TO MARKET A NEW DRUG*. https://www.govinfo.gov/content/pkg/CFR-2022-title21-vol5/pdf/CFR-2022-title21-vol5-part314.pdf. Accessed April 24, 2023.

Goh, K.-Y., Heng, C.-S., and Lin, Z. 2013. "Social Media Brand Community and Consumer Behavior: Quantifying the Relative Impact of User- and Marketer-Generated Content," *Information Systems Research* (24:1), pp. 88-107.

Gosal, G. P. S. 2015. "Opinion Mining and Sentiment Analysis of Online Drug Reviews as a Pharmacovigilance Technique," *International Journal on Recent and Innovation Trends in Computing and Communication* (3:7), pp. 4920-4925.

Gregor, S., Hevner, A. R. 2013. "Positioning and Presenting Design Science Research for Maximum Impact," *MIS Quarterly* (37:2), pp. 337-355.

Gräßer, F., Kallumadi, S., Malberg, H., and Zaunseder, S. 2018. "Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning," *Proceedings of the 2018 International Conference on Digital Health*, Lyon, France, pp. 121-125.

Gu, X., Gu, Y., and Wu, H. 2017. "Cascaded Convolutional Neural Networks for Aspect-Based Opinion Summary," *Neural Processing Letters* (46), pp. 581-594.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75-105.

Honnibal, M., Montani, I., Van Landeghem, S, Boyd, A. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. https://github.com/explosion/spaCy. Accessed February 15, 2023.

Huang, K., Altosaar, J., and Ranganath, R. 2019. "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," *arXiv :1904.05342*.

Imani, M., and Noferesti, S. 2022. "Aspect extraction and classification for sentiment analysis in drug reviews," *Journal of Intelligent Information Systems* (59), pp. 613-633.

Johnson, A. J., and Bootman, J. L. 1997. "Drug-related morbidity and mortality and the economic impact of pharmaceutical care," *American Journal of Health-System Pharmacy* (54:5), pp. 554-558.

Kaas-Hansen, B. S., Gentile, S., Caioli, A., and Andersen, S. E. 2022. "Exploratory pharmacovigilance with machine learning in big patient data: a focused scoping review," *Basic & Clinical Pharmacology & Toxicology* (132:3), pp. 233-241.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. 2014. "A Convolutional Neural Network for Modelling Sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, pp. 655-665.

Laboreiro, G., Sarmento, L., Teixeira, J., and Oliveira, E. 2010. "Tokenizing micro-blogging messages using a text classification approach," in *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, Toronto, Canada, pp. 81-88.

Landis, J. R., and Koch, G. G. 1977. "The Measurement of Observer Agreement for Categorical Data," *Biometrics* (33:1), pp. 159-174.

Lazarou, J., Pomeranz, B. H., and Corey, P. N. 1998. "Incidence of Adverse Drug Reactions in Hospitalized Patients: A Meta-analysis of Prospective Studies," *JAMA* (279:15), pp. 1200-1205.

Leaman, R., Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. "Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp. 117-125.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. 2019. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics* (36:4), pp. 1234-1240.

Li, X., and Lam, W. 2017. "Deep Multi-Task Learning for Aspect Term Extraction with Memory Interaction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 2886-2892.

Liu, D., Li, Y., and Thomas, M. A. 2017. "A Roadmap for Natural Language Processing Research in Information Systems," *Proceedings of the 50th Hawaii International Conference on System Sciences*, Honolulu, HI, pp. 1112-1121.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., and Ge, B. 2023. "Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models," *arXiv:2304.01852*.

Liu, Y., Ott, M., Goyal, N., Du Jingfei, Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv:1907.11692*.

Locke, S., Bashall, A., Al-Adely, S., Moore, J., Wilson, A., and Kitchen, G. B. 2021. "Natural language processing in medicine: A review," *Trends in Anaesthesia and Critical Care* (38), pp. 4-9.

Lombard, M., Snyder-Duch, J., and Bracken, C. C. 2002. "Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability," *Human Communication Research* (28:4), pp. 587-604.

Lovett, M., Bajaba, S., Lovett, M., and Simmering, M. J. 2018. "Data Quality from Crowdsourced Surveys: A Mixed Method Inquiry into Perceptions of Amazon's Mechanical Turk Masters," *Applied Psychology* (67:2), pp. 339-366.

Ma, Y., Peng, H., Khan, T., Cambria, E., and Hussain, A. 2018. "Sentic LSTM: A Hybrid network for targeted aspect-based sentiment analysis," *Cognitive Computation* (10:4), pp. 639–650.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. 2022. "Deep Learning-based Text Classification: A Comprehensive Review," *ACM Computing Surveys* (54:3), pp. 1-40.

Na, J.-C., and Kyaing, W. Y. M. 2015. "Sentiment Analysis of User-Generated Content on Drug Review Websites*," Journal of Information Science Theory and Practice* (3:1), pp. 6-23.

Nahler, G. 2009. *Dictionary of Pharmaceutical Medicine*, 2nd ed. Vienna: Springer Vienna.

Nazir, A., Rao, Y., Wu, L., and Sun, L. 2022. "Issues and Challenges of Aspect-based Sentiment Analysis: A Comprehensive Survey," *IEEE Transactions on Affective Computing* (13:2), pp. 845-863.

Nikfarjam, A., and Gonzalez, G. H. 2011. "Pattern Mining for Extraction of Mentions of Adverse Drug Reactions from User Comments," in *AMIA Annual Symposium Proceedings*, pp. 1019-1026.

O'Connor, K., Pimpalkhute, P., Nikfarjam, A., Ginn, R., Smith, K. L., and Gonzalez, G. 2014. "Pharmacovigilance on Twitter? Mining Tweets for Adverse Drug Reactions," in *AMIA Annual Symposium Proceedings*, pp. 924-933.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M. 2011. "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research* (12), pp. 2825-2830.

Pilipiec, P., Liwicki, M., and Bota, A. 2022. "Using Machine Learning for Pharmacovigilance: A Systematic Review," *Pharmaceutics* (14:2).

Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., and Gonzalez, G. 2015. "Utilizing social media data for pharmacovigilance: a review," *Journal of Biomedical Informatics* (54), pp. 202-212.

Schouten, K., and Frasincar, F. 2016. "Survey on Aspect-Level Sentiment Analysis," *IEEE Transactions on Knowledge and Data Engineering* (28:3), pp. 813-830.

Sharif, H., Zaffar, F., Abbasi, A., and Zimbra, D. 2014. "Detecting adverse drug reactions using a sentiment classification framework," in *Proceedings of the ASE/IEEE International Conference on Social Computing*, Stanford, CA.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. 2008. "Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, HI, pp. 254-263.

Tubishat, M., Idris, N., and Abushariah, M. A. 2018. "Implicit aspect extraction in sentiment analysis: Review, taxonomy, opportunities, and open challenges," *Information Processing & Management* (54:4), pp. 545-563.

Unnikrishnan, R., Kamath, S., and Ananthanarayana, V. S. 2023. "Efficient parameter tuning of neural foundation models for drug perspective prediction from unstructured socio-medical data," *Engineering Applications of Artificial Intelligence* (123), 106214.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. 2017. "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems*, Long Beach, CA, pp. 5998-6008.

Xia, L., Wang, G. A., and Fan, W. 2017. "A Deep Learning Based Named Entity Recognition Approach for Adverse Drug Events Identification and Extraction in Health Social Media," in *International Conference on Smart Health,* Hong Kong, China, pp. 237-248.

Yan, Z., Xing, M., Zhang, D., and Ma, B. 2015. "EXPRS: An extended pagerank method for product feature extraction from online consumer reviews," *Information & Management* (52:7), pp. 850-858.

Yang, M., Wang, X., and Kiang, M. 2013. "Identification of Consumer Adverse Drug Reaction Messages on Social Media," in *Proceedings of the Pacific Asia Conference on Information* Systems, Jeju Island, Korea, 193.

Zhu, D. H., Ye, Z. Q., and Chang, Y. P. 2017. "Understanding the textual content of online customer reviews in B2C websites: A cross-cultural comparison between the U.S. and China," *Computers in Human Behavior* (76), pp. 483-493.

Zhu, X., Sobihani, P., and Guo, H. 2015. "Long short-term memory over recursive structures," in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, pp. 1604-1612.

Züllig, K., Erlebach, S., Kupfer, A., and Zimmermann, S. 2023. "Bargain Hunting on Black Friday–Making Great Deals and Bragging About Them," in *Proceedings of the 56th Hawaii International Conference on System Sciences*, Honolulu, HI, pp. 3952-3961.