BiGRU-Multi-Head Self-Attention Network for Chinese sentiment classification

Shitao Wang\textsuperscript{1,a}, Jiangfeng Li\textsuperscript{1,b,}\ast, Defeng Hu\textsuperscript{1,c}

\textsuperscript{1}School of Software Engineering, Tongji University, Shanghai, China
\textsuperscript{a}1831578@tongji.edu.cn, \textsuperscript{b}lijf@tongji.edu.cn, \textsuperscript{c}1831596@tongji.edu.cn

Abstract—The rapid development of Internet and the extensive user base have led to a significant increase in the propagation speed of information and public opinion. Quick and accurate sentiment analysis of public opinion can help analyze and manage public opinion. Since BiGRU can simplify network parameters, and the Multi-Head Self-Attention (MHSA) mechanism can learn long-distance features more accurately and quickly. In this study, we propose a BiGRU-MHSA network to perform Chinese sentiment classification. The innovation of the model lies in the combination of BiGRU and MHSA in order to focus on several semantic centers and comprehend the whole text more precisely and efficiently. At the same time, we use the Chinese comment datasets on Weibo to verify the effectiveness and accuracy of the model. From the experimental results, we find that our proposed model improves the accuracy by about 1% on different comments datasets without consuming too much computing resources.

1. INTRODUCTION
With the development of network technology and the popularization of mobile devices, the comments made by netizens on mobile social apps (such as Weibo and WeChat) often represent their opinions on a certain event, thus forming online public opinion. The accelerating mobile Internet era has greatly accelerated the propagation of both information and public opinion. Internet public opinion is playing an increasingly important role in recent years. Therefore, there is a need to analyze public opinion quickly and accurately, and sentiment analysis is part of it [1]. At present, there are already some fee-paying opinion analysis platforms about Weibo.

Sentiment analysis mainly refers to the analysis of subjective text and judgment of its emotional tendency. In terms of analysis granularity, sentiment analysis falls into two categories. One is coarse-grained sentiment analysis, which only classifies text into two (three) basic sentiment polarities: positive, negative, (or neutral). The other is fine-grained sentiment analysis, which can be subdivided into various emotions, corresponding to the two-classification and multi-classification problems. Sentiment analysis has three methods: based on linguistic rules (sentiment dictionary), based on traditional machine learning, and based on deep learning. With the rapid development of software and hardware, deep learning methods are currently the mainstream research direction.

Due to the relatively simple architecture of the BiGRU (Bidirectional Gated Recurrent Unit), it has low requirements on computer hardware. But the ability to screen features with extremely long-distance sequences is limited. At the same time, the Multi-Head Self-Attention (MHSA) mechanism can easily obtain long-distance high-dimensional features. Although the architecture is complex, it does not consume much computing resources for a small amount of use. Therefore, this paper adopts the
architecture of BiGRU-MHSA. For the sentiment classification task of Chinese text, the traditional BiGRU network is used to extract short-distance features, and then the MHSA mechanism is used to extract features on a global scale to learn the cross-sequence complex semantic features of the entire text. Taking the advantage of BiGRU and MHSA, the extracted features are then classified into emotions. We propose this model to improve the classification accuracy of the complex multi-classification task of emotion with limited computing resources.

2. RELATED WORK

2.1. Traditional Method
Since sentiment analysis is essentially the mining of opinions and can be classified as a kind of data mining. Text sentiment analysis is to dig out the opinions or sentiments behind the text, and the traditional method is to use sentiment dictionary. The sentiment dictionary is very effective in fine-grained sentiment analysis of Weibo text. However, since the construction of the sentiment dictionary requires a large amount of labeled sentiment word corpus and grammatical structure prepared in advance, and it is impossible to analyze new words and sentence patterns, and there are some limitations in comprehensive application. Zhu et al. [2] proposed a semantic orientation computation of Chinese words based on HowNet. Reference [3] constructed a word sense disambiguation (WSD) sentiment classifier based on WordNet and SentiWordNet to improve the accuracy of sentiment analysis.

Pang et al. [4] used three traditional machine learning methods to classify the sentiment of articles, namely Naive Bayes, Maximum Entropy Classification, and Support Vector Machine. Liu et al. [5] studied the scalability of naive Bayes classifiers on large datasets and designed a big data analysis system. Dey et al. [6] compared the effects of K-NN and Naive Bayes on sentiment classification tasks. Machine learning methods are relatively mature, but the disadvantage is that users need to have considerable domain knowledge.

2.2. Word Embedding
In order to solve the dimensional explosion problem of traditional text vector representation, Bengio et al. [7] proposed a distributed representation (word vector) for the first time and used a neural network to build a language probabilistic model. In the language probability model, the similarity of word vectors can represent the similarity of semantics. To solve the huge amount of calculation caused by the neural network probabilistic language model, Mnih [8] improved the feature-based Log-Bilinear model. The word2vec model proposed by Mikolov et al. [9-10] based on previous research and the Glove model proposed by Pennington [11] have become the most widely used word vector model due to their excellent effects.

2.3. Deep Learning Method
In 1997, Hochreiter et al. [12] proposed and implemented LSTM (long and short-term memory network), which improved the problem of gradient disappearance and gradient explosion caused by long-term information transmission by adding a gating mechanism. Cho et al. [13] proposed GRU (Gated Recurrent Unit), which integrates two gated switches into one, which simplifies the network architecture and shortens the training time.

Kim et al. [14] used word vectors convoluted by CNN (Convolutional Neural Networks) to obtain local semantic features and then implemented text classification using pooling and fully connected layers. Jin et al. [15] proposed a CNN-LSTM model for multi-dimensional sentiment analysis (valence-arousal). Zhang et al. [16] proposed a neural network based on the combination of CNN and GRU to test hate language on Twitter. The method of deep learning has greatly promoted the development trend of the NLP (Natural Language Processing) because of the simple entry threshold, but for some difficult tasks, basic neural network models such as CNN and RNN will appear at a loss.
2.4 Attention Mechanism
Bahdanau et al. [17] first proposed the attention mechanism and applied it to the field of machine translation. Mnih et al. [18] used the attention mechanism for image classification for the first time and achieved obvious results. Luong et al. [19] proposed a local attention mechanism. The self-attention mechanism has led the development trend in the field of Computer Visual and NLP because of its reasonable interpretability and good results. Google [20] proposed a transformer architecture and a BERT (Bidirectional Encoder Representations from Transformers) model based on a Multi-Head Attention mechanism in 2017. Perez [21] proved that the multi-layer self-attention architecture with position coding is turing complete under some theoretical assumptions (such as infinite precision operations). Cordonnier [22] proposed that a MHSA layer with enough heads can reset the parameters to achieve the effect of any convolutional layer and capture the same pattern of grid pixels.

However, the excessive depth of the standard BERT model leads to excessive parameters. Top-of-the-line GPU equipment, massive datasets, and a large amount of training time for training are required, daunting for small research institutions and enthusiasts in the NLP field. Therefore, some simplified BERT models have appeared in recent years.

3. METHOD
In this paper, we propose a model named BiGRU-MHSA Network to make the emotion classification. The process is mainly composed of data preprocessing, model constructing, and training.

3.1 Data Preprocessing
Due to the arbitrariness of comments from the Internet, the first step after obtaining the text data is to clean up the data and remove the messy code and expressions. For the emotion classification, punctuations other than exclamation marks are of little value. Removing redundant symbols can effectively shorten the text sequence and simplify the training.

Stop Words refer to certain words that can be filtered without affecting the classification results in order to save memory and improve efficiency in information retrieval, which can also be used in some coarse-grained tasks, such as classification. Stop Words requires a manual definition for a particular task, and we should take care so as not to miss key lexeme.

Chinese word segmentation is the most basic step in Chinese NLP. Due to the characteristics of Chinese, there is no separator between words, and the raw text needs to be divided into a sequence of words before encoding the text. At present, Chinese word segmentation algorithms can be divided into three categories: dictionary-based methods, statistical methods, such as N-Gram and hidden Markov model, and rule-based methods. Many mature word segmentation tools have been widely applied worldwide, such as Spacy in English and Jieba in Chinese. As an open-source toolkit, Jieba realizes efficient word graph scanning based on a prefix dictionary to generate Directed Acyclic Gram (DAG) of all possible segmentation of
Chinese words in sentences. After the word segmentation, we can segment a sentence into a sequence of words.

For the uniform input size of the neural network, the word list should be aligned. The method adopted is padding the short text with '0', trimming the long text, and filtering out the samples that are too long. Finally, a mutual-mapping dictionary of index-word is established to correspond words with indexes one by one, so that the word list is simplified into a list of word indexes as the input of the sentiment classification model.

3.2. BiGRU-MHSA Network
BiGRU-MHSA Network is a kind of encoder-decoder architecture, in which BiGRU is an encoder, MHSA is used as a decoder, and a softmax layer for multi-classification or a sigmoid layer for two-classification is connected terminally. The architecture of the BiGRU-MHSA is shown in Fig. 1.

3.2.1. Word embedding layer
Word embedding is the first step of the sentiment classification network. It is necessary to map the discrete word to a continuous word vector that contains certain semantic information with appropriate dimensions. Word2Vec includes skip-gram and continuous bag-of-words (CBOW). The skip-gram model assumes that background words are generated based on the central word, while the CBOW model just opposites the assumption. The essence of the word embedding layer is a vector transformation function, which converts a certain input "word A" into a vector x with a fixed dimension. Take advantage of the semantic information automatically detected during training, word embedding can be generally utilized for other tasks. The process of word embedding can be expressed as (1)

$$ w_i = W(x_i) $$ (1)

where $x_i$ is the input word, $w_i$ is the word vector with fixed dimension and $W(\ )$ is the transformation function.

3.2.2. BiGRU layer
We firstly use BiGRU to extract short-term feature on the sequence of word vectors. GRU is a slight variation of LSTM, which reduces the three memory gates to two. The specific calculation process of GRU is illustrated by (2).
where $x_t$ is the input vector and $h_t$ is the hidden state vector at time $t$.

BiGRU is composed of hidden neuron sequences in two directions respectively. This architecture enables each output neuron to incorporate the semantics of both the previous and the following text. The architecture of the BiGRU layer is shown in Fig. 2 which can also be expressed by (3).

\[
\begin{align*}
\tilde{h}_t &= GRU(h_{t-1}, x_t) \\
\tilde{h}_t &= GRU(h_{t+1}, x_t) \\
h_t &= \tilde{h}_t \oplus \tilde{h}_t 
\end{align*}
\]

(3)

### 3.2.3. Self-attention mechanism

The attention mechanism comes from computer visual processing, which can obtain the focus of attention by browsing all the information. Recently, it has also been fully used in natural language processing to extract the main part of the text sequence corresponding to a target. The most common method to calculate attention score is dot multiplication as (4).

\[
\begin{align*}
e_i &= h_i \cdot s^T \\
\alpha_i &= \text{softmax}(e_i) \\
\alpha &= \sum_{i=1}^{N} \alpha_i h_i
\end{align*}
\]

(4)

where $e_i$ is attention score of word $i$, $\alpha_i$ is the attention weight and $\alpha$ is the weighted attention.

Attention mechanism is essentially a lookup table of a set of key-value pairs. Find the correlation coefficient of all Keys with a Query. The correlation coefficient can be treated as weight and multiplied by the value corresponding to each key to calculate a weighted sum. A method of matrix multiplication is adopted to simplify the formula and parallelize calculation.

To further abstract the equation, the concept of key-value pairs can be used to replace the state and features in the hidden layer in (5). The query feature (Q), the query feature (K), and the expression feature (V) of a vector can be different.

\[
\begin{align*}
e_i &= q_i \cdot k^T \\
\alpha &= \sum_{i=1}^{N} \alpha_i v_i
\end{align*}
\]

(5)

Scaled dot-product attention is the key component of Transformer which is initially proposed by Google. The main change is the scaling factor of $1/(d_k)^{0.5}$ in order to decrease the parametric and stabilize the gradient as shown in (6).
Here X is the input matrix and \( W_{\text{qry}}, W_{\text{key}}, W_{\text{val}} \) are transformation matrices.

As we can see in (6), it doesn’t distinguish the order of the input sequence. That is to say, in theory, any rearrangement of the input sequence can train out the corresponding network to produce the identical and correct output (the weight of the network only needs to change the order according to the change of the input). To solve this phenomenon, position encoding is superimposed on the input vector. In this way, the ordering information of input has been embedded in the input vector, so that the corresponding relationship between input and output becomes unique. The position encoding process is shown in (7).

\[
A = \text{softmax}(\frac{XW_{\text{qry}}W_{\text{key}}W_{\text{val}}(X+P)^T}{\sqrt{d_k}})XW_{\text{val}}
\]

The MHSA mechanism further improves the self-attention layer. The performance of the attention layer can be improved in two ways. On the one hand, it expands the model’s ability to focus on different locations; On the other hand, it provides multiple “representation subspaces” for the attention layer. With MHSA, we have multiple sets of query/key/value weight matrices. Each of these sets is initialized randomly, and after training, each set is used to project the input embedding (or vector from lower encoder/decoder) into a different representation subspace. Eventually, the MHSA is organized by concatenating multiple heads as shown in (8).

\[
\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_h)\mathcal{W}^o
\]

\[
\text{head}_i = \text{Attention}(Q_i, K_i, V_i)
\]

3.2.4. Softmax (Sigmoid) classification layer

Finally, a fully connected layer is used to concatenate the features of the first and last position selected from the output of the decoder and connect to \( n \) neurons corresponding to \( n \) kinds of results. When two classes, the sigmoid function is used to directly represent the classification probability; when more than 3 classes, the softmax function is used to normalize the value to the classification probability.

4. EXPERIMENT AND ANALYSIS

4.1. Experiment Datasets

We adopt the method of supervised learning which requires labeled sentiment analysis datasets. At present, there are not so many Chinese Weibo sentiment analysis datasets. We managed to download 3 datasets from the Internet, namely simplifyweibo_4_moods, NLPCC2014 Task1, and weibo_senti_100k. The details are shown in Table 1.

| Datasets          | Classes | Samples | Balance |
|-------------------|---------|---------|---------|
| simplifyweibo_4_moods |         |         |         |
| NLPCC2014 Task1    |         |         |         |
| weibo_senti_100k  |         |         |         |
The text of datasets has not been preprocessed which remains such noises as English, symbols, and user names. The experiment uniformly uses 80% of the data as the training set, 10% as the validation set, and the remaining 10% as the test set.

4.2. Experiment Settings
To simplify the training process, we omitted the training of Word2Vec in this experiment. The 300-dimensional pre-trained Word2Vec published on the Internet by Li et al. [23] is uniformly used in this experiment. The Word2Vec is pre-trained from the Weibo text using skip-gram and down-sampling, and the parameters of the word embedding layer are frozen during training. In addition, we chose a Chinese stop vocabulary containing only 1208 words.

The hyper-parameters of the network are selected through multiple experiments to obtain the optimal values considering both training speed and accuracy. The largest text length is set to 100, the learning rate is 0.0005, the batch size is 64, the hidden size of BiGRU is 100, dropout rate is 0.25, the layer of MHSA encoder is 1, the number of multi-head is 4, the dimension of feed-forward network is 256, the optimizer algorithm to backpropagate gradient is Adam, and the loss function is Cross-Entropy.

4.3. Results and Analysis
The models compared in this experiment are BiLSTM, BiGRU, MHSA-BiGRU, and BiGRU-MHSA. The hyperparameters of each model are consistent.

From the experiment results shown in Table 2, the model combining MHSA and BiGRU can generally achieve better results on each dataset. Comparing the two models of BiLSTM and BiGRU, it can be seen that the difference in effect was not significant. But considering that BiGRU has fewer model parameters and training time, BiGRU is a better choice. Comparing the two methods of MHSA-BiGRU and BiGRU-MHSA, it can be found that the effect of BiGRU-MHSA is better. Maybe MHSA, as a non-serialized model, can more accurately capture the information of the entire sequence during decoding. In addition, we also find that for relatively easy binary classification tasks, BiGRU-MHSA offers little improvement, but for datasets with more categories and unbalanced samples, the advantage of MHSA is more obvious.

| Datasets | BiLSTM | BiGRU | BiGRU-MHSA | MHSA-BiGRU |
|----------|--------|-------|------------|------------|
| simplifyweibo_4_moods | 59.69 | 59.76 | 61.4 | 60.2 |
| NLPCC2014 Taskl | 66.1 | 66 | 66.8 | 66 |
| weibo_senti_100 | 91.1 | 91.22 | 91.32 | 91.18 |

5. Conclusion
Sentiment analysis of Chinese text has always been the basis of online public opinion analysis. In this regard, this paper proposes a model that combines the ability of BiGRU to learn short-distance sequence features with the ability of MHSA to extract long-distance sequence features. Through comparative experiments on multiple Chinese Weibo datasets, the effectiveness of the BiGRU-MHSA model is verified, and the accuracy rate is increased by about 1% compared with the baseline model.

\[^1\] http://files.cnblogs.com/files/pinard/stop_words.zip
Due to the high noise of Weibo comments, the classification accuracy of the model is still not ideal for datasets that are not well preprocessed.

In the future, further research can be carried out in many aspects, such as the hierarchical MHSA architecture of text combined with the idea of hierarchical attention.

REFERENCES

[1] R. Piryani, D. Madhavi, V.K. Singh. “Analytical mapping of opinion mining and sentiment analysis research during 2000–2015,” in Information Processing & Management, vol.53, no.1, 2017, pp.122-150.

[2] Y.L. Zhu, J. Min, Y.Zhou, X.J. Huang, W.U. Li-De, “Semantic orientation computing based on HowNet,” in Journal of Chinese information processing, vol.20, no.1, 2006, pp.14-20.

[3] R. Jose, VS. Chooralil, “Prediction of election result by enhanced sentiment analysis on Twitter data using Word Sense Disambiguation,” International Conference on Control Communication & Computing, India, pp.638-641, November 2015.

[4] B. Pang, L. Lee, S. Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques,” Empirical Methods in Natural Language Processing, USA, pp.79-86, July 2002.

[5] B. Liu, E. Blasch, Y. Chen, S. Dan, G. Chen, “Scalable sentiment classification for big data analysis using naïve bayes classifier,” IEEE international conference on big data, USA, pp.99-104, October 2013.

[6] L. Dey, S. Chakraborty, A. Biswas, B. Bose, S. Tiwari, “Sentiment analysis of review datasets using naïve bayes and k-nn classifier,” unpublished.

[7] Bengio, Yoshua, Ducharme, Régine, Vincent, Pascal, Janvin, Christian, “A neural probabilistic language model,” in Journal of machine learning research, vol.3 2003, pp.1137-1155.

[8] A. Mnih, G. E. Hinton, “A scalable hierarchical distributed language model,” in Advances in neural information processing systems, vol.21, 2008, pp.1081-1088.

[9] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient estimation of word representations in vector space,” in Computer Science, 2013.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in neural information processing systems, vol.26, 2013, pp.3111-3119.

[11] J. Pennington, R. Socher, C. Manning, “Glove: Global Vectors for Word Representation,” Empirical Methods in Natural Language Processing, Qatar, pp.1532-1543, October 2014.

[12] S. Hochreiter, J. Schmidhuber. “Long short-term memory,” in Neural computation, vol.9, no.8, 1997, pp.1735-1780.

[13] K. Cho, B. Van Merriënboer, C. Gulcehre, F. Bougares, H. Schwenk, Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” unpublished.

[14] Y. Kim, “Convolutional neural networks for sentence classification,” unpublished.

[15] W. Jin, L.C. Yu, K. R. Lai, X. Zhang, “Dimensional sentiment analysis using a regional CNN-LSTM model,” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Germany, vol. 2, pp.225-230, August 2016.

[16] Z.Q. Zhang, D. Robinson, J. Tepper, “Detecting hate speech on twitter using a convolution-gru based deep neural network,” European semantic web conference, Greece, pp.745-760, June 2018.

[17] D. Bahdanau, K. Cho, Y. Bengio. “Neural machine translation by jointly learning to align and translate,” unpublished.

[18] V. Mnih, N. Heess, A. Graves, K. Kavukcuoglu, “Recurrent models of visual attention,” in Advances in neural information processing systems, vol.27, 2014, pp.2204-2212.

[19] M.T. Luong, H. Pham, C.D. Manning, “Effective Approaches to Attention-based Neural Machine Translation,” in Computer Science, 2015.
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, vol.30, 2017, pp.5998-6008.

[21] J. Pérez, J. Marinković, P. Barceló, “On the Turing Completeness of Modern Neural Network Architectures,” International Conference on Learning Representations. Canada, April 2018.

[22] J. B. Cordonnier, A. Loukas, M. Jaggi, “On the Relationship between Self-Attention and Convolutional Layers,” International Conference on Learning Representations. United States, May 20.

[23] S. Li, Z. Zhao, R.F. Hu, W.S. Li, T. Liu, X.Y. Du, Analogical Reasoning on Chinese Morphological and Semantic Relations, ACL 2018, pp. 209-221.