IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 17, 2020, accepted May 4, 2020, date of publication May 7, 2020, date of current version May 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2993169

An Effective Emotional Expression and Knowledge-Enhanced Method for Detecting Adverse Drug Reactions

ZHENGGUANG LI[1,2], HONGFEI LIN[1], AND WEI ZHENG[2]
[1]College of Computer Science and Technology, Dalian University Of Technology, Dalian 116023, China
[2]Software Institute, Dalian Jiaotong University, Dalian 116028, China

Corresponding author: Hongfei Lin (hongfeilin@dlut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772103, Grant 61572102, Grant 61702080, and Grant 61771087, in part by the Natural Science Foundation of Liaoning Province, China, under Grant 20180551078, and in part by the Open Project Foundation of Intelligent Information Processing Key Laboratory of Shanxi Province under Grant CICIP2018006.

ABSTRACT Discovering more time-effective and a wider range of adverse drug reactions (ADRs) from social texts related to feelings concerning taking medication has recently received significant interest in pharmacovigilance research. Recognizing the posts that include ADRs is an important step for detecting ADRs from social texts. The existing systems show the unsatisfactory performance due to the insufficient expression of emotions and the inadequacy of information expression in short social texts. Although these systems exploit emotional features to improve the performance of their methods, the representation of word-level emotional scores is insufficient for emotional expression. Moreover, most of the systems make less use of medical knowledge to enhance the detection of the potential relationship between drugs and adverse reactions in posts. Therefore, enough expression of emotion and medical knowledge in sparse medical social texts may be explored to improve system performance. This paper proposed an effective method integrating sufficient emotional expression and medical knowledge to detect ADRs from medical tweets. First, the proposed method utilized sentence-level emotional context and word-level emotional score to learn sufficient emotional information for distinguishing between ADR and non-ADR tweets. Furthermore, a co-occurrence dictionary of each drug and its relevant ADRs was constructed by means of a medical resource (MedDRA) and drug site (www.drug.com) to help the proposed model focus on posts containing drugs and ADRs. Finally, a convolutional neural network (CNN) model on the basis of bidirectional encoder representations from transformers (BERT) performed the classification task. The proposed model achieved better overall performance than the other existing methods on two Twitter datasets (F1-scores of 72.64% and 64.98% on PSB2016 and SMM4H, respectively).

INDEX TERMS Adverse drug reaction, medical knowledge, emotional context, co-occurrence dictionary, social text.

I. INTRODUCTION

More than 50 million posts are published every day according to Twitter's official reports. Therefore, Twitter provides rich large-scale multimedia data for various research opportunities [1] involving ADR detection, which focuses on automatically classifying ADRs (positive and negative) given the post content. ADR detection from social texts is an important task for discovering ADRs [2] due to the limitations of clinical experiments. Since ADRs may be exposed when people share their feelings concerning taking medication on social media, social texts may contain more time-effective and a wider range of ADRs. However, due to the colloquialism of social texts and the sparseness of posts including descriptions of ADRs or drugs, some approaches that perform well in other written biomedical texts such as PubMed cannot be directly used in social texts. Hence, researchers have attempted to find ADRs in social texts. Text mining and partially supervised learning methods [3] are integrated to classify ADR (positive instances) and non-ADR messages (negative instances), and researchers employ various features such as word embedding [4], position feature [5] and medical knowledge [6] to

The associate editor coordinating the review of this manuscript and approving it for publication was Kin Fong Lei.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

promote the whole performance of their methods. Moreover, researchers utilize attention mechanisms [7], transfer learning [8], co-training learning [9], broad learning [10] and multi-task learning [11] to learn these deep dominant features [12]. Medical resources and emotional score are merged into features that represent the semantic meaning of the text segments of different methods. However, it is difficult to automatically capture the semantic representation of short social texts. Thus, it is more important for short social texts to enhance the ability of information representation.

People often express their abundant emotions and feelings in social media posts. Therefore, the innate emotional elements that are implicated in social texts are an important cue in detecting ADRs. Some studies introduce the emotional analysis of social texts collected by the crawler method [13]. In addition, term frequency-inverse document frequency (TF-IDF) as an emotional feature [14] is used for the ADR detection task. The score of emotional words [15] is exploited to find posts containing ADRs from social texts. However, the extant experimental results show that word-level emotional scores are insufficient for capturing richer emotional expression [16]. Moreover, researchers [14], [17] have suggested that sentimental analysis is effective in extracting ADRs from social texts. In fact, some emotions are often implied in the whole semantic representation of posts. For instance, one post stated, "I reaaaallly need to take my Paxil, but it makes me feel so delirious and just messed up"; the obvious negative emotions may be found even if the emotional words do not exist in the post. The whole emotional representation of posts may contribute to further distinguishing between ADR and non-ADR posts.

Moreover, a standard medical knowledge base such as the Unified Medical Language System (UMLS) has been employed in prior studies to detect mentions of ADRs. Adverse event entities have been extracted from patient forums [2] using drug safety databases [18] such as MedEffect and COSTART. Recent studies have also adopted MedDRA and SIDER to better understand and match users' expressions of drugs and ADRs in social media [19]. These methods only utilize the medical resource to supplement features rather than enhance information representation for short social texts, but they overlook the characteristics of insufficient information representation due to text length limitation.

To tackle the aforementioned limitations, we propose an effective emotional expression and knowledge-enhanced method, which integrates the word-level emotional score and sentence-level emotional context information. Moreover, the proposed model enhances the potential relationship between drugs and adverse reactions via medical resources. Inspired by BioBERT [20], a pre-trained biomedical language representation model for biomedical text mining, we pre-train a new BERT using a large-scale sentimental analysis corpus to extract sentence-level emotional context information from tweets. The word-level emotional score is calculated by the sentimental dictionary and regarded as the weight coefficient of the subsequent input. The tweets considered for discovering ADRs generally contain at least one drug name. Then, posts with co-occurrence drug names and adverse reactions are the main objectives in the extraction of adverse reactions. Medical resources such as MedDRA and DrugBank [21] facilitate the construction of co-occurrence pairs for drugs and their adverse reactions. In the paper, we mainly build the drug-ADR co-occurrence pairs dictionary via MedDRA and supplementary ADR, crawling the drug-related data from the drug website. In addition, the extracted drug-ADR pairs generated by the established co-occurrence dictionary as the co-occurrence sub-sentences are fed into the model, contributing to focusing on the key drug names and adverse reactions, thus improving model performance. The experimental results demonstrate that the drug-ADR co-occurrence pairs increase the recall rate while guaranteeing precision as much as possible. In addition, the word-level emotional score and sentence-level emotional context information help the model promote its overall performance.

The main contributions of the paper are summarized as follows:

- The emotional context information extracted by our pre-trained BERT is introduced into our neural network architecture. This contributes to extracting the positive or negative emotions for distinguishing ADR posts from non-ADR posts, resulting in the promotion of the whole performance. Furthermore, the word-level emotional score as the weight coefficient of words contributes to discovering the ADRs associated with potential emotional words.
- The co-occurrence sub-sentences generated by the drug-ADR co-occurrence dictionary clearly specify on what the model should focus. These co-occurrence pairs improve the accuracy of positive example classification and lead to an increase in recall rate.

State-of-the-art results are obtained on two real-world Twitter datasets (PSB2016 and SMM4H, with F1-scores of 72.64% and 64.98%, respectively) compared to other methods in pharmacovigilance.

II. RELATED WORK

Social texts contain not only abundant emotions but also people's feelings after taking medicines. Researchers use social texts to conduct emotional analysis and detect ADRs.

A. SENTIMENTAL ANALYSIS IN SOCIAL MEDIA

Sentimental analysis involves various research fields such as product recommendation [22], flight service [17] and opinion mining [23]. The data used in sentiment analysis are collected from online networks such as micro-blogs [24] and health forums [25]. The methods for sentiment analysis are roughly divided into the pattern- and machine learning-based approaches. Researchers extract a small number of features from domain knowledge [26] and contextual semantics [27] to train their classifier. Although these methods achieve good

results on different corpora, their limited by domain dependence. Other researchers have recently turned to studying sentimental analysis via machine learning-based [3], [28] methods. Combining CNN and LSTM [2], an attention mechanism [29], BERT [30] and Emoji embedding [1] are successively applied in the sentimental analysis research, greatly improving performance. Researchers also find a potential relationship between ADRs and emotional analysis in social texts [31]. They employ such features as emotional score and emotional word frequency [2]–[4] to classify and extract ADRs. Moreover, researchers also analyse in depth the contribution of sentimental analysis to ADRs [14], [16], [32]. Therefore, deep potential emotional analysis features may enhance the performance of the detection of ADRs from social texts.

B. AUTOMATIC ADR DETECTION FROM SOCIAL TEXTS

In addition to traditional feature-/kernel-based approaches [33], [34], several neural models are proposed to detect ADRs from social texts in PSB Tasks 1 and 2 [4], including embedding-based models, semi-supervised CNN-based models [35] and RNN-based models [36]. Recently, attentive RNN [29], [37] has also been used to improve the performance of identified ADRs. Multi-head self-attention with various features [38] has some advantages over CNN, CRNN and CNN with an attention mechanism on ADR tweet classification. Transfer learning [8], co-training [9] and multi-task learning [39] are adopted to extract ADRs, classify tweets mentioning ADRs and normalize ADRs concept, and multi-task learning achieves the state-of-the-art result. With BERT performing well in many NLP tasks, researchers introduce the knowledge base and conditional random field (CRF) into BERT for the automatic classification of ADRs (text classification) and extraction of ADRs (NER) on SMM4H Shared Task 2019 [19], respectively.

III. METHODS

In this study, the collected social texts for detecting ADRs usually contain at least one drug name, which co-occurs with some symptoms regarded as ADRs, which is different from other text classification datasets. Therefore, the drug-ADR co-occurrence sub-sentence as the auxiliary sentence is fed into basic BERT to enhance sentence representation (Section A). Social media posts contain abundant emotions and feelings. Hence, emotional elements are an important cue for detecting ADRs. The sentimental score of words multiplies the output features of basic BERT, and the product is fed into a transformer component to further extract a deep representation of sentences with the co-occurrence drug and ADRs (Section B). Moreover, our pre-trained BERT, which is obtained via pre-training a large number of emotional analysis corpus collected from Twitter, is employed to fully express emotional information (Section C). Finally, the concatenation of the output of the transformer and the [CLS] output of our pre-trained BERT are used as the input of the convolutional neural network, and the final classification result is obtained via Softmax operation (Section D). The architecture of our model is illustrated in Figure 1.

A. INPUT OF BASIC BERT

The input of basic BERT consists of the masked tweet and the drug-ADR co-occurrence sub-sentence. The drug-ADR co-occurrence sub-sentence is employed to enhance sentence representation, focusing on tweets containing drugs and ADRs. The reason for building the co-occurrence sub-sentence is that the tweets for detecting ADRs are collected from a large number of social media posts according to a pre-defined drug dictionary, and drugs usually co-occur with some symptoms regarded as ADRs in positive tweets. Therefore, first, a co-occurrence dictionary using mainly the MedDRA database containing approximately 1,430 drugs and their known side effects is constructed to extract drug-ADR co-occurrence pairs from tweets. Second, the drug name list provided by MedDRA does not fully contain the drug name list used for the experimental data when we analyse the experimental data. Hence, we crawl ''more common'' and ''less common'' content from the drug site (www.drug.com) in the form of ''https://www.drugs.com/sfx/#drug-side-effects.html'' (where #drug will be replaced with the actual crawling drug) to obtain the drug-ADR co-occurrence pairs as a supplement to the co-occurrence dictionary. Eventually, a list of 1494 drugs and their adverse reactions are obtained, and more than 33,000 drug-ADR co-occurrence pairs are extracted, as shown in Figure 2. Then, co-occurrence pairs are obtained from tweets using the above-mentioned dictionary. After the drug name is accurately found, we use the greedy algorithm to match the maximum words, which appear in the corresponding ADR part of the co-occurrence word in a tweet. For instance, ''fluoxetine #ac(h)e'' and ''citalopram #ac(h)e'' are extracted from the tweet, namely, ''@notquite-real yeah, I mean, fluoxetine made me feel like shit, and citalopram makes me feel ac(h)e, so worth considering if you ever have to''. The sub-sentence is represented as ''fluoxetine #_ac(h)e, citalopram #ac(h)e'', as taking Fluoxetine or Citalopram can cause headaches.

The final input of basic BERT is denoted as $s_1 = (w_1, w_2, \ldots w_n)$ and $s_2 = (d_1, co_1^1, co_1^2 \ldots co_1^k, \ldots d_m, co_m^1, co_m^2, \ldots, co_m^p)$. Here, s_1 is a piece of text corresponding to a social media post consisting of a sequence of n words, and each w_i represents a word in the vocabulary of size V. Moreover, s_2 is a sequence of m drug and its corresponding ADR co-occurrence pairs, called the co-occurrence sub-sentence, and each d_m and co_m^p represents a drug and its pth ADR in the vocabulary of size $m + Diff(\sum_{i=1}^{m}\sum_{k=1}^{p} co)$, respectively, where $Diff$ denotes the number of co-occurrence pairs after removing the repeated pairs in the experimental dataset.

B. WORD-LEVEL EMOTIONAL SCORE AND TRANSFORMER

Social texts usually contain more or less positive or negative emotions. Researchers utilize emotional features for such social NLP tasks as sentimental classification [40] and public

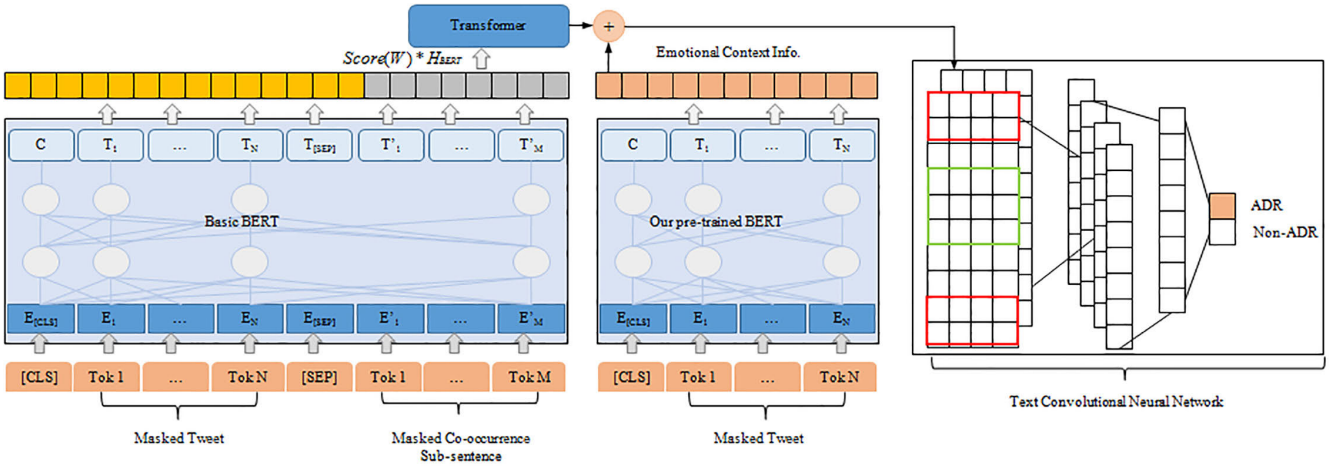


FIGURE 1. The model architecture. The BERT layer consists of basic BERT and our pre-trained BERT. The left BERT is provided by Google, while the right BERT is obtained via pre-training a large number of emotional analysis corpus collected from Twitter. Moreover, [SEP] is the separator between a tweet and the drug-ADR co-occurrence sub-sentence in basic BERT. The product between the sentimental score of words and output features of basic BERT is fed into a transformer component. Then, the output of the transformer component and the [CLS] output of our pre-trained BERT are concatenated as the input of the convolutional neural network.

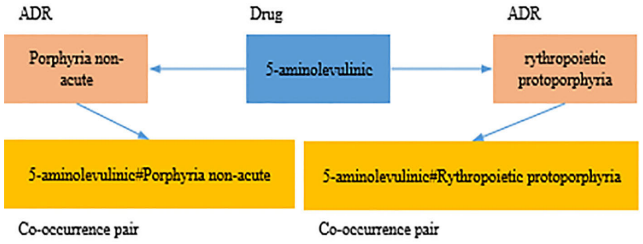


FIGURE 2. Representation of drug-ADR occurrence word.

opinion analysis [41]. Therefore, the emotional score of each word is calculated using SentiWordNet3.0 and described as follows.

$$Score(w) = \begin{cases} 1 + Score_{dict}^{neg.} & Score_{dict}^{neg.} > Score_{dict}^{pos.} \\ 1 - Score_{dict}^{pos.} & Score_{dict}^{neg.} < Score_{dict}^{pos.} \\ 1 & other \end{cases} \quad (1)$$

where $Score_{dict}^{neg.}$ and $Score_{dict}^{pos.}$ are negative and positive scores, respectively, in SentiWordNet3.0. Then, the product of $Score(w)$ and the sequence output of BERT are fed into the transformer component to further enhance the semantic representation of tweets. The output of the transformer component serves as a partial input to the downstream model.

C. SENTENCE-LEVEL EMOTIONAL CONTEXT

The innate emotional elements that are implicated in social texts are useful for social NLP tasks [32]. Nevertheless, it is insufficient that only word-level emotional scores are used to capture richer emotional expression due to some emotions implied in the whole semantic representation of posts. Hence, the emotional context information is extracted from tweets via BERT to compensate for the deficiency of word-level emotion. However, the performance of BERT mainly depends on the size and quality of the corpora on which they are pre-trained. Since BERT provided by Google is designed as a general-purpose language model that is pre-trained on the English Wikipedia and Books Corpus, the texts are more official and almost unemotional. Conversely, the datasets of our task consist of tweets, which contain richer emotion. Moreover, users' inputs are freer, more irregular and dirtier on twitter than in official texts, resulting in grammatical errors, spelling mistakes and the manual abbreviation of words. Therefore, basic BERT designed for general-purpose natural language understanding is not suitable for extracting emotional context information from tweets. To obtain the required sentence-level emotional context, we pre-train our BERT using the Sentiment140 dataset (https://www.kaggle.com), which contains 1,600,000 automatically tagged tweets (half positive and half negative). Then, the final hidden output of the BERT is taken as the sentence-level context information, which is also the partial input of our downstream model.

D. DOWNSTREAM MODEL

The downstream model, the textual convolutional neural network (CNN), is employed to further extract the dominant features, owing to the shortness of social texts and the effectiveness of CNN [35] for detecting ADRs. As shown in Figure 1, the input of the downstream model consists of two parts, namely, the output of the transformer component (left part) and the emotional context (right part). The input of the transformer is defined as $H_L = [h_1^* w_1, h_2^* w_2, \ldots, h_n^* w_n]$, where h_i is the final hidden state of basic BERT, and w_i is the emotional score of the corresponding word calculated by equation 1. Moreover, sentence-level emotional context is represented by $H_R = [h_1^{ctx}, h_2^{ctx}, \ldots, h_n^{ctx}]$. Then, H_L and H_R are concatenated as the embedding of the k^{th} tweet, i.e., $s_k = [H_L; H_R]$. Similar to TextCNN[42], multiple

convolutional kernels are used, and the output of the k[th] kernel is denoted as $\hat{h}_k = conv1d(W^*[H_L; H_R] + b)$. Then, all the outputs of different kernels are concatenated, namely, $h = concat(\hat{h_1}, \hat{h_2}, \ldots, \hat{h_m})$, where m is the number of kernels. Finally, the output of the max pooling operation is fed into the output layer, denoted as $h_o = \max pooling(h)$. The final vector can be regarded as a high-level representation of the tweet and is used as a feature for the ADR classification task:

$$y = soft\max(W_{class}h_o + b_{class}) \quad (2)$$

where W_{class} and b_{class} are learnable parameters.

The imbalance problem is a general problem in social NLP tasks. Therefore, according to the results of Wang *et al.* [28] and our analysis on the number of positive and negative examples, the imbalance ratios of (the number of negative examples vs the number of positive examples) both datasets are approximately 10:1 [43]. Inspired by Lin *et al.* [44], the balanced factor is used to make the model more focused on the unbalanced positive example. The loss function for ADR detection is described in equation 3:

$$J = -\frac{\gamma}{1+\gamma}\frac{1}{|S^+|}\sum_{i=1}^{m}(y_i^+ \log(p(y_i|w_i)) - \frac{1}{1+r}\frac{1}{|S^-|}\sum_{j=1}^{n} y_j^- \log(p(y_j|w_j))) \quad (3)$$

where $|S^+|$ and $|S^-|$ are the number of positive and negative examples, respectively, andγ is a balanced factor.

IV. EXPERIMENTS

A. DATASETS

In our experiments, two Twitter corpora are employed to verify the effectiveness of the proposed model and perform comparison experiments of the baselines. The twitter ADR dataset from PSB2016-Task1 [4] contains 10,822 tweets in the original dataset, while the dataset from Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task 2018-Task3 [19], which is also an extension of the PSB2016-Task1 dataset, consists of approximately 15,000 annotated tweets as training data and 9000 annotated test tweets, respectively. These tweets related to drugs prescribed for chronic diseases and the prevalence of drug use were annotated by two domain experts under the guidance of a pharmacology expert [2]. Both experimental corpora only provide the tweet and user IDs but do not allow for the sharing of actual raw tweet text for the purpose of protecting user privacy. Hence, we have to re-crawl the original texts using the tweet and user ID via Twitter's Service Streaming API; only 6,700 (61.9%) and 17,000(70.8%) tweets are still publicly available in PSB2016 and SMM4H2018, respectively. The dataset and source code of PSB2016 and SMM4H2018 is available at https://github.com/dllzg2012/ Co-Senti-BERTCNN.git.

B. DATA FOR CROSS-VALIDATION

There may be no positive examples in the training, validation or test sets when generating cross-validation data on the imbalanced datasets directly using the general generation method for cross-validation data. Models will lack fairness and generalization if we use the dataset without positive examples to verify the validity of the model. To solve this problem, the general cross-validation data-generation method is used to generate the training, validation and test data of the positive and negative examples. Then, the data of the positive and negative examples are combined and shuffled. The creation process of the 5-fold cross-validation data is illustrated in Figure 3.

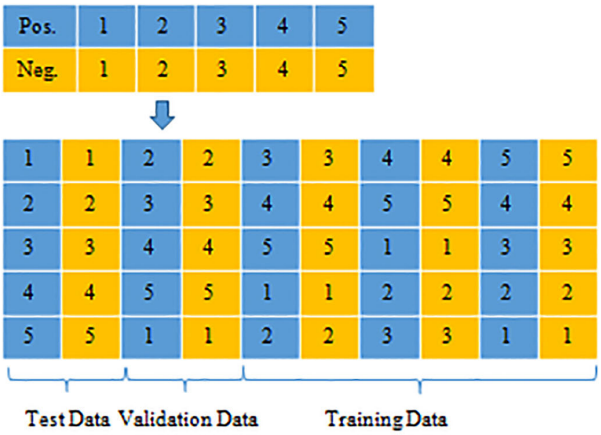


FIGURE 3. The creation process of the 5-fold cross-validation data. Two consecutive datasets are used as test and validation data, respectively, and the rest of the data are used as training data, where the positive and negative data are in the same division position.

C. TRAINING DETAILS

All models are implemented with the open-source deep learning package TensorFlow and Python3.6. Pre-trained BERT, "uncased_L-12_H-768_A-12", is provided by Google and is downloaded from https://github.com/google-research/bert via the keywords "BERT-Base, Uncased". Moreover, the word embedding, which is used for our baseline TextCNN, BiLSTM and BiGRUm can be taken from Godin, Weissenbacher *et al.* [43]. In every epoch, we perform batch training on our ADR corpus with a batch size of 12. L2 regularization is used on the non-recurrent connections, with a dropout rate of 0.5 and a scale parameter of 0.01. The proposed model adopts the Adam optimization method, with a learning rate of 0.00001 during training, and trains our BERT with an epoch number of 5. When training our model and TextCNN, kernel sizes of 2, 3 and 4 and a filter size of 32 are set.

V. RESULTS AND DISCUSSION

A. EVALUATION

Precision, recall, F1-score and AUC are used as the evaluation measures of ADR classification, as shown in equations 4-7.

$$P_{ADR} = \frac{TP_{ADR}}{TP_{ADR} + FN_{ADR}} \quad (4)$$

$$R_{ADR} = \frac{TP_{ADR}}{TP_{ADR} + FP_{ADR}} \quad (5)$$

$$F1 - score_{ADR} = \frac{2^* P^*_{ADR} R_{ADR}}{P_{ADR} + R_{ADR}} \quad (6)$$

$$AUC_{ADR} = \frac{\sum I(P_{ADR}, P_{non-ADR})}{M^* N} \quad (7)$$

where TP_{ADR} is the number of true ADR tweets, FN_{ADR} is the number of false non-ADR tweets, FP_{ADR} is the number of false ADR tweets, and M and N are the number of ADR and non-ADR tweets, respectively. $I(P_{ADR}, P_{non-ADR})$ is described as shown in equation 8.

$$I(P_{ADR}, P_{non-ADR}) = \begin{cases} 1 & P_{ADR} > P_{non-ADR} \\ 0.5 & P_{ADR} = P_{non-ADR} \\ 0 & P_{ADR} < P_{non-ADR} \end{cases} \quad (8)$$

B. MODELS

To demonstrate the effectiveness of our proposed model, we compare it against multiple baseline methods and state-of-the-art approaches for the ADR classification task.

1) TextCNN

This is the classic convolutional neural network model for performing the sentence classification task. TextCNN [42] consists of input, convolution, max pooling and full connection, and a Softmax (output) layer. This serves as a downstream model for our entire model.

2) BiLSTM AND BiGRU

Bidirectional long short-term memory networks (BiLSTM, with LSTM as the basic RNN unit) and bidirectional recurrent neural network (BiGRU, with GRU as the basic RNN unit) as a natural language process model are applied for the pharmacovigilance task [45].

3) CRNN AND CNNA

CRNN and CNNA [46] are both proposed by Trung Huynh *et al.* for ADR classification. CRNN is a convolutional neural network concatenated with a recurrent neural network with GRU as the basic RNN unit and RLU for the convolutional layer. CNNA is a CNN integrated with an attention mechanism.

4) SEMI-MULTI-CNN

Lee *et al.* [35] trains multi-model with self-collected various tweets and then uses majority vote to classify ADR and non-ADR tweets.

5) MT-ATTEN-COV

This is a state-of-the-art model for preforming ADR-related tasks on the PSB2016 corpus. MT-Atten-Cov [39] is a multi-task neural network model, learning ADR-classification, ADR-labelling and ADR-indication tasks with different levels of supervision collectively.

6) BERT+KNOWLEDGE

This is a state-of-the-art model for performing the ADR classification task on SMM4H [19]. The model builds <drug, ADR> pairs, generates binary features and then integrates the features with the output of BERT.

7) BERTCNN

This is our base model for the ADR detection task integrating BERT and CNN. We use the output of BERT as the input of the TextCNN.

8) CO-SENTI-BERTCNN

This is our proposed framework for ADR classification with the concatenation of the drug-ADR co-occurrence sub-sentence and sentence-level emotional context information. We use the output of our pre-trained BERT as the sentence-level emotional context information and the word-level emotional score as the weight of the influence of words' overall classification. Moreover, drug-ADR co-occurrence sub-sentences allow the model to pay attention to the dominant part of tweets for distinguishing between ADR and non-ADR.

C. PERFORMANCE COMPARISON WITH OTHER EXISTING METHODS

To show the validity of the proposed model, we report the results on official divided data and our divided cross-validation data of SMM4H. Table 1 demonstrates the performance comparison between our Co-Senti-BERTCNN method and other state-of-the-art methods on the PSB2016 or SMM4H corpus. Note that in our experiments, the number of tweets crawled down is not consistent with the number of tweets used by the existing methods, but the proportion of positive and negative examples remains basically unchanged. Therefore, the results are comparable to some extent. First, on PSB2016, TextCNN only achieves 42.74% precision, 50.00% recall, 46.08% F1-score and 0.7127 for AUC. The recall obtained by CNNA is increased due to the attention mechanism that contributes to focusing on positive ADR tweets. However, Semi-Multi-CNN achieves state-of-the-art results in 2017, owing to a variety of tweets, which are useful for improving the precision rate.Furthermore, MT-Atten-Cov achieves the state-of-the-art performance for employing the attention mechanism and multi-task learning in 2018, mainly because the ADR labelling task in MT-Atten-Cov contributes to promoting the recall rate, resulting in an improvement of the F1-score. Compared with MT-Atten-Cov, the proposed method reduces the precision by 2 percentage points, while the recall rate increases by 4 percentage points, which makes the F1-score reach 72.64%. The reasons may be that the co-occurrence sub-sentence helps the model focus on the positive tweets, and sentence-level context information is useful for promoting precision; namely, the two components balance the overall performance. Second, on SMM4H, the proposed model gains 0.6373, 0.6628 and 0.6498 in precision, recall and F1-score, respectively. However, compared with

TABLE 1. Performance comparison of our method and other existing methods on the PSB2016 and SMM4H corpa.

	PSB2016				SMM4H			
	P(%)	R(%)	F1(%)	AUC	P(%)	R(%)	F1(%)	AUC
TextCNN [21]	42.74	50.00	46.08	0.7127	42.86	32.57	37.01	0.6494
BiLSTM [7]	60.29	41.00	48.81	0.6900	38.68	30.28	33.97	0.6365
BiGRU [7]	58.02	47.00	51.93	0.7161	38.93	33.14	35.80	0.6496
CRNN [17]	49.00	55.00	51.00	-	-	-	-	-
CNNA [17]	40.00	66.00	49.00	-	-	-	-	-
Semi-Multi-CNN [25]	70.21	59.64	64.5	-	-	-	-	-
MT-Atten-Cov [6]	**72.88**	70.54	70.69	-	-	-	-	-
BERT+Knowledge [5]	-	-	-	-	60.79	**68.85**	64.57	-
Co-Senti-BERTCNN	70.87	**74.49**	**72.64**	**0.8408**	63.73	66.28	**64.98**	**0.8198**

TABLE 2. Average performance comparison of our method and classic existing methods on the cross-validation data of the SMM4H corpus.

	P(%)	R(%)	F1(%)	AUC
TextCNN [21]	50.38	34.19	40.06	65.56
BiLSTM [7]	44.90	40.66	42.42	68.17
BiGRU [7]	46.02	38.74	41.65	65.56
BERTCNN	59.46	**65.22**	62.05	**82.54**
Co-Senti-BERTCNN	**64.13**	64.79	**64.24**	82.10

TABLE 3. Performance of our model on 5-fold cross-validation data of the SMM4H corpus.

Fold	P(%)	R(%)	F1(%)	AUC
1.	60.38	**70.15**	64.90	**0.8669**
2.	**72.14**	61.08	**66.15**	0.8057
3.	58.56	66.32	62.20	0.8147
4.	64.83	62.07	63.42	0.8050
5.	64.75	64.33	64.54	0.8125
Avg.	64.13	64.79	64.24	0.8209

BERT+Knowledge, the model decreases by 2.6% in recall and increases by 3% in precision. We suspect that the co-occurrence sub-sentence introduces noise when the dataset contains more non-ADR tweets (note that more noise data and less positive tweets are contained in SMM4H [28]). Nevertheless, sufficient emotional expression (word-level emotion score and sentence-level emotional context) promotes the performance of the F1-score.

To verify the generalization of our method, the performance comparison between our Co-Senti-BERTCNN method and other baseline methods on the cross-validation data of SMM4H is shown in Tables 2 and 3. From Table 2, we observe that BERTCNN achieves the best recall rate and AUC value, whereas Co-Senti-BERTCNN obtains the best precision and F1-score. Moreover, the precision and recall rate of our method are close to each other, while the recall rate of BERTCNN is higher than precision. This suggests that it will reduce the recall rate and improve the precision rate to achieve better overall performance if our method is generalized and tells us that emotional expression has a certain impact on the recall rate in generalization. Table 3 presents the results of each fold using our method, and we find that on 5-fold cross-validation data, the maximum difference is 13.58, 9.07 and 3.95 percentage points in precision, recall rate and F1-score, respectively. This shows that on the premise of keeping the positive-to-negative ratio unchanged, our method fluctuates greatly in precision, followed by recall rate, and remains unchanged in F1-score and AUC. Therefore, the F1-score of our method is stable in generalization performance.

D. THE EFFECT OF WORD-LEVEL EMOTIONAL SCORE, SENTENCE-LEVEL EMOTIONAL CONTEXT AND CO-OCCURRENCE SUB-SENTENCE

The effect of three key components on the performance of our model is investigated through the PSB2016 and SMM4H datasets, namely, word-level emotional score (WEmoS), sentence-level emotional context (SEmoCTX) and the drug-ADR co-occurrence sub-sentence (CoSen) mentioned in sections III.A, III.B and III.C, as shown in Table 4. BERTCNN feeds the final hidden state of the BERT provided by Google into the downstream TextCNN, which is regarded as the baseline. Then, three key components are gradually introduced into the baseline model.

When CoSen is added into the baseline, the recall rate on PSB2016 and SM44H increases obviously first, while the precision decreases by at least 10%. The result shows that CoSen can truly help the proposed model focus on the ADR tweets, but it misleads the model to concentrate on the tweets containing co-occurrence pairs to a certain extent. Then, SEmoCTX is also introduced into the baseline, the recall rate decreases by 10%, and the precision increases by 10% on SMM4H. However, SEmoCTX can help our model promote 15% precision and 1% recall increases on PSB2016. The

TABLE 4. The effect of the sentimental score, drug-ADR co-occurrence pair and sentence-level emotional context on performance of the PSB2016 and SMM4H corpa.

	PSB2016				SMM4H			
model	P(%)	R(%)	F1(%)	AUC	P(%)	R(%)	F1(%)	AUC
BERTCNN	67.31	67.96	67.63	0.8214	**82.69**	49.14	61.65	0.7426
BERTCNN+ CoSen	56.59	70.87	62.93	0.8239	56.67	**68.00**	61.82	0.8240
BERTCNN+ WEmoS	58.82	67.96	63.06	0.8132	67.11	58.29	62.38	0.7827
BERTCNN+ SEmoCTX	63.72	63.10	63.41	0.7954	73.71	55.36	63.23	**0.8503**
BERTCNN+ CoSen+ SEmoCTX	**72.55**	71.84	72.19	**0.8440**	76.56	56.00	64.68	0.7747
BERTCNN+ CoSen+ SEmoCTX+ WEmoS	70.87	**74.49**	**72.64**	0.8408	63.73	66.28	**64.98**	0.8198

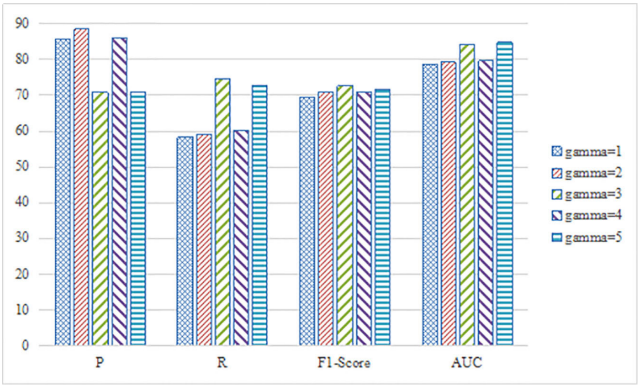


FIGURE 4. The precision, recall, F1-score and AUC of different gamma on PSB2016.

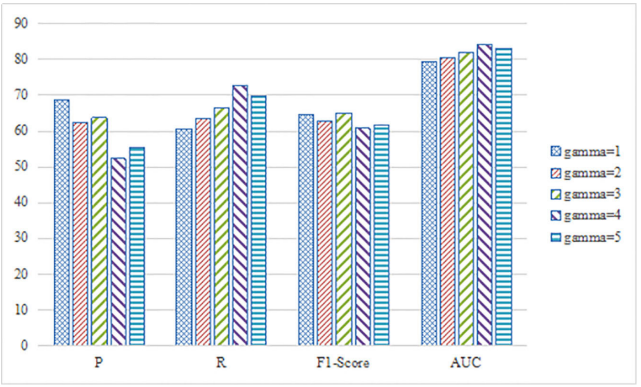


FIGURE 5. The precision, recall, F1-score and AUC of different gamma on SMM4H.

result shows that SEmoCTX mainly contributes to improving precision. Moreover, SEmoCTX on the dataset with a little noise and relatively balanced data has some advantages over that on other datasets. This implies that SEmoCTX contains abundant global context information, which can compensate for the limitations concerning misleading, improving overall performance. Finally, WemoS highlights the dominant emotional word to further balance the precision and recall, which can lead to a small increase in the recall rate. As a result, F1-score is promoted.

E. DIFFERENT PERFORMANCE OF DIFFERENT BALANCED FACTORS

The experimental datasets are unbalanced,, and the loss function of the proposed model introduces a balanced factor γ. This section demonstrates the performance of different balanced factors γ. The experiments are conducted on PSB2016 and SMM4H using our method when gamma is 1, 2, 3, 4 and 5, as shown in Figures 4 and 5. The model obtains the best F1-score when $\gamma = 3$, which is equivalent to a positive-to-negative ratio of 1:3. A similar conclusion is reached by Liu *et al.* [47].

First, on PSB2016, Figure 4 shows that the proposed model obtains the best precision when gamma is equal to 2. Furthermore, the best recall rate and AUC are achieved when gamma is 3. Second, on SMM4H, Figure 5 shows that when gamma is set to 1, the model obtains the best precision but achieves the best recall rate and AUC when gamma is equal to 4. The proposed method obtains different results when we set different gammas on the two datasets due to the inconsistency of the positive (ADR tweets) and negative (non-ADR tweets) proportions of the two datasets (PSB2016 is 1:9.6 and SMM4H is 1:16) because gamma itself serves to balance the proportions of positive and negative examples.

VI. ERROR ANALYSIS

To quantitatively analyse the effect of emotional expression on ADR classification, we download the code of the combined CNN and LSTM model in [48] from https://github.com/pmsosa/CS291K and put the test set into the emotional classification model to obtain the corresponding emotional label. As shown in Table 5, the sentimental labels of the second, 5th and 7th tweets are gained by the model, as well as other tweets for which our best method had prediction errors or there are prediction disagreements between our method and the baseline BERTCNN.

In the first section of Table 5, we show two examples, for which our method (Co-Senti-BERTCNN) and the baseline have a disagreement in their predictions. The proposed method predicts the first tweet as a true ADR

TABLE 5. Examples of false negatives, false positives and prediction disagreement between our model and the baseline.

Example Tweet	ADR Label	Sentimental Label	Co-occurrence	Prediction Label (our model/baseline)
Prediction disagreement (our method/baseline)				
glad that worked humira made me sick remicade stopped working nothing works cimzia on now but not working well either	ADR	-	■	ADR/Non-ADR
sorry to hear that i was on humira for yrs now remicade for mo we must shake this fatigue	ADR	Negative	-	ADR/Non-ADR
Labelling error				
How do guys ejaculate on paxil?? #antidepressants	ADR	-	-	Non-ADR
False Positives				
also does humira make you throw up remicade voms for me	ADR	-	■	Non-ADR
i get my headaches when my humira wears off my dad is on remicade	ADR	Negative	-	Non-ADR
False Negatives				
doc ordered all tests for humira or remicade	Non-ADR	-	■	ADR
date april am it is the bingo socialist tramadol water and very paxil levitra can free ve	Non-ADR	Positive	-	ADR

tweet, but BERTCNN gives it a non-ADR label. The reason for this difference is that our model utilizes the drug-ADR co-occurrence pair, "humira#ad red sick sic vomiting sting hot". In addition, the sentimental label of the second tweet is predicted as negative, and the tweet is an ADR tweet. The proposed method gains the right label owing to capturing the negative emotions hidden in the tweet.

In the second part, we first present a tweet that the model predicts wrong due to mislabelling. Although rich emotional expression and the co-occurrence sub-sentence are useful for identifying ADR posts containing co-occurrence or negative emotions partly, some noise or excessive focusing on co-occurrence pair is also brought in, which results in classification errors of such posts such as the 4th (containing a drug-ADR co-occurrence pair) and 5th (containing negative emotions) tweets. In fact, not all tweets containing co-occurrence pairs or negative emotions contain ADR when also containing a drug. Nevertheless, our model labels them as ADR tweets, which are false negative examples, such as the 6th and 7th tweets.

VII. CONCLUSION

Discovering ADRs on social media has become a major research trend recently due to the widespread and real-time nature of social media, but not due to the limitations and lags of clinical experiments. However, due to insufficient expression of emotions and inadequacy of information expression in short social texts, the existing methods do not achieve unsatisfactory performance. In this paper, we propose a neural network model for the ADR detection task. The model uses the word-level emotional score and sentence-level emotional context gained by our pre-trained BERT to capture the tweets that contain the negative emotions. These negative emotions may be an inherent clue of ADR posts. Moreover, we generate the co-occurrence sub-sentence using the drug-ADR medical dictionary. These sub-sentences help the proposed method extract the dominant hidden feature for distinguishing between ADR and non-ADR tweets, resulting in an increase in the recall rate and overall performance. The experimental results and analysis show that word-level emotional score and sentence-level emotional context contribute to promoting precision and the overall performance of ADR classification. In addition, the co-occurrence sub-sentence reduces the precision in part, but it achieves the improvement of the recall rate and promotes the F1-score. However, further improvement is needed on SM44H datasets containing more non-ADRs tweets. Therefore, improved BERT and additional features will be considered in future work. In addition, greater medical knowledge may be combined into our model.

REFERENCES

[1] Y. Chen, J. Yuan, Q. You, and J. Luo, "Twitter sentiment analysis via bi-sense emoji embedding and attention-based LSTM," in *Proc. ACM Multimedia Conf. Multimedia Conf. (MM)*, 2018, pp. 117–125, doi: 10.1145/3240508.3240533.

[2] A. Sarker and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," *J. Biomed. Inf.*, vol. 53, pp. 196–207, Feb. 2015.

[3] X. Wang and M. Kiang, "Identification of consumer adverse drug reaction messages on social media," in *Proc. PACIS*, Jeju-do, South Korea, 2013, p. 193.

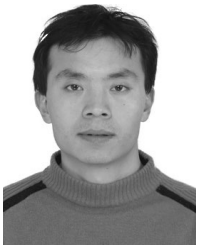
[4] W. Wang, "Mining adverse drug reaction mentions in Twitter with word embeddings," in *Proc. Pacific Symp. Biocomput. Social Media Mining Shared Task Workshop*, vol. 1, 2016, pp. 1–5. Accessed: Aug. 1, 2016. [Online]. Available: http://diego.asu.edu/psb2016/acceptedpapers/DLIR.pdf

[5] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *J. Amer. Med. Inform. Assoc.*, vol. 22, pp. 671–681, May 2015, doi: 10.1093/jamia/ocu041.

[6] H. Sampathkumar, X. Chen, and B. Luo, "Mining adverse drug reactions from online healthcare forums using hidden Markov model," *BMC Med. Inform. Decis. Making*, vol. 14, Oct. 2014, Art. no. 91, doi: 10.1186/1472-6947-14-91.
[7] X. Zhao, D. Yu, and V. G. V. Vydiswaran, "Identifying adverse drug events mentions in tweets using attentive, collocated, and aggregated medical representation," in *Proc. 4th Social Media Mining Health Appl. (SMM4H) Workshop Shared Task*, 2019, pp. 62–70.
[8] A. Dirkson and S. Verberne, "Transfer learning for health-related Twitter data," in *Proc. of 4th Social Media Mining Health Appl. (SMM4H) Workshop Shared Task*, 2019, pp. 89–92.
[9] S. Gupta, M. Gupta, V. Varma, S. Pawar, N. Ramrakhiyani, and G. K. Palshikar, "Co-training for extraction of adverse drug reaction mentions from tweets," 2018, pp. 1–6, *arXiv:1802.05121*. [Online]. Available: http://arxiv.org/abs/1802.05121
[10] H. Zhao, J. Zheng, W. Deng, and Y. Song, "Semi-supervised broad learning system based on manifold regularization and broad network," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 3, pp. 983–994, Mar. 2020.
[11] S. Gupta, M. Gupta, V. Varma, S. Pawar, N. Ramrakhiyani, and G. K. Palshikar, "Multi-task learning for extraction of adverse drug reaction mentions from tweets," 2018, *arXiv:1802.05130*. [Online]. Available: http://arxiv.org/abs/1802.05130
[12] W. Deng, H. Liu, J. Xu, H. Zhao, and Y. Song, "An improved quantum-inspired differential evolution algorithm for deep belief network," *IEEE Trans. Instrum. Meas.*, early access, Mar. 25, 2020, doi: 10.1109/TIM.2020.2983233.
[13] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentimental reviews using machine learning techniques," *Procedia Comput. Sci.*, vol. 57, pp. 821–829, Jan. 2015.
[14] D. S. Sahana and L. Girish, "Automatic drug reaction detection using sentimental analysis," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 4, no. 5, pp. 2163–2170, 2015.
[15] M. Guerini, L. Gatti, and M. Turchi, "Sentiment analysis: How to derive prior polarities from SentiWordNet," in *Proc. EMNLP*, Oct. 2013, pp. 1259–1269.
[16] D. Egger, F. Uzdilli, and M. Cieliebak, "Adverse drug reaction detection using an adapted sentiment classifier," in *Proc. Social Media Mining Shared Task Workshop Pacific Symp. Biocomput.*, 2016, pp. 1–5.
[17] D. Dutta Das, S. Sharma, S. Natani, N. Khare, and B. Singh, "Sentimental analysis for airline Twitter data," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 263, Nov. 2017, Art. no. 042067, doi: 10.1088/1757-899X/263/4/042067.
[18] R. Ginn, P. Pimpalkhute, A. Nikfarjam, A. Patki, K. O'Connor, A. Sarker, K. Smith, and G. Gonzalez, "Mining Twitter for adverse drug reaction mentions: A corpus and classification benchmark," in *Proc. 4th Workshop Building Evaluating Resour. Health Biomed. Text Process.*, no. 1, 2014, pp. 1–8.
[19] S. Chen, Y. Huang, X. Huang, H. Qin, J. Yan, and B. Tang, "HITSZ-ICRC: A report for SMM4H shared task 2019-automatic classification and extraction of adverse effect mentions in tweets," in *Proc. 4th Social Media Mining Health Appl. (SMM4H) Workshop Shared Task*, 2019, pp. 47–51.
[20] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. Ho So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," 2019, pp. 1–8, *arXiv:1901.08746*. [Online]. Available: http://arxiv.org/abs/1901.08746
[21] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart, "DrugBank 3.0: A comprehensive resource for 'Omics' research on drugs," *Nucleic Acids Res.*, vol. 39, pp. D1035–D1041, Jan. 2011, doi: 10.1093/nar/gkq1126.
[22] R. Dong, M. P. O'Mahony, M. Schaal, K. McCarthy, and B. Smyth, "Sentimental product recommendation," in *Proc. 7th ACM Conf. Recommender Syst. (RecSys)*, 2013, pp. 411–414, doi: 10.1145/2507157.2507199.
[23] H. Bagheri and M. J. Islam, "Sentiment analysis of Twitter data," 2017, *arXiv:1711.10377*. [Online]. Available: http://arxiv.org/abs/1711.10377
[24] X. Zou, J. Yang, and J. Zhang, "Microblog sentiment analysis using social and topic context," *PLoS ONE*, vol. 13, no. 2, Feb. 2018, Art. no. e0191163, doi: 10.1371/journal.pone.0191163.
[25] N. Douali, and P. Staccini, "Social media and patient health outcomes: Findings from the yearbook 2014 section on consumer health informatics," *Yearb Med Inf.*, vol. 23, no. 1, pp. 195–198, Aug. 2014, doi: 10.15265/IY-2014-0038.
[26] G. Katz, N. Ofek, and B. Shapira, "ConSent: Context-based sentiment analysis," *Knowl.-Based Syst.*, vol. 84, pp. 162–178, Aug. 2015.
[27] H. Saif, Y. He, M. Fernandez, and H. Alani, "Contextual semantics for sentiment analysis of Twitter," *Inf. Process. Manage.*, vol. 52, no. 1, pp. 5–19, Jan. 2016.
[28] C.-K. Wang, H.-J. Dai, F.-D. Wang, and E. C.-Y. Su, "Adverse drug reaction post classification with imbalanced classification techniques," in *Proc. Conf. Technol. Appl. Artif. Intell. (TAAI)*, Nov. 2018, pp. 5–9, doi: 10.1109/TAAI.2018.00011.
[29] S. Ramamoorthy and S. Murugan, "An attentive sequence model for adverse drug event extraction from biomedical text," 2018, *arXiv:1801.00625*. [Online]. Available: http://arxiv.org/abs/1801.00625
[30] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," 2019, *arXiv:1903.09588*. [Online]. Available: http://arxiv.org/abs/1903.09588
[31] S. Karimi, C. Wang, A. Metke-Jimenez, R. Gaire, and C. Paris, "Text and data mining techniques in adverse drug reaction detection," *ACM Comput. Surv.*, vol. 47, no. 4, pp. 1–39, Jul. 2015.
[32] I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, and G. H. Gonzalez, "Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts," *J. Biomed. Informat.*, vol. 62, pp. 148–158, Aug. 2016.
[33] R. Leaman and L. Wojtulewicz, "Towards Internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks," in *Proc. Workshop Biomed. Natural Lang. Process. (ACL)*, Jul. 2010, pp. 117–125.
[34] A. Nikfarjam and G. H. Gonzalez, "Pattern mining for extraction of mentions of adverse drug reactions from user comments," in *Proc. AMIA Annu. Symp.*, 2011, pp. 1019–1026.
[35] K. Lee, A. Qadir, S. A. Hasan, V. Datla, A. Prakash, J. Liu, and O. Farri, "Adverse drug event detection in tweets with semi-supervised convolutional neural networks," in *Proc. of 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 1–10, doi: 10.1145/3038912.3052671.
[36] V. Echeverria, G. Falcones, J. Castells, R. Granda, and K. Chiluiza, "Semi-supervised recurrent neural network for adverse drug reaction mention extraction," in *Proc. CEUR Workshop*, vol. 1828, 2017, pp. 94–98. [Online]. Available: https://arxiv.org/abs/1709.01687
[37] C. Pandey, Z. Ibrahim, H. Wu, E. Iqbal, and R. Dobson, "Improving RNN with attention and embedding for adverse drug reactions," in *Proc. Int. Conf. Digit. Health (DH)*, Jul. 2017, pp. 67–71, doi: 10.1145/3079452.3079501.
[38] C. Wu, F. Wu, J. Liu, S. Wu, Y. Huang, and X. Xie, "Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tweet representation and multi-head self-attention," in *Proc. EMNLP Workshop SMM4H, 3rd Social Media Mining Health Appl. Workshop Shared Task*, 2018, pp. 34–37.
[39] S. Chowdhury, C. Zhang, and P. S. Yu, "Multi-task pharmacovigilance mining from social media posts," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 117–126, doi: 10.1145/3178876.3186053.
[40] M. Vadivukarassi, N. Puviarasan, and P. Aruna, "Sentimental analysis of tweets using naive Bayes algorithm," *World Appl. Sci. J.*, vol. 35, no. 1, pp. 54–59, 2017.
[41] V. Patel, G. Prabhu, and K. Bhowmick, "A survey of opinion mining and sentiment analysis," *Int. J. Comput. Appl.*, vol. 131, no. 1, pp. 24–27, 2012.
[42] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: http://arxiv.org/abs/1408.5882
[43] D. Weissenbacher, A. Sarker, A. Magge, A. Daughton, K. O'Connor, M. Paul, and G. Gonzalez-Hernandez, "Overview of the fourth social media mining for health (#SMM4H) shared task at ACL 2019," in *Proc. 4th Social Media Mining Health Appl. (SMM4H) Workshop Shared Task*, 2019, pp. 21–30.
[44] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2999–3007.
[45] A. Cocos, A. G. Fiks, and A. J. Masino, "Deep learning for pharmacovigilance: Recurrent neural network architectures for labeling adverse drug reactions in Twitter posts," *J. Amer. Med. Inform. Assoc.*, vol. 24, no. 4, pp. 813–821, Jul. 2017.
[46] T. Huynh, Y. He, A. Willis, and R. Stefan, "Adverse drug reaction classification with deep neural networks," in *Proc. 26th Int. Conf. Comput. Linguist.*, 2016, pp. 877–887.

[47] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Computer Vision ECC*. Cham, Switzerland: Springer, 2016, doi: 10.1007/978-3-319-46448-0_2.

[48] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "A combined CNN and LSTM model for arabic sentiment analysis," in *Machine Learning and Knowledge Extraction* (Lecture Notes in Computer Science), vol. 11015. Cham, Switzerland: Springer, 2018, pp. 179–191, doi: 10.1007/978-3-319-99740-7_12.



ZHENGGUANG LI received the master's degree in computer science from Dalian Jiaotong University, in 2007. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Dalian University of Technology. His research interests include text mining and natural language processing in social media and biomedical fields.

HONGFEI LIN received the B.Sc. degree from Northeast Normal University, in 1983, the M.Sc. degree from the Dalian University of Technology, in 1992, and the Ph.D. degree from Northeastern University, in 2000. He is currently a Professor with the School of Computer Science and Technology, Dalian University of Technology, where he is also the Director of the Information Retrieval Laboratory. He has published over 100 research articles in various journals, conferences, and books. His research projects are funded by the National Natural Science Foundation of China and the National High-Tech Development Plan. His research interests include information retrieval, text mining, natural language processing, effective computing, text mining for biomedical literature, biomedical hypothesis generation, information extraction from large-biomedical resources, learning to rank, sentimental analysis, and opinion mining.

WEI ZHENG received the Ph.D. degree in computer application technology from the Dalian University of Technology, China, in 2018. She is currently a Lecturer with the School of Software Engineering, Dalian Jiaotong University, Dalian, China. Her research interests include text mining for biomedical literature and social media, knowledge graph construction, and natural language processing.