A Causality Driven Approach to Adverse Drug Reactions Detection in Tweets

Humayun Kayesh, Md. Saiful Islam(✉), and Junhu Wang

School of Information and Communication Technology,
Griffith University, Gold Coast, Australia
humayun.kayesh@griffithuni.edu.au,
{saiful.islam,j.wang}@griffith.edu.au

Abstract. Social media sites such as Twitter is a platform where users usually express their feelings, opinions, and experiences, e.g., users often share their experiences about medications including adverse drug reactions in their tweets. Mining and detecting this information on adverse drug reactions could be immensely beneficial for pharmaceutical companies, drug-safety authorities and medical practitioners. However, the automatic extraction of adverse drug reactions from tweets is a nontrivial task due to the short and informal nature of tweets. In this paper, we aim to detect adverse drug reaction mentions in tweets where we assume that there exists a cause-effect relationship between drug names and adverse drug reactions. We propose a causality driven neural network-based approach to detect adverse drug reactions in tweets. Our approach applies a multi-head self attention mechanism to learn word-to-word interactions. We show that when the causal features are combined with the word-level semantic features, our approach can outperform several state-of-the-art adverse drug reaction detection approaches.

Keywords: Adverse drug reaction detection · Causality · Neural network · Multi-head self attention

1 Introduction

Adverse Drug Reaction (ADR), which is considered to be responsible for millions of fatalities every year, is the harmful and unwanted reaction caused by the usage of a medical product [17]. The pharmaceutical companies use ADR information to inform patients about any potential side effects of their products. The identification of ADRs is critical not only for pharmaceutical companies but also for health care management authorities. The accurate and automatic identification of ADRs can save a huge amount of money spent every year by the health care management authorities for treating patients affected by ADRs.

Ideally, adverse drug reactions are identified by lab experiments. Patients are also encouraged to report ADRs through self-reporting systems. However, both of these techniques have limitations. Lab tests are often conducted on a

© Springer Nature Switzerland AG 2019
J. Li et al. (Eds.): ADMA 2019, LNAI 11888, pp. 316–330, 2019.
https://doi.org/10.1007/978-3-030-35231-8_23

Fig. 1. ADRs in tweets with the cause-effect relationship

limited number of subjects and run for a limited period whereas an ADR may accrue after a long time of using a drug. This issue is partially solved by the self-reported ADR collections where patients voluntarily report adverse reactions of drugs. However, the steps of self-reporting systems are often too complex and time-consuming, hence the patients who experience severe adverse reactions on medicines do not feel the urge to go through all standard steps to report an ADR. This contributes to the rise of unknown or unidentified ADRs.

Nowadays, social media platform such as Twitter has become a popular source of ADRs-related information. Users often tweet for seeking information about different health conditions or just to share their medication experiences. Tweets often include drug names and ADR mentions. Consider the two example tweets as visualized in Fig. 1. The first example mentions an ADR where the name of the drug is *Olanzapine* and the adverse reaction is *tired*. The second example shows that the patient is unable to cry after taking *Effexor*. Due to the availability of a huge volume of data and potentiality of patient-reported hidden ADRs, pharmacovigilance researchers are now interested in automatic ADR detection from tweets. However, automatic detection of ADRs from tweets is a nontrivial problem. There are several challenges in the task of detecting ADRs from tweets. Firstly, tweets are highly informal; sometimes, the posts contain grammatically incorrect sentences, misspellings, and emojis. Secondly, the posts are often short and the adverse reactions are vaguely described. Thirdly, a variety of expressions are used by different twitter users to explain the same thing.

The existing approaches [16,17] that detect ADRs from tweets exploit co-occurrence information as a signal to detect the ADR relationship between a drug and an adverse reaction. However, the co-occurrence of a word with a drug name may not always suggest that the word is an ADR. For example in the tweet "Thank god for vyvanse or I would be sleeping on the cash register right now", the word *sleeping* is used with a drug name *vyvanse*, but it is not an ADR. Other approaches such as [3] apply word embeddings as features but word embeddings cannot detect the relationship between drug names and ADRs as the drug names are not regular English words. Hence, it is important to take the causal relationship between drug and ADRs into consideration rather than co-occurrence of words or word embeddings only.

In this paper, we aim to detect ADRs from tweets by extracting causal features between drug name (cause) and adverse reaction (effect). Causal features are extracted for every word by splitting a tweet into segments such as prefix, midfix, and postfix depending on the position of the word and calculating tf-idf

from each segment. The causal features are then combined with word features, which are extracted from word embeddings. Finally, we propose to apply a multi-head self attention (MHA) mechanism on the combined features to detect ADRs from tweets. The main contributions of this paper are given below.

- We propose a novel method to extract causal features for a sequence labeling-based ADR detection.
- We propose a causality driven ADR detection framework that combines causality features and word-level semantic features and applies multi-head self attention to detect ADRs in tweets.
- We compare our proposed ADR detection approach with several existing ADR detection approaches and show that our approach produces relatively stable and robust results on two benchmark datasets.

The rest of the paper is organized as follows. Section 2 reviews the existing approaches from the literature. Section 3 presents our causality driven approach to adverse drug reaction detection. Section 4 demonstrates the superiority of our approach by conducting extensive experiments with two benchmark datasets. Finally, Sect. 5 concludes our work and presents future research direction.

2 Related Work

The existing approaches to automatic ADR detection from social media short text including tweets can be grouped into two broad categories: ADR Signal detection and sequence labeling-based ADR detection.

2.1 ADR Signal Detection

The approaches in this category aim to detect signals that indicate an ADR mention in a text. Yang et al. [17] propose an approach to detect ADR signals from tweets. In this approach, the authors apply the association mining technique to detect ADRs. In association rule mining, the frequently co-occurring words are considered to be an indication of association with each other. The authors also used proportional reporting ratios (PPR) [4,9] which is a statistical approach to calculate the association strength of a drug and an ADR. A major drawback of the approach is that it only considers the co-occurrence of drugs and adverse reactions irrespective of any semantic relation conveyed between them. Another association rule mining-based approached is proposed in [13]. This approach proposes an end-to-end multi-drug ADR detection system. Firstly, the authors apply an association rule mining technique to generate drug-ADR signals. Then, a pruning technique is used to discard unimportant associations. Finally, a contextual association clustering technique is used to detect multi-drug ADRs.

Some recent approaches apply machine learning (ML) techniques to detect ADR signals. Huynh et al. [8] propose a neural network (NN) based approach to detect ADRs from the text. The authors propose Convolutional Recurrent NN

and Recurrent Convolutional NN, where word embeddings are used as features for learning. Bollegala et al. [1] propose another ML-based technique to extract causality patterns to detect ADR signals in Tweets. In this approach, a set of lexical patterns are extracted from a set of manually annotated tweets using the skip-gram technique. The lexical patterns are then used to generate feature vectors. Finally, a linear SVM model is trained on the feature vectors to detect the causal relationship between a drug and an adverse reaction. However, this approach cannot detect more than one drug-ADR relations in a single post.

2.2 Sequence Labeling-Based ADR Detection

Some approaches in the literature aim to detect the exact locations of the ADR words in a text. These approaches consider ADR detection as a sequence labeling problem. Song et al. [14] propose a sequence labeling-based ADRs detection approach. The authors consider adverse reactions as named entities and they apply conditional random field (CRF) [10] to label ADRs in tweets. The CRF model is trained on the lexical features and the contextual features such as n-grams and part-of-speech (POS) tags. However, this approach does not identify the drug-ADR relationships.

Many recent approaches apply a neural network-based technique to detect ADRs. Chowdhury et al. [2] propose a multi-task framework to detect both ADR and indications. The framework, which is based on Recurrent Neural Network (RNN), contains a binary classifier to predict whether a text has ADR or not. Additionally, the framework contains two sequence labeling models to label ADRs and indications. Another neural network-based approach [3] aims to label ADRs on tweets. The proposed method in this approach includes a Bidirectional Long Short-Term Memory (BLSTM) [6], which uses a forward RNN and a reverse RNN in its core. However, the experiments are performed on a small dataset.

3 Our Approach

This section presents our causality driven approach to adverse drug reaction detection from tweets in detail.

3.1 Problem Formulation

We consider an adverse drug reaction (ADR) as an event that is caused by another event of taking one or many drugs. We assume that there exists a causal relationship between taking drugs and ADRs. In the following example: "I was sucking on this Lozenge that's supposed to numb sore throats and now I can't feel my mouth", *can't feel my mouth* is an adverse reaction of the drug *Lozenge*. We aim to detect such ADRs in a sequence of text in a tweet by applying causal inference. Our research question is as follows:

RQ: *How can we apply causal inference to extract ADRs from tweets?*

Words	Ugh	the	Olanzapine	makes	me	so	tired	:(
Labels	O	O	O	O	O	O	I-ADR	O

Fig. 2. The words in a tweet are labeled with 'I-ADR' and 'O' labels

Let us denote a tweet as τ which contains a sequence of words $W = [w_0, w_1, ..., w_{n-1}]$, where n is the number of words in τ. We aim to develop a function f that will take W as input and detect corresponding labels for each word. Formally, we define the problem studied in this paper as follows.

$$f(W) = L \tag{1}$$

$$\forall L_i \in L : L_i = \begin{cases} \text{I-ADR} & \text{if } w_i \text{ is an ADR,} \\ \text{O} & \text{otherwise} \end{cases} \tag{2}$$

where L is an output sequence generated by f that contains a sequence of labels of the corresponding words in W. The labels 'I-ADR' (Inside-ADR) and 'O' (Outside-ADR) denote the ADR words and the non-ADR words respectively. For example, consider the words in the first tweet as given in Fig. 1, the corresponding labeling of the words in this tweet is illustrated in Fig. 2.

3.2 Proposed ADR Detection Model

In this section, we describe our ADR detection model that combines both causal features and word-level features and apply a multi-head self attention (MHA) mechanism in four modules including the label prediction module as follows.

- **Causal Features Extraction Module:** This module captures the causal features for each candidate ADR word in a tweet. Here, we assume that a drug name in a tweet is the *cause* and other words are candidate ADR words (i.e., effect of the cause).
- **Word Features Extraction Module:** This module extracts the local contextual features by applying BLSTM on the word-level semantic features. We exploit word embeddings to extract word-level semantic features from the words in a tweet.
- **Attention Mechanism Module:** The features extracted by the first two modules are combined and used as input to the third module. This module aims to detect ADR words by capturing the interactions between them via multi-head self attention mechanism.
- **Label Prediction Module:** The labels for a sequence of words in a tweet is generated by this module. The output of the attention mechanism module is used as the input to this module.

The internal procedures of each of the module are described in detail below.

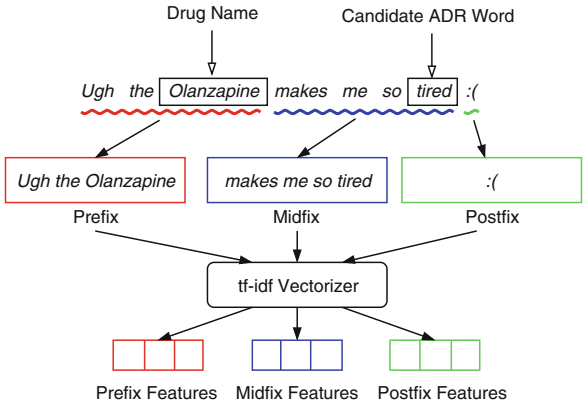


Fig. 3. Split of a tweet into segments: prefix, midfix and postfix, and the corresponding prefix tf-idf based prefix, midfix and postfix features extraction

Causal Features Extraction Module. We assume that when a tweet has a drug name, any other word in the tweet can be an ADR word for that drug. In other words, we consider every word except the drug name as a candidate ADR word. Also, a drug name is considered to be the cause and all other words are considered to be the potential effect in our approach. To detect such relationships, we aim to extract patterns from the occurrence of words around drug names and the candidate ADR words.

- First, we split a tweet into three segments: prefix, midfix, and postfix. Prefix contains the words starting from the beginning of the tweet until the drug name or the candidate ADR word whichever comes first. Midfix contains the words between the drug name and the candidate ADR words. Postfix contains the words after the drug name or the candidate ADR words, whichever comes last, until the end of the tweet. Assume that w_i is a drug name and w_j is a candidate ADR word in a tweet τ with n number of words, where the words are represented as $W = [w_0, w_1, ..., w_{n-1}]$. For the candidate ADR word w_j, the prefix $W_{pre} = [w_0, w_1, ..., w_i]$, midfix $W_{mid} = [w_{i+1}, w_{i+2}, ..., w_j]$, and postfix $W_{post} = [w_{j+1}, w_{j+2}, ..., w_{n-1}]$. For example, consider the words in the first tweet as given in Fig. 1, the corresponding prefix, midfix and postfix segmentation of this tweet is illustrated in Fig. 3.
- Then, we convert W_{pre}, W_{mid}, and W_{post} into tf-idf feature vectors v^d_{pre}, v^d_{mid}, and v^d_{post} respectively, where d is the size of the dictionary. The vectors are then concatenated together to prepare a single feature vector $v^m_{c_j}$, where $m = 3d$. Similarly, every word in a tweet τ is represented as a vector v^m_c. The vectors are then combined together to prepare a matrix $M_c \epsilon R^{n' \times m}$, where n' is the maximum length of word sequence in the dataset. If the number of words $n < n'$, then we add padding to avoid generating variable length matrices. We pass M_c to a one dimensional convolutional neural network (CNN) followed by a ReLU activation layer, which generate a causal feature

matrix $M'_c \epsilon R^{n' \times m'}$ as follows:

$$M'_{c\{i,j\}} = ReLU(I_{i,j} \times K_r + b_r) \tag{3}$$

where I is the output from CNN; m', K_r and b_r are the dimension of the causal features, the kernel and the bias parameters, respectively.

Word Features Extraction Module. We use word embedding features generated by a publicly available pre-trained Word2vec model [5] to capture the word-level semantic features. The model is trained on more than 400 million domain-independent tweets.

- First, a tweet τ is tokenized into word sequences. The word sequences are then padded by a default token if $n < n'$, where n is the number of words in τ and n' is the maximum length of the sequence in the dataset. The padding is essential to avoid generating a variable-length feature matrix. The word sequences are then replaced by their corresponding indices by looking up in a dictionary. The dictionary contains the words in our dataset and the corresponding indices. Using this dictionary, τ is converted to a vector $v_w^{n'}$. In the next step, the index vector is used to extract word embeddings from a pretrained Word2vec model [5] and the tweet τ is converted into a matrix $M_w \epsilon R^{n' \times l}$, where l is the dimension of the word embeddings.
- The word embeddings feature matrix M_w is then passed to a BLSTM to extract local contextual features from the word sequence. The output of the BLSTM layer is another matrix $M'_w \epsilon R^{n' \times l'}$, where l' is the length of word-level features.

Attention Mechanism Module. In our model, we use a Multi-head self attention (MHA) layer to detect ADR labels. We apply the MHA mechanism proposed by Vaswani et al. [15] that can capture word to word interactions in a word-sequence. We combine the causal features M'_c, word-level contextual features M'_w to create an augmented matrix by concatenating them together as follows:

$$M \epsilon R^{n' \times k} = (M'_c \,|\, M'_w) \tag{4}$$

where $k = m' + l'$. This combined feature matrix M is passed to the MHA layer. In general, not every word in a word sequence contributes equally to decide whether a particular word is an ADR or not. MHA mechanism allows us to capture the interactions between words. This feature allows us to learn what other words to focus on while predicting the label for a word. This layer produces another matrix $M' \epsilon R^{n' \times k}$, which is passed to the label prediction module where the ADR labels are predicted.

Label Prediction Module. This module consists of a fully connected layer with softmax as the activation function. The matrix M' of the attention mechanism module is passed to this layer where the softmax function finally outputs

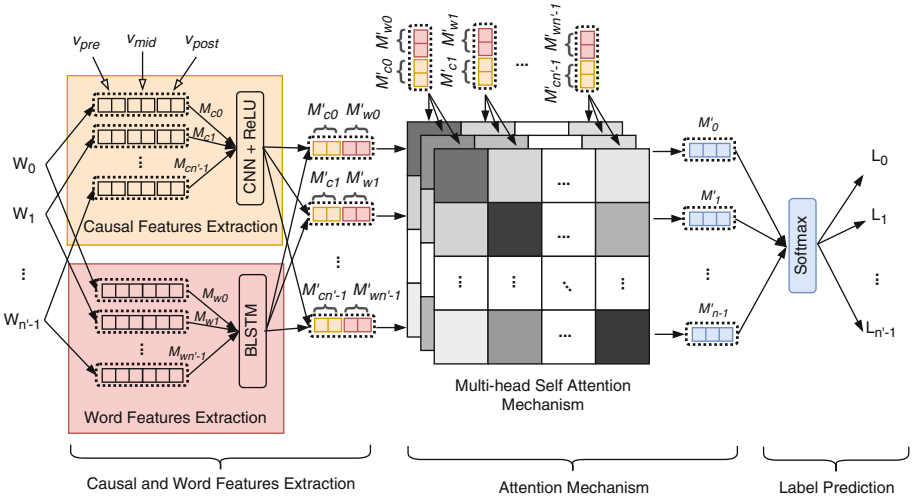


Fig. 4. Architecture of the proposed ADR detection system

label L_i of each word w_i as follows:

$$L_i = softmax(M_i' k_s + b_s) \tag{5}$$

where k_s and b_s are the learnable parameters.

The overall architecture of our proposed ADR detection system in tweets consisting of the above modules is illustrated in Fig. 4.

4 Experiment

This section presents our experimental results and demonstrates the superiority of our approach to ADR detection from tweets.

4.1 Dataset and Experimental Settings

We use two benchmark datasets to evaluate our approach. The first dataset, we refer to it as ASU_CHUP, is published by Cocos et al. [3]. This dataset is an updated version of the Twitter ADR Dataset (v1.0) [12]. The other dataset, we refer to it as SMM4H, is used in Shared Task 2 of Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task 2019[1]. Both of the datasets are labeled by human annotators. We use 75% data for training and 25% data for the test as experimented in [3]. We also use 10% of the training data as the validation data similar to [3]. The validation data is used for the hyperparameter optimization of the ADR detection models experimented in this paper. Table 1 presents the summarized statistics of the tested datasets.

[1] https://healthlanguageprocessing.org/smm4h/challenge/.

Table 1. Test datasets statistics

Datasets	Tweets(ADRs)	
	Training set	Test set
ASU_CHUP dataset	585(492)	206(172)
SMM4H dataset	1487(1368)	496(464)

In the causal feature extraction, we apply a CNN with 120 filters and kernel size 5. We use ReLU as the activation function in the CNN. The CNN layer is followed by a fully collected layer of 16 nodes where we use ReLU as the activation function in each node. In the word-level features extraction, we use 400-dimension pretrained word embeddings [5] and a BLSTM with 80 units. We set tanh as the activation function and dropout to 0.1 in the BLSTM model. The multi-head self attention model has 41 heads, which is the maximum number of words in a tweet in our dataset ($n' - 1$). We optimize the combined model using the RMSprop optimizer [7] function and the categorical crossentropy loss function [11]. We also use *accuracy* as the metric to optimize the model during training. We evaluate our approach by calculating the approximate match score [12], which is used for evaluating sequence labeling-based methods [3]. We report average precision, recall, and f1-score for 5 runs with different random seeds.

4.2 Performance Evaluation

We evaluate our causality-driven ADR detection approach against several benchmarks methods. The benchmark approaches that we have implemented in our experiments are described below.

- **CRF** [10]: We implement a conditional random fields (CRF) model on 400-dimension word-embedding features.
- **Cocos et al.** [3]: This approach applies a BLSTM to predict ADR labels, where the word embedding is used as the feature set.
- **BLSTM**: This benchmark approach applies a BLSTM on the word embeddings features to generate the word-level features. Additionally, character features and causal features are used in this approach. To generate the character features, we convert each word into a one-hot-vector of characters and generate embeddings of 100 dimensions. We apply a CNN on the word embeddings and the hidden state output of the CNN model is used as the character features. The schematic diagram of the approach is illustrated in Fig. 5. The causal features are extracted as described in Sect. 3.2. Finally, the character features, word features, and causal features are concatenated together and passed to a fully connected layer.
- **BLSTM+CharMHA**: This approach concatenates word features and character features and then applies an MHA layer.
- **BLSTM+CharCausalMHA**: In this approach, word features, character features, and causal features are concatenated together and then an MHA layer is applied.

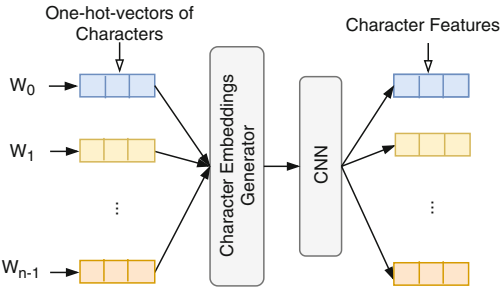
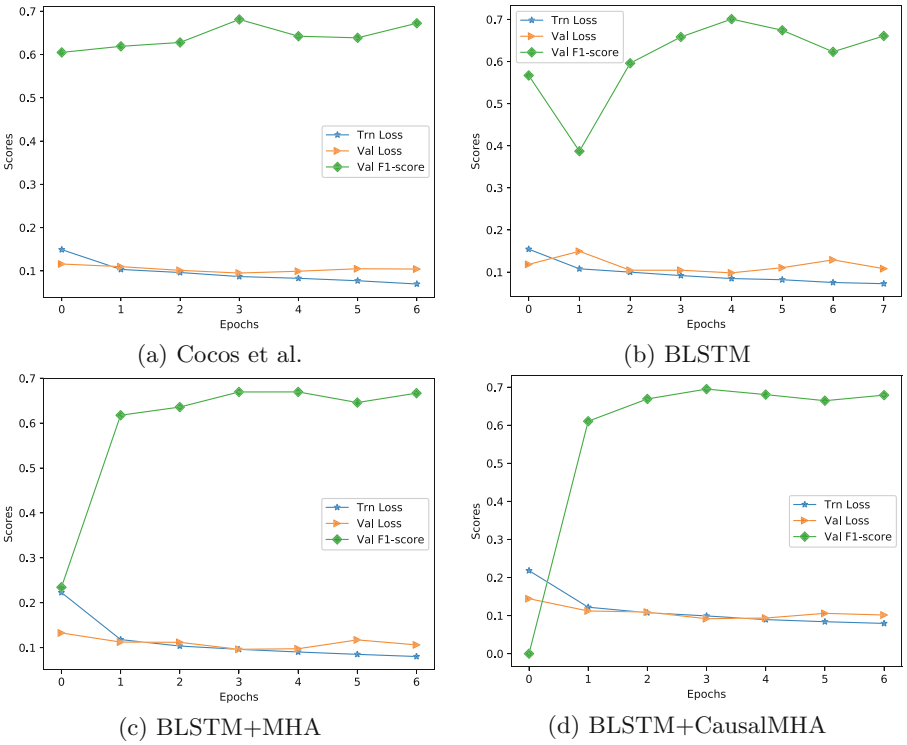


Fig. 5. Generating character features from each word in a tweet

- **BLSTM+MHA**: This is the closest variant of our proposed model, which applies an MHA model on the word features but it does not use the causality features.
- **BLSTM+CausalMHA**: This is our approach as described in Sect. 3.2.

Parameter Optimization. We optimize loss function while training using the training data and the validation data. We run our models for a maximum of 30 epochs and store each model but stop training as soon as the model starts to overfit. The best model in terms of f1-scores on the validation data is used on the test data. Figure 6 displays training loss, validation loss and f1-scores for different epochs of Cocos et al., BLSTM, BLSTM+MHA and BLSTM+CausalMHA approaches. For Cocos et al., the model starts overfitting after the $3rd$ epoch, whereas for all of the other approaches, the model starts overfitting after the $4th$ epoch as the validation loss and training loss starts diverging.

Results. Table 2 shows that our proposed method BLSTM+CausalMHA outperforms the benchmark approaches on the SMM4H dataset in terms of recall and f1-scores while keeping the precision comparable. We report that our approach achieves more than 5% improvement in terms of recall and more than 1% improvement in terms of f1-score over the closet performing benchmark BLSTM+MHA. In the smaller datasets ASU_CHUP, our approach is comparable to the top-performing existing approach *Cocos et al.* in terms of f1-score. However, this approach performs poorly on the SMM4H dataset as its f1-score is significantly less than our approach. We also compare our proposed approach BLSTM+CausalMHA with *Cocos et al.* and BLSTM+MHA on the reduced training data of the SMM4H dataset to show that the improvement of recall and f1-score is consistent. While keeping the test set the same, we reduce the number of training data to 25% and 50% of the original training data. Figure 7 demonstrates that our approach BLSTM+CausalMHA outperforms these two benchmark approaches in terms of both recall and f1-score.



(a) Cocos et al. (b) BLSTM

(c) BLSTM+MHA (d) BLSTM+CausalMHA

Fig. 6. Optimization of loss function

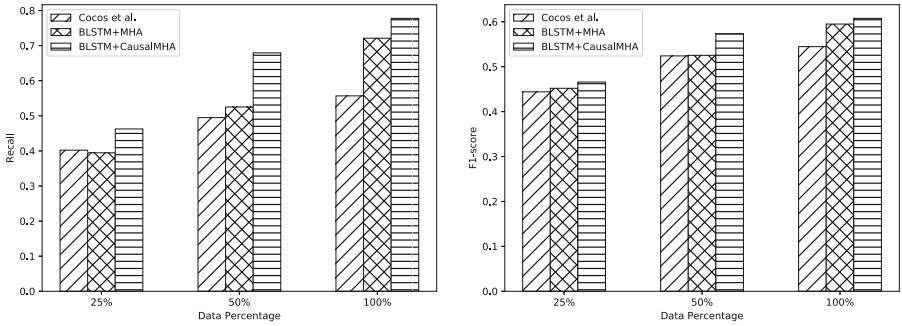
Effectiveness of Causal Features. We find that the application of causal features improves ADR detection performances on both ASU_CHUP and SMM4H datasets. The results between BLSTM+MHA and BLSTM+CausalMHA, where the only difference is causal features, are reported in Table 2. Table 2 demonstrates that we get higher f1-scores on both of the tested datasets if causal features are used with word-level semantic features. Additionally, Fig. 7(b) shows that the f1-scores of BLSTM+CausalMHA is higher than BLSTM+MHA for 25%, 50% and 100% training sets. These verify the consistency of the effectiveness of causal features in different sizes of the training sets.

We also investigate the effectiveness of causal features when combined with both character features and word features in ADR detection. In Table 2, we compare BLSTM+ CharCausalMHA and BLSTM+CausalMHA, where BLSTM+ CharCausalMHA applies character features in addition to causal and word features. We observe that the character features negatively impact the results. This is because, character features seems to add more noise than useful information.

Case Study. To demonstrate the effectiveness of our proposed approach BLSTM+ CausalMHA which applies causal features in addition to word-level

Table 2. Comparison of results of our approach BLSTM+CausaMHA with existing approaches

Approaches	ASU_CHUP dataset			SMM4H dataset		
	Precision	Recall	F1-score	Precision	Recall	F1-score
CRF	**0.8224**	0.5175	0.6348	**0.5510**	0.4218	0.4771
Cocos et al.	0.6610	**0.8313**	0.7242	0.5459	0.5568	0.5447
BLSTM	0.6869	0.7500	0.7119	0.5466	0.5851	0.5600
BLSTM + CharMHA	0.7226	0.5725	0.6037	0.5357	0.6561	0.5882
BLSTM + CharCausalMHA	0.7324	0.6688	0.6563	0.5337	0.6600	0.5829
BLSTM + MHA	0.7197	0.7550	0.7162	0.5091	0.7211	0.5950
BLSTM + CausalMHA	0.7287	0.7250	**0.7241**	0.5037	**0.7777**	**0.6077**



(a) Recall (b) F1-score

Fig. 7. Evaluation of Cocos et al., BLSTM+MHA and BLSTM+CausalMHA (our approach) on the reduced training sets

features to detect ADRs from tweets, we present several tweets labeled by BLSTM+MHA, Cocos et al. and our approach BLSTM+CausalMHA. We preserve user privacy by replacing usernames by the '<user>' tag. We also replace the names of the medicine by the '<medicine>' tag. We present two tweets below for which our approach BLSTM+CausalMHA can detect the ADR words correctly but BLSTM+MHA and Cocos et al. cannot capture the ADRs.

– **Tweet 1:** *<user>: playing league while on <medicine> is too damn stressful. back to public i go!*
True Tags: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'I-ADR', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
Cocos et al.: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
BLSTM+MHA: ['O', 'O', 'O', 'I-ADR', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O']

BLSTM+CausalMHA: [‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘I-ADR’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
ADR: *stressful*

– **Tweet 2:** *<medicine> gives me a head high but it makes me feel like throwing up after. ugh . #thestruggle*
True Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘I-ADR’, ‘I-ADR’, ‘I-ADR’, ‘I-ADR’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
Cocos et al.: [‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘I-ADR’, ‘I-ADR’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
BLSTM+MHA: [‘O’, ‘O’, ‘O’, ‘O’, ‘I-ADR’, ‘I-ADR’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘I-ADR’, ‘I-ADR’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
BLSTM+CausalMHA: [‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘I-ADR’, ‘I-ADR’, ‘I-ADR’, ‘I-ADR’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
ADR: *feel like throwing up*

In Tweet 1, our approach can correctly detect the ADR word *stressful* because we applied causal features to capture the causal relationship between the drug name and the ADR words. However, the approaches that apply only word features, cannot detect the relationship, hence these approaches labeled the ADR words incorrectly. In Tweet 2, our approach can detect the whole phrase *feel like throwing up* as the ADR but BLSTM+MHA and Cocos et al. can only detect the ADR phrase partially.

– **Tweet 3:** *bedtime. hopefully the headache and nausea from the <medicine> will be gone leaving me with only the more common side effects of antibiotics.*
True Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘I-ADR’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
Cocos et al.: [‘O’, ‘O’, ‘O’, ‘I-ADR’, ‘O’, ‘I-ADR’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
BLSTM+MHA: [‘O’, ‘O’, ‘O’, ‘O’, ‘I-ADR’, ‘O’, ‘I-ADR’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
BLSTM+CausalMHA: [‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘I-ADR’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’]
ADR: *nausea*

In Tweet 3, all three approaches were able to correctly label the ADR word *nausea*. However, BLSTM+MHA and Cocos et al. also labeled *headache* as an ADR word. Although headache is a word that describes an undesirable health condition, in this tweet, it is not used as an ADR word. Rather *headache* was the reason for undertaking the drug. In this scenario, only our approach was able to predict a word as a non-ADR word. The next two tweets are examples where our approach struggles to capture ADRs.

– **Tweet 4:** *<user> <user> ugh. my partner was on <medicine> for a while. it made him hide in the apartment. :/*
True Tags: [‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘O’, ‘I-ADR’, ‘I-ADR’, ‘I-ADR’, ‘I-ADR’, ‘O’, ‘O’]

Cocos et al.: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'I-ADR', 'O', 'O', 'O', 'O', 'O']
BLSTM+MHA: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'I-ADR', 'O', 'O', 'I-ADR', 'O', 'O']
BLSTM+CausalMHA: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'I-ADR', 'O', 'O', 'I-ADR', 'O', 'O']
ADR: *hide in the apartment*

– **Tweet 5:** <*medicine*> *is definitely kicking my ass but my skin lools great.*
True Tags: ['O', 'O', 'O', 'I-ADR', 'I-ADR', 'I-ADR', 'O', 'O', 'O', 'O', 'O', 'O']
Cocos et al.: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'I-ADR', 'O', 'O', 'O']
BLSTM+MHA: ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O','I-ADR', 'O','O', 'O']
BLSTM+CausalMHA: ['O', 'O', 'O', 'I-ADR', 'O', 'O', 'O', 'O', 'I-ADR', 'O', 'O', 'O']
ADR: *kicking my ass*

Tweet 4 contains a sequence *hide in the apartment* as ADRs which contains a preposition *in* and an article *the*. Both BLSTM+CausalMHA and BLSTM+MHA can detect the words *hide* and *apartment* as the ADR while Cocos et al. can only detect *hide* as the ADR word. One reason behind this is that causal features and word features for prepositions and articles are less informative, which makes it difficult for the models to decide in which context the words should be labeled as ADR. In Tweet 5, the ADR words are *kicking my ass*. Only our approach detected *Kicking* as an ADR, although it misses the other two words, where there is a pronoun *my*. Whereas, all three approaches incorrectly label *skin* as an ADR. From this analysis, we can conclude that capturing the long ADR phrases with less meaningful words, such as articles, prepositions and pronouns is a challenging task and it could be a possible future extension of this work.

5 Conclusion

In this paper, we propose a sequence labeling based ADR detection approach which applies causal features with word-level semantic features to detect ADRs from tweets. The causal features are extracted for every word after splitting a tweet into prefix, midfix, and postfix. The causal features are combined with word-level semantic features, which is extracted from word embeddings, and then a Multi-head self attention mechanism is applied in our approach to detect ADRs from tweets. We show that the proposed approach works for both small and large size of datasets. The proposed approach can only label whether a word is an ADR but it cannot identify the boundary or hierarchical structure of ADRs in text, e.g. whether an ADR word inside another ADR. We consider this an important research direction for the future work of this paper.

References

1. Bollegala, D., Maskell, S., Sloane, R., Hajne, J., Pirmohamed, M.: Causality patterns for detecting adverse drug reactions from social media: text mining approach. JMIR Public Health Surveill. **4**(2), e51 (2018)
2. Chowdhury, S., Zhang, C., Yu, P.S.: Multi-task pharmacovigilance mining from social media posts. In: WWW (2018)
3. Cocos, A., Fiks, A.G., Masino, A.J.: Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. JAMIA **24**(4), 813–821 (2017)
4. Evans, S., Waller, P.C., Davis, S.: Use of proportional reporting ratios (PRRS) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol. Drug Saf. **10**(6), 483–486 (2001)
5. Godin, F., Vandersmissen, B., De Neve, W., Van de Walle, R.: Multimedia lab @ acl wnut ner shared task: named entity recognition for Twitter microposts using distributed word representations. In: Proceedings of the Workshop on Noisy User-Generated Text, pp. 146–153 (2015)
6. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
7. Hinton, G., Srivastava, N., Swersky, K.: Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Neural networks for machine learning, Coursera lecture 6e (2012)
8. Huynh, T., He, Y., Willis, A., Rüger, S.: Adverse drug reaction classification with deep neural networks. In: COLING, pp. 877–887 (2016)
9. Ji, Y., et al.: A potential causal association mining algorithm for screening adverse drug reactions in postmarketing surveillance. IEEE Trans. Inf Technol. Biomed. **15**(3), 428–437 (2011)
10. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: ICML, pp. 282–289 (2001)
11. LeCun, Y., et al.: Handwritten digit recognition with a back-propagation network. In: Advances in Neural Information Processing Systems, pp. 396–404 (1990)
12. Nikfarjam, A., Sarker, A., O'Connor, K., Ginn, R., Gonzalez, G.: Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. JAMIA **22**(3), 671–681 (2015)
13. Qin, X., Kakar, T., Wunnava, S., Rundensteiner, E.A., Cao, L.: Maras: signaling multi-drug adverse reactions. In: KDD, pp. 1615–1623 (2017)
14. Song, Q., Li, B., Xu, Y.: Research on adverse drug reaction recognitions based on conditional random field. In: International Conference on Business and Information Management, pp. 97–101 (2017)
15. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
16. Yang, C.C., Jiang, L., Yang, H., Tang, X.: Detecting signals of adverse drug reactions from health consumer contributed content in social media. In: Proceedings of ACM SIGKDD Workshop on Health Informatics. ACM (2012)
17. Yang, C.C., Yang, H., Jiang, L., Zhang, M.: Social media mining for drug safety signal detection. In: International Workshop on Smart Health and Wellbeing, pp. 33–40 (2012)