IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 6, 2020, accepted January 12, 2020, date of publication January 17, 2020, date of current version January 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2967103

Sentiment Classification Based on Part-of-Speech and Self-Attention Mechanism

KEFEI CHENG[1], YANAN YUE[2], AND ZHIWEN SONG[2]

[1]College of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2]College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Corresponding author: Yanan Yue (s170231063@stu.cqupt.edu.cn)

This work was supported by the National Social Science Foundation of China under Grant 17XFX013.

ABSTRACT Currently, various attention-based neural networks have achieved successes in sentiment classification tasks, as attention mechanism is capable of focusing on those words contributing more to the sentiment polarity prediction than others. However, the major drawback of these approaches is that they only pay attention to the words, the sentimental information contained in the part-of-speech(POS) is ignored. To address this problem, in this paper, we propose Part-of-Speech based Transformer Attention Network(pos-TAN). This model not only uses the Self-Attention mechanism to learn the feature expression of the text but also incorporates the POS-Attention, which uses to capture sentimental information contained in part-of-speech. In addition, our innovative introduction of the Focal Loss effectively alleviates the impact of sample imbalance on model performance. We conduct substantial experiments on various datasets, and the encouraging results indicate the efficacy of our proposed approach.

INDEX TERMS Part-of-speech, self-attention mechanism, sentiment classification, focal loss.

I. INTRODUCTION

The task of sentiment classification is to divide the text into two or more types of praise or derogatory according to the meaning and sentimental information expressed by the text. Sentiment classification is the division of the author's sentiment orientation, viewpoint and attitude, which can solve the disorder of various commentary information on the Internet to a certain extent, and it is convenient for users to accurately locate the required information. At present, the mainstream sentiment classification methods are mainly divided into three categories: 1) lexicon-based methods; 2) machine learning based methods; 3) deep learning based methods. The key part of the lexicon-based method is '' lexicon + rule'', that is, the sentimental lexicon is used as the main basis for judging the sentimental polarity, and the corresponding judgment rules are designed [1]–[3]. Machine learning based sentiment classification methods are broadly divided into supervised, unsupervised and semi-supervised. Traditional machine learning methods are dedicated to manually extracting an abundance of features like bag-of-words and TF-IDF which are used to train a sentiment classifier such as SVM [4], [5], NB [6]. As we all know, deep learning models are becoming more popular because they can automatically learn semantic representations from high-dimensional original data without carefully designed feature engineering. As a powerful technique for text modeling, Convolutional Neural Network (CNN) and Recurrent Neural Networks (RNNs) have been widely applied to sentiment analysis tasks [7]–[11]. However, these methods can automatically learn text features, but cannot focus on the parts that are important to sentiment classification.

As a simulation of human attention, the attention mechanism can focus on specific parts of the text, so it is applied to the neural network to achieve various tasks, including machine translation [12], [13], reading comprehension [14], [15], image processing [16]–[18], etc. Also, there are already some works using attention mechanism to deal with sentiment classification [19]–[21]. However, these approaches only pay attention to the words, the part-of-speech features are ignored. In fact, part-of-speech contains sentimental information that is helpful for classification. Tang *et al.* [22] compared the classification effects of nouns, verbs, adjectives, and adverbs as features, and found that all four parts-of-speech have different degrees of sentimental color. Some previous works regarding the POS and sentiment classification have been implemented such as [23]–[26]. Based on this insight, in this paper, we put forward the POS-Attention.

The associate editor coordinating the review of this manuscript and approving it for publication was Claudio Cusano.

Each part-of-speech is integrated into the network as a vector expression, and different parts-of-speech are weighted through attention mechanism, so as to learn different features.

But, most existing attention-based approaches rely on CNN and RNNs. CNN is able to capture the local features, but ignores long-term dependencies between words. The difference between RNNs and CNN based model is that RNNs are better at modeling long-distance semantics in text and capturing contextual information. Inspired by Google's Transformer model [27], which completely abandon RNNs and achieve good results in machine translation tasks only using attention mechanism, we design a new Part-of-Speech based Transformer Attention Network(pos-TAN). This model uses the Self-Attention mechanism to learn the dependencies between words in different positions and capture feature information at different spatial levels. And it also incorporates part-of-speech information to mine the modified relations between words, acquire richer sentimental features. On the other hand, most of the studies on sentiment classification are based on the assumption that the samples are balanced. While in practice, the number of samples in different categories varies greatly, and imbalances are common, so it is necessary to take appropriate methods to deal with it.

The work of this paper is summarized as follows.

- In view of the sample imbalance problem, this paper introduces Focal Loss [28] as the loss function of the sentiment classification model. By reducing the weight of the samples in a larger number of classes, the model pays more attention to the samples in a small number of classes in training, thus alleviate the impact of sample imbalance on classification to some extent.
- We design the POS-Attention to capture the sentimental information. Each part-of-speech is integrated into the network as a vector expression, and different parts-of-speech are weighted through attention mechanism, so as to learn different sentimental features.
- We propose a new Part-of-Speech based Transformer Attention Network(pos-TAN), which combines part-of-speech and Self-Attention mechanism. The effectiveness of the model on the sentiment classification task is verified by the experiments.

The rest of this paper is organized as follows. We first briefly review sentiment classification methods and introduce the Self-Attention mechanism in Section II. Afterwards, the proposed sentiment classification model (pos-TAN) is presented in Section III. In Section IV, extensive comparison experiments are conducted to prove the superiority of our proposed model. At last, we draw a conclusion and envision the future in Section V.

II. RELATED WORK

A. SENTIMENT CLASSIFICATION

Sentiment classification is one of the main research topics in Natural Language Processing(NLP), it can be treated as traditional text classification, and solved by some general classification models [29]–[31]. Traditional feature engineering-based models usually focus on extracting efficient features such as lexical features [32], [33], topic-based features [34], [35]. With the rapid development of deep learning, CNN and RNNs have been applied to obtain better representations of sentences for sentiment classification. Especially, Long Short-term Memory Networks (LSTMs) have shown a striking promise in sentiment analysis [11], [36]. Cai and Xia [37] used two individual CNN architectures to learn textual features and visual features, which can be combined as the input of another CNN architecture for exploiting the internal relation between text and image, so as to realize multimedia sentiment analysis. Dong *et al.* [9] put forward an adaptive Recursive Neural Network by modeling syntactic relations on tweet data. Lai *et al.* [38] proposed a Recurrent Convolution Neural Network (RCNN) which uses recurrent structures in the convolution layer to classify texts. Wang *et al.* [39] proposed a regional CNN-LSTM model for multi-dimensional sentiment analysis.

More recently, a new research direction in deep learning has emerged, which introduces an attention mechanism to neural network models. The attention mechanism is capable of focusing on the parts of text that are more important to the current task. Hence, various attention-based approaches have been proposed to solve the sentiment analysis task [40]–[42]. Wang *et al.* [43] proposed an Attention-based Long Short-Term Memory Network for aspect-level sentiment classification. Hu *et al.* [44] proposed constrained attention networks (CAN) for multi-aspect sentiment analysis, and introduce orthogonal and sparse regularizations to constrain the attention weight allocation, helping learn better aspect-specific sentence representations.

B. SELF-ATTENTION MECHANISM

The attention mechanism was first proposed in the field of visual images, and then applied by Bahdana et al. to machine translation tasks [12]. Then the attention mechanism is widely used in various NLP tasks based on neural network models such as RNN/CNN, which can help the model select the features that are more important to the current task from many features. Most of the attention models are attached to the Encoder-Decoder framework. The Encoder-Decoder framework can be thought of as a general-purpose processing model for generating another sentence from a sentence. For the sentence pair <Source, Target>, our goal is to give the source sentence, and generate the target sentence through the Encoder-Decoder framework.

Source and Target are respectively composed of their respective word sequences. The nature of Attention can be described as a mapping of a Query to a series of Key-Value pairs to an output, which is shown in Figure 1.

The calculation of Attention is divided into three steps. First, we perform a similarity calculation on Query and each Key, so as to obtain the weight coefficient of the Value corresponding to each Key.

$$f(Q, K_i) = Q^T K_i \tag{1}$$

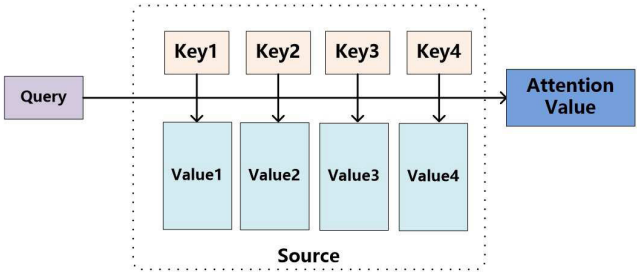

FIGURE 1. The abstract graph of attention.

Then we use Softmax to normalize the weights,

$$a_i = Softmax(f(Q, K_i)) = \frac{\exp(f(Q, K_i))}{\sum_{j=1}^{L} \exp(f(Q, K_j))} \tag{2}$$

Finally, the weights and corresponding values of Value are weighted and added up to obtain the final Attention value.

$$Attention(Q, K, V) = \sum_{i}^{L} a_i * Value_i \tag{3}$$

where L is the length of the input sentence (the length of Source), and Query is a sequence of words in Target. Currently, Key and Value are usually the same in the NLP task.

The Self-Attention mechanism can be seen as a special form of ordinary Attention, which is the Attention between word sequences within Source. Q, K, and V have the same input, that is, each word in the sentence must be calculated for Attention with other words. Self-Attention can ignore the distance between words and calculate the dependencies directly, learn the internal structure of sentences, and better capture the syntactic and semantic information of sentences.

III. THE PROPOSED APPROACH

The overall framework of the model is shown in Figure 2. It is divided into four parts: Embedding Layer, Text Representation Layer, POS-Attention Layer and Classification Layer.

A. EMBEDDING LAYER

Given a sentence $S = \{w_1, w_2, w_3, \ldots, w_n\}$ and the part-of-speech $P = \{p_1, p_2, p_3, \ldots, p_n\}$ of each word w_i, both S and P have a length of n. Each word is mapped into a low-dimensional real-valued vector called Word Embedding. For each word w_i, the vector $e_i \in R^{d_w}$ is obtained after Word Embedding. Since the model in this paper uses

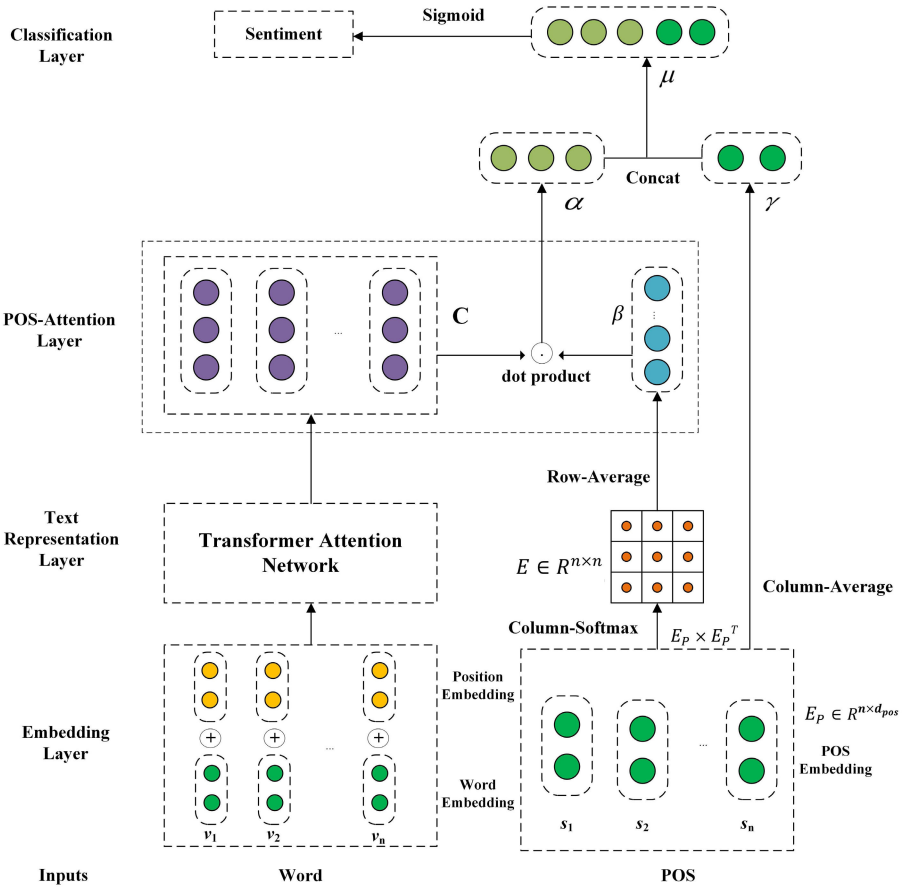

FIGURE 2. Architecture of the part-of-speech based transformer attention network(pos-TAN).

Self-Attention mechanism instead of RNNs, it is impossible to model the sequence of words like RNNs. As a result, relevant position information is added to the input sequence. The vector representation of position information draws on the Transformer [27] model, and the position vector $z_i \in R^{d_w}$ of each word is obtained by Positional Embedding. The word vector has the same dimensions as the position vector, and they are added up to get the final embedding vector v_i of the word:

$$v_i = [e_i + z_i] \in R^{d_w} \tag{4}$$

Here, the parameter matrix in the embedding process is $M^{V \times d_w}$, where d_w is the dimension of the word vector, and V is the vocabulary size.

In linguistics, part-of-speech is a basic grammatical attribute of vocabulary, and also has certain semantic information. Therefore, we introduce part-of-speech features into the model and each part-of-speech is learned as a vector expression. Each part-of-speech p_i obtains the corresponding part-of-speech vector $s_i \in R^{d_{pos}}$ by POS Embedding. The parameter matrix $W^{V_{pos} \times d_{pos}}$, where d_{pos} is the dimension of the part-of-speech vector, and V_{pos} is the size of the part-of-speech table.

B. TEXT REPRESENTATION LAYER

The Transformer model proposed by Google is a machine translation model entirely based on attention mechanism. It abandons the structure of CNN, RNNs, and can learn semantic relations from words far away by means of Self-Attention mechanism, which is not only has a groundbreaking use of a new model structure, but also a significant improvement in mission performance, parallelism and ease of training. The Encoder structure in the Transformer, also called Transformer Attention Network(TAN), which has become an important benchmark model in text feature representation, as shown in Figure 3. TAN mainly uses the Multi-Head Attention mechanism. The core part is Scaled Dot-product Attention. Its calculation formula can be described as:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

where d_k is the dimension of matrix Q and K. Multi-head Attention first performs h times linear transformation on Q, K, and V respectively, where the parameter matrix of each linear transformation is different. Then it conducts the Scaled Dot-product Attention in parallel for h times, and splices the results. Finally, the output of the Multi-head Attention is obtained by linear transformation. This mechanism is similar to convolution operations, allowing the model to learn different information in different representation subspaces, and to characterize the semantic relationships of sentences as much as possible.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{6}$$

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \tag{7}$$

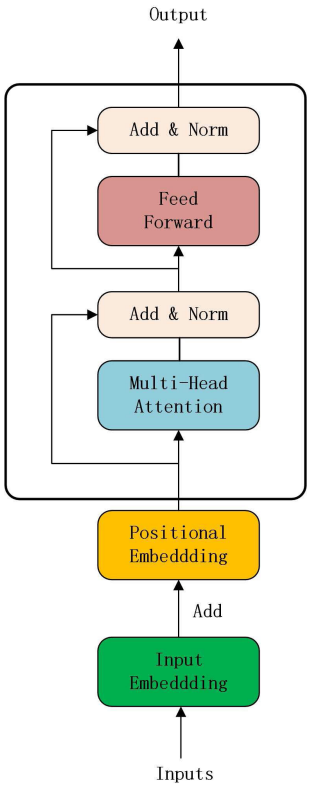

FIGURE 3. Architecture of the transformer attention network (TAN).

Here, the parameter matrix $W_i^Q \in R^{d_{model} \times d_k}$, $W_i^K \in R^{d_{model} \times d_k}$, $W_i^V \in R^{d_{model} \times d_v}$, $W^O \in R^{d_{model} \times hd_v}$, where d_v is the dimension of V, d_{model} is the dimension of the output of Multi-head Attention. In this work, we employ h = 4 parallel attention layers, or heads. For each layer we use $d_k = d_v = d_{model}/h$.

C. POS-ATTENTION LAYER

Part-of-speech is the basic grammatical attribute of words. Words with different parts-of-speech represent different components in sentences. The existing works regarding the POS and sentiment classification can be divided into two categories. One is to distinguish the different roles of each word in sentence structure or semantics by part-of-speech tagging(POS tagging), which is helpful for the model to extract appropriate features [23]–[25]. The other is to map the part-of-speech into a vector representation by embedding, and then directly input it into the model with the word vector for training [45]. However, most of the previous part-of-speech based approaches only learn the shallow part-of-speech features, while the connection between words and parts-of-speech has been ignored. In this paper, part-of-speech is integrated into different layers of the network, and sentimental information contained in part-of-speech is learned at multiple levels.

Here, the part-of-speech is mapped to the part-of-speech vector by embedding layer, then all part-of-speech vectors

form a part-of-speech vector-matrix $E_P \in R^{n \times d_{pos}}$. Especially, we use POS-Attention to fuse the semantic information and part-of-speech information of words. The matrix E_P is multiplied by its transpose matrix to get a new matrix $H = E_P \times E_P^T \in R^{n \times n}$. The values in the matrix H represent the correlation between each part-of-speech. After a column softmax, we get a new matrix $E \in R^{n \times n}$,

$$E_{ij} = \frac{\exp^{(H_{ij})}}{\sum_i \exp^{(H_{ij})}} \tag{8}$$

indicating the degree of correlation between each part-of-speech.

After row averaging matrix E, we get a vector $\beta \in R^n$, which contains the attention weight corresponding to each part-of-speech.

$$\beta = \frac{1}{n} \sum_j E_{ij} \tag{9}$$

The final part-of-speech attention vector $\alpha \in R^n$ is obtained from matrix C and β by the dot product operation, where the matrix C is the output of TAN.

$$\alpha = C\beta^T \tag{10}$$

The introduction of POS-Attention strengthens the extraction of part-of-speech features, making the model fully consider the contribution of part-of-speech to the sentiment classification, hence is good for learning deeper sentimental features in the text.

D. CLASSIFICATION LAYER

After column averaging E_P, we get a global part-of-speech vector $\gamma \in R^{d_{pos}}$.

$$\gamma = \frac{1}{n} \sum_i E_{Pij} \tag{11}$$

And the final vector expression μ of the text is composed of part-of-speech attention vector α and γ.

$$\mu = [\alpha, \gamma] \tag{12}$$

Here, γ can be regarded as a global part-of-speech vector containing multiple part-of-speech information. The purpose is to reduce the influence of sentence components which are close in position or similar in content but have no modified relations on classification.

The vector expression μ of text is a feature vector that combines text semantic information and part-of-speech information. Then the representation vector is fed to a one-layer fully connected network by

$$\hat{y} = \sigma(W\mu + b) \tag{13}$$

At last, it is mapped by the Sigmoid activation function into a numerical probability between [0,1]. When this probability is greater than 0.5, the sentiment tendency of the sample is predicted to be positive; when the probability is less than 0.5, the prediction is negative.

E. FOCAL LOSS

Focal Loss [28] was first used to address the imbalance problems between easy examples and hard examples in one-stage object detection. The Cross Entropy (CE) loss is usually used in the binary classification. For a single sample,

$$L_{CE} = -y log\hat{y} - (1-y) log(1-\hat{y}) = \begin{cases} -log\hat{y}, & y = 1 \\ -log(1-\hat{y}), & y = 0 \end{cases} \tag{14}$$

where y is 0 or 1, representing the true category of the sample, and $\hat{y}$ is the predicted category of the model. For notational convenience, we define P:

$$P = \begin{cases} \hat{y}, & y = 1 \\ 1-\hat{y}, & y = 0 \end{cases} \tag{15}$$

and rewrite $L_{CE} = -log(P)$.

The Focal Loss is obtained by adding the modulation factor $(1-P)^\lambda$ to the Cross Entropy loss. Its purpose is to reduce the weight of easy samples and make the model pay more attention to the learning of hard samples.

$$L_{FL} = -(1-P)^\lambda log(P) \begin{cases} -(1-\hat{y})^\lambda log(\hat{y}), & y = 1 \\ -\hat{y}^\lambda log(1-\hat{y}), & y = 0 \end{cases} \tag{16}$$

Here, $\lambda \geq 0$ is a modulation parameter, which can smoothly adjust the ratio of weight reduction of some easily classified samples. When the sample is misclassified and the P is small, the modulation factor is close to 1, and the loss is not affected much. As P is close to 1, the modulation factor tends to be zero, and the loss of the easy samples is reduced in weight.

F. MODEL TRAINING

The sample imbalance problem existing in the sentiment classification is similar to that in the object detection. It is not difficult to find that most of the samples in a larger number of classes are easily classified. As the model can learn more about the characteristics of samples in this category, the features of the samples in a small number of classes will become less obvious such that these samples are difficult to classify. Therefore, in this work, we introduce Focal Loss into the sentiment classification, and uses it instead of Cross Entropy as the loss function in model training. The loss function of the model is:

$$loss = -\sum_i^n (y_i)(1-\hat{y}_i)^\lambda log(\hat{y}_i) + (1-y_i)\hat{y}_i^\lambda log(1-\hat{y}_i) \tag{17}$$

where n is the number of samples, $\lambda \geq 0$ is the modulation parameter. The sample label y_i is 0 or 1, and the corresponding sentimental polarity is negative and positive. In the experiment, the stochastic gradient descent algorithm is used to minimize the loss function.

IV. EXPERIMENTS

In this section, we present our experimental setup and evaluate the performance of our proposed pos-TAN model on various sentiment classification datasets.

TABLE 1. Dataset information.

Dataset	Positive		Negative	
	Train	Test	Train	Test
TSB	6300	700	2700	299
Waimai	3600	400	7188	799
Weibo	53994	5999	10789	1200
NLPCC2014	4500	500	1499	167
Yelp 2013	107450	11939	25726	2858
Amazon	160204	17800	160251	17806

A. DATESETS

The statistics of the used datasets are summarized in Table 1. Each dataset is split into train and test set in a ratio of 9:1. Each sample is labeled with two sentiment polarities: positive and negative.

- **TSB** is a Chinese hotel reviews dataset collected by Tan [46], which including 2999 negative reviews and 7000 positive reviews.
- **Waimai** dataset contains 11987 user comments collected from a takeout platform in China.
- **Weibo** dataset is a collection of comments from Weibo, China's largest social platform. There are 11998 negative comments and 59993 positive comments.
- **NLPCC2014** is a binary sentiment classification dataset from NLPCC 2014 Evaluation Task ''Sentiment Classification with Deep Learning Technology'' [47].
- **Yelp 2013 reviews** [48] are obtained from the Yelp Challenge Dataset, which removes the reviews labeled with 2-star, 3-star, and 4-star, only containing binary labels(1-star represents negative and 5-star represents positive).
- **Amazon** [29] contains 3,650,000 reviews. We randomly select 10% of the original dataset for testing. The reviews are labeled with 0 or 1, represent negative and positive respectively. 0 corresponds to 1-star and 2-star reviews, and 2 corresponds to 4-star and 5-star reviews. 3-star reviews with neutral sentiment were not included in the original.

B. EXPERIMENTAL SETUP

In our experiments, we randomly initialize the word embeddings, position embeddings and pos embeddings from $U(-\epsilon, \epsilon)$, where $\epsilon = 0.02$. The dimensions of word embeddings and position embeddings are 128. The dimension of pos embeddings is 8. Our experiments are conducted with a batch size of 50, and initial learning rate of 0.001. The number of iterations is 10 and the dropout rate is set to 0.1.

The evaluation metric used here is classification accuracy. Accuracy measures the overall sentiment classification performance, is formalized as:

$$Accuracy = \frac{T}{N} \tag{18}$$

where T is the number of samples correctly predicted and N is the total number of samples.

C. BASELINES

We compare our proposed pos-TAN with the following baselines:

- **SVM** has strong generalization ability and high classification accuracy, which is suitable for the binary classification problem.
- **LR** with L1-regularization not only avoids over-fitting, but also has the function of feature selection.
- **BernoulliNB** is a commonly used text classification algorithm with stable classification efficiency.
- **Bi-LSTM** is a bidirectional recurrent neural network composed of two LSTM in opposite directions. It stitches together the output of the two LSTMs at the last time step as a text feature representation and is used for sentiment classification.
- **TextCNN** [7] adopts convolutional neural network and uses different sizes of convolution kernels to extract key information from sentences, so as to better obtain local features of text.
- **HAN** [30] is proposed for document classification, which uses a hierarchical structure of ''word-sentence'' to represent a document. In addition, the model has two levels of attention mechanisms, which exist on the word-level and the sentence-level, respectively.
- **TCN** [49] is a sequence modeling benchmark, which combines causal convolution, residual connection, and dilation convolution. The training speed is faster than the recurrent neural network model. It is not only good at capturing temporal dependencies, but also can capture local information.

D. RESULTS AND ANALYSIS

The classification accuracy results of our model compared with other competitive models are shown in Table 2. We can see that TextCNN, Bi-LSTM and other deep learning methods are superior to traditional machine learning methods. Because deep learning methods are capable of effectively generating feature representations without handcrafted feature engineering. Among all the deep learning methods, TextCNN, Bi-LSTM performs worse. This is largely due to their relatively simple network structure and poor representation of text features. Bi-LSTM can capture long-distance dependencies, but cannot learn local features like CNN. HAN, TCN and our proposed pos-TAN all represent the features of text at different levels.

Compared with HAN, the classification accuracy of our proposed pos-TAN+ Focal Loss improves on most datasets. Especially, it has increased by about 1.3% on the TSB dataset and Weibo dataset. HAN adopts word-level and sentence-level hierarchical structure, and applies word attention and sentence attention respectively on these two levels. It can enable the network to extract important words and sentences from documents, so it performs well on document classification. However, when it is applied to the sentiment classification task, it is not as good as our model. Sentiment

TABLE 2. Sentiment classification accuracy of our proposed pos-TAN against baselines. Accuracy is the evaluation metric. Best results are in bold.

Model	TSB	Weibo	Waimai	NLPCC2014	Yelp 2013	Amazon
SVM	0.8947	0.9311	0.8574	0.7697	0.8972	0.9034
LR	0.8841	0.8654	0.8590	0.7444	0.8680	0.8547
BernoulliNB	0.8380	0.8401	0.8292	0.7466	0.8673	0.8379
Bi-LSTM	0.8976	0.9706	0.8860	0.7753	0.9398	0.9329
TextCNN	0.8990	0.9720	0.8874	0.7811	0.9523	0.9334
HAN	0.9047	0.9647	0.8884	**0.7991**	0.9678	0.9412
TCN	0.9122	0.9738	0.8791	0.7932	0.9514	0.9429
pos-TAN+Focal Loss	**0.9170**	**0.9771**	**0.8916**	0.7946	**0.9686**	**0.9454**

TABLE 3. An ablation test on our proposed model to evaluate the effects of focal Loss and part-of-speech. pos-TAN+Focal Loss incorporates part-of-speech information into transformer attention network (TAN) and replaces CE Loss with focal loss.

Model	TSB	Weibo	Waimai	NLPCC2014	Yelp 2013	Amazon
TAN+CE Loss	0.8991	0.9532	0.8790	0.7834	0.9532	0.9313
TAN+Focal Loss	0.9078	0.9593	0.8782	0.7916	0.9589	0.9365
pos-TAN+Focal Loss	**0.9170**	**0.9771**	**0.8916**	**0.7946**	**0.9686**	**0.9454**

classification tasks are usually sentence-level rather than document-level. Although HAN uses a hierarchical attention structure, when it is used in sentiment classification tasks, the hierarchical network degenerates into the word attention network, and the role of sentence attention is not obvious. TCN focuses on sequence modeling, which not only good at capturing temporal dependencies, but also can capture local information. This is the main reason why TCN exceeds TextCNN and Bi-LSTM. Compared with TCN, our pos-TAN+ Focal Loss achieves absolute increases with 1.25% and 1.72% on the Waimai dataset and Yelp 2013 dataset respectively. It also improved slightly on the TSB dataset and Weibo dataset by 0.48% and 0.33% respectively.

What accounts for such increases in performance is that we not only use the Self-Attention mechanism to learn the feature expression of the text but also incorporates part-of-speech information. Self-Attention can capture the dependencies between words in different positions. The POS-Attention can assign weights to different parts-of-speech, so that the network pays more attention to the learning of sentimental information contained in part-of-speech. In addition, Focal Loss also makes some contribution to the improvement on classification accuracy. Therefore, our proposed pos-TAN+ Focal Loss achieves the best performance among most datasets.

E. ABLATION EXPERIMENT

To evaluate the effects of Focal Loss and part-of-speech, we present an ablation test on the proposed model. Table 3 shows the results of TAN+CE Loss, TAN+Focal Loss, pos-TAN+ Focal Loss. From the results, we can see that TAN+Focal Loss performs a litter better than TAN+CE Loss but far below pos-TAN+ Focal Loss. The classification accuracy of pos-TAN+ Focal Loss is generally higher than TAN+ Focal Loss by about 1%.

Although the use of Focal Loss on the original TAN does not significantly improve the classification accuracy, it also alleviates the impact of imbalanced data to a certain extent. Part-of-speech contains sentimental information that is helpful for classification. More sentimental features can be obtained by effectively modeling part-of-speech information. After we incorporate part-of-speech information into different layers of the TAN, the classification accuracy is greatly improved. It validates the advantages of our proposed method on modeling part-of-speech information.

F. HYPERPARAMETER SENSITIVITY ANALYSIS

A new parameter λ is introduced in Focal Loss to smoothly adjust the ratio of weight reduction of the samples in a larger number of classes. Intuitively, the value of λ will affect the performance of the model. Therefore, we compare the classification accuracy of the model on various datasets under different λ values. The results are shown in Figure 4.

When $\lambda = 0$, Focal Loss is equivalent to CE Loss; when $0 < \lambda < 1.2$, the classification accuracy will change with λ and fluctuate greatly; When $\lambda > 1.2$, the classification accuracy of the model generally shows a downward trend, so the situation when $\lambda > 2$ is not shown in the Figure 4. Combined with Table 3, we can see that the performance of the model is improved after the introduction of Focal Loss, which is determined by the value of λ to some extent. Although the values of λ with the best classification results on different datasets are different, the performance of the model

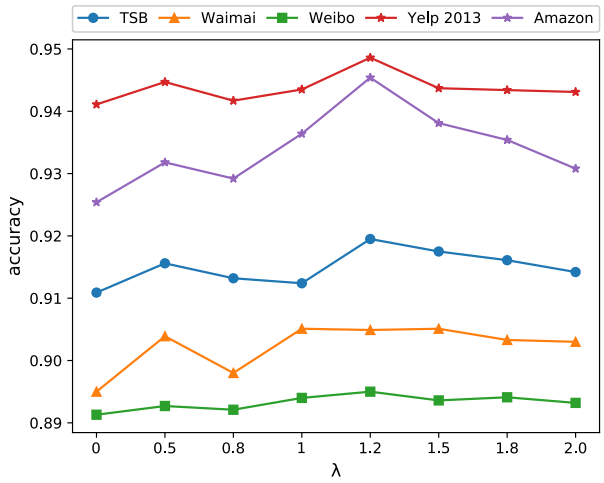

FIGURE 4. The classification accuracy of our proposed pos-TAN on various datasets under different λ values. Hyperparameter λ is introduced in focal loss to smoothly adjust the ratio of weight reduction of the samples in a larger number of classes.

is relatively stable when $\lambda = 1.2$, and the classification accuracy is better overall. Therefore, the default λ is set to 1.2.

G. VISUALIZATION OF POS-ATTENTION WEIGHT

In order to reflect the contribution of part-of-speech information to the sentiment classification, we select four samples from the datasets to visualize the weight of the POS-Attention of each word. The resluts are shown in Figure 5. We can see that adjectives such as "差","黑", "硬", "舒服", and "便利", have the highest weights. The weights of the adverbs such as "太", and the verbs such as "堵塞", "出行", "值得", "推荐"are ranked second. And the nouns such as "毛巾", "床垫", "下水道", "交通", "饮食" have the lowest weight. In linguistics, Adjectives are usually used to indicate the nature, state, and characteristics of people or things, while adverbs are used to modify adjectives, to indicate degree, range, etc. For example, "差" in Figure 5(a) is a negative comment, while "太"deepens the degree of "'差'and further expresses negative sentiment. Verbs often indicate actions or attitudes. As shown in Figure 5(b), verbs such as "值得"and "推荐" indicate a positive attitude.

According to the above analysis, we can know that the user's sentiment is often reflected by the corresponding part-of-speech, and different parts-of-speech represent different sentimental intensity. As can be seen in figure 5(c), 5(d), the introduction of part-of-speech information can make the model focus on the parts with sentiment orientation in the

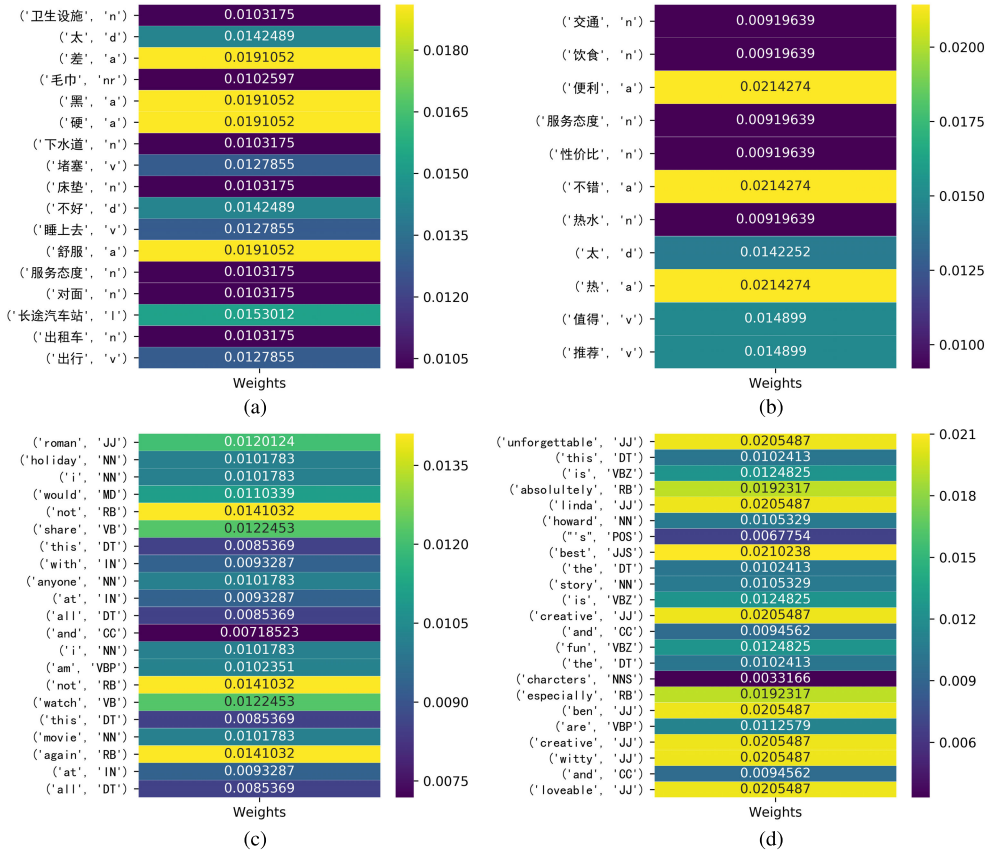

FIGURE 5. Visualization of POS-Attention weights derived from four samples in different datasets. (a),(b) TSB. (c) Yelp 2013. (d) Amazon. In each subfigure, the words and corresponding parts-of-speech are shown in the brackets on the left, and the right shows the POS-Attention weights of the parts-of-speech.

TABLE 4. The Chinese–English bilingual table of Figure 5(a),(b).

Chinese	English	Part-of-Speech	Chinese	English	Part-of-Speech
“卫生设施”	sanitary facilities	noun	“对面”	opposite site	noun
“太”	too	adverb	“长途汽车站”	bus station	noun
“差”	awful	adjective	“出租车”	taxi	noun
“毛巾”	towel	noun	“出行”	trip	verb
“黑”	black	adjective	“交通”	traffic	noun
“硬”	tough	adjective	“饮食”	diet	noun
“下水道”	sewer	noun	“便利”	convenient	adjective
“堵塞”	block	verb	“性价比”	price performance	noun
“床垫”	mattress	noun	“不错”	fine	adjective
“不好”	bad	adjective	“热水”	hot water	noun
“睡上去”	sleep	verb	“热”	boiling	adjective
“舒服”	comfortable	adjective	“值得”	deserve	verb
“服务态度”	service attitude	noun	“推荐”	recommend	verb

sentence. Such as ‘‘best’’, ‘‘creative’’, ‘‘fun’’ and other adjectives, these words will be given a higher weight, and ‘‘is’’, ‘‘the’’, ‘‘and’’, etc. as a part of grammar, it actually has little effect on sentiment classification. Through the attention interaction between part-of-speech and words, the model can focus on learning the sentimental information expressed by words such as ‘‘best’’, ‘‘creative’’, and ‘‘fun’’, while ignoring other words that have less impact on sentiment classification.

V. CONCLUSION

In this paper, we propose a sentiment classification model, pos-TAN. This model not only uses the Self-Attention mechanism to learn the feature expression of the text but also incorporates the POS-Attention, which uses to capture sentimental information contained in part-of-speech. In addition, we introduce the Focal Loss to alleviate the impact of sample imbalance on the classification effect. In the end, we have conducted extensive experiments on both Chinese and English datasets and observe from the experimental results that: (1) our model is capable of extracting sentimental information from parts-of-speech; (2) our model can combine the merits of POS-Attention and Focal Loss to improve the sentiment classification accuracy; (3) the visualization of POS-Attention weight shows different parts-of-speech have sentimental intensity in different degrees, so that the network can pay more attention to the learning of words with a high weight.

Although our model has great potentials in sentence-level sentiment classification, this work ignores the aspect information. There may be multiple aspects in a sentence, and sometimes it is necessary to identify the sentiment polarity of a specific aspect. In the future, we intend to extend the proposed part-of-speech based attention model to the aspect-level sentiment classification.

REFERENCES

[1] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*. New York, NY, USA: ACM, 2004, pp. 168–177.

[2] Y. Rao, J. Lei, L. Wenyin, Q. Li, and M. Chen, “Building emotional dictionary for sentiment analysis of online news,” *World Wide Web*, vol. 17, no. 4, pp. 723–742, Jul. 2014.

[3] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis,” *Comput. Linguistics*, vol. 37, no. 2, pp. 267–307, Jun. 2011.

[4] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” in *Proc. ACL Conf. Empirical Methods Natural Lang. Process.* Stroudsburg, PA, USA: Association for Computational Linguistics, vol. 10, 2002, pp. 79–86.

[5] T. Mullen and N. Collier, “Sentiment analysis using support vector machines with diverse information sources,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2004, pp. 412–418.

[6] S. Y. Chung and H. J. Yoon, “Affective classification using Bayesian classifier and supervised learning,” in *Proc. 12th Int. Conf. Control, Automat. Syst. (ICCAS)*, Oct. 2012, pp. 1768–1771.

[7] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.

[8] G. Cai and B. Xia, “Convolutional neural networks for multimedia sentiment analysis,” in *Natural Language Processing and Chinese Computing*. Nanchang, China: Springer, 2015, pp. 159–167.

[9] L. Dong, F. Wei, C. Tan, D. Tang, M. Zhou, and K. Xu, “Adaptive recursive neural network for target-dependent Twitter sentiment classification,” in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2014, pp. 49–54.

[10] T. H. Nguyen and K. Shirai, “PhraseRNN: Phrase recursive neural network for aspect-based sentiment analysis,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2509–2514.

[11] D. Tang, B. Qin, X. Feng, and T. Liu, “Effective LSTMs for target-dependent sentiment classification,” in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers (COLING)*, 2016, pp. 3298–3307.

[12] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” Sep. 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473

[13] T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.

[14] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu, “Attention-over-attention neural networks for reading comprehension,” in *Proc. 55th Annu. Meeting Assoc. for Comput. Linguistics*, vol. 1, 2017, pp. 593–602.

[15] B. Dhingra, H. Liu, Z. Yang, W. Cohen, and R. Salakhutdinov, “Gated-attention readers for text comprehension,” in *Proc. 55th Annu. Meeting Assoc. for Comput. Linguistics*, vol. 1, 2017, pp. 1832–1846.

[16] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, “DRAW: A recurrent neural network for image generation,” in *Proc. 32th Int. Conf. Mach. Learn. (ICML)*, vol. 37. Lille, France: JMLR.Org, 2015, pp. 1462–1471.

[17] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, “Recurrent models of visual attention,” in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Cambridge, MA, USA: MIT Press, 2014, pp. 2204–2212.

[18] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, vol. 37. Lille, France: JMLR.Org, 2015, pp. 2048–2057.
[19] J. Zeng, X. Ma, and K. Zhou, "Enhancing attention-based LSTM with position context for aspect-level sentiment classification," *IEEE Access*, vol. 7, pp. 20462–20471, 2019.
[20] J. Du, L. Gui, Y. He, R. Xu, and X. Wang, "Convolution-based neural attention with applications to sentiment classification," *IEEE Access*, vol. 7, pp. 27983–27992, 2019.
[21] Z. Li, Y. Wei, Y. Zhang, and Q. Yang, "Hierarchical attention transfer network for cross-domain sentiment classification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5852–5859.
[22] H.-F. Tang, S.-B. Tan, and X.-Q. Cheng, "Research on sentiment classification of Chinese reviews based on supervised machine learning techniques," *J. Chin. Inf. Process.*, vol. 21, no. 6, pp. 88–94, Jun. 2007.
[23] J. Tian, D. Zhu, and H. Long, "Chinese short text multi-classification based on word and part-of-speech tagging embedding," in *Proc. Int. Conf. Algorithms, Comput. Artif. Intell.* New York, NY, USA: ACM, 2018, p. 62.
[24] G. Wang, Z. Zhang, J. Sun, S. Yang, and C. A. Larson, "POS-RS: A random subspace method for sentiment classification based on part-of-speech analysis," *Inf. Process. Manage.*, vol. 51, no. 4, pp. 458–479, 2015.
[25] P. Kalarani and S. S. Brunda, "Sentiment analysis by pos and joint sentiment topic features using SVM and ANN," *Soft Comput.*, vol. 23, no. 16, pp. 7067–7079, 2019.
[26] C. Nicholls and F. Song, "Improving sentiment analysis with part-of-speech weighting," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 3, Jul. 2009, pp. 1592–1597.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.
[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
[29] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, vol. 1. Cambridge, MA, USA: MIT Press, 2015, pp. 649–657.
[30] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1480–1489.
[31] L. Xiang, X. Jin, L. Yi, and G. Ding, "Adaptive region embedding for text classification," May 2019, *arXiv:1906.01514*. [Online]. Available: https://arxiv.org/abs/1906.01514
[32] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 90–94.
[33] D. Bespalov, B. Bai, Y. Qi, and A. Shokoufandeh, "Sentiment classification based on supervised latent n-gram analysis," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.* New York, NY, USA: ACM, 2011, pp. 375–382.
[34] F. Xianghua, L. Guo, G. Yanyan, and W. Zhiqiang, "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and hownet lexicon," *Knowl.-Based Syst.*, vol. 37, pp. 186–195, Jan. 2013.
[35] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proc. 18th ACM Conf. Inf. Knowl. Manage.* New York, NY, USA: ACM, 2009, pp. 375–384.
[36] J.-H. Wang, T.-W. Liu, X. Luo, and L. Wang, "An LSTM approach to short text sentiment classification with word embeddings," in *Proc. 30th Conf. Comput. linguistics speech Process. (ROCLING)*, 2018, pp. 214–223.
[37] G. Cai and B. Xia, "Convolutional neural networks for multimedia sentiment analysis," in *Proc. 4th CCF Conf. Natural Lang. Process. Chin. Comput. (NLPCC)*, vol. 9362. Berlin, Germany: Springer-Verlag, 2015, pp. 159–167.
[38] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 2267–2273.
[39] X. Wang, W. Jiang, and Z. Luo, "Combination of convolutional and recurrent neural network for sentiment analysis of short texts," in *Proc. 26th Int. Conf. Comput. Linguistics, Tech. Papers (COLING)*, 2016, pp. 2428–2437.
[40] Z. Wu, X.-Y. Dai, C. Yin, S. Huang, and J. Chen, "Improving review representations with user attention and product attention for sentiment classification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5989–5996.
[41] S. Gu, L. Zhang, Y. Hou, and Y. Song, "A position-aware bidirectional attention network for aspect-level sentiment analysis," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 774–784.
[42] A. Rozental and D. Fleischer, "Amobee at SemEval-2018 task 1: GRU neural network with a CNN attention mechanism for sentiment classification," Apr. 2018, *arXiv:1804.04380*. [Online]. Available: https://arxiv.org/abs/1804.04380
[43] Y. Wang, M. Huang, and L. Zhao, "Attention-based lstm for aspect-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 606–615.
[44] M. Hu, S. Zhao, L. Zhang, K. Cai, Z. Su, R. Cheng, and X. Shen, "CAN: Constrained attention networks for multi-aspect sentiment analysis," Dec. 2018, *arXiv:1812.10735*. [Online]. Available: https://arxiv.org/abs/1812.10735
[45] H. Hongye, Z. Jin, and Z. Zuping, "Text sentiment analysis combined with part of speech features and convolutional neural network," *Comput. Eng.*, vol. 44, no. 11, pp. 215–220, 2018.
[46] S. Tan and J. Zhang, "An empirical study of sentiment analysis for chinese documents," *Expert Syst. Appl.*, vol. 34, no. 4, pp. 2622–2629, May 2008.
[47] Y. Cao, R. Xu, and T. Chen, "Combining convolutional neural network and support vector machine for sentiment classification," in *Proc. Chin. Nat. Conf. Social Media Process.* Singapore: Springer, 2015, pp. 144–155.
[48] D. Tang, B. Qin, and T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 1014–1023. [Online]. Available: https://www.aclweb.org/anthology/P15-1098
[49] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," Mar. 2018, *arXiv:1803.01271*. [Online]. Available: https://arxiv.org/abs/1803.01271

KEFEI CHENG is currently a doctor. He is also a Professor with the College of Cyber Security and Information Law, Chongqing University of Posts and Telecommunications, China. His current research interests include cloud computing, network security, and natural language processing.

YANAN YUE is currently pursuing the M.Eng. degree with the College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. Her current research interests include sentiment analysis and deep learning.

ZHIWEN SONG is currently pursuing the M.Eng. degree with the College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. Her current research interests include text classification and machine learning.