Journal of Biomedical Informatics 123 (2021) 103896

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Adversarial neural network with sentiment-aware attention for detecting adverse drug reactions

Tongxuan Zhang [a,b], Hongfei Lin [b,*], Bo Xu [b], Liang Yang [b], Jian Wang [b], Xiaodong Duan [c]

[a] Tianjin Normal University, Tianjin, China
[b] Dalian University of Technology, Dalian, China
[c] Dalian Minzu University, Dalian, China

ARTICLE INFO

Keywords:
Adverse drug reactions
Social media
Text classification
Attention mechanism
Adversarial training

ABSTRACT

Adverse drug reaction (ADR) detection is an important issue in drug safety. ADRs are health threats caused by medication. Identifying ADRs in a timely manner can reduce harm to patients and can also assist doctors in the rational use of drugs. Many studies have investigated potential ADRs based on social media due to the openness and timeliness of this resource; however, they have ignored the fine-grained emotional expression in social media text. In addition, the benchmark datasets from social media are usually small, which can result in the problem of over-fitting. In this paper, we propose the Adversarial Neural Network with Sentiment-aware Attention (ANNSA) model, which enhances the sentimental element in social media and improves the performance of neural networks via data augmentation. Specifically, a sentiment-aware attention mechanism is proposed to extract the word-level sentiment features associated with sentiment words and learn task-related information by optimizing a task-specific loss. For low-resource datasets, we use an adversarial training approach to generate perturbations of the word embeddings via an implicit regularization technique. ANNSA was tested on three social media ADR detection datasets, namely, Twitter, TwiMed (Twitter) and CADEC. The experimental results indicated the ability to achieve F1 values of 48.84%, 64.18% and 83.06%, respectively, comparable to the best results reported for state-of-the-art methods. Our study demonstrates that sentiment words are highly correlated with ADRs and that word-level sentiment features can assist in detecting ADRs from social media datasets.

1. Introduction

Adverse drug reactions (ADRs) refer to the harmful reactions that occur when normal doses of drugs are used. ADRs seriously endanger the health of individuals and cause enormous economic losses to the medical system and society. Adverse events resulting from the use of marketed drugs are a major public health problem, accounting for 28% of emergency room visits, 5% of hospital admissions, and 5% of hospital deaths [1,2]. Annually, the expenses due to ADRs are up to $75 billion [3]. Therefore, the timely and accurate detection of ADRs is essential to prevent adverse reactions caused by medication and reduce medical costs [4].

Generally, before a drug is marketed, a large number of clinical trials are performed to identify ADRs. However, it is difficult to identify all potential ADRs due to time and cost limitations. Thus, there is still a need to identify ADRs from multiple sources after a drug is marketed [5–8]. In early works, researchers mainly relied on a spontaneous reporting system consisting of compulsive and voluntary reporting of suspected adverse drug events (ADEs) by pharmaceutical companies, consumers, and health care professionals. The US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) is one of the most prominent spontaneous reporting systems [4]. However, due to the complicated approval steps involved, the update speed and accuracy of these systems are usually poor. Therefore, researchers prefer to identify ADRs from other, relatively real-time resources.

With the development of the Internet, social media has provided people with a platform for sharing information with each other. Some patients want to communicate their feelings after taking medication with other patients through social media. Although information from medical systems (e.g., FAERS) is more authoritative than that from social media, these open sources (e.g., Twitter) can also serve as a reference to provide a direction for ADR research. In recent years, social media has been widely used as a data source for ADR detection.

A person's emotional state may be affected by physical discomfort.

* Corresponding author.
E-mail address: hflin@dlut.edu.cn (H. Lin).

https://doi.org/10.1016/j.jbi.2021.103896
Received 28 February 2021; Received in revised form 22 August 2021; Accepted 23 August 2021
Available online 4 September 2021

The social media posts of patients are related to their opinions and feelings and contain rich emotional components. Numerous studies have shown that there are strong emotional components in social media, such as microblogs [9] and Twitter [10], as they are often resources for sentiment analysis [11], user stance detection [12], and opinion mining [13]. Moreover, researchers have proposed that sentiment analysis should be effective in extracting ADRs from social media text [14]. Accordingly, previous studies [10,14] have integrated sentiment information by directly concatenating the representations of sentences and the scores of sentence-level sentiment polarity. However, sentence-level sentiment score features provide only a rough basis for learning sentiment information from different sentiment words in sentences.

Another issue with ADR detection is that the datasets extracted from social media tend to be small. Previous works have handled this problem by introducing external resources [15], adding annotated training sets to the original data [16], or using multi-task learning methods [17,18] to jointly train a model on the named entity recognition (NER) and ADR detection tasks. Although these approaches have enabled great improvements in ADR detection, they usually require external data support and extra annotation workloads. Hence, we hope to increase the diversity of training sets without the use of any additional resources.

To address the above problems, we propose the Adversarial Neural Network with Sentiment-aware Attention (ANNSA) model for ADR detection on social media. Instead of simply concatenating the scores of sentence-level sentiment polarity with the representations of sentences, we apply a sentiment-aware attention mechanism to extract word-level sentiment features by learning a weight matrix of sentiment words for sentences. In this way, the information from sentiment words can directly influence the computation of sentence representations. The resulting sentiment-aware representations of sentences also help the model learn task-related information. Additionally, regarding the statistics of the sentiment words of sentences, we find that sentiment words obviously overlap with ADR mentions. Thus, through the introduction of word-level sentiment features, ANNSA is able to pay more attention to ADR mentions.

Inspired by adversarial training [19], we introduce an adversarial perturbation method for word embeddings that not only improves model robustness without requiring additional training sets but also effectively prevents over-fitting when using a small-scale corpus. In our adversarial perturbation approach, adversarial samples are created by making small modifications to word embeddings. In the training phase, we minimize an additional regularization cost to resist such perturbations to make the model more robust to unseen datasets. Additionally, to address the misspelling problem in social media text, we introduce a convolutional neural network (CNN) that operates at the character level to learn lexical patterns. Finally, we report a massive experiment conducted on multiple social media ADR corpora to assess the performance of ANNSA. The significance of our research is that ANNSA can identify some potential adverse reactions in real-world scenarios and can help support pharmaceutical research and development.

Based on the above discussion, our contributions are as follows.

1. We present a sentiment-aware attention mechanism that can capture sentiment features at the word level and help to learn task-related information for ADR detection. Experimental results show that this mechanism helps the resulting model pay more attention to ADR mentions.

2. To address the issue of small social media datasets, we apply an adversarial perturbation mechanism in ANNSA. This mechanism can improve the generalization ability of ANNSA via data augmentation without any external resources.

3. We report an experiment conducted on three different widely used social media ADR datasets. The results demonstrate that our proposed method makes significant progress in solving the previously presented problems.

2. Related work

2.1. ADR detection based on social media

Early studies on ADR detection were often based on data such as those from spontaneous reporting systems (SRSs) [20] and clinical reports [21]. Due to the advantages of social network data [22,23] concerning ADRs, social media data present both unique challenges and interesting opportunities for natural language processing (NLP) methods for ADR detection [16]. The initial research mainly used lexicon-based approaches [24] to identify ADRs in text. When there were no labelled data available, researchers tended to use unsupervised methods [25] for statistical analysis. More recently, a growing trend towards implementing ADR detection through supervised methods has emerged. In these methods, researchers [26–29] combine various machine learning classifiers with various features to automatically extract and classify messages. These methods focus on shallow language features and do not capture deep semantic features or contextual representations of sentences.

In recent years, with the development and application of deep learning methods, many neural network methods have been applied to ADR detection tasks. Wu et al. [30] developed an approach with multi-head self-attention and hierarchical tweet representation to detect ADRs. Lee et al. [31] used a semi-supervised deep learning model for ADR detection in tweets. Li et al. [15] introduced an adversarial transfer learning method for ADR detection, which can improve the results obtained on small datasets.

Some studies have suggested that sentiment information is helpful for ADR detection [32]. Inspired by this, many approaches for integrating sentiment information have been proposed, such as introducing sentiment score features or sentiment word frequency into ADR detection [16]. Shen et al. [14] utilized a multi-channel CNN to identify ADRs using a sentiment score. They concatenated the sentence representation and the sentiment score as the input to the final prediction layer to classify text. Li et al. [10] used a model integrating medical knowledge and sentiment expression to detect ADRs. They obtained sentiment scores using Bidirectional Encoder Representations from Transformers (BERT) via pre-training on a large number of sentiment analysis datasets. Most previous works have used a sentiment score as a sentence-level feature but have not explored the relationship between the sentiment words and sentences. Alhuzali et al. [33] classified the sentiment polarity of affairs in tweets and applied transfer learning to detect ADRs in tweets. They measured the word coverage between the sentiment analysis corpus and the ADR corpus on all words. Since both datasets contained a number of common words, the statistical results were insufficient to show the relationship between sentiment words and ADRs. In this paper, we consider a sentiment-aware representation without noise instead of an affective polarity score and explore the relation between sentiment words and adverse reactions.

2.2. Adversarial training

The concept of adversarial training [34] originates from generative adversarial networks (GANs) [35], which are widely applied in computer vision. Generally, the goal of adversarial training is to use unknowable perturbations to interfere with neural networks. In recent years, some studies have applied adversarial training to NLP tasks, such as text classification, aspect-based sentiment analysis (ABSA), NER, and part-of-speech (POS) tagging. Miyato et al. [36] added adversarial perturbations to word embeddings in a semi-supervised text classification model using adversarial training. Yasunaga et al. [37] applied adversarial training to POS tagging and found that adversarial training can effectively prevent over-fitting for low-resource languages. Zhou et al. [38] used an adversarial transfer network with adversarial training to address low-resource NER. Karimi et al. [39] proposed that data similar to the training set can be produced through adversarial training, an

approach that can be applied to the embedding space to make neural networks more robust.

These works indicate that adversarial training can allow significant results to be obtained on NLP tasks via data augmentation. Therefore, we apply adversarial training to improve the robustness and effectively improve the generalization ability of ADR detection when only a small amount of annotated data is available for training.

3. Method

In this paper, we regard the ADR detection task as a binary classification task. The architecture of ANNSA is illustrated in Fig. 1. ANNSA consists of five components: (1) an embedding layer that concatenates word-level embeddings with character-level embeddings, (2) a sentiment-aware attention mechanism that extracts word-level sentiment features by learning a compatibility matrix between sentences and sentiment words, (3) a feature extractor that captures contextual information from the original and sentiment-aware representations using two bidirectional long short-term memory (Bi-LSTM) [40] layers, (4) an adversarial training mechanism that adds adversarial perturbations to the embedding layer for data augmentation, and (5) a predictor that predicts the results of ADR detection. We first list the symbols used and their definitions and then describe each part of the proposed model in detail.

3.1. Problem formulation and definitions

Given a sentence sequence X, $\{x_1, x_2, \cdots, x_n\}$, ANNSA estimates whether this sentence contains evidence of ADRs. x_i denotes the vector representation of the i^{th} word, and n denotes the length of the sequence. The utilized symbols and their descriptions are listed in Table 1.

3.2. Embedding layer

ANNSA takes word-level embeddings $x_i^{word} \in R^{ew}$ and character-level embeddings $x_i^{char} \in R^{ec}$ as inputs, where ew and ec represent the dimensions of the word-level and character-level embeddings, respectively. Word-level embeddings are obtained by searching a pre-trained embedding matrix. For rare and out-of-vocabulary (OOV) words arising through colloquial expression on social media, we employ a character-level CNN (char-CNN) [41] to extract character-level features. The final word representation of word x_i is the concatenation of the word-level and character-level embeddings, which is expressed as $x_i = [x_i^{word}; x_i^{char}] \in R^{ew+ec}$.

Word-level embeddings. Word-level embeddings have been widely used for NER [38], sentiment analysis [42], text classification [43], and other NLP tasks. The most representative approaches for extracting word embeddings are Word2Vec [44] and Global Vectors for Word Representation (GloVe) [45]. For the experiment reported in this paper, we downloaded a total of 2,680,617 MEDLINE abstracts from PubMed by using the query string "drug". Then, these abstracts were used to train word embeddings by using Word2Vec [44] to convert one-hot encodings into continuous values in low dimensions and pre-train the word embeddings. In the vector space, words can be mapped to similar positions if they have similar meanings. The word embeddings for OOV words were randomly initialized.

Character-level embeddings. The texts posted on social media are usually informal and colloquial and often even contain spelling errors. These attributes of social media may cause OOV problems such that vector representations cannot be found from pre-trained word embeddings. Regarding the characteristics of word-based morphology, words with the same structure tend to have similar meanings (such as suffixes or prefixes). The character-level embeddings we used were initialized randomly and updated during training. Previous studies [46] have shown that a char-CNN method is an effective approach for extracting morphological information. Therefore, we captured character-level embeddings using a char-CNN to solve the OOV problem.

3.3. Sentiment-aware attention mechanism

To capture the sentiment words from social media datasets, we chose a dictionary with a large number of sentiment words: SenticNet[1]. It is a publicly available resource for opinion mining built by exploiting artificial intelligence (AI) and semantic web techniques [47]. To build this dictionary, a method was applied to create a polarity for nearly 14,000 concepts using NLP techniques [48]. It enables the use of semantics and linguistics to address tasks such as political topic analysis [49] and sentiment analysis [50,51] rather than simply relying on word co-occurrence frequency.

To make full use of sentiment features, we introduce a sentiment-aware attention mechanism to assign a sentiment-sensitive weight to each word in a sentence, with the aim of learning fine-grained sentiment features and task-related information. In this mechanism, Word2Vec is used to initialize the word embeddings of sentiment words, expressed as $X^s \in R^{m\times d}$, where m represents the number of sentiment words. The main process of the sentiment-aware attention mechanism is to learn a compatibility matrix G between the representation of a sentence $X^w \in R^{n\times d}$ and the representation of the sentiment words $X^s \in R^{m\times d}$. The matrix $G \in R^{m\times n}$ is computed as follows:

$$G = \tanh(X^w U X^{sT}) \tag{1}$$

where $U \in R^{d\times d}$ represents a trainable parameter matrix.

Then, a weight vector $g \in R^m$, which is the sentiment score for each word, is obtained through row-wise max pooling over G. Finally, the attention weight and the sentiment-aware representation $X^{s'}$ are computed as follows:

$$\alpha tt = \text{softmax}(g) \tag{2}$$

$$X^{s'} = \alpha tt X^w \tag{3}$$

3.4. Feature extractor

To encode the representations of the original and sentiment-aware inputs, Bi-LSTM is utilized, which has been successfully used for several NLP tasks [38,52]. LSTM [40] is a powerful variant of the recurrent neural network (RNN) architecture designed to address the vanishing and exploding gradient problems [53] and to extract contextual information from sentences. A Bi-LSTM layer contains both forward and backward contexts. Therefore, it can capture information from both past contexts and future contexts. The hidden state of an LSTM unit is expressed as follows:

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \tag{4}$$

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \tag{5}$$

$$C_t = tanh(W_C \bullet [h_{t-1}, x_t] + b_C) \tag{6}$$

$$C_t = f_t * C_{t-1} + i_t * C_t \tag{7}$$

$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_t = o_t * tanh(C_t) \tag{9}$$

An LSTM unit contains three gates: a forget gate f_t, an input gate i_t and an output gate o_t. These gates determine the information at the current time step t. C_t denotes the memory cell. $W_f \in R^{d\times(ew+ec+d)}$,

[1] https://www.sentic.net/

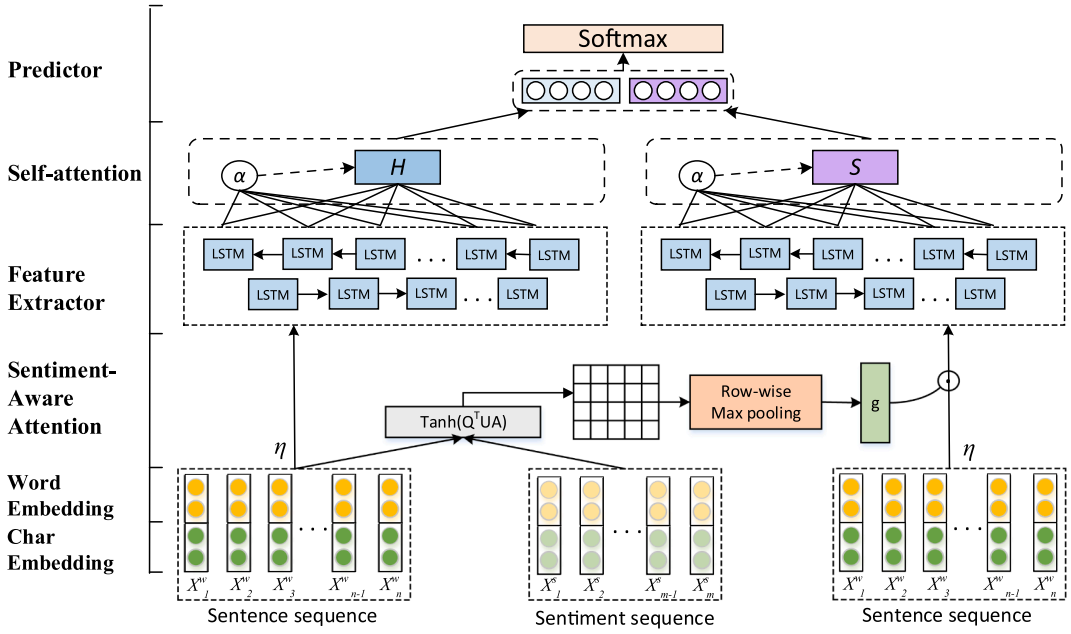

Fig. 1. The architecture of ANNSA.

Table 1
Definitions of notations.

Notation	Explanation
X, x_i	A sequence representation obtained by concatenating word embeddings with character embeddings; the vector representation of the i^{th} word.
y	A label.
X^w	A sentence sequence representation.
X^s	A sentiment word representation.
$X^{s'}$	A sentiment-aware sentence representation.
g	An attention weight vector.
C_t	A memory cell.
h_t, h_{t-1}	The hidden representations of the cells at time steps t and t-1.
$\overrightarrow{h}_t$, $\overleftarrow{h}_t$	A forward hidden state and a backward hidden state.
U, W_f, W_i, W_C, W_o,W_a	Trainable weight matrices.
b_f, b_i, b_C, b_o, b_s, b_a,w_s	Trainable parameters.
H, S	The hidden representation of a sentence after Bi-LSTM; a sentiment-aware hidden representation after Bi-LSTM.
η	A perturbation.
Ω, ε	The disturbance space and a small norm.
θ	The parameters of the model.

$W_i \in R^{d\times(ew+ec+d)}$, $W_C \in R^{d\times(ew+ec+d)}$ and $W_o \in R^{d\times(ew+ec+d)}$ are weight matrices. $b_f \in R^d$, $b_i \in R^d$, $b_C \in R^d$ and $b_o \in R^d$ are bias parameters. $\sigma()$ and $tanh()$ represent activation functions. $h_t \in R^d$ represents the hidden state. d represents the dimensionality of the hidden state of the LSTM unit. Finally, two sub-networks are obtained through equations (4)-(9) and are concatenated to form the final hidden state representation for the sentence. The formula is as follows:

$$H = (h_1', h_2', \cdots, h_n') = \begin{bmatrix} \overrightarrow{h}_1 \overrightarrow{h}_2 \cdots & \overrightarrow{h}_n \\ \overleftarrow{h}_1 \overleftarrow{h}_2 \cdots & \overleftarrow{h}_n \end{bmatrix} \tag{10}$$

where $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ represent the forward and backward hidden states, respectively, and $h_t' \in R^{2d}$ represents the final hidden state representation after the Bi-LSTM layer at time step t. Finally, we obtain the contextual representations H and S for the original and sentiment-aware inputs, respectively.

3.5. Adversarial training mechanism

Recently, many studies have shown that deep neural networks are fragile against adversarial examples [34,38]. Adversarial training is a powerful regularization tool for improving the generalization ability of a model and is widely used in NLP. Adversarial training can also prevent a model from falling into a local minimum. Generally, adversarial training can be expressed in the following form [54]:

$$\eta_x = \min_{\theta} E_{(x,y)\ D}[\max_{\eta\in\Omega} L(x+\eta, y; \theta)] \|\eta\| \le \varepsilon \tag{11}$$

Here, D represents the training data. x and y represent the input hidden state and label, respectively. θ denotes the parameters of the model. $L()$ represents the loss function of the model. η is a perturbation, Ω is the disturbance space, and ε is a small norm. Specifically, we generate an adversarial sample by adding a perturbation η. The purpose of adding η is to make the loss function as large as possible. Finally, we use the adversarial sample $x+\eta$ as data to minimize the loss to update the parameters θ. The optimization process is performed by alternately maximizing and minimizing, which is known as the fast gradient method (FGM). The process can be computed as follows:

$$\eta_x = \varepsilon \frac{g}{\|g\|_2} g = \nabla L(x+\eta, y; \theta) \tag{12}$$

where ε is determined on the validation set.

In the study reported in this paper, we generated adversarial samples in the embedding layer.

3.6. Predictor

The predictor is used to judge whether a social media text is related to ADRs. First, self-attention pooling is applied to reduce the dimensionality of the outputs of the two Bi-LSTM layers while considering the global features of sentences. The final sentence representation is the concatenation of the two outputs. Then, the probability of each predicted class is obtained via the softmax function. Finally, the cross-entropy loss is applied as the training objective function of ANNSA.

The classification process and the training objective function are expressed as follows:

$$a = softmax(v^T \tanh(W_a H + b_a)) \quad (13)$$

$$h = \sum_{i=1}^{N} a_i H_i \quad (14)$$

where $W_a \in R^{2d \times 2d}$ is the weight matrix, $b_a \in R^{2d}$ and $v \in R^{2d}$ are trainable parameters, and h denotes the contextual original features output by the self-attention mechanism. Here, we use *selfattention*() to denote the process represented by formulas (13)-(14). Furthermore, the sentiment-aware output s is obtained via the self-attention mechanism:

$$s = selfattention(S) \quad (15)$$

$$p = softmax(w_s \bullet [h; s] + b_s) \quad (16)$$

$$L = \frac{1}{N} \sum_{i} -[y_i \bullet \log(p_i) + (1 - y_i) \bullet \log(1 - p_i)] \quad (17)$$

Here, $W_s \in R^{2d}$ and $b_s \in R^{2d}$ are parameters to be trained. [;] represents the concatenation operation. L represents the loss of ADR classification. y_i is the true label, and p_i is the predicted label.

4. Results

4.1. Datasets

To evaluate the performance of ANNSA, we conducted experiments on three widely adopted social media ADR datasets.

Note that because of Twitter's privacy policy, actual tweets cannot be shared. We obtained the tweet text by using web crawlers based on unique tweet IDs. The original Twitter dataset contains a total of 10,822 tweets, including 1,238 positive samples and 9,584 negative samples. However, we could obtain only 6,471 tweets, including 744 positive samples and 5,727 negative samples. The original TwiMed (Twitter) dataset contains a total of 1,000 tweets, including 390 positive samples and 610 negative samples. However, we could obtain only 625 tweets, including 232 positive samples and 393 negative samples.

1) **Twitter** [55]: The Twitter dataset consists of tweets collected using the brand and generic names of drugs as well as phonetic misspellings thereof because these are common spelling errors in user posts on Twitter. 2) **TwiMed (Twitter)** [8]: The TwiMed corpus consists of two parts, TwiMed (PubMed) and TwiMed (Twitter), which contain sentences extracted from PubMed and Twitter, respectively. In this experiment, we used only TwiMed (Twitter), which was reacquired using tweet IDs. 3) CSIRO Adverse Drug Event Corpus (**CADEC**) [56]: The CADEC dataset was extracted from medical forum posts. The sentences often deviate from punctuation rules and formal English grammar since they are written in colloquial language. Descriptions of the three datasets used in this experiment are given in Table 2. Since the number of long sentences in these datasets is sparse, we set a fixed length n and cut off each sentence at n. Note that sentences longer than n account for only 1% of all sentences.

Moreover, we analysed the proportions of sentiment words that were searched in SenticNet, i.e., the maximum proportion of sentiment word coverage of the texts, the mean proportion of sentiment word coverage of the texts, and the proportion of sentiment word coverage of the ADR mentions. The results are provided in Table 3. The statistical results show that the coverage of sentiment words and ADR mentions was quite high. This indicates a strong dependence between sentiment words and ADR mentions.

Table 2
Brief descriptions of the social media ADR datasets.

Dataset	Positive	Negative	Total	Max sentence length	Experimental data length
Twitter	744	5,727	6,471	46	46
TwiMed (Twitter)	232	393	625	135	65
CADEC	2,478	4,996	7,474	241	70

4.2. Experimental settings

We used Keras to perform model training. The dimensionality of the word-level embeddings was 200, and the dimensionality of the character-level embeddings was 30. We set the maximum word length to 30. The dimensionality of the Bi-LSTM hidden units was 100. The filter size of the char-CNN was 3, and the number of filters was 30. The batch size in our experiment was 16, and the number of epochs was 30.

We compared ANNSA with other methods in terms of the precision (P), recall (R) and F-score (F1). F1 quantifies the overall performance of a model by balancing P and R. To verify the validity of ANNSA, we performed 10-fold cross-validation on all social media datasets since these datasets were not separated into training and test sets.

4.3. Baseline methods

Due to privacy concerns, social media datasets provide only tweet IDs, not tweet texts. Because some of the original tweets could not be found, we obtained the text of fewer tweets than in the original datasets. The datasets we used were similar to those used by Li et al. [15]. To ensure fair comparisons, we compared ANNSA with the baselines considered in Li et al. [15], which included the following methods: 1) **RCNN** [57]: This method combines a CNN and an LSTM network. 2) **HTR-MSA** [30]: This is a model with multi-head self-attention and hierarchical tweet representation. 3) **CNN + corpus** [15]: This method adds an extra annotated corpus to the datasets used to train a CNN, which can improve the ADR detection performance. 4) **CNN-transfer** [15]: The CNN-transfer framework is based on a transfer learning method. 5) **ATL** [15]: The adversarial transfer learning framework combines the transfer learning mechanism with an adversarial training strategy.

4.4. Comparisons with baseline methods

Table 4 provides the results of comparisons of ANNSA with the other considered methods on the three datasets. Table 4 shows that ANNSA achieved new state-of-the-art results on all datasets, which suggests that our proposed method was able to effectively improve ADR detection performance. HTR-MSA [30] is a relatively complex model that requires a large amount of annotated data for parameter optimization. Thus, the performance of this model was poorer than that of RCNN on these small-scale datasets.

The NN + corpus [15], CNN-transfer [15], and ATL [15] methods improve ADR detection performance by introducing additional annotated data. CNN-transfer and ATL introduce features from other domains into ADR detection. However, excessive reliance on large-scale annotated data from other domains might restrict the generalization ability of the models due to noise from different sources. Compared to ATL, ANNSA improved the F1 scores by 2.60%, 0.64%, and 0.3% on Twitter, TwiMed (Twitter), and CADEC, respectively. Note that our proposed sentiment-aware attention and adversarial learning mechanisms

Table 3
The proportions of sentiment words.

Dataset	Max (%)	Mean (%)	ADR (%)
Twitter	75.00	15.27	38.49
TwiMed (Twitter)	66.67	15.84	39.27
CADEC	80.00	17.04	42.93

Table 4
Comparison of ANNSA with state-of-the-art methods.

Dataset	Twitter			TwiMed (Twitter)			CADEC		
(%)	P	R	F1	P	R	F1	P	R	F1
RCNN [57]	50.00	42.88	46.17	61.26	65.96	63.52	81.99	76.63	79.22
HTR-MSA [30]	37.06	**58.33**	45.33	60.67	61.70	61.18	81.77	77.64	79.65
CNN + corpus [15]	47.94	43.82	45.79	52.75	61.28	56.69	**85.40**	75.99	80.42
CNN-transfer [15]	60.23	35.62	44.76	61.84	60.00	60.91	84.75	79.38	81.98
ATL [15]	56.26	39.25	46.24	**63.68**	63.40	63.54	84.30	81.28	82.76
ANNSA	49.10	50.46	**48.84**	58.82	**73.34**	**64.18**	82.73	**83.52**	**83.06**

achieved noticeable improvements on all datasets without any additional resources.

Furthermore, the ratio between positive and negative examples is high in the Twitter dataset, approximately 1:7.7. ANNSA still achieved good performance on this dataset, which suggests that ANNSA is not sensitive to the problem of unbalanced labels. Although the language used on social media often deviates from punctuation rules and formal English grammar, ANNSA was not affected by such language expression.

4.5. Ablation study

To further verify the validity of our proposed method, we conducted an ablation study. Tables 5, 6 and 7 provide the results obtained on the three social media corpora. The ablation tests included the following: 1) without the sentiment-aware attention mechanism (removing the sentiment-aware attention mechanism); 2) without adversarial perturbation (discarding the adversarial perturbation in the embedding layer); and 3) baseline (removing the sentiment-aware attention mechanism and the adversarial perturbation in the embedding layer).

We found that the F1 values on all three datasets were lower with the elimination of either the sentiment-aware attention mechanism or adversarial perturbation. This finding demonstrates the contribution of the sentiment-aware attention mechanism and adversarial perturbation to ADR detection on social media. However, the decline in performance caused by the absence of the sentiment-aware attention mechanism was larger than that caused by the absence of adversarial perturbations on the Twitter and CADEC corpora. We can conclude that sentiment-aware attention is more important than adversarial perturbation for ADR detection. Furthermore, the results show that the sentiment-aware attention mechanism mainly contributed to improving the recall. Moreover, the datasets with relatively balanced data (namely, TwiMed (Twitter) and CADEC) showed some advantages over the Twitter dataset (in which the ratio of negative to positive examples is approximately 7.7:1). However, the adversarial perturbation method could somewhat compensate for the limitations presented by unbalanced data, improving the F1 performance.

4.6. Effectiveness of the adversarial perturbation

To evaluate the effectiveness of the adversarial perturbation method on a low-resource dataset, we conducted numerous experiments with different sized training sets based on the three datasets. We randomly selected subsets from each social media dataset with varying data proportions of 0.2, 0.4, 0.6, and 0.8. The models implemented for comparison included the following: "baseline", which was a model with Bi-LSTM and self-attention, and "baseline (+adv)", which was a baseline model with adversarial perturbation. We show the experimental results of ANNSA and the other models on the three datasets in Figs. 2, 3, and 4, with F1 as the evaluation index.

Figs. 2, 3, and 4 show that the performance on all three datasets improved significantly as the size of the training set increased, especially on TwiMed (Twitter). Moreover, ANNSA, which includes sentiment-aware attention and adversarial perturbation, achieved a higher F1 than the "baseline" model on all three datasets, thus demonstrating the effectiveness of these components on small datasets. Compared with the "baseline" model, "baseline (+adv)" achieved better results. These findings support the data augmentation power of adversarial training for cases with few training examples, indicating that adversarial training can facilitate the learning of information from a small-scale training set. Even on the CADEC dataset, the result of the baseline model that

Table 5
Ablation study on Twitter.

Dataset	Twitter			
(%)	P	R	F1	ΔF1
ANNSA	49.10	50.46	48.84	–
w/o sentiment	50.10	46.02	47.39	−1.45
w/o adversarial	46.14	49.16	47.09	−1.75
Baseline	48.26	46.80	46.85	−1.99

Table 6
Ablation study on TwiMed (Twitter).

Dataset	TwiMed (Twitter)			
(%)	P	R	F1	ΔF1
ANNSA	58.82	73.34	64.18	–
w/o sentiment	55.63	70.70	61.04	−3.14
w/o adversarial	55.92	72.59	62.06	−2.12
Baseline	57.24	65.43	60.24	−3.94

Table 7
Ablation study on CADEC.

Dataset	CADEC			
(%)	P	R	F1	ΔF1
ANNSA	82.73	83.52	83.06	–
w/o sentiment	80.10	84.06	81.97	−1.09
w/o adversarial	79.37	84.41	81.77	−1.29
Baseline	78.25	83.89	80.91	−2.15

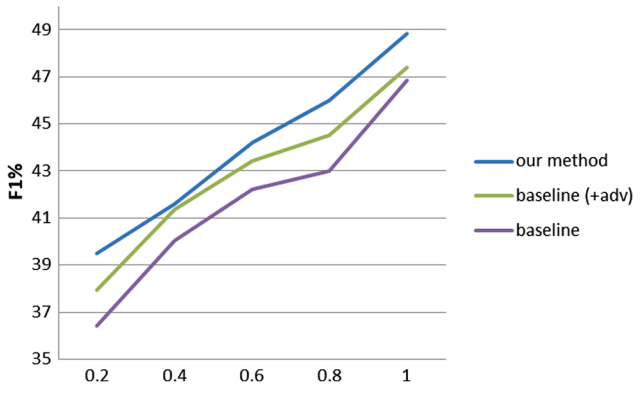

Fig. 2. Experimental results on the scaled Twitter dataset.

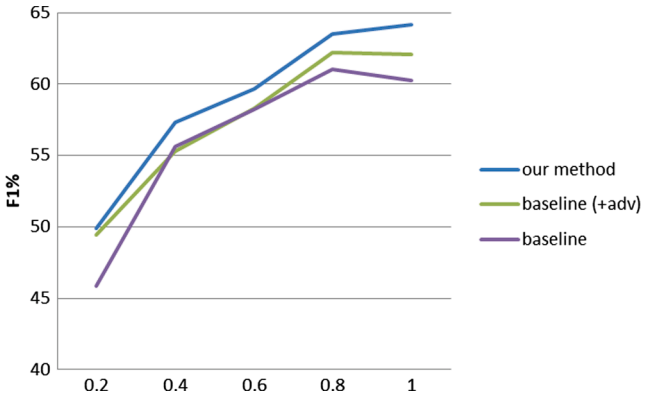

Fig. 3. Experimental results on the scaled TwiMed (Twitter) dataset.

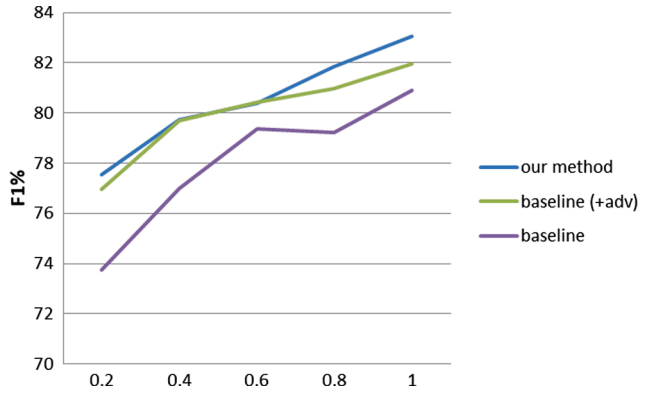

Fig. 4. Experimental results on the scaled CADEC dataset.

included adversarial data was the same as that of ANNSA for data scales of 0.4 and 0.6.

4.7. Case study

In this section, we introduce several case studies to visually explore the influence of the sentiment-aware attention mechanism on ANNSA. The visualization results are illustrated in Fig. 5. We list the visualization results, sentiment words, and ADR mentions. The results show that the incorporation of sentiment-aware attention can help the model to focus on ADR mentions that contribute to the detection of ADRs.

In the example shown in Fig. 5(a), ANNSA strongly considered the ADR mention "tendon damage" and the sentiment words "treated", "tendon" and "achilles", which were important for inferring whether this sentence mentions an ADR. In the example shown in Fig. 5(b), the words "sleepp" and "dirty" gained high attention scores, where "sleepp" is an ADR mention and "dirty" is a sentiment word. Although the ADR mention "sleepp" was misspelled (it should be "sleep"), ANNSA could still recognize this word and assign a high attention score to it via character-level embeddings. This shows that ANNSA is robust to nonstandard descriptive texts from social media. Furthermore, not all sentiment words played an important role, such as "wonder" and "win". The visualization results show that the model could effectively focus on ADR mentions via the sentiment-aware attention mechanism.

5. Discussion

Our paper provides a new look (sentiment factor) at ADR detection on social media. The ADRs usually cause physical or mental distress to patients after drugs. When patients describe their feelings, personal emotion will be included in their words. To investigate the relationship between the ADR mentions and sentiment words, we compared the two types of words in the same sentence (section 4.1). We observed that there is a large amount of overlap between the ADR mentions and sentiment words, which indicated that the ADR mentions contained rich sentiment information. Hence, we proposed to introduce sentiment features into the detection of ADRs.

Previous studies [10,14,33] have found that sentiment information is useful for the ADR detection task on social media. However, these methods only considered the coarse-grained sentiment features of text, did not gain some insights into the relationship between ADRs and emotions. To address these issues, we used a sentiment-aware attention mechanism to incorporate sentiment features into the ADR task. Experimental results demonstrate that the fine-grained sentiment features can effectively obtain high-quality ADR information. Additionally, to improve the robustness of the model, we utilized adversarial perturbation to generate adversarial examples. This method can provide an additional regularization benefit for examples that are slightly different from the training examples.

The sentiment-aware attention proposed by our paper is base on word-level features. This method can be extended to other NLP tasks to achieve specific feature fusion, such as domain features or lexical features. Furthermore, this method can not only effectively integrate a single feature, but also realize the effective integration of multiple features.

This study has the following limitations: (1) Although ANNSA has achieved higher F1 scores, in this work, only sentiment words were considered without distinguishing positive or negative emotions. From the visualization of the sentiment-aware attention mechanism (Section 4.7), we observed that not all sentiment words played a promoting role in the judgment of ADRs. Therefore, it is promising to consider different emotional polarities and perform specific operations in the detection of ADRs. (2) The judgment of ADRs belongs to the category in the biomedical field. The analysis of drugs or adverse reactions may be more accurate when combined with biomedical knowledge. However, in this paper, we only considered the sentiment information and the deep semantic information of the text without biomedical information. (3) We evaluated the effectiveness of our method on a relatively small dataset in this paper. Because of the Twitter's privacy policy, the tweet text can only be obtained based on unique tweet IDs. However, the tweets may

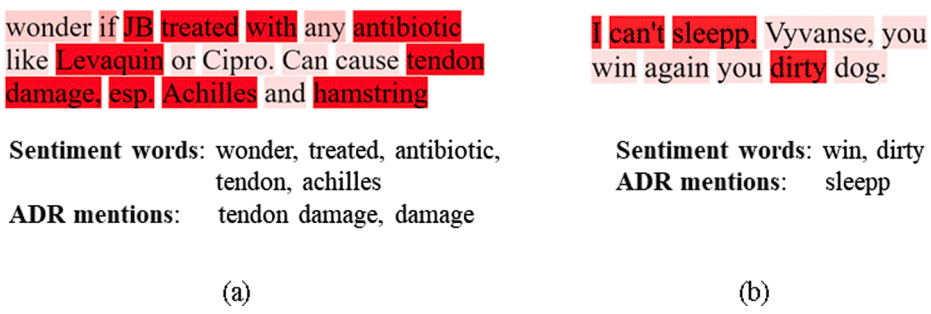

Fig. 5. Visualization of the sentiment-aware attention mechanism. Darker colours represent higher attention weights.

disappear after a while, we can't obtain all tweets by tweet IDs. Since the generalization of the method is crucial to the over-all viability of this task, it is promising to verify the validity of the model on larger-scale datasets. (4) Patients often take multiple drugs at the same time. However, the method proposed by our paper is effective for the discovery of adverse reactions caused by a single drug, ignoring drug-drug interactions.

Our future work will include the following: (1) We will consider the role of emotional polarities in ADR detection and explore different methods to fuse sentiment features with textual information. (2) We will explore how to integrate biomedical knowledge into social media to enhance specialist guidance for ADR detection. (3) We will further verify the effectiveness and generalization of our model on larger-scale accessable datasets. (4) Since the occurrence of adverse reactions is not only caused by a single drug, we want to explore the interaction relations between multiple drugs for ADR detection.

6. Conclusion

In this paper, we propose an adversarial network with a sentiment-aware attention mechanism that can effectively integrate sentiment features into a model for ADR detection by learning attention scores from sentiment words from social media texts, thereby improving the robustness of the model. Experiments demonstrate that our proposed ANNSA model achieves significant results in the ADR detection task on three social media datasets. Analyses suggest that the sentiment-aware attention mechanism can help the model to focus on ADR mentions and that adversarial learning can further enhance the performance of the model for a limited amount of data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been supported by the Natural Science Foundation of China (No. 62076046, 62006034),Natural Science Foundation of Liaoning Province(No. 2021-BS-067).

References

[1] J. Lazarou, B.H. Pomeranz, P.N. Corey, Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies, JAMA 279 (15) (1998) 1200–1205.
[2] D.C. Classen, S.L. Pestotnik, R.S. Evans, et al., Adverse drug events in hospitalized patients: excess length of stay, extra costs, and attributable mortality, JAMA 277 (4) (1997) 301–306.
[3] S.R. Ahmad, Adverse drug event monitoring at the Food and Drug Administration, J. Gen. Intern. Med. 18 (1) (2003) 57–60.
[4] R. Xu, Q.Q. Wang, Large-scale combining signals from both biomedical literature and the FDA Adverse Event Reporting System (FAERS) to improve post-marketing drug safety signal detection, BMC Bioinf. 15 (1) (2014) 17.
[5] T. Zhang, H. Lin, Y. Ren, et al., Adverse drug reaction detection via a multihop self-attention mechanism, BMC Bioinf. 20 (1) (2019) 479.
[6] H. Gurulingappa, A.M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, L. Toldo, Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports, J. Biomed. Inform. 45 (5) (2012) 885–892.
[7] S. Santiso, A. Casillas, A. Pérez, The class imbalance problem detecting adverse drug reactions in electronic health records, Health Inform. J. 25 (4) (2019) 1768–1778.
[8] N. Alvaro, Y. Miyao, N. Collier, TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations, JMIR Public Health Surveillance, 2017, 3(2): e24.
[9] X. Zou, J. Yang, J. Zhang, Microblog sentiment analysis using social and topic context, PloS one 13 (2) (2018) e0191163.
[10] Z. Li, H. Lin, W. Zheng, An effective emotional expression and knowledge-enhanced method for detecting adverse drug reactions, IEEE Access 8 (2020) 87083–87093.
[11] S.U. Hassan, N.R. Aljohani, N. Idrees, et al., Predicting literature's early impact with sentiment analysis in Twitter, Knowledge-Based Syst. 192 (2020), 105383.
[12] K. Darwish, P. Stefanov, M. Aupetit, et al., Unsupervised user stance detection on twitter, Proceedings of the International AAAI Conference on Web and Social Media, 2020, 14: 141-152.
[13] A. Agarwal, B. Xie, I. Vovsha, et al., Sentiment analysis of twitter data, Proceedings of the workshop on language in social media (LSM 2011), 2011, 30-38.
[14] C. Shen, H. Lin, K. Guo, K. Xu, Z. Yang, J. Wang, Detecting adverse drug reactions from social media based on multi-channel convolutional neural networks, Neural Comput. Appl. 31 (9) (2019) 4799–4808.
[15] Z. Li, Z. Yang, L. Luo, et al., Exploiting adversarial transfer learning for adverse drug reaction detection from texts, J. Biomed. Inform., 2020: 103431.
[16] A. Sarker, G. Gonzalez, Portable automatic text classification for adverse drug reaction detection via multi-corpus training, J. Biomed. Inform. 53 (2015) 196–207.
[17] S. Yadav, A. Ekbal, S. Saha, et al., A unified multi-task adversarial learning framework for pharmacovigilance mining, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 5234-5245.
[18] S. Chowdhury, C. Zhang, P.S. Yu, Multi-task pharmacovigilance mining from social media posts, Proceedings of the 2018 World Wide Web Conference (2018) 117–126.
[19] W.E. Zhang, Q.Z. Sheng, A. Alhazmi, C. Li, Adversarial attacks on deep-learning models in natural language processing: A survey, ACM Transactions on Intelligent Systems and Technology (TIST) 11 (3) (2020) 1–41.
[20] R. Harpaz, W. DuMouchel, N.H. Shah, D. Madigan, P. Ryan, C. Friedman, Novel data-mining methodologies for adverse drug event discovery and analysis, Clin. Pharmacol. Ther. 91 (6) (2012) 1010–1021.
[21] H. Gurulingappa, A. Mateen-Rajpu, L. Toldo, Extraction of potential adverse drug events from medical case reports, J. Biomed. Semantics 3 (1) (2012) 1–10.
[22] R. Ginn, P. Pimpalkhute, A. Nikfarjam, et al., Mining Twitter for adverse drug reaction mentions: a corpus and classification benchmark, Proceedings of the fourth workshop on building and evaluating resources for health and biomedical text processing, 2014, 1-8.
[23] A. Nikfarjam, A. Sarker, K. O'connor, et al., Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features, J. Am. Med. Inform. Assoc. 22 (3) (2015) 671–681.
[24] M. Kuhn, M. Campillos, I. Letunic, et al., A side effect resource to capture phenotypic effects of drugs, Mol. Syst. Biol. 6 (1) (2010) 343.
[25] J. Bian, U. Topaloglu, F. Yu, Towards large-scale twitter mining for drug-related adverse events, Proceedings of the 2012 international workshop on Smart health and wellbeing, 2012, 25-32.
[26] M. Yang, X. Wang, M.Y. Kiang, Identification of consumer adverse drug reaction messages on social media, PACIS, 2013, 193.
[27] A. Patki, A. Sarker, P. Pimpalkhute, et al., Mining adverse drug reaction signals from social media: going beyond extraction, Proc. BioLinkSig, 2014, 2014: 1-8.
[28] Z. Zhang, J.Y. Nie, X. Zhang, An ensemble method for binary classification of adverse drug reactions from social media, Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing (2016) 1.
[29] M. Rastegar-Mojarad, R.K. Elayavilli, Y. Yu, et al., Detecting signals in noisy data-can ensemble classifiers help identify adverse drug reaction in tweets. Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing, 2016.
[30] C. Wu, F. Wu, J. Liu, et al., Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tweet representation and multi-head self-attention, Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task. 2018: 34-37.
[31] K. Lee, A. Qadir, S.A. Hasan, et al., Adverse drug event detection in tweets with semi-supervised convolutional neural networks, Proceedings of the 26th International Conference on World Wide Web, 2017, 705-714.
[32] C. Sun, L. Huang, X. Qiu, Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence, arXiv preprint arXiv:1903.09588, 2019.
[33] H. Alhuzali, S. Ananiadou, Improving classification of adverse drug reactions through using sentiment analysis and transfer learning, Proceedings of the 18th BioNLP Workshop and Shared Task, 2019: 339-347.
[34] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572, 2014.
[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., Generative adversarial nets, Adv. Neural Inform. Process. Syst. 27 (2014) 2672–2680.
[36] T. Miyato, A.M. Dai, I. Goodfellow, Adversarial training methods for semi-supervised text classification, arXiv preprint arXiv:1605.07725, 2016.
[37] M. Yasunaga, J. Kasai, D. Radev, Robust multilingual part-of-speech tagging via adversarial training, arXiv preprint arXiv:1711.04903, 2017.
[38] J.T. Zhou, H. Zhang, D. Jin, et al., Dual adversarial neural transfer for low-resource named entity recognition, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 3461-3471.
[39] A. Karimi, L. Rossi, A. Prati, et al., Adversarial training for aspect-based sentiment analysis with BERT, arXiv preprint arXiv:2001.11316, 2020.
[40] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[41] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, arXiv preprint arXiv:1509.01626, 2015.
[42] Y. Wang, A. Sun, J. Han, et al., Sentiment analysis by capsules, Proceedings of the 2018 World Wide Web Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2018: 1165-1174.
[43] W. Zhao, J. Ye, M. Yang, et al., Investigating capsule networks with dynamic routing for text classification, arXiv preprint arXiv:1804.00538, 2018.

[44] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781, 2013.
[45] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014: 1532-1543.
[46] A.N. Samatin Njikam, H. Zhao, Chartec-net: An efficient and lightweight character-based convolutional network for text classification, J. Electr. Comput. Eng. 2020 (2020).
[47] E. Cambria, R. Speer, C. Havasi, et al., Senticnet: A publicly available semantic resource for opinion mining, AAAI fall symposium: commonsense knowledge, 2010, 10(0).
[48] P. Gonçalves, M. Araújo, F. Benevenuto, et al., Comparing and combining sentiment analysis methods, Proceedings of the first ACM conference on Online social networks (2013) 27–38.
[49] S. Rill, D. Reinel, J. Scheidt, et al., Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowl.-Based Syst. 69 (2014) 24–33.
[50] X. Li, H. Xie, L. Chen, et al., News impact on stock price return via sentiment analysis, Knowl.-Based Syst. 69 (2014) 14–23.
[51] A. Muhammad, N. Wiratunga, R. Lothian, Contextual sentiment analysis for social media genres, Knowl.-Based Syst. 108 (2016) 92–101.
[52] P. Cao, Y. Chen, K. Liu, et al., Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 182–192.
[53] J.L. Elman, Finding structure in time, Cognitive Sci. 14 (2) (1990) 179–211.
[54] A. Madry, A. Makelov, L. Schmidt, et al., Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083, 2017.
[55] A. Sarker, A. Nikfarjam, G. Gonzalez, Social media mining shared task workshop, Biocomputing 2016: Proceedings of the Pacific Symposium (2016:) 581–592.
[56] S. Karimi, A. Metke-Jimenez, M. Kemp, et al., Cadec: A corpus of adverse drug event annotations, J. Biomed. Inform. 55 (2015) 73–81.
[57] T. Huynh, Y. He, A. Willis, et al., Adverse drug reaction classification with deep neural networks, Coling (2016).