



Adverse drug event detection and extraction from open data: A deep learning approach

Brandon Fan^a, Weiguo Fan^{b,*}, Carly Smith^c, Harold “Skip” Garner^{d,e}

^a Blacksburg High School, Blacksburg, VA

^b Tippie College of Business, University of Iowa, Iowa City, IA

^c Stanford University, Stanford, VA

^d Center for Bioinformatics and Genetics, Edward Via College of Osteopathic Medicine, Blacksburg, VA

^e Gibbs Cancer Center and Research Institute, Spartanburg, SC

ARTICLE INFO

Keywords:

Information extraction
Deep learning
Pharmacovigilance
Drug side effects
Open data
BERT
Natural language processing

ABSTRACT

Drug prescription is a task that doctors face daily with each patient. However, when prescribing drugs, doctors must be conscious of all potential drug side effects. In fact, according to the U.S. Department of Health and Human Services, adverse drug events (ADEs), or harmful side effects, account for 1/3 of total hospital admissions each year. The goal of this research is to utilize novel deep learning methods for accurate detection and identification of professionally unreported drug side effects using widely available public data (open data). Utilizing a manually-labelled dataset of 10,000 reviews gathered from WebMD and Drugs.com, this research proposes a deep learning-based approach utilizing Bidirectional Encoder Representations from Transformers (BERT) based models for ADE detection and extraction and compares results to standard deep learning models and current state-of-the-art extraction models. By utilizing a hybrid of transfer learning from pre-trained BERT representations and sentence embeddings, the proposed model achieves an AUC score of 0.94 for ADE detection and an F1 score of 0.97 for ADE extraction. Previous state of the art deep learning approach achieves an AUC of 0.85 in ADE detection and an F1 of 0.82 in ADE extraction on our dataset of review texts. The results show that a BERT-based model achieves new state-of-the-art results on both the ADE detection and extraction task. This approach can be applied to multiple healthcare and information extraction tasks and used to help solve the problem that doctors face when prescribing drugs. Overall, this research introduces a novel dataset utilizing social media health forum data and shows the viability and capability of using deep learning techniques in ADE detection and extraction as well as information extraction as a whole. The model proposed in this paper achieves state-of-the-art results and can be applied to multiple other healthcare and information extraction tasks including medical entity extraction and entity recognition.

1. Introduction

The global presence of chronic diseases is growing rapidly. In the United States, around 140 million Americans, half of the total U.S. population, suffer from one or more chronic medical conditions (Lin, Chen, Brown, Li & Yang, 2017). According to the World Health Organization (WHO), people with chronic illnesses such as high blood pressure, diabetes, heart disease, stroke, cancer,

* Corresponding author.

E-mail addresses: brandonfan1256@gmail.com (B. Fan), weiguo-fan@uiowa.edu (W. Fan).

<https://doi.org/10.1016/j.ipm.2019.102131>

Received 12 August 2019; Received in revised form 19 September 2019; Accepted 20 September 2019

Available online 22 October 2019

0306-4573/ © 2019 Elsevier Ltd. All rights reserved.

coronary artery disease, and HIV rely on long-term medications to function daily (Brown & Bussel, 2011; Kisa, 2003). As a result, it's imperative that doctors are aware of all potential drug side effects during every day decision-making (Aron, Dutta, Janakiraman & Pathak, 2011).

As a field, pharmacovigilance is defined as the analysis of drug effects after a drug has been put to market (World Health Organization, 2002). Pharmacovigilance wishes to analyze the problems and any side effects a drug may cause. Adverse drug events, or ADEs, are defined as any injury or side effect caused by taking a specific, drug-related medical treatment. Current statistics show that ADEs constitute over 3.5 million physician visits, 1 million emergency department visits, and more than 2 million injuries, hospitalizations and deaths (Liu & Chen, 2015). Along with a significant impact on health, ADEs also have a widespread economic impact, incurring costs of over \$75 Billion (Harpaz, DuMouchel, Ryan & Friedman, 2012). ADEs are also considered the most common and preventable medical errors (Seeger, Jha & Bates, 2007). Current governmental systems in place include MedWatch and mandatory reporting to the FDA's Adverse Event Reporting System (FAERS) database (Sarker et al., 2015). Though these systems are in place, they are far from perfect and often fail to record all adverse reactions (Sarker et al., 2015). According to the U.S. Department of Health and Human Services, ADEs cause relatively 1/3 of all total hospital visits, which corresponds to 45 million hospitalizations per year in the United States alone. This problem roots from the fact that ADEs are not being taken into proper consideration during prescription (Kohli & Kettinger, 2004). It is known that drug side effects can be classified into two primary types: reported and unreported. Reported side effects are adverse reactions listed by manufacturers on drug labels and reported to the Federal Drug Administration (FDA). Unreported drug side effects are ADEs that occur but are not reported by the manufacturer, subsequently unknown to doctors, which has indirectly caused the current increase in hospitalizations. Due to this problem, pharmacovigilance research has popularized to diligently and accurately detect and extract drug side effects for drug professionals and manufacturers alike.

We view ADE extraction from text as an information extraction problem (Papagiannopoulou & Tsomakas, 2018; Peng & McCallum, 2006; Yuan & Yu, 2019; Zhang, Boons & Batista-Navarro, 2019), which has witnessed significant growth in recent years due to the advancement of AI and machine learning, especially deep learning (LeCun, Bengio & Hinton, 2015; Mikolov, Chen, Corrado, & Dean, 2013; Peters et al., 2018; Rokach, 2010). Despite the growth in and potential of deep learning, there has been little use of it for ADE detection and extraction. Additionally, current extraction methods are far from optimal and oftentimes inaccurate, infeasible, and ungeneralizable. In this paper, we design a novel approach to solve the challenging problem of automated ADE extraction from online open data. Our research is primarily driven by the increased usage of online social media for ADE detection and extraction (Cocos, Fiks & Masino, 2017; Harpaz et al., 2012; Korkontzelos et al., 2016; Liu & Chen, 2015). As mentioned in Harpaz et al., online media provides new venues for pharmacovigilance due to the consistent presence of large amounts of recent and unaltered patient-data (Harpaz et al., 2012). This paper investigates the use of the latest developments in deep learning as accurate models for ADE extraction compared to current approaches.

We make several major contributions to the pharmacovigilance research and information extraction from open data literature:

- 1) We propose the use of a new social media dataset utilizing health social media forums instead of conventional social media datasets such as twitter and medical case reports. We tasked domain experts to manually tag 10,000 reviews from two major data sources: WebMD and Drugs.com, data sources that have not been utilized in previous pharmacovigilance research (extant research has utilized twitter data (Cocos et al., 2017; Ginn et al., 2014; Korkontzelos et al., 2016), and medical articles like Medline (Gurulingappa, Mateen-Rajpu & Toldo, 2012)). Each domain expert was tasked with labeling the presence of an ADE within a review, paying special attention to potential false positives like "I used Tylenol to treat headache", in this case headache is not an ADE.
- 2) We propose a novel deep learning approach encompassing latest development in NLP for information extraction – BERT model, along with sentence embedding, for ADE detection and extraction from online open data.
- 3) We implement a multi-faceted experiment that compares current state-of-the-art statistical models, standard deep learning models, and deep learning with novel BERT word embeddings and sentence embeddings (our proposed solution) on the task of ADE detection and extraction of unstructured reviews. Using the large-scale data set we developed, we are able to show the major advantages of our newly proposed approach. Our model's implementation allows for automation of ADE discovery from open data sources, while allowing more drug side effects to be identified for doctors. In addition, this model can be applied to other problems like entity extraction in electronic medical records (Qian et al., 2016), clinical narratives (Ferraro et al., 2013), and medical event extraction (Li, Liu, Antieau, Cao & Yu, 2010), as well as general information extraction problems including named entity recognition, and concept extraction.

The rest of our paper is structured as follows: we first discuss in Section 2 related works in the field of pharmacovigilance, current research methods including the lexicon and statistic-based approaches. We then discuss artificial intelligence and deep learning and their application in natural language processing, information extraction, and pharmacovigilance. Section 3 details the novel framework introduced in this paper for ADE detection and extraction. We discuss the methodology details including motivation, model architecture, and data collection. Section 4 provides details about the experiments implemented (ADE detection and ADE extraction). We discuss the benchmarks, datasets, and results for each task. We discuss the results, their implications for future research, and the limitations of our framework in Section 5. Section 6 concludes the paper with an overall summary of the research and the research's major highlights.

2. Related works and conceptual backgrounds

2.1. Pharmacovigilance research

Pharmacovigilance research has seen widespread growth in both the social media spectrum (Ginn et al., 2014; Harpaz et al., 2012; Korkontzelos et al., 2016; Liu & Chen, 2015), medical report spectrum (Gurulingappa et al., 2012), and electronic health records and messaging (Chee, Berlin & Schatz, 2011). Despite the datasets, current pharmacovigilance research can be divided into two primary types of methods: lexicon-based approaches and statistic-based approaches.

2.1.1. Lexicon-based approach

Standard lexicon-based approaches rely on the syntactical schematics of language. In pharmacovigilance, lexicon-based extraction methods primarily use large data sets of medical and drug-related terms. These dictionaries are standard among all lexicon-based approaches and include sources such as the Unified Medical Language System (UMLS), and the Consumer Health Vocabulary (CHV). Some research has even attempted to construct their own dictionaries, opting to create a vocabulary based on word or phrase frequencies (Adams, Gruss & Abrahams, 2017). In Adams et al., the researchers propose the use of a "smoke word" dictionary conglomerated from Amazon Reviews. The dictionary is specific towards the problems attempting to be solved: Joint and Muscle Pain Relief treatments. These records were crawled and selected for a total of 32,000 records. Adams et al. utilizes this dictionary to improve on traditional sentiment analysis algorithms to reach new state-of-art results. This research shows the efficacy of lexicon-based approaches and the validity of social media to properly identify drug side effects. Other research attempts to use feature engineering, which utilizes a combination of UMLS and CHV to create semantic feature groups that can be used to identify drug side effects (Liu & Chen, 2015). By coupling the external information from UMLS and CHV, systems are able to identify a greater number of side effects than when not using such dictionaries.

However, key limitations exist. Lexicon-based methods require constant updates therefore may fail to recognize new drugs on the market, which are typically dangerous due to their unknown side effects. Additionally, many concepts of language, such as negation, severity, and prepositional phrasing, cannot be properly captured by a lexicon. Thus, the limitations of a lexicon-approach call for improvements to solve these problems.

2.1.2. Statistical learning approach

Statistical learning approaches couple lexicon-based approaches with frequency analysis, machine learning, and probability – all of which attempt to improve the detection of ADEs. Statistical learning has the capability to add on context and frequency of words among the overall corpus to improve prediction. Such approaches utilize similar datasets as seen in the previous section (Adams et al., 2017; Liu & Chen, 2015). Thus, it is feasible for a statistical approach to properly identify and mitigate preventable ADEs. A primary study in the use of statistical learning is found in a study done by Gurulingappa et al. (2012). In this paper, the researchers utilize a support vector machine model to identify ADEs from medical case reports, specifically utilizing the Medline database. They commence a statistical approach by adopting an ontological framework and coupling the resulting data with a support vector machine method to properly classify drug side effects. The data was manually labeled by professionals. Another statistical, SVM-based approach was implemented on the twitter database (Bian, Topaloglu & Yu, 2012). This approach attempts to utilize statistical learning at a larger scale with over 2 billion tweets. Bian et al. first filtered the tweets for drug-related terms matched with an adverse reaction database and then utilized support vector machines to classify ADEs from the associated reviews. Ginn et al. utilize a similar process by crawling twitter for drug-related tweets and then manually labeling drug side effects and matching drug side effect terms to common side effect databases like SIDER and FAERS. They then implement an SVM and naïve-bayes approach for the ADE detection task (detecting the presence of drug side effects in the tweet), showing the viability of pharmacovigilance through the use of statistical learning (Ginn et al., 2014).

Though statistical learning is a considerable improvement, it still has many of the limitations as its counterpart does. Due to this, statistical models are severely limited in their scopes of knowledge and struggle with being able to adapt to new datasets and languages. Additionally, statistical methods only understand syntactical relationships among words (i.e. noun-verb relationships, entity recognition, etc.) but do not understand relative word meaning. Furthermore, as the statistical approach relies upon the initial lexicon-based approach, they are subject to the same limitations and criticisms as the lexicon approaches.

2.2. Artificial intelligence/deep learning and NLP

Modern developments in artificial intelligence has seen widespread use across multiple industries and practices. The impact of artificial intelligence, more specifically deep learning, has seen major growth and impact across many fields and industries (LeCun, Touresky, Hinton & Sejnowski, 1988). Deep learning is a new class of artificial intelligence techniques that use multiple layers of interconnected nodes to model relationships among input data in various machine learning tasks (He, Zhang, Ren & Sun, 2016; Hochreiter & Schmidhuber, 1997; LeCun et al., 2015; Mikolov, Sutskever et al., 2013; Pennington, Socher & Manning, 2014; Pereira, Pinto, Alves & Silva, 2016). Since its creation, deep learning has progressed in fields of medicine and health informatics with nodule detection models (Gruetzmacher, Gupta & Paradice, 2018), medical tagging (Xiao, Choi & Sun, 2018), and medical scan analysis. In modern research, the intersection of health informatics and natural language processing has become a rich field for the application of deep learning techniques (Cocos et al., 2017; Korkontzelos et al., 2016; Serban, Thapen, Maginnis, Hankin & Foot, 2019; Yuan & Yu, 2019). This intersection was greatly assisted by the development of word embeddings (Mikolov, Chen et al., 2013),

and later transformer representations of words (Pennington et al., 2014; Pereira et al., 2016; Vaswani et al., 2017; Yang et al., 2019), enabling greater word and language processing. Both word embeddings and transformers seek to find the most relevant representation for converting words into vectors. Due to the major advancements in deep learning as a general field and specific to natural language processing, it is a prime subject of analysis to help solve the pharmacovigilance problem.

2.2.1. Deep learning in natural language processing

Much of the limitations in extant research are a result of limited representation based on syntax and word frequencies rather than semantics. These limitations, coupled with the inability of current algorithms to understand and process data in their original representations (Bengio & LeCun, 2007), calls for new avenues to solve the pharmacovigilance problem, where deep-learning is most promising (LeCun et al., 2015). Primary natural language processing algorithms incorporate modern deep learning approaches, including recurrent neural networks (RNNs) (Shertstinsky, 2018). A more specific type of RNNs are Long-Short Term Memory Networks (LSTMs) (Hochreiter & Schmidhuber, 1997). LSTMs are units that recurrently compute nonlinear transformations on a piece of data. Equations with LSTM calculations are shown below. In this example, an input timestep x_t is inputted and goes through a series of nonlinear transformations that combines previously “memorized” information (i.e. long-term memory) to output a timestep result h_t .

LSTM Unit Calculations

$$\begin{aligned} z_t &= \sigma(W_z[h_{t-1}, x_t]) \\ r_t &= \sigma(W_r[h_{t-1}, x_t]) \\ \tilde{h}_t &= \tanh(W_{\tilde{h}}[r_t * h_{t-1}, x_t]) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \end{aligned} \quad (1)$$

For each timestep, a piece of data will be passed to the next. Thus, one can immediately recognize that the recurrent neural network architecture and the LSTM architecture can produce key proponent labels of dynamic length inputs. Additionally, bidirectional LSTMs compute an output for a current word token w_i and computes an output for the word token $w_n - 1 - i$ (Graves & Schmidhuber, 2005). These two LSTM outputs are then averaged in an element-wise fashion to produce a final output vector that is then passed onto later LSTM units.

For each timestep (i.e. each word) an output is calculated, and the entire sequence is returned. This sequence is then added in an element-wise operation with the feed forward neural network that repeats across timesteps. By applying this type of network to the pharmacovigilance task, we can capture more complex relationships between words and improve extraction.

2.2.2. Word embeddings and transformers

Word embeddings are the method of converting a set of words into an N by D-dimensional matrix, where D is the dimensional output and N is the number of words. Early research used the Word2Vec algorithm proposed by Mikolov, Chen et al. (2013). In their research, a feed forward neural network was trained to predict the next word given a specific word and its preceding context. Then, the hidden layer within the feed forward neural network is used to convert words into associated vectors. This Word2Vec algorithm was applied to train multiple large-scale word embeddings focused on large datasets such as Twitter and Wikipedia, producing pre-trained embeddings like GloVe, a set of word vectors trained on over 5 billion articles from Wikipedia and other online data sources (Pennington et al., 2014).

Despite the prevalence of pre-trained embeddings, recent research has introduced new methods of computing word embeddings: transformers. Transformers have popularized with the introduction of attention mechanisms and sequence-to-sequence algorithms (Vaswani et al., 2017). A transformer, rather than using recurrent neural networks, utilizes an extensive amount of multi-headed attention mechanisms (a method of gathering the context of a word) and feed forward neural networks combined to improve on current sequence to sequence tasks. A transformer consists of encoders and decoders, where data is first “encoded” by the encoders and then further “decoded” into the desired sentence. The primary difference between the encoder and the decoder is the use of masked multi-headed attention in the decoder steps, masked meaning that all information after a word and all similarities between words are removed so only preceding words are considered for prediction. As these transformers are trained on an array of tasks, embeddings are computed and have been shown to produce better results (Devlin, Chang, Lee & Toutanova, 2018; Vaswani et al., 2017; Yang et al., 2019). BERT is a specific type of transformer that was proposed by Google (Devlin et al., 2018). They are learned by introducing a novel mechanism to bidirectionally train transformers to ultimately produce more complex characteristics to compare word semantics. Because of the theoretical and empirical benefit of contextual embedding learned through transformers over traditional word embeddings, we employ and compare the contextual embeddings in the problems of ADE detection and extraction. We also introduce the novel use of sentence embeddings (concatenation of word embeddings) to help introduce greater context on a per-word basis during detection and extraction tasks.

2.2.3. Deep learning in information extraction

As mentioned in the previous section, deep learning has seen widespread use in the field of natural language processing. Deep learning has also been used extensively for the information extraction, concept extraction, and information retrieval tasks (Papagiannopoulou & Tsomakias, 2018; Peng & McCallum, 2006; Yuan & Yu, 2019; Zhang et al., 2019). For example, Yuan & Yu utilized a bag of words model combined with linguistic and syntactical features to uncover health claims found in news headlines (Yuan & Yu, 2019). Zhang et al. utilized a neural network to commence named entity recognition (a similar problem to ADE

extraction) and attribute extraction to discover different accounts of the same news, (therefore attributing the information to a certain person) (Zhang et al., 2019). Zhang et al.'s use of deep learning neural networks outperformed the previous state of the art by approximately 11.96%, showing the viability and potential to use deep learning for NLP tasks that include tasks like named entity recognition and ultimately extraction. Papagiannopoulou & Tsmoakas introduce the use of GloVe vectors proposed by Mikolov, Sutskever, Chen, Corrado, and Dean (2013) to identify key phrases within a document and achieving state of the art results (Papagiannopoulou & Tsmoakas, 2018). Another paper proposed by Hoang & Mothe employ NLP techniques to identify and detect geography-related terms within tweets, combining this information with other tweet metadata to extract the certain words that help identify location within tweets (Hoang & Mothe, 2018). By proposing the problem of ADE detection and extraction that is similar to the problem of named entity recognition or concept extraction, namely extracting certain words of relevance or classifying words into various categories, we can promptly utilize deep learning techniques to solve the problem of pharmacovigilance.

2.2.4. Deep learning in pharmacovigilance

Despite the prevalence of deep learning, it has hardly been applied to pharmacovigilance and ADE extraction. A primary study in the use of deep learning for pharmacovigilance was published in 2017 (Cocos et al., 2017). Cocos et al. utilize a bidirectional LSTM trained upon a sample twitter dataset of 844 tweets to identify drug side effects, achieving a final F1 score of 0.77. In addition, preliminary research in the use of sentiment analysis and its effects in ADE extraction from tweets and forum posts was published in 2016 (Korkontzelos et al., 2016). Here, Korkontzelos et al. utilize a novel sentiment algorithm combined with an SVM to improve upon ADE detection and extraction in twitter data (Korkontzelos et al., 2016). However, no work has attempted to specifically analyze health social media forum data like WebMD and Drugs.com at a large scale or utilized modern natural language processing techniques like transformers (including BERT). In addition, previous works using social media often use very limited amounts of data, reaching only about 500–1000 datapoints, significantly reducing the generalizability of the model. Due to the limited scope of data and model simplicity, extant research shows limited representation of the theoretical benefit of deep learning. Therefore, the motivation behind our research is to improve upon the current model approaches by incorporating novel deep learning-based features, and to utilize a novel dataset incorporating health social media forum data from WebMD and Drugs.com to ultimately improve upon the current state-of-the-art in ADE detection and extraction.

A summarization of the literature review in Tables 1. Works selected for this literature review primarily pertain to the most recent important developments in ADE extraction and detection and were therefore selected as part of the review.

Table 1
Literature review in pharmacovigilance.

| Paper name | Dataset used | Primary approach | Features & contribution |
|----------------------------|------------------------------|--------------------------------------------------|-----------------------------------------------------------------------------------------------------------|
| Adams et al. (2017) | Twitter, CHV, FAERS, Medline | Lexicon-Based Approach | Created a novel “smoke” word dictionary to identify drug side effect words |
| Liu and Chen (2015) | UMLS, CHV, FAERS, MeSH | Lexicon-Based Approach with Statistical Learning | Created unique semantic groups implemented from UMLS |
| Gurulingappa et al. (2012) | Medline | Statistical learning with SVMs | Utilize support vector machines to identify adverse reactions from medical case reports |
| Bian et al., 2012) | Twitter | SVMs | Earliest approach to using SVMs on twitter dataset |
| Ginn et al. (2014) | Twitter, SIDER | SVM, Naïve-Bayes | Created custom twitter dataset for ADE detection |
| Korkontzelos et al. (2016) | Twitter | Sentiment Analysis, SVM | Introduced the use of Sentiment Analysis to improve upon previous methods |
| Cocos et al. (2017) | Twitter | Deep Learning | First use of recurrent neural networks for twitter ADE extraction. However, dataset was severely limited. |

3. A novel transformer-based framework for ADE detection and extraction

3.1. Overview of the framework

The objectives of this framework is to (1) take a review R and detect the presence of ADEs (i.e. ADE detection), and (2) for each word that is in R , w_R , we wish to classify into three classes: unimportant, drug name, or drug side effect (i.e. ADE extraction). The overall methodology for the research is seen in the figure below but is described as follows (Fig. 1). We take data from WebMD and Drugs.com, two novel social media forums that have not been utilized for the ADE detection or extraction task. We then get medical students to help identify the ADEs within each review to create the labels for the data. In order to commence training and prediction, the reviews are converted into BERT embeddings that are then passed into our deep learning model to compute probabilities across words as to detect the presence of ADEs. This is then compared to the true label and used to improve the model during training. Each box will be expanded upon in the following sections. The blue boxes in the ADE extraction methodology are our primary contribution to the research literature.

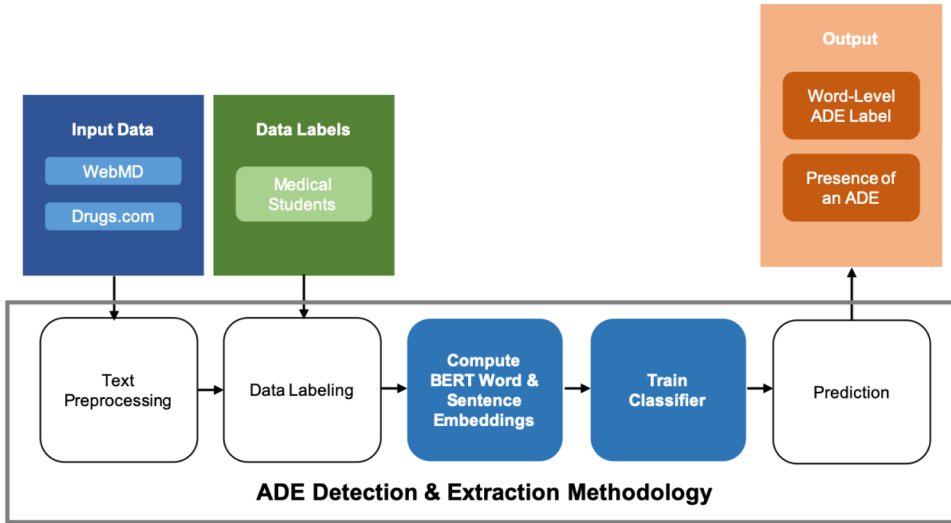


Fig. 1. Proposed ADE detection and extraction framework. Our primary contribution lies in the introduction of novel datasets (WebMD and Drugs.com). By utilizing health social media forum data, we are getting the most relevant information relating to drugs and can guarantee a larger number of posts related to ADEs and drugs themselves. In addition, we utilize novel BERT word and sentence embeddings (Vaswani et al., 2017) to ultimately detect and extract ADEs.

3.1.1. Traditional deep learning approach

Traditionally, detection and extraction methods compute word embeddings and pass it through a vanilla recurrent neural network, or LSTM. Despite the effectiveness and simplicity of this approach, the results wrought from the traditional approach are far from optimal (as seen by the experiments discussed later) and call for novel introduction of different methods for improved accuracy outside of just changing the different types of word embeddings (i.e. GloVe or BERT). This is caused by the fact that the word embeddings may not be able to completely capture the entire context of a sentence or review especially as the review size becomes larger (Devlin et al., 2018; Vaswani et al., 2017). As a result, this can lower performance and calls for another direction to introduce a greater context.

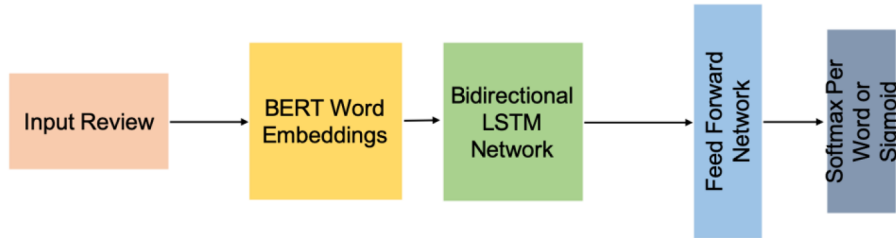


Fig. 2. Traditional deep learning approach w/ BERT.

3.1.2. New deep learning approach

In order to solve this problem of limited contextualization, we introduce the use of novel sentence embeddings. These sentence embeddings are computed as a concatenation of the word embeddings and then appended to each word before acted upon by the LSTM. We first pass this sentence embedding through a feed forward neural network as to “encode” it and learn necessary features to be passed to the LSTM, not only does this learn the necessary features of a sentence as if to “summarize” or learn the sentence, but it also standardizes the sentence embeddings to be appended to each word before the LSTM. By adding the BERT sentence embeddings, we enable the model to utilize a greater context (at the sentence level) to more accurately determine whether not a word is an ADE (extraction) and whether a review contains an ADE (detection). This context, as a result of the training of the model, wrought from both the sentence embedding and the contextual word embeddings inherently allows for more accurate detection of negation and sentiment. Thus, a revised version of the model architecture is shown in Fig. 3. More specifically we take a review R and compute a sentence embedding S_R that is equal to the length of R and pass S_R into a feed forward network $D(x)$ and concatenate $D(S_R)$ with each word W_R thereby passing $S_R + W_R$ into the LSTM.

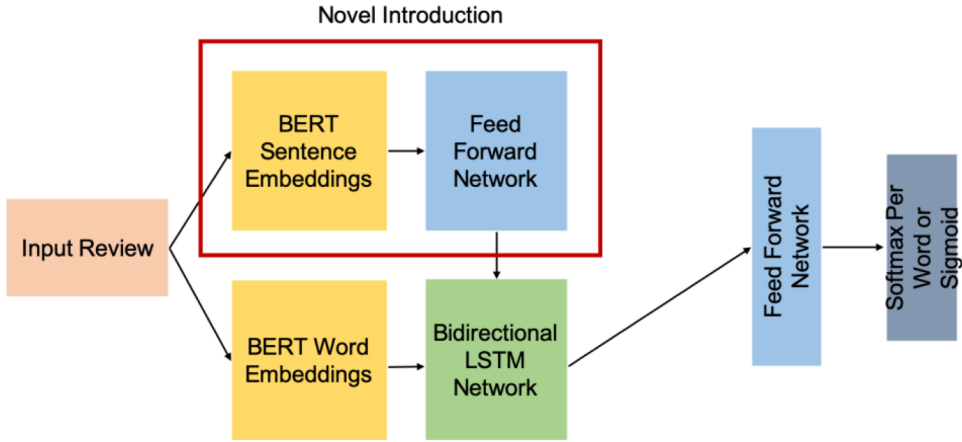


Fig. 3. New deep learning framework w/ sentence embeddings.

3.2. Proposed ADE detection algorithmic model architecture

The goal of this algorithm is to be able to detect the presence of ADEs within a review. We first take the input review and compute BERT, contextual embeddings; we then utilize the BERT word embeddings to compute a sentence embedding that is then concatenated with each word embedding to provide a greater context and improve performance. Specifically, for each word W_R with dimension 1×768 (the dimensions of the word vector), we concatenate a sentence vector S_R $N \times 768$ producing an $S_R + W_R$ vector with dimensions $(N + 1) \times 768$ that is passed to the LSTM. We then compute a final sigmoid probability for the detection of ADEs. Our primary contribution lies in the use of novel contextual embeddings, specifically BERT-Large embeddings pretrained on a large language dataset. These word embeddings have deeper, richer embeddings that create better vector representations for each word. By utilizing contextual embeddings in comparison to normal word embeddings, we provide models more information to act upon on a per-word basis. We also utilize novel BERT sentence embeddings. By concatenating the BERT sentence and word, it enables each timestep to get the surrounding words' context and the entire sentence context that includes other features computed by BERT itself. This improves prediction and results in comparison to standard NLP models that only utilize word embeddings. Thus, the use of the BERT embeddings is another primary contribution to pharmacovigilance literature. An algorithmic approach is shown in Fig. 4.

Algorithm 1: ADE Detection Algorithm

Data: Input Review
Result: Output probability of the presence of ADEs within review
for each word $w_F \in \text{review } F$ **do**
 if word w_F is not in stopwords S **then**
 lemmatize word w_F convert w_F to BERT vector v_w
 end
end
compute feed-forward BERT sentence embedding s_F
for each vector $v_w \in \text{word vectors } V$ **do**
 compute bidirectional LSTM result $L(v_w)$
 $D(v_w) = s_F + L(v_w)$
end
compute sigmoid probability $S = \frac{1}{1+e^{-D}} \in [0, 1]$

Fig. 4. Algorithmic ADE detection approach.

3.3. Proposed ADE extraction algorithmic model architecture

For ADE extraction, we utilize BERT word and sentence embeddings to compute adverse drug event labels at each time step of the input review. Notice the difference between this approach and the ADE detection approach. Rather than computing an overall sigmoid (as seen in Fig. 4), we compute a softmax per word, thus computing a result for each word and detecting whether or not a word is a drug, ADE or unimportant. Again, we utilize novel BERT word and sentence embeddings to utilize a greater amount of context to improve and enhance prediction at the word level, a primary contribution to the pharmacovigilance and information extraction literature. An algorithmic approach for the ADE extraction is shown below.

Algorithm 2: ADE Extraction Algorithm

Data: Input Review
Result: Output drug side effect label at word level

```

for each word  $w_F \in \text{review } F$  do
  if word  $w_F$  is not  $\in \text{stopwords } S$  then
    lemmatize word  $w_F$  convert  $w_F$  to BERT vector  $v_w$ 
  end
end
compute feed-forward BERT sentence embedding  $s_F$ 
for each vector  $v_w \in \text{word vectors } V$  do
  compute bidirectional LSTM result  $L(v_w)$ 
   $D(v_w) = s_F + L(v_w)$ 
  compute softmax  $S_w(D(v_w)) = \frac{\exp D(v_w)}{\sum_{[0,1,2]} \exp D(v_w)}$  for word.
end

```

Fig. 5. Algorithmic ADE extraction approach.

3.4. Datasets

Two primary datasets were used for this research: WebMD and Drugs.com. Both websites have never been used in previous research and were chosen primarily for their large drug data source as well as their strong and abundant user base. Each website was organized per drug. Thus, reviews were written for each drug. On average, each drug contained approximately 30 reviews from various users. Approximately 2/3 of these reviews contained drug side effects. The reviews were labeled by local medical students. A total of 10,000 reviews were labeled with approximately 6037 drug side effects labeled in total. The exact numbers are seen in the table below. In order to maintain consistency and reduce confounding variables, stratified random sampling was done at both the dataset and drug level to create training, testing, and validation datasets.

3.4.1. WebMD

WebMD is one of the top online health social media forums on the current market. WebMD's data encompasses a wide variety of healthcare-based information including health conditions, doctor consultations, as well as a wide source of drugs and their associated uses. For the purpose of this research, we utilized WebMD's large drug database that includes a multitude of drugs. As seen in Table 2, utilizing data crawlers, approximately 14,000 total drugs were crawled along with their associated review totaling 241,980. This, on average, is approximately 17.36 reviews per drug. The reviews crawled from WebMD are created by its community members and moderated by hired doctors and professionals. In addition, each review is provided a review rating from 1 to 5 to determine its authenticity and usefulness. Also, the length of the reviews varied, ranging from 10 words to 100 words depending on the drug, with an average length being 37.5 words. In addition, the drug description and drug purpose were crawled for further reference during drug profiling. Screenshot of a drug profile (Fig. 6) and its associated reviews (Fig. 7) on WebMD are shown below.

3.4.2. Drugs.com

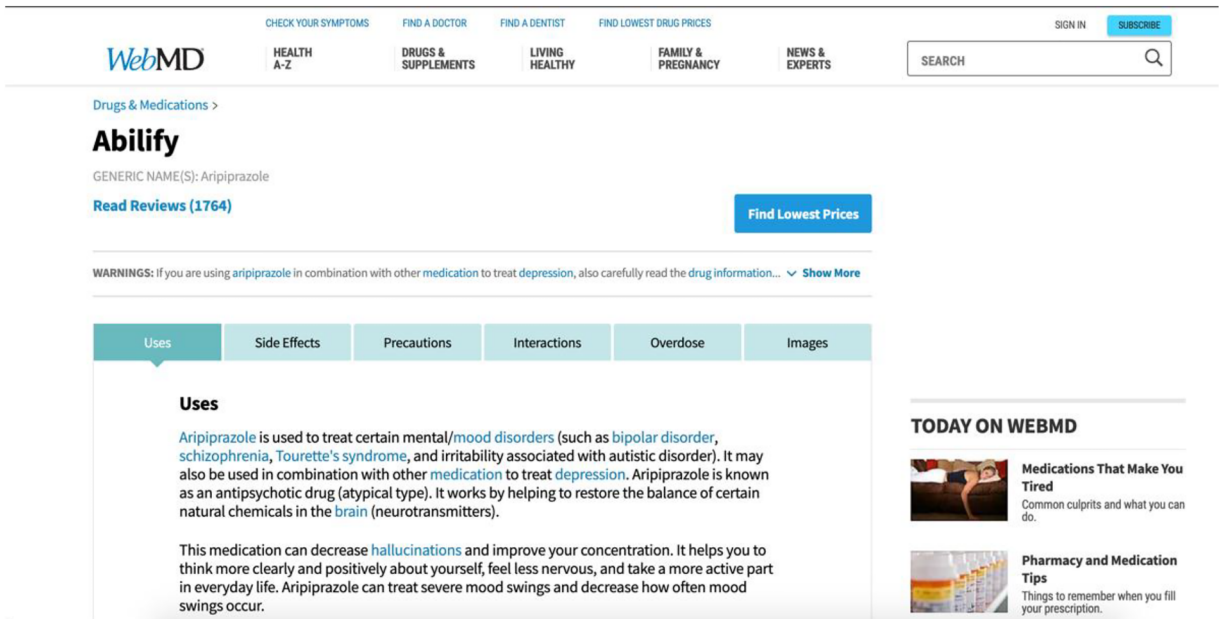
Drugs.com offers information including drug description, purpose, and common side effects extracted by physicians, practitioners, and pharmacologists. In addition to simply drug side effects, Drugs.com offers an easy way to categorize drug side effects based on the anatomical structure a specific side effect attack. This research utilizes the drug reviews reported by users on the site. Similar to WebMD, associated reviews, definitions, and uses for each drug were crawled. In comparison to WebMD, drugs.com reviews were considerably longer in length and contained varying levels of comprehension. Drugs.com contained reviews not only from normal consumers, but also doctors and pharmacologists who provided more detailed reviews. Review ratings were also provided and crawled. Drugs.com had less data labeled as it was added on at a later stage in the project.

3.4.3. Data labeling

Once the data was crawled from both WebMD and Drugs.com, 15 medical students from Edward Via College of Osteopathic Medicine were called to label a total of 10,000 data points: 6000 from WebMD and 4000 from Drugs.com. For each review, medical students were tasked with classifying each word in to three primary categories: unimportant, drug name, and drug side effect. Unimportant words were given a value of 0, drug names were given a value of 1, and drug side effects were given a value of 2. Labelers were instructed to be careful when labeling adverse drug events as to reduce the possibility of false positives such as taking the drug "Zofran" for "throw up" rather than obtaining "throw up" because of "Zofran". By ensuring that the data is properly labeled, we can ensure that the deep learning model can properly identify which reactions are actual adverse reactions and which are not. A cross-annotator score (calculated by a Cohen's kappa) of 0.95 was achieved. An example labeled review is seen in Fig. 8.

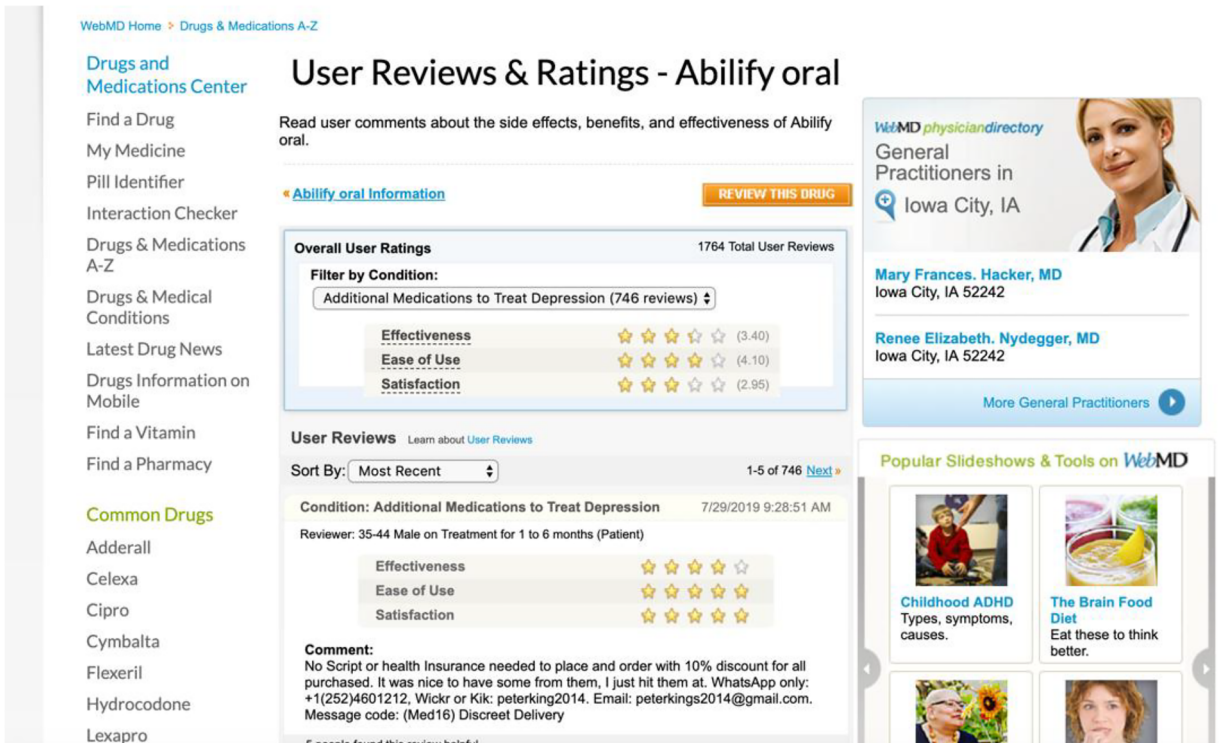
Table 2
Dataset of review distribution and numbers.

| Dataset | Drugs | Total reviews | Reviews labeled |
|-----------|--------|---------------|-----------------|
| WebMD | 13,935 | 241,980 | 6000 |
| Drugs.com | 13,935 | 179,368 | 4000 |
| Totals: | N/A | 421,348 | 10,000 |



The image shows the WebMD drug profile for Abilify (Aripiprazole). At the top, there are navigation links: CHECK YOUR SYMPTOMS, FIND A DOCTOR, FIND A DENTIST, FIND LOWEST DRUG PRICES, SIGN IN, and SUBSCRIBE. Below these are category links: HEALTH A-Z, DRUGS & SUPPLEMENTS, LIVING HEALTHY, FAMILY & PREGNANCY, and NEWS & EXPERTS. A search bar is on the right. The main heading is 'Abilify' with the generic name 'Aripiprazole' and a link to 'Read Reviews (1764)'. A 'Find Lowest Prices' button is also present. A warning section states: 'WARNINGS: If you are using aripiprazole in combination with other medication to treat depression, also carefully read the drug information... Show More'. Below this are tabs for 'Uses', 'Side Effects', 'Precautions', 'Interactions', 'Overdose', and 'Images'. The 'Uses' tab is active, showing that Aripiprazole is used for bipolar disorder, schizophrenia, Tourette's syndrome, and depression. It also mentions that it helps with hallucinations and improves concentration. On the right, there is a 'TODAY ON WEBMD' section with links to 'Medications That Make You Tired' and 'Pharmacy and Medication Tips'.

Fig. 6. Example WebMD drug profile.



The image shows the WebMD user reviews for Abilify oral. The left sidebar contains navigation links: WebMD Home, Drugs & Medications A-Z, Drugs and Medications Center, Find a Drug, My Medicine, Pill Identifier, Interaction Checker, Drugs & Medications A-Z, Drugs & Medical Conditions, Latest Drug News, Drugs Information on Mobile, Find a Vitamin, Find a Pharmacy, Common Drugs, Adderall, Celexa, Cipro, Cymbalta, Flexeril, Hydrocodone, and Lexapro. The main heading is 'User Reviews & Ratings - Abilify oral'. Below this is a section for 'Abilify oral Information' with a 'REVIEW THIS DRUG' button. The 'Overall User Ratings' section shows 1764 Total User Reviews. The 'Filter by Condition' dropdown is set to 'Additional Medications to Treat Depression (746 reviews)'. The ratings are: Effectiveness (3.40), Ease of Use (4.10), and Satisfaction (2.95). The 'User Reviews' section shows a list of reviews. The first review is from a 35-44 Male on Treatment for 1 to 6 months (Patient), dated 7/29/2019 9:28:51 AM. The review text is: 'No Script or health Insurance needed to place and order with 10% discount for all purchased. It was nice to have some from them, I just hit them at. WhatsApp only: +1(252)4601212, Wickr or Kik: peterking2014. Email: peterkings2014@gmail.com. Message code: (Med16) Discreet Delivery'. The review is rated 5 stars. On the right, there is a 'WebMD physician directory' section for Iowa City, IA, listing Mary Frances. Hacker, MD and Renee Elizabeth. Nydegger, MD. Below this is a 'Popular Slideshows & Tools on WebMD' section with links to 'Childhood ADHD Types, symptoms, causes.' and 'The Brain Food Diet Eat these to think better.'.

Fig. 7. Example review list.

3.4.4. Data preprocessing

In order to maintain consistency during vocabulary creation, all data was first converted into lowercase, tokenized, stemmed, and then lemmatized to guarantee word consistency and normalization during vocabulary and embedding creation. Once all 10,000 data points were properly preprocessed, a stratified sampling split of 0.7, and 0.3 was implemented based upon drug to split the data into training, and testing datasets respectively.

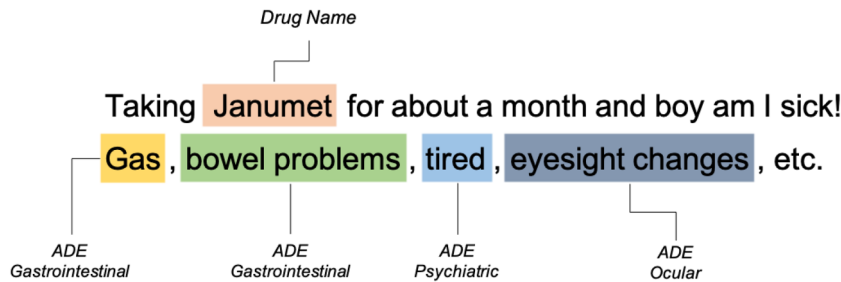


Fig. 8. Example labeled review. In this example, Janumet would have been given a label of 1, and the various ADEs (i.e. Gas, bowel problems, tired, eyesight changes) would be given a label of 2. All words that were unrelated to the task would be given a label of 0.

4. Experiments

4.1. Experiment 1: ADE detection

We implement an ADE detection experiment to compare the various embeddings used within deep learning. ADE detection data was calculated by the presence of an ADE. Thus, if a review was labeled to have an ADE, that review was given a label of 1. If no ADE was present in the review, a label of 0 was given. Here, we utilize our proposed BERT word embeddings coupled with BERT sentence embeddings (a concatenation of the BERT word embeddings) to predict the presence of ADEs within a review. The key difference in model architecture between the ADE detection and the ADE extraction algorithm is the difference in the final layer. Rather than trying to compute a value for each word, we are predicting a value for the entire review, thus a different final layer that is a single dense unit, rather than a recurrent unit like an LSTM. Here, we utilize a sigmoid activation for our final dense layer to receive a probability between 0 and 1.

4.1.1. ADE detection benchmarks

Three benchmarks were used for the ADE detection experiment: non-pretrained word embeddings, pre-trained word embeddings proposed by Cocos et al. (2017), and BERT word embeddings. Each are elaborated below.

4.1.1.1. Non-pretrained word embeddings. A standard N by 400-dimensional word embedding was implemented for this benchmark. The word embeddings were randomly initialized and were learned during runtime with the model implementation discussed below. This is often the standard baseline that is used to compare against deep learning models but because it's not finetuned on a specific corpus, does not produce the best results.

4.1.1.2. Pre-trained word embeddings (Cocos et al., 2017). These word embeddings were proposed by Cocos et al. A standard N by 400-dimensional word embedding was first pretrained on a corpus of medical articles from Medline. The word embeddings were then used as the weights during training and were finetuned on the drug dataset.

4.1.1.3. BERT word embeddings. BERT word embeddings are an upgrade from traditional word embeddings (Devlin et al., 2018; Mikolov, Sutskever et al., 2013) and are able to capture the context around a word rather than simply the immediate context (i.e. the word behind and in front of a word). As a result, theoretically, the BERT word embeddings should outperform previous word embeddings. Thus, we use this as a baseline to not only compare against our model, but to compare against the previous baselines as to prove that BERT outperforms current embedding techniques.

4.1.2. AUC metric for ADE detection

For the ADE detection comparison with Cocos et al. (2017), we utilized the area under the curve of the receiver operating characteristic (AUC) as in Cocos et al. (2017). It is commonly used for imbalanced data set evaluation. The AUC measures the ability of a binary classifier to properly distinguish between two primary groups (in the ADE detection problem, the presence and absence of drug side effects). The receiver operating characteristic is calculated by plotting the true positive rate against the false positive rate at different threshold settings. A higher AUC (near 1) is desired while a low AUC (near 0.5) means that the model cannot discriminate effectively.

4.1.3. Model training

To maintain consistency, the same bidirectional LSTM proposed in Cocos et al. was used (a 100-unit bidirectional LSTM, trained for 10 epochs) (Cocos et al., 2017).

4.1.4. ADE detection results

In order to compare the various word embeddings, we take the liberty of commencing a more general experiment, a binary detection of the presence of an adverse drug event within a review. Thus, if an ADE is present in a review, that review is given a label

Table 3
ADE detection results.

| Model | AUC |
|---------------------------------|-------------|
| No Pre-Trained Embeddings | 0.82 |
| Pre-Trained Embeddings | 0.85 |
| BERT Word Embeddings | 0.91 |
| BERT Word + Sentence Embeddings | 0.94 |

of 1 and the review is given a label of 0 if no ADE is present in the review. To test these models, we train the models on the detection task mentioned above and evaluate the models utilizing the AUC metric discussed in [Section 4.1.2](#). Results are shown for the ADE detection task comparing the deep learning with non-pretrained word embeddings, deep learning with pretrained word embeddings ([Cocos et al., 2017](#)), and deep learning with BERT word embeddings are shown below in [Table 3](#).

The results show that the BERT-based model significantly outperforms both models, showing improved classification accuracy of the BERT word embedding and its ability to capture meaningful word semantics and contextual elements.

4.2. Experiment 2: ADE extraction

We implement the ADE extraction experiment to detect the ability of deep learning to outperform current state-of-the-art approaches (lexicon and statistical) in the task of ADE extraction. The extraction benchmarks are elaborated on below.

4.2.1. ADE extraction benchmarks

4.2.1.1. Statistics-based approach ([Liu & Chen, 2015](#)). A core benchmark that we utilize in this research is state-of-the-art statistical-based approaches. In this paper, we utilize Liu & Chen's approach of support vector machines with a custom kernel. To keep results and methods consistent, we utilize the same datasets that Liu & Chen tested, including the Unified Medical Language Service, MetaMap and the CHV vocabulary. This approach was implemented and run on our dataset. The SVM was implemented using the scikit-learn library.

4.2.1.2. Decision tree approach. This compares another standard machine learning approach found in machine learning. Here, we utilize a Decision Tree Classification model as a method to compare to the deep learning models proposed in this paper. This approach also compares the efficacy of machine learning in comparison to deep learning. We restrict the depth of the tree to a maximum of 10 with a criterion of the Gini impurity. The Decision Tree was implemented using the scikit-learn library.

4.2.1.3. Deep learning with normal word embeddings ([Cocos et al., 2017](#)). We utilize this benchmark as the primary basis of comparison for our model. Cocos et al.'s model for pharmacovigilance utilized a non-pretrained word embedding combined with a bidirectional long-short-term memory network to identify drug side effects. Non-pretrained embeddings mean that word relationships will be learned during training time rather than pre-learned on a dataset. We replicate their model as it is presented in their paper.

4.2.1.4. Deep learning with pre-trained word embeddings ([Cocos et al., 2017](#)). This baseline is a comparison of the current best state-of-the-art model in ADE extraction. The model was trained by Cocos et al. using pretrained embeddings from a non-domain specific Twitter dataset. We replicate the model for this experiment as a comparison of the effectiveness of both the model architecture and the different word embeddings.

4.2.1.5. Deep learning with BERT word-level embeddings. This baseline is a comparison of only using BERT word embeddings for the deep learning model and not using the BERT sentence embeddings. BERT word embeddings were pretrained on a language dataset and fine-tuned on Medline.

4.2.2. F1 metric for ADE extraction

For the ADE extraction experiment, the weighted F1 metric shown in the calculation below was used. The F1 metric is another method of representing a model's accuracy. We utilize F1 as a means of measuring the ADE extraction model's effectiveness of identifying drug side effect labels and detecting non-drug side effect labels. F1 scores are calculated on a per class basis (i.e. unimportant, drug name, and drug side effect) and the weighted F1 combines the results into a formal measurement. The F1 metric calculation is shown below. In the case of ADE extraction, it is more important that we get accurate classification of drug side effects (i.e. true positives) instead of false positives. Thus, in the case of the problem, precision is more important than recall and is also reflected in the use of a weighted F1 measurement. Because the F1 is weighted, if precision is low for a particular class (i.e. drug side effect), then this will negatively impact the weighted F1 score.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (2)$$

F1 Score Calculation

Table 4
ADE extraction results.

| Model | F1-score | Precision | Recall |
|---------------------------------------------|-------------|--------------|--------------|
| Statistic-Based Approach (Liu & Chen, 2015) | 0.42 | 0.39 | 0.455 |
| Lexicon-Approach (Decision Tree) | 0.33 | 0.35 | 0.31 |
| No Pre-trained Embeddings | 0.78 | 0.82 | 0.74 |
| Pre-Trained Embeddings (Cocos et al., 2017) | 0.87 | 0.89 | 0.85 |
| BERT Word Embeddings | 0.92 | 0.93 | 0.91 |
| BERT Word + Sentence Embeddings | 0.97 | 0.974 | 0.966 |

4.2.3. ADE extraction results

Results for the ADE extraction comparing the current state-of-the-art lexicon-based approach, a standard machine learning rule-based approach, deep learning with non-pretrained embeddings, deep learning with pretrained word embeddings, and our proposed model that combines BERT word embeddings and BERT sentence embeddings are shown below in Table 4.

Again, as we see by the results, the proposed model significantly outperforms current state-of-the-art lexicon-based approaches, machine learning approaches, and the deep learning approaches proposed by Cocos et al. Because current lexicon-based systems and statistic-based systems often rely upon medical-based dictionaries, many layman terminologies are often not identified, producing poor results in the experiment. Our proposed model achieved a significantly higher F1 score and shows that created a new state-of-the-art extraction model. This is most likely caused by the feature rich word embeddings found utilizing the BERT embeddings. In addition to just the word embeddings, the new mechanism of computing a sentence embedding and adding it to the BERT word embeddings is a novel deep learning architecture for the pharmacovigilance task. This increases performance in comparison to previous models, adds context to the prediction task, and provides multiple different features in combination to improve deep learning performance. Sample predictions on the unlabeled dataset by the model are shown in Table 5. ADEs are extracted for every word that is labeled a 2 by the model.

Despite the results, there are still cases where the model fails to identify certain drug side effects, including abbreviated side effects and misspelled words. However, this problem can be solved by expanding upon abbreviations and attempting to program a spellchecker. In addition, the extraction of side effects, it is difficult for the model to recognize multi-word drug side effects. For example, many drug side effects are a conjunction of multiple words together, so it is still necessary to recognize whether an ADE is a single word or phrase.

Table 5
Example extracted drug side effects.

| |
|--------------------------------------------------------------------|
| The gel worked but it was painful , however I am free of BV |
| Propanolol did not help at all with head pain |
| I was severely constipated and had to discontinue use |
| I have had a very bad allergic reaction to this medication |

4.3. Model implementations

Models were implemented using the Keras Python Library utilizing a Tensorflow Backend. We set the sentence embedding feed forward network output to 100 units and the LSTM unit output to 512 units. The word reviews were each padded with zeros (equal to the longest review) to ensure sequence length consistency. We trained the model for 15 epochs.

5. Discussion

5.1. Results analysis

Upon analysis of the results, one can recognize that our model achieves significant results in both ADE detection and ADE extraction. This represents the increasing viability and application of deep learning techniques to problems like pharmacovigilance, and the ability of BERT word embeddings to outperform standard word embeddings. Because pretrained word embeddings are implemented through feed forward neural networks, they pale in comparison to the rich BERT word embeddings that are trained utilizing transformers, a coalition of encoders and attention mechanisms. In addition, the combination of both BERT word and sentence embeddings enables each word classification to utilize a greater context of the entire sentence, making our model one of the most effective models for the ADE detection and extraction tasks.

5.2. Implications for research

Our research introduces the use of BERT word and sentence embeddings that can capture greater contextual elements in comparison to normal word embeddings. Finally, our model and its findings can be applied and transferred across multiple information

extraction tasks. Because ADE extraction is simply an information extraction task, any other named entity recognition problems can be solved by the model proposed in this paper, showing the capability of BERT and the use of contextual embeddings to improve current models.

5.3. Limitations and future research

There are many avenues for further research. An analysis of drug side effect discovery across languages should also be considered and a potential multilingual model should be created but was not available during the training and collection of data for the model. Though our model did achieve high performance, there is still room to improve. Further research should attempt to analyze the implementation of ensemble-based approaches, pooling results from lexicon-based approaches, and deep learning approaches to compute a final ADE extraction label. In addition, other features should be combined with BERT embeddings such as sentiment analysis (Korkontzelos et al., 2016) to more accurately identify drug side effects from reviews. This research proves the viability of deep learning with the use of BERT, however further research should continue to investigate the use of new deep learning methods such as XLNet (Yang et al., 2019) to improve performance and accuracy and generalize it to more information processing and extraction tasks. This research introduces the use of social media datasets WebMD and Drugs.com, but only goes to the extent of labeling 10,000 total points. Future research should implement methods of online and semi-supervised learning methods to label and utilize the rest of the reviews found in the dataset.

6. Conclusion

This study introduces a novel use of social media health forum data for ADE extraction from websites like WebMD and Drugs.com where extant research utilizes data from twitter, electronic medical records, and medical case reports. This study also produces results that show that deep learning is a viable option for the pharmacovigilance task and the utilization of novel natural language processing techniques with BERT sentence and word embeddings significantly outperform previously implemented statistical, lexicon, and deep learning approaches in the ADE extraction task. Our proposed model achieves new state-of-the-art F1 results that can become the basis for improvement. This study also shows the feasibility of BERT word embeddings over other types of non-pretrained word embeddings and pretrained embeddings by comparing ADE detection results with our proposed model achieving an AUC of 0.94. Not only can the model be used for ADE extraction, but it can also be used for general information extraction tasks including named entity recognition and content extraction, generalizing the model to a wider array of tasks. Despite the success in the ADE extraction task, further research should analyze ways to classify drug side effects into manageable categories for doctor viewing and drug professionals. Furthermore, a method for recognizing semantic similarity between drug side effects is an important consideration in order to map similar drug side effects into one “formal definition”. However, this research provides a new frontier to explore the use of deep learning in a wide array of pharmacovigilance tasks.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.ipm.2019.102131](https://doi.org/10.1016/j.ipm.2019.102131).

References

- Adams, D. Z., Gruss, R., & 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, 108–120.
- Aron, R., Dutta, S., Janakiraman, R., & Pathak, P. A. (2011). The impact of automation of systems on medical errors: Evidence from field research. *Information Systems Research*, 22(3), 429–446.
- Bengio, Y., & LeCun, Y. (2007). Scaling learning algorithms towards AI. *Large-Scale Kernel Machines*, 34(5), 1–41.
- Bian, J., Topaloglu, U., & Yu, F. (2012). Towards large-scale twitter mining for drug-related adverse events. *International workshop on smart health and wellbeing*.
- Brown, M. T., & Bussell, J. K. (2011). Medication adherence: Who cares? *Mayo clinic proceedings*.
- Chee, W. B., Berlin, R., & Schatz, B. (2011). Predicting adverse drug events from personal health messages. *AMIA annual symposium proceedings*.
- Cocos, A., Fiks, A. G., & Masino, A. J. (2017). Deep learning for pharmacovigilance: Recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *Journal of American Medical Informatics Association*.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint*.
- Ferraro, J. P., Daume, H., I.I., DuVall, S. L., Chapman, W. W., Harkema, H., & Haug, P. J. (2013). Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *Journal of the American Medical Informatics Association*, 20(5), 913–939.
- Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., O'Connor, K., Sarker, A., et al. (2014). Mining Twitter for adverse drug reaction mentions: A corpus and classification benchmark. *Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing*.
- Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 602–610.
- Gruetzmacher, R., Gupta, A., & Paradice, D. (2018). 3D deep learning for detecting pulmonary nodules in CT scans. *Journal of the American Medical Informatics Association*, 25(10), 1301–1310.
- Gurulingappa, H., Mateen-Rajpu, A., & Toldo, L. (2012). Extraction of potential adverse drug events from medical case reports. *Journal of Biomedical Semantics*.
- Harpaz, R., DuMouchel, W., Shah, N. H., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology & Therapeutics*, 91(6), 1010–1021.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Hoang, T. B. N., & Mothe, J. (2018). Location extraction from tweets. *Information Processing & Management*, 54(2), 129–144.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 1735–1780.
- Kisa, A., Sabate, E., & Nuno-Solinis, R. (2003). *Adherence to long-term therapies: Evidence for action*. World Health Organization.
- Kohli, R., & Kettinger, W. J. (2004). Informing the clan: Controlling physicians' costs and outcomes. *MIS Quarterly*, 363–394.
- Korkontzelos, I., Nikfarjam, A., Shallow, M., Sarker, A., Ananiadou, S., & Gonzalez, G. (2016). Analysis of the effect of sentiment analysis on extracting adverse drug

- reactions from tweets and forum posts. *Journal of Biomedical Informatics*.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- LeCun, Y., Touresky, D., Hinton, G., & Sejnowski, T. (1988). A theoretical framework for back-propagation. *Proceedings of the 1988 connectionist models summer school*.
- Li, Z., Liu, F., Antieau, L., Cao, Y., & Yu, H. (2010). Lancet: A high precision medication event extraction system for clinical text. *Journal of the American Medical Informatics Association*, 17(5), 563–567.
- Lin, Y.-K., Chen, H., Brown, R. A., Li, S.-H., & Yang, H.-J. (2017). Healthcare predictive analytics for risk profiling in chronic care: A bayesian multitask learning approach. *MIS Quarterly*, 41(2).
- Liu, X., & Chen, H. (2015). A research framework for pharmacovigilance in health social media: Identification and evaluation of patient adverse drug event reports. *Journal of Biomedical Informatics*, 58, 268–279.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space, *arXiv preprint*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *NIPS*.
- Papagiannopoulou, E., & Tsomakas, G. (2018). Local word vectors guiding keyphrase extraction. *Information Processing & Management*, 54(6), 888–902.
- Peng, F., & McCallum, A. (2006). Information extraction from research papers using conditional random fields. *Information Processing & Management*, 42(4), 963–979.
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Conference on empirical methods in natural language processing*.
- Pereira, S., Pinto, A., Alves, V., & Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging*, 35(5), 1240–1251.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). Deep contextualized word representations. *NAACL*.
- Qian, L., Guan, Y., Dong, X., Huang, L., Yu, Q., & Yang, J. (2016). A multiclass classification method based on deep learning for named entity recognition in electronic medical records. *2016 New York scientific data summit (NYSDS)*.
- Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1–2), 1–39.
- Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., et al. (2015). Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54, 202–212.
- Seger, A. C., Jha, A. K., & Bates, D. W. (2007). dverse drug event detection in a community hospital utilising computerised medication and laboratory data. *Drug Safety*, 30(9), 817–824.
- Serban, O., Thapen, N., Maginnis, B., Hankin, C., & Foot, V. (2019). Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing & Management*, 1166–1184.
- Shertstinsky, A. (2018). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network, *arXiv*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *NIPS*.
- World Health Organization, The importance of pharmacovigilance (2002).
- Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, 25(10), 1419–1428.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining, *arxiv*.
- Yuan, S., & Yu, B. (2019). HClaimE: A tool for identifying health claims in health news headlines. *Information Processing & Management*, 1220–1233.
- Zhang, H., Boons, F., & Batista-Navarro, R. (2019). Whose story is it anyway? Automatic extraction of accounts from news articles. *Information Processing & Management*, 56(5), 1837–1848.