Classification of Adverse Drug Reactions in Social Media

Rong Zhou
School of Economics and Management, Xi’an 710071, China.

Abstract
There are a large number of tweets with adverse drug reactions in social media, and the semantics of different levels of tweets are different. This article proposes a deep learning network method based on supervised learning, starting from different levels of tweets, find the special semantics contained in different levels, then merge them, and use the BiLSTM method to build a classification model to find tweets related to adverse drug reactions. This article starts from the three levels of sentences; the first level: the character level, using the charCNN model to extract features at the character level; the second level of words, using the BERT model to extract features at this level Vector, third level: sentence level, use the syntactic dependency toolkit developed by Stanford to extract the relationship pairs between words in the sentence and vectorize them; after obtaining the three levels of feature vectors, use the BiLSTM model Build the classifier and finally complete the model.

Keywords
Social Media; Dependency Parsing; ADR; BERT.

1. Introduction
At present, Adverse Drug Reactions (ADR) is one of the hot issues that the medical community and the public are concerned about. In the past, people were also concerned about the safety of medications. Some drugs may have reactions that are different from the results of clinical trials. However, due to the limitations of communication channels, most people can only use their own experience and the experience of those around them, and they cannot form information or knowledge. However, the emergence of the Internet has broken the communication restrictions between people. People can post their thoughts on social media. Of course, adverse drug reactions will also appear on public social media including Twitter, or Those professional health websites (such as MedHelp) that focus on health issues. How to find information about adverse drug reactions in the vast number of social media information has become the only way to study drug safety through social media.
Social media brings both opportunities and challenges for ADR related work. Social media contains a large number of user-published data, which can truly reflect and portray users’ lives, but because these texts are non-standardized languages, data mining tasks become more difficult. In addition, finding content related to medication safety is also a very difficult task.
Based on this, this paper proposes a social media adverse drug reaction classification model based on multi-level vectorized representation of sentences with different granularities and BiLSTM network. This article starts from the three granularities of sentences. At the character level, charCNN is used to extract text features; at the word level, the BERT (bidirectional encoder representation from transformers) model is used to extract text features; at the sentence level, syntactic dependence is used to extract the grammar of the sentence. Relationship, because BERT can extract semantic features at the sentence level. Then connect the three features and input them into the BiLSTM model for further coding and classification.
The follow-up content of this article is as follows: Section 2 introduces related work, Section 3 describes in detail the model and related methods constructed in this study, Section 4 is the
specific experimental process and experimental results, Section 5 summarizes the full text, and gives to sum up.

2. Literature Reference

The popularity of the Internet enables it to provide patients with a platform for communicating with each other and sharing their medication experience. The openness and convenience of the Internet compared to other platforms makes it very convenient to collect texts related to adverse drug reactions. Based on this, research on adverse drug reactions on social media has gradually become popular. This includes many models based on machine learning methods, such as Naive Bayes [1], SVM [2], Maximum Entropy Classifier [3], Hidden Markov Model [4] and other methods, which have achieved certain research results.

In 2013, Jiang K et al. crawled from Twitter and related tweets of five drugs with a large number of users, such as Duloxetine, which were earlier on the market, and constructed naive Bayes, SVM, and maximum entropy classifiers to identify tweets related to adverse reactions. Finally, a professional medical text dictionary was used to identify adverse reaction entities from these tweets, thereby improving the progress of the experiment. In 2015, Sarker and Gonzalez [5] used support vector machine classifier (SVM) to conduct adverse drug reaction experiments by manually extracting word features. In 2016, Huynh et al. used CNN to extract features, but the effect was average [6]. In the 2017 competition The Social Media Mining for Health (SMM4H), many participating teams used CNN and SVM to test the classification of adverse drug reactions on social media. The LSTM model has also been used in related competitions in 2018. In 2019, the BERT model has also become the model favored by the participating teams in the competition. In 2019, Xinyan Zhao et al. used the LSTM model based on attention and aggregated representation to achieve the task, showing that contextual text can help discover useful local information and improve overall medical concepts [7]. In 2020, Lin Hongfei and others used a method based on the combination of a capsule network and a long and short-term memory network to detect adverse drug reactions in social media, and proved that the method has good performance.

The development of deep learning models in the field of natural language processing is rapid, from the earliest RNN (Recurrent neural network) to the LSTM (Long Short Term Memory Network) model. Kim proposed the TextCNN method for text classification in 2014, and it has good results on multiple data sets [8]. In 2015, Xiang Zhang et al. proposed a text classification based on character-based convolutional neural networks, and the effect is better than traditional methods [9]. In 2015, Tang et al. proposed a sentiment classification model that uses the GRU model to model documents [10]. In 2015, Lai et al. proposed a recurrent neural network classification method without artificial features, referred to as RCNN [11]. In 2017, Felbo et al. proposed the DeepMoji model and achieved good results. The model uses a two-layer Bi-LSTM to capture contextual features, and then uses a line attention mechanism proposed by the author to use the embedding layer and the 2-layer BiLSTM respectively as Input and get the vector representation of the document [12]. Google proposed the Transformer model in 2017, which opened up a new method in the field of NLP [13]. And in 2019, the pre-training model BERT was proposed and achieved excellent results [14].

Syntactic dependency analysis has always been a hot issue in the field of NLP. Melchuk comprehensively and systematically studied the dependency grammar theory of English in 1988. Eisner proposed the idea of transforming syntactic dependency into dependency book in 1997 [15]. Stanford University has developed a corresponding tool library for it-Stanford CoreNLP. In 2016, Alexandros Komnions et al. proved that word embedding based on syntactic dependency is superior to word embedding based on context [16]. In 2016, Timothy Dozat et al. proposed a regularized interpreter, using a biaffine classifier model to achieve syntactic
dependence, and achieved the most advanced or nearly the most advanced performance on the standard libraries of six different languages [17]. In 2017, Hao Peng et al. proposed a deep neural network model that parses sentences into three semantic dependency graph forms. The model is based on the BiLSTM model and found to have significant effects [18]. In 2020, Juntao Yu et al. used the idea of graph-based dependency analysis to score the start tag and end tag pair in a sentence through the biaffine model, and use it to explore all ranges so that the model can accurately predict named entities. After evaluating 8 corpora, the accuracy rate increased by 2.2% [19].

3. Research Method

With the in-depth study of deep learning network models in the NLP field, more and more researchers use deep learning network models for text classification research. This article is divided into three granularities of sentences-characters, words, and sentences. After modeling, the three-dimensional feature vectors are combined, and then the BiLSTM classifier is used to form the adverse drug reaction classification model on social media in this article. First, based on the feature extraction of character-level information, charCNN is used for feature extraction; secondly, the word-level feature vector is extracted using the BERT model; then, the syntactic dependency toolkit provided by Stanford University is used to obtain the dependency relationship between words used to represent sentence-level features; finally, the three feature vectors are fused, and the vector sequence is input into BiLSTM for further semantic coding and classification, and finally the classification result is obtained.

3.1. CharCNN

charCNN is a character-level text classification model proposed by Xiang Zhang and Junbo Zhao et al. in 2015 [9]. Studies have proved that when the training set is large enough, the convolutional network does not require meaning at the word level, nor does it require information such as the grammatical and syntactic structure of the language to achieve excellent results.

Figure 1. CharCNN model

The network structure of charCNN is shown in Figure 1, which is mainly composed of three parts: convolutional layer, pooling layer and fully connected layer, and its approach is similar to general CNN operations. The focus of charCNN is on character quantization. The input of the model is continuous characters. First, the characters need to be converted into a sequence of numbers that can be processed. An alphabet of size m is specified in the original charCNN article:

```
abcdefghijklmnopqrstuvwxyz0123456789
-,.!?"'\@#$%ˆ&*˜’+\-=\[\]\{\}
```

305
Convert each character into a one-hot vector, and then convert the character sequence into a sequence with a fixed length of dimension m, and convert the characters or empty characters that are not in the alphabet into an m-dimensional all-zero vector [9].

Next, it will go through multi-layer convolution and pooling operations. In the original paper, 6 convolutional layers are designed. In this study, in order to improve the execution efficiency of the model, the number of convolutional layers is reduced to 3 layers, and finally following the fully connected layer, there are 3 fully connected layers in the original paper. The output of the first two layers is set to 1014, and the output of the last layer is set according to its own task. In this research, a layer of fully connected The characteristics of the connection layer.

3.2. Bi-LSTM

Long Short Term Memory (LSTM) is a special RNN proposed by Hochreiter and Schmidhuber. LSTM controls the transmission state through the gated state, remembering what needs to be remembered for a long time, forgetting is not important Information; instead of like RNN, there can only be a superposition of memories. LSTM can effectively solve the problem of gradient dispersion and gradient explosion, and it has significant effects for many tasks that require "long-term memory".

The cell structure of LSTM is shown in the figure 2:

![LSTM cell model](image)

The core part of the LSTM structure is the uppermost line in the figure above---C, called the cell state, which has always existed in the entire LSTM system, among which:

\[ C_t = f_t \cdot C_{t-1} + i_t \cdot C_t \]  

Among them, it is called the forget gate, which expresses the needs of used as a feature for calculation. Is a vector with each element in the range [0,1]. The sigmoid function is usually used as the activation function. The \( \otimes \) in the above figure represents the most important gate mechanism in LSTM, and the forgetting gate represents the unit multiplication relationship between and, which is represented by the following formula:

\[ f_t = \sigma(Wf \cdot [h_{t-1}, x_t] + b_f) \]  

Among them, \( C_t \) represents the unit status update value, which is obtained by the following function:

\[ C_t = \tanh(Wc \cdot [h_{t-1}, x_t] + b_c) \]  

\( i_t \) is called the input gate, which is calculated by \( x_t \) and \( h_{t-1} \) via the sigmoid activation function, as shown below:
\[ i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \]  \hspace{1cm} (4)

Some of the features \( C_t \) are used to control \( i_t \) are used to update \( C_t \), the same as, as shown below:

\[ C_t = f_i * C_{t-1} + i_t * C_t \]  \hspace{1cm} (5)

Finally, in order to calculate the predicted value and generate the complete input for the next time slice, the hidden node output needs to be calculated, as shown below:

\[ o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \]  \hspace{1cm} (6)

\[ h_t = o_t * \tanh(C_t) \]  \hspace{1cm} (7)

The core content of LSTM includes two information streams and three gate structures. Through the structure diagram of LSTM, it is found that there is a fatal problem in modeling sentences with LSTM: it is impossible to encode information from back to front. We can see from the two information streams of unit state and hidden node output that LSTM is a one-way information propagation model, but for some text output tasks, the latter information can still affect the preceding words and sentences. This is a problem that LSTM cannot solve.

In order to solve this problem, the BiLSTM model is proposed. The main idea is to add a backward LSTM model to the forward LSTM model. At the same time, the forward and backward layers jointly connect the output layer and share weights among them. Its network structure is shown in the figure 3:

![Bi-LSTM model](image)

Figure 3. Bi-LSTM model

Calculate the forward direction from time 1 to \( t \) time \( t \) in the forward layer, and obtain and save the output of the hidden node at each time. The backward layer calculates backward from time \( t \) to time 1, and obtains and saves the hidden node output at each time. Finally, at each moment, combine the output results of the forward layer and the backward layer at the corresponding time to obtain the final output result, as shown below:

\[ h_t = f(w_1x_t + w_2h_{t-1}) \]  \hspace{1cm} (8)

\[ \hat{h}_t = f(w_3x_t + w_4h_{t+1}) \]  \hspace{1cm} (9)

\[ o_t = g(w_5h_t + w_6\hat{h}_t) \]  \hspace{1cm} (10)
4. Experimental Results and Analysis

In this paper, the BiLSTM, charCNN, and charCNN-BiLSTM methods are selected as controlled experiments, and the performance of the model designed in this study in the identification of adverse drug reactions on social media has been tested. The data set used in this study is the data set of the 2017 SMM4H shared task evaluation. This data set mainly provides the ID numbers of Twitter users posting posts. However, due to factors such as user deletion of posts, 7,168 Twitter tweets were finally collected.

Due to the small number of samples in the data set in this study, in order to fully train, the data set division of 6:2:2 was not adopted, but the data set division of 8:2 was adopted. The data set of the final test set is at the end of training later, randomly sampled on the entire data set.

In order to verify the effectiveness of the model proposed in this research, we have done a number of comparative experiments. Through comparative analysis with the classic model, the performance of this model is reflected, and the test results are shown in the table 1:

<table>
<thead>
<tr>
<th>Method</th>
<th>F1(percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiLSTM</td>
<td>65.19</td>
</tr>
<tr>
<td>CNN</td>
<td>64.99</td>
</tr>
<tr>
<td>CNN-BiLSTM</td>
<td>84.25</td>
</tr>
</tbody>
</table>

Through comparative experiments, it can be seen that the deep learning model network based on the fusion of BERT based on different sentence granularities proposed in this paper is helpful to improve the classification performance compared with traditional deep learning algorithms. Because traditional deep learning models start from a granularity of sentences to do feature extraction, and most of them are based on word2Vec word vectors as model input. Although the word2Vec word vector table is powerful enough to represent some sentence-level information, it is lacking in the entire model. Although the BERT model is particularly powerful when it is pre-trained, and it is characterized at the word level and semantic level of the sentence, it lacks character-level feature representation and lacks the representation of the relationship at the sentence grammar level. In this study, starting from the three granularities of social media tweets, feature extraction and model construction can effectively improve the identification efficiency of adverse drug reactions on social media.

5. Summary

With the development and growth of the Internet, the Internet world and the real world have more intersections, and social media is a concentrated expression of the intersections. In social media, people can share what they have seen and heard in real life, and because of the nature of social media, people can get rid of some of the shackles in reality, and can express their opinions and feelings more freely. Adverse drug reactions are a topic that can never be avoided with the production and use of drugs. Although a lot of related tests have been done before leaving the factory, various problems will still occur during use. The emergence of social media makes Analyzing the medication experience of ordinary people has become one of the feasible ways.

This paper uses natural language processing technologies such as charCNN, and BiLSTM to characterize the different granular features of Twitter tweets to construct the model of this study. The research in this paper shows that through the model construction of feature vectors with different granularities, it is possible to achieve effective drug adverse reaction detection tasks, which is of great significance for improving the medication experience and ensuring
medication safety. Although the results of the experiment are objective, the F1 value for the
detection of adverse drug reactions based on social media is generally low. This is related to the
imbalance of positive and negative cases in social media and tweets, and it is also related to the
excessive colloquial expression on social media.

References

extraction [C]// AAAI-99 workshop on machine learning for information extraction. 1999: 37-42.
Coling. 2016.
[7] Zhao X, Yu D, Vydiswaran V G V. Identifying adverse drug events mentions in tweets using attentive,
collocted, and aggregated medical representation [C]// Proceedings of the Fourth Social Media
[8] Yoon Kim.Convolutional Neural Networks for Sentence Classification [C]// Conference on empirical
methods in natural language processing, vol. 3: Conference on empirical methods in natural
classification [C]// Proceedings of the 2015 conference on empirical methods in natural language
processing. 2015: 1422-1432.
Proceedings of the AAAI Conference on Artificial Intelligence. 2015, 29(1).
representations for detecting sentiment, emotion and sarcasm[J]. arXiv preprint arXiv:1708.00524,
2017.
2017.
[16] Komninos A, Manandhar S. Dependency based embeddings for sentence classification tasks[C]//
Proceedings of the 2016 conference of the North American chapter of the association for
1611.01734, 2016.