Dual-Channel CNN-BiGRU Sentiment Analysis Method Based on Part of Speech

Xin XU

School of Computer, Electronics and Information Guangxi University, Nanning, Guangxi 530004, China

xx0881@foxmail.com

Keywords: Part of speech, Cnn, Bigru, Dual channel

Abstract: For less research on lexical to emotional tendency in sentiment analysis, traditional CNN model sentiment classification method will have classification inaccuracy problem, this paper proposes a dual-channel CNN-BiGRU sentiment analysis method research method based on part of speech, can use rules to extract the corresponding lexical text combined with the original text to construct a dual-channel input to the neural network, the CNN can extract local semantic features, the Bi-GRU is used to extract global features containing context, which are fused with local features to complement them, and finally the fused features are input to the classifier for sentiment tendency determination. The experimental results on Chinese dataset show that the proposed-to-model approach in this paper has improved in accuracy, recall, and F-value compared to traditional neural networks.

1. Introduction

With the rapid development of Internet and information technology, people are more and more inclined to express their opinions on online shopping, news media, reading and watching movies, etc. These emotionally inclined evaluation information may seem insignificant, but they actually contain rich emotional information. Rapidly and accurately analyzing and extracting sentiment expressions from massive comment data is of great reference and research value for government opinion monitoring, enterprise market research and personal consumption choices. Sentiment Analysis mainly refers to the use of natural language processing and computer linguistics to identify and extract subjective information from the original material and find out the polarized opinions and attitudes of opinion makers on certain topics.

Text sentiment analysis is mainly a coarse-grained sentiment analysis that classifies the positive and negative sentiment polarity of a known passage in a text, and three mainstream research approaches have been generated: based sentiment lexicon, machine learning, and deep learning approaches. Based on sentiment lexicon sentiment analysis method, Hu et al [1] used adjectives as the main indicator of sentiment polarity judgment by constructing a WordNet-based sentiment lexicon, first selecting sentiment words with known sentiment polarity as seed words, then expanding the sentiment lexicon, and finally calculating the sentiment tendency. Liu et al [2] proposed a method to improve Word2Vec's sentiment lexicon construction in the photography domain by constructing a sentiment lexicon of reviews in the photography domain. Based on the machine learning approach, Pang et al [3] proposed an experimental study on text sentiment classification task using machine learning methods and using three classical classifiers, SVM, Parsimonious Bayesian and Maximum Entropy, and concluded that machine learning for text sentiment classification is better. Paltoglou et al [4] first calculated the feature weights of all words in the text, and then input the weight information of the words into the model to complete the analysis of text sentiment tendency.

In the process of conducting large-scale data analysis, deep learning[5] is found to be significantly better than traditional machine learning. Bengio et al [6] earlier proposed neural network language models for sentiment analysis. Kim et al [7] used CNN to model short texts and verified experimentally that CNN have some advantages over sentiment analysis. Although CNNs

  DOI: 10.25236/iiiece.2022.029

are outstanding in the field of sentiment analysis, they focus more on local features and completely ignore the influence of lexical contextual environment on word meaning. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) introduce the gate mechanism based on the traditional RNN, which better overcomes the problems of RNN such as not being able to rely on information for a long time , will have gradient explosion and gradient dispersion during the training process [8].Tai et al [9] used LSTM model to effectively solve the temporal connection problem in sentiment analysis and achieved good classification results. Cho et al [10] proposed GRU, which is a little simpler and spends less time cost compared to LSTM, and can effectively improve the model training time.Rozental et al [11] proposed a joint model of BiGRU and CNN to extract text feature information, and achieved good results in the experimental results.Bai et al [12] proposed a hybrid neural network structure using BiLSTM-CNN for the classification of two feature fusions.

Summarizing and analyzing the content of existing to research results, the following problems exist in the field of sentiment analysis.

1)Too little research on lexicality in textual sentiment analysis and relatively little influence of partial lexicality on the study of affective tendencies.

2) In deep learning models for sentiment analysis, CNN is able to learn spatial features in the data, but have the problem of not being able to learn temporal features of the data. LSTM is able to learn the temporal features in the network traffic data, but it cannot learn the spatial features in the data. Using CNN or GRU alone can only extract some of the features in the network traffic data, which has some limitations.

In view of the above many shortcomings, among many deep learning models, this paper fully considers the superiority of CNN in extracting local features of text, and secondly considers the good performance of GRU evolved from LSTM based on the unique memory and selection characteristics possessed by processing sequence problems, proposes dual-channel CNN-BiGRU sentiment analysis method achieve automatic feature learning and sentiment tendency determination. In Chinese words, the four word types that can express emotional tendencies are mainly nouns, verbs, adjectives and adverbs, which are chosen to build a data set with relatively few noise points.

2. Dual-Channel CNN-BiGRU Based Sentiment Analysis Model

2.1 Overall Model

The overall model structure proposed in this paper is shown in Figure 1.

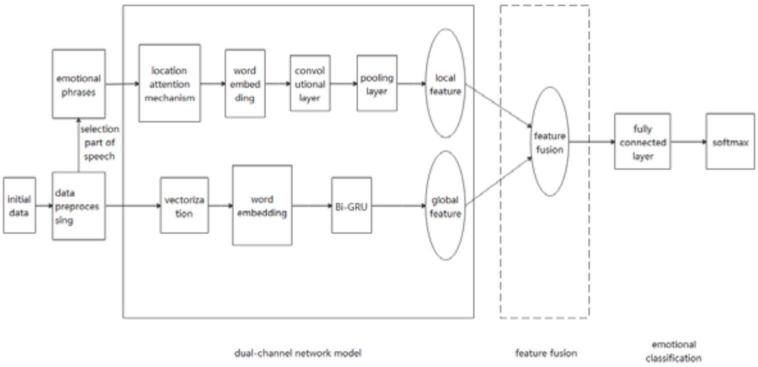

Fig.1 CNN-BiGRU Dual-Channel Neural Network Model Incorporating Part of Speech Attention Mechanism

Channel 1 takes the pre-processed Chinese text through lexical selection and positional attention mechanism trained into word vector form through embedding layer as the input layer of CNN layer to extract local features; Channel 2 vectorizes the pre-processed Chinese text as the input layer of Bi-GRU to extract global features. Then it goes through the feature fusion layer and finally the classification is performed by Softmax.

2.2 Part of Speech Selection

In Chinese commentary, not all words can express emotional tendencies, and among Chinese words, the four types of words that can express emotional tendencies are mainly nouns, verbs, adjectives, and adverbs [16]. Therefore, this paper mainly uses these four types of words with emotional tendency as text features for sentiment analysis. In this way, not all the parts of speech are taken into account in the calculation of text sentiment tendency analysis, but only some parts of speech that have a large impact on sentiment tendency are preserved.

Table 1 Parts of Speech of Jieba Participle

part of speech	express	part of speech	express	part of speech	express
noun	n	adverbial	d	stative word	z
time word	t	preposition	p	pronoun	r
word in which	s	conjunction	c	numeral	m
noun of locality	f	auxiliary word	u	measure word	q
verb	v	exclamation	e	prefix	h
adjectives	a	modal particle	y	suffix	k
differentiating word	b	onomatopoeia	o	punctuation	w

The raw unprocessed text is represented as$w = \{x_1, x_2, \cdots x_n\}$, where n is the number of all comments and x_i is the i-th comment. By the lexical selection method proposed in this paper, the text containing only noun, verb, adjective, and adverb lexis is obtained$W = \{x_1, x_2, \cdots x_n\}$.

2.3 Attention Mechanisms

Considering the different relative positions of words in the text, the relative positions can indicate to some extent the importance of words for evaluating attributes. Attention is calculated in the sentiment analysis model using lexical and relative position information, and the attention weights corresponding to each word are calculated.

(1) Calculation of the relative position of words

The relative position of the word w with subscript i in the comment is given by$_i$

$$l(w_i) = \text{abs}(i - ai)$$

Where: $l(w_i)$ denotes the relative position distance of word w_i, i denotes the subscript value of wi of the word, ai denotes the subscript value of the evaluation central word in the sentence, and abs() denotes the operation of taking the absolute value. After obtaining the relative position distance of the word $l(w_i)$, it is converted into an n-dimensional feature vector using binary encoding to represent.

(2) Lexical information vector of words

The noun, verb, adjective and adverbial natures are represented by vectors using binary coding. The vectors corresponding to different lexical properties are shown in the following table.

Table 2 the Vectors Corresponding to Different Part of Speech

part of speech	corresponding vector	part of speech	corresponding vector
noun	[0,0]	adjectives	[1,0]
verbs	[0,1]	adverbial	[1,1]

(3) Attention computation based on part of speech and location information

The computed relative position vectors and word vectors are stitched together as part of speech and positional information vectors of words. The attention mechanism model receives the set of part of speech and positional information vectors P:[p1,p2,...,pi] and calculates to generate the attention weight value α. The formula for calculating the attention weight is as follows.

$$e = \tanh(Wp + b)$$

$$\alpha = \text{softmax}(e) = \frac{exp(e)}{\sum exp(e)}$$

Where $e \in R^{di \times T}$,di denotes the dimension size of the lexical and positional feature vectors, T denotes the number of words in the text, tanh denotes the sine activation function, and$w \in R^{d_i \times d_n}$.$b \in R^{d_n \times T}$, generates the intermediate variables$e \in R^{d_n \times T}$.$u \in R^{d_n \times 1}$,the generated attention weights $\alpha \in R^{1 \times T}$.

3. Dual Channel Inputs

3.1 Convolutional Neural Networks

Convolutional Neural Networks(CNN) are characterized by local perception characteristics, each neuron can only perceive locally, and in the local connection, the parameters of each neuron are the same, and the convolutional operation is performed by actually bringing up one local information. Therefore, extracting a small number of lexical utterances, using CNN can effectively extract local information features.

In CNN model training, the text needs to be fed into it in the form of a matrix, which is processed by the above$W = \{x_1, x_2, \cdots x_n\}$, in which$x_n \in R$, the corresponding text word vector sentences are obtained by Word2Vec technique

$$X = x_1 + x_2 \cdots + x_n$$

where x_i corresponds to w_i the denoted word vector, + is the splicing operation.

The convolution layer performs the convolution operation by connecting the input layer with three convolution kernels of the same window size, and extracts the static local features between h adjacent words, denoted by Ci , as shown in equation (1)

$$C_i = f(w \cdot x_{i:i+h-1} + b) \qquad (1)$$

where: b denotes the offset and f denotes the nonlinear function.

Applying this convolution kernel to every neighboring word vector window of size h in the text data $\{x_{1:h}, x_{2:h+1}, \ldots, x_{n-h+1:n}\}$, a feature map is generated as shown in equation (2).

$$C = [c_1, c_2, \cdots c_{n-h+1}] \qquad (2)$$

The pooling layer is responsible for secondary filtering of the features obtained from the convolutional layer to improve the overall network training efficiency while extracting important features. The convolutional kernel slides over the input data to extract a set of local feature information, and the role of the pooling layer is to filter out the most important features among them. In this paper, we use the maximum pooling operation, when the height of the convolutional kernel h, the feature values filtered by the pooling layer on the feature vector c obtained after the convolutional kernel operation $c_{cnn} = \max(c_1, c_2, \cdots, c_{n-h+1})$.

3.2 Bi-GRU Channel

Since the part of speech processed text may omit parts of it, resulting in the loss of contextual textual information, the Bi-GRU model is used to obtain global features containing contextual information.

Gated recurrent neural network (GRU) is an improvement of long short-term memory network (LSTM), GRU has a memory state unit that is constant from beginning to end, in this memory unit, the input gate and forgetting gate in the original LSTM are replaced by update gate, which makes GRU simpler than LSTM in network structure, and GRU needs less tensor operations and requires less, so GRU is simpler to train and faster to train than LSTM. The state calculation formulas for each GRU unit are presented as formulas (3)–(6).

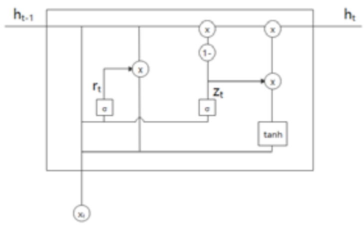

Fig.2 GRU Structure

$$r_t = \sigma(W_r \cdot [h_{t-1}, v_t]) \qquad (3)$$
$$z_t = \sigma(W_z \cdot [h_{t-1}, v_t]) \qquad (4)$$
$$\tilde{h}_t = tanh(W \cdot [r_t \cdot h_{t-1}, v_t]) \qquad (5)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \qquad (6)$$

Where r_t, z_t represent the reset gate, the update gate; W_r, W_z, and W are the weight matrices; h_{t-1} is the hide layer sequence information at time t-1; v_t is the unprocessed text word vector; σ and tanh are the activation function; $*$ denote matrix multiplication; [] denote vector concatenation.

But GRU only learns the information in front of the current word, and does not learn the information behind the current word, for a word, the semantic understanding of this word is closely related to the words before and after it. Therefore, in this paper, a bidirectional gated recurrent neural network Bi-GRU is selected to build the model, which fully considers the contextual information of the current word.

$$c_{bgru} = [\overrightarrow{h}_t, \overleftarrow{h}_t] \qquad (7)$$

c_{bgru} is the concatenation of the forward and backward hidden layer output of feature information; $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ denote the forward GRU and the backward GRU feature information.

3.3 A Dual-Channel Chinese Sentiment Analysis Model with Part of Speech

Although the part of speech processed text enables the convolutional neural network to obtain local features with more explicit sentiment tendency c_{cnn}, there is also the problem of ignoring contextual information, and the global features extracted using Bi-GRU are considered as a remedy for the local feature ignoring problem, so this paper c_{bgru} are fused. The fusion formula is shown as formula.

$$\mathrm{c} = c_{cnn} \oplus c_{bgru}$$

The fusion is to splice c_{cnn} and c_{bgru} together and use them together as the input of the fully connected layer, and introduce the dropout mechanism, which can effectively avoid the model from relying on some features, finally input them into the softmax classifier.

4. Experiment and Analysis

4.1 Experimental Data

In this paper, a corpus of Chinese comments is constructed by using a web crawler to select a total of 16735 comments on Douban, and 803 texts expressing positive sentiment and 8703 texts expressing negative sentiment are respectively labeled and distinguished by manual means. In the training process, 80% of the data are randomly selected as the training set and the remaining 20% as the test set.

4.2 Model Hyperparameter Settings

In the training process of deep learning model, parameter tuning is crucial for the final classification effect. In order to select the best network training parameters, the final selected model hyperparameter configuration is shown in Table 3 and 4.

Table 3 Word2vec Word Vector Training Parameters

Parametric indicators	value	Parameter Description
sg	0	CBOW
sample	1e-5	Sampling Threshold
window	5	Context window size
min_count	5	Baseline word frequency, Words smaller than min_cout are discarded
size	150	Feature vector dimension

Table 4 Model Hyperparameter Configuration

Parameter Name	Parameter values	Parameter Description
loss	binary_crossentropy	Cross-entropy loss function
optimizer	Adam.	optimizer
epochs	50	Number of iterations of network training
batch_size	32	Number of samples selected for each iteration of training

lr	0.0001	training speed
activation	sigmoid	fully connected layer classifier
dropout	0.3	Preventing model overfitting

4.3 Evaluation Indicators

In this paper, the correlation evaluation of the model is completed by using three parameters, accuracy (P), recall (R) and F1 value. The calculation formula is.

$$P = \frac{TP}{TP+FP} \qquad (8)$$

$$R = \frac{TP}{TP+FN} \qquad (9)$$

$$F1 = \frac{2TP}{2TP+FP+FN} \qquad (10)$$

where data points with TP as a positive example are marked as positive, data points with FP as a negative example are marked as positive, data points with TN as a negative example are marked as negative, and data points with FN as a positive example are marked as negative. the F1 value is the summed mean of accuracy and recall.

4.4 Experimental Results and Analysis

To verify the effectiveness of the method proposed in this paper, it is compared with the traditional SVM, single-channel CNN, single-channel BiGRU, and CNN-LSTM, which is currently a more effective combined deep learning model.

Table 5 Comparison Experimental Results of Different Methods

experimental model	P	R	F1
SVM	0.8	0.770073	0.790262
Single-channel CNN	0.81	0.76	0.784204
Single-channel BiGRU	0.83	0.823	0.826485
CNN-LSTM	0.861	0.827	0.843658
Dual-channel CNN-BiGRU	0.916	0.893	0.9043

From the results in Table 5, it can be seen that the dual-channel sentiment analysis method using fused part of speech in this paper can effectively improve the classification accuracy, recall rate and F1 value of the model. Comparing with the machine learning algorithm SVM, the machine learning method is based on feature engineering and uses the word or syntactic structure as the classification special, which is difficult to obtain the deep semantic information of the sentence and the recognition degree is low. Single-channel CNN, single-channel BiGRU models obtain semantic information of sentences from one side, contextual information is missing in CNN, and BiGRU is missing local information.

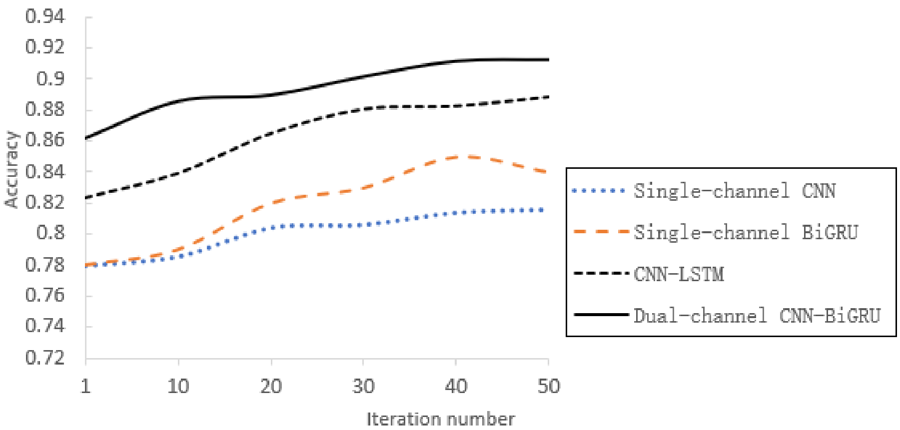

Fig.3 Effect of the First 50 Iterations of the Experiment

Figure 3 shows the effect of the first 50 iterations of each neural network in the experiment. It can be seen that: the dual-channel CNN-BiGRU outperforms the other neural networks in each round of iteration, and the CNN-LSTM outperforms the single-channel CNN, and the single-channel BiGRU. because the dual-channel CNN-BiGRU has more input information, more information can be extracted, and the classification is better as the number of iterations increases.

5. Conclusion

Sentiment analysis research is of great significance in this era of information explosion. In this paper, the problems of traditional CNN sentiment classification model are proposed to fuse lexical features into CNN, and the obtained features are more sentiment inclined, use BiGRU to solve the problem of ignoring context information and finally fusion is performed by the extracted features. In the experiments of sentiment analysis based on Chinese comments, there are corresponding improvements in accuracy, recall, and F1, verifying that the model in this paper has good sentiment classification effect.

References

[1] Hu M, Liu B. Mining and summarizing customer reviews[C]// Tenth Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. acm, 2004.

[2] Liu Yaqiao, Lu Xiangyan, Deng Kaikai, Ruan Kaidong, Liu Jun. A method for constructing sentiment dictionaries for reviews in photography [J]. Computer Engineering and Design, 2019,40(10):3037-3042.

[3] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[4] Paltoglou G, Thelwall M. A study of information retrieval weighting schemes for sentiment analysis[C]// Acl, Meeting of the Association for Computational Linguistics, July, Uppsala, Sweden. dblp, 2010.

[5] Dos Santos C, Gatti M. Deep convolutional neural networks for sentiment analysis of short texts [C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014: 69-78.

[6] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model [J]. Journal of machine learning research, 2003, 3(Feb): 1137-1155.

[7] Kim Y. Convolutional neural networks for sentence classification [C]//Proceedings of Conference on EMNLP,2014:1746 -1751.

[8] Hochreiter S, Bengio Y, Frasconi P, et al. Gradient flow in recurrent nets: the difficulty of learning long term dependencies [M]. Wiley - IEEE Press, 2001: 237 -243.

[9] Tai, Kai Sheng, Richard Socher, and Christopher D. Manning. “Improved semantic representations from tree-structured long short-term memory networks.” arXiv preprint arXiv:1503.00075 (2015).

[10] Cho K , Van Merrienboer B , Gulcehre C , et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation [J]. Computer ence, 2014.

[11] Rozental A, Fleischer D. Amobee at SemEval-2018 Task 1: GRU Neural Network with a CNN Attention Mechanism for Sentiment Classification[C]// Proceedings of The 12th International Workshop on Semantic Evaluation. 2018.

[12] Jing B, Fei L, Donghong J. ATTENTION BASED BILSTM-CNN CHINESE MICROBLOGGING POSITION DETECTION MODEL [J]. Computer Applications and Software, 2018 (3): 52.