

Hierarchical Hybrid Neural Networks With Multi-Head Attention for Document Classification


Weihao Huang, School of Electronics and Information Engineering, School of Physics and Telecommunication Engineering, South China Normal University, China

Jiaojiao Chen, School of Electronics and Information Engineering, South China Normal University, China

Qianhua Cai, School of Electronics and Information Engineering, South China Normal University, China*

Xuejie Liu, School of Electronics and Information Engineering, South China Normal University, China

Yudong Zhang, School of Informatics, University of Leicester, UK

 <https://orcid.org/0000-0002-4870-1493>

Xiaohui Hu, School of Electronics and Information Engineering, South China Normal University, China

ABSTRACT

Document classification is a research topic aiming to predict the overall text sentiment polarity with the advent of deep neural networks. Various deep learning algorithms have been employed in the current studies to improve classification performance. To this end, this paper proposes a hierarchical hybrid neural network with multi-head attention (HHNN-MHA) model on the task of document classification. The proposed model contains two layers to deal with the word-sentence level and sentence-document level classification respectively. In the first layer, CNN is integrated into Bi-GRU and a multi-head attention mechanism is employed in order to exploit local and global features. Then, both Bi-GRU and attention mechanism are applied to document processing and classification in the second layer. Experiments on four datasets demonstrate the effectiveness of the proposed method. Compared to the state-of-the-art methods, the model achieves competitive results in document classification in terms of experimental performance.

KEYWORDS

BiGRU, CNN, Document Level, Hierarchical Characteristics, Hybrid Attention Network, Multi-Head Self-Attention Mechanism, NLP, Text Classification

INTRODUCTION

There has been research interest towards the improvements in textual information processing in the past decade (Ali et al., 2017). Accompanying the evolution of computer technology, the volume of text data available online has had strong growth in recent years. As an important branch in the field of natural language processing (NLP), document classification aims to determine the sentiment

DOI: 10.4018/IJDWM.303673

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

polarity of a document, which plays a pivot role in a variety of tasks, including spam email filtering (Liu & Wang, 2010), topic extraction (Sarioglu, 2014), sentiment analysis (Hu et al., 2015), social public opinion mining (Guan et al., 2009) and more. In most cases, the discussed documents are ranked with different scores or stars representing the corresponding sentiment, while a higher score generally indicates more positive sentiment. With an accurate comprehension of the sentiment results and a deep understanding of the given document, the performance of document classification can be improved accordingly.

Advances in deep learning algorithms give rise to new opportunities to promote the efficacy of NLP tasks significantly. State-of-the-art document classification approaches are typically dominated by two distinguishing neural networks: the convolutional neural network (CNN), and the recurrent neural network (RNN). Recent publications report the superiority of the RNN in dealing with sequential inputs of various lengths. That is, the RNN models are capable of not only modeling the long-term dependencies (Habimana et al., 2020), but can also capture the semantics within contextual information (Du et al., 2019). More specifically, the two most well-known RNNs, namely long short-term memory (LSTM) and gated recurrent unit (GRU), are employed as a key module for tackling such issues in miscellaneous NLP methods (Li et al., 2019). On the other hand, the CNN is more effective in extracting the sentiment-related features from word sequences in comparison with the RNN (Du et al., 2019). The main reason is that the CNN can make full use of the textual data to collect the feature vectors with minimum parameters. In such a manner, the local importance from salient parts is thus captured (Zhao et al., 2021).

In order to improve document classification accuracy, the absence of sentiment information within long-distance texts needs to be considered comprehensively. To address this issue, the attention mechanism is employed, which supplements and enhances the long-distance sentiment information delivering in either the CNN or the RNN models. To be more specific, the attention mechanisms identify the significance in exploiting the hidden states and computing the class distributions, based on how to determine the attentive weights of different words (Liang et al., 2017). In this way, the models that integrate the attention mechanism into the RNN/CNN are able to model the textual data regardless of the distances. While restricted to the architecture of deep-learning algorithms, the attention mechanism still shows its distinctiveness by precisely capturing the sentiment of a specific part from the document. In fact, an attention model proposed by Google, namely Transformer, is implemented based solely on the attention mechanisms, without using the neural network structure (Vaswani et al., 2017). This work allows for multi-attention layers running in parallel, and outperforms all previously reported ensembles, which sets the foundation of the multi-head attention network.

Among existing document classification methods, two major concerns remain challenging. One concern is that most models, in spite of incorporating the hierarchical structure, fail to resolve the distinctions between sentences and the document. Another concern is that the relationship between words, which also makes a contribution to the document classification, attracts less attention. In light of the above discussion, an ideal document classification model should precisely identify the sentiment by exploiting all sources of textual information. Inspired by Vaswani et al. (2017) and Yang et al. (2016), the objective of this work is to propose a hierarchical hybrid neural network with a multi-head attention (HHNN-MHA) model for document classification. Aiming to identify the significances of different words and different sentences, two distinguishing attention mechanisms are employed on each level for modeling (Huang et al., 2021). Furthermore, a convolution module is integrated into Bi-GRU via a learnable gating mechanism, while a gated linear unit (GLU) is taken to obtain both local and global features. In line with the multi-head attention network, the relations of different components are determined based merely on the input document. The contributions of this paper are threefold, which are summarized as follows:

1. According to the hierarchical structure within the document, a hybrid attention network is dedicatedly designed to deal with the hierarchical characteristics.

2. An improved Bi-GRU is carried out with the integration of the CNN module, which aims to capture the local relationship among words.
3. To improve the model generalization and compute the similarity between words directly, the multi-head self-attention mechanism is employed for semantic encoding. In this way, more informative semantic features can be extracted and the classification accuracy can be improved.

RELATED WORK

Document Classification Methods

Previous works on document classification highlighted the remarkable achievements of the CNN- and the RNN-based methods. In the domain of NLP, the employment of the CNN greatly facilitates the processing of multiple tasks (Collobert & Weston, 2008). For the purpose of document classification, Kim et al. (2014) devised a method, namely TextCNN, based on the reversed convolutional layer, which brings about a simpler structure and less computation. Likewise, Kalchbrenner et al. (2014) proposed a dynamic convolutional neural network (DCNN) to effectively deal with the interaction among different words. Johnson and Zhang (2017) established a deep pyramid convolutional neural network by studying and deepening word-level CNNs. Moreover, hierarchical neural networks attract a great deal of interest due to their ability to tackle different levels of textual information. Zheng et al. (2019) proposed a hierarchical neural network derived from TextCNN, which is known as TextHCNN. In TextHCNN, the document representation is generated via the feature map convolution from sentence level, while the sentence representation is generated by using the CNN within the word level.

By contrast, the utilization of the RNN specializes in the modeling of long-term dependency sequences (Habimana et al., 2020). Concretely, the RNN is able to work on input data of different lengths, based on where to capture the context information and identify long-distance sentiment (Du et al., 2019). Compared to the classical RNN, the widely-applied RNN models, LSTM and GRU, are originally proposed to enhance long-term memory delivery. One can observe that Wang et al. (2015) employed LSTM to deal with the data from social media for sentiment analysis. Besides, there is an ongoing trend to integrate The RNN and The CNN for mutual supplementary and enhancement. For example, Wang et al. (2016) presented a jointed CNN and RNN architecture, which exploits both local features and long-distance dependencies for sentiment analysis. In addition, Guggilla et al. (2016) proposed LSTM- and CNN-based deep neural networks and obtained impressive outcomes in argumentative claim classification.

Multi-Head Attention Mechanism

The attention mechanism is typically an integral part to characterize the component dependencies with attentive weights in deep-learning methods (Vaswani et al., 2017). To the best of the authors' knowledge, the integration of the attention mechanism into the RNN/CNN is currently the most common approach to identify the significance in sequential data processing. Gao et al. (2018) proposed a two-layer attention network based on classic CNN, with the goal of capturing the importance of different words and sentences. Similarly, Liu et al. (2020) fused the attention mechanism into the ELMo (Embedding from Language Models) neural network to obtain comprehensive semantic information.

Since the attention mechanism has the potential to be enormously beneficial to draw the relation, the Transformer model, which relies entirely on an attention mechanism by eschewing recurrence or convolution, is thereby built up (Vaswani et al., 2017). With such a simple network architecture, the transformer is trained significantly faster and exceeds the performance of state-of-the-art methods. Notably, the multi-head attention presented in Transformer is further applied to other NLP tasks. The working principle of the multi-head attention mechanism is presented in Figure 1 and Figure 2, and is described as follows.

Figure 1. Scaled-dot-product attention

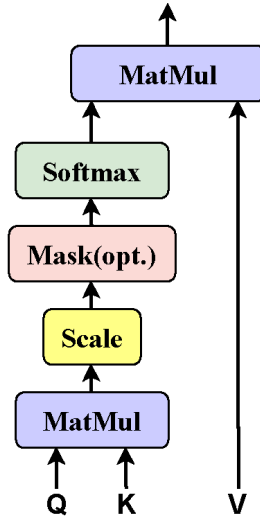
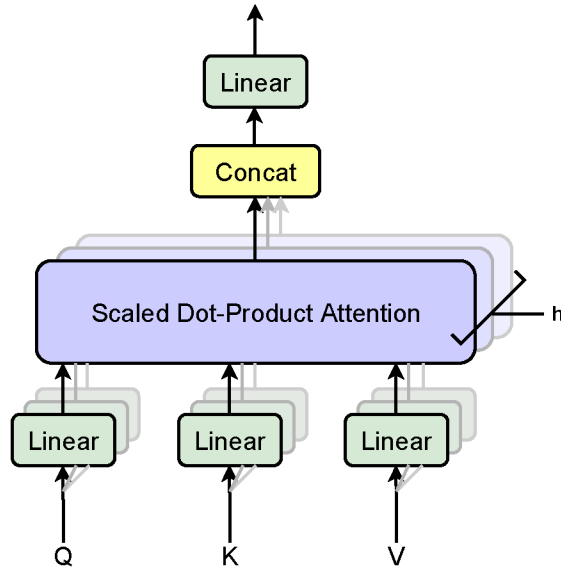


Figure 2. The multi-head attention mechanism



Scaled-dot-product attention is proposed by Vaswani et al. (2017), and performs well on machine translation. It is defined as Equation 1:

$$Attention(Q, K, V) = softmax \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (1)$$

where Q denotes query, K denotes key, and V denotes value, $\sqrt{d_k}$ is scale factor to avoid the large results.

The multi-head attention mechanism (MHA) (Vaswani et al., 2017) comprises zoom dot multiplying attention combined with a parameter matrix, which can process information in different representation subspaces from different locations in parallel. First, the queries, keys, and values are mapped through the parameter matrix, and then the parallel operation of the attention function is performed. Since the values of q and k change constantly, the parameters of different heads are detached. As such, these results are concatenated as the final outcome value (Xiao et al., 2020). The specific calculation is defined as follows:

$$\begin{cases} MultiHead(Q, K, V) = Concat(head_1, \dots, head_h) \\ where : head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \end{cases} \quad (2)$$

where W^Q , W^K , and W^V are the appropriate parameters learned by the authors' model training process.

THE APPROACH

In this section, the general framework of the model based on a hierarchical hybrid neural network with a multi-head attention mechanism is described. As shown in Figure 3, the proposed model consists of a word encoder, a word-level attention layer, a sentence encoder, and a sentence-level attention layer, in which a multi-head self-attention mechanism and a convolutional module are used at word-level attention.

Word Encoder

Suppose that there are L sentences in a document, the authors need to define the i^{th} sentence contains T_i words and w_{it} with $t \in [0, T]$ representing the words in the i^{th} sentence. First, the authors vectorized the words by embedding the matrix W_e , $x_{it} = W_e w_{it}$. Then, using a bidirectional GRU to connect the information from both directions for words, a Bi-GRU unit related to the sequence contextual information was obtained. This Bi-GRU contained the forward GRU \vec{f} that reads sentence s_i from w_{i1} to w_{iT} and a backward GRU \overleftarrow{f} that reads sentence s_i from w_{iT} to w_{i1} :

$$\begin{cases} \vec{h}_{it} = \overrightarrow{GRU}(x_{it}), t \in [1, T] \\ \overleftarrow{h}_{it} = \overleftarrow{GRU}(x_{it}), t \in [T, 1] \\ h_{it} = [\vec{h}_{it}, \overleftarrow{h}_{it}] \end{cases} \quad (3)$$

where h_{it} is the output of Bi-GRU, containing both \vec{h}_{it} and \overleftarrow{h}_{it} , with context information included.

The CNN algorithms originate from biological vision principles, and local features of the input can be extracted by different tools (e.g., multi-networks, convolution, and down sampling). In addition, without losing context information, using the output of Bi-GRU as the input of the CNN can maintain context and local relevance.

The model of this paper further enhances the sequential context representation with a convolutional module, shown in Figure 4 (Cai et al., 2019). Given an input hidden state sequence H , three convolutional operations are utilized to obtain three output vectors $D_{k=1}$, $D_{k=3}$, and $D_{k=5}$, where K is the convolutional kernel size.

The authors perform the optimization operation by concatenating the three outputs to extract different n-gram features with better learning ability:

Figure 3. General framework of the model based on the hierarchical hybrid neural

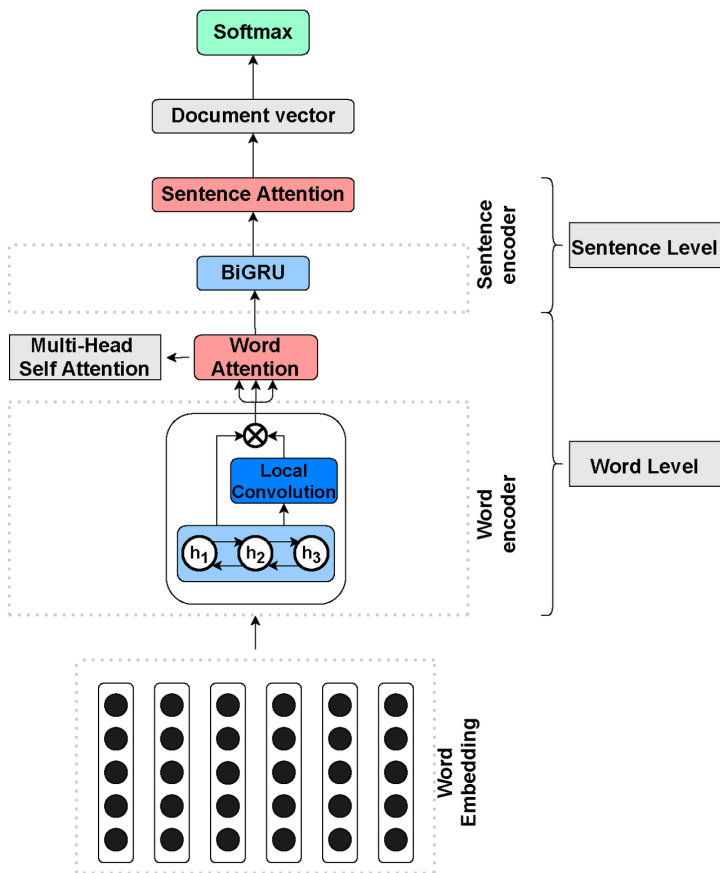
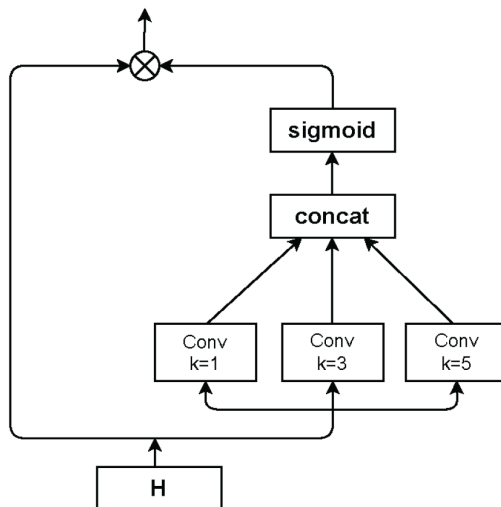


Figure 4. Schematic diagram of the local convolution module with gated linear unit



$$D = [D_{k=1}, D_{k=3}, D_{k=5}] \quad (4)$$

A learnable threshold mechanism is set, instead of using the result of the convolution operation as the convolution module's output, which is able to better filter the sequential context based on local importance. Additionally, the gated linear unit (GLU) uses a function to control the information selection of the features. The authors introduce a similar architecture (Figure 4) to select how much sequential context information should be retained, as follows:

$$R = \sigma(W_d D + b_d) \odot (W_h H + b_h) \quad (5)$$

where H is a series of hidden states $H = \{h_1, \dots, h_l\}$ of the source text mapped by a bidirectional GRU. Specifically:

$$H = BiGRU(x_{it}) \in R^{l \times d_{h_t}} \quad (6)$$

where x_{it} is the embedding representation of the document and d_{h_t} is the output dimension.

Word Attention

The attention mechanism was proposed in the use of encoder-decoder structures for neural machine translation (NMT) (Bahdanau et al., 2014). The attention mechanism is currently very common in deep learning models, but it is not limited to the encoder-decoder hierarchy. It is worth mentioning that the attention mechanism can be applied only on the encoder to solve tasks such as text classification or representation learning.

The multi-head self-attention mechanism (MHSA) is a special case of MHA, where Q , K , and V of the self-attentive layer all come from the output of the previous encoder layer, i.e., the input $Q = K = V$. To be specific, the outcome from the CNN-BiGRU is taken as a fixed value and sent to the self-attention network. By computing the similarity between Q and K , the attention coefficient is obtained via normalization. Consequently, sentence representation is computed by using the weighted summation of the attention weight and the input vector. The authors use the multi-head self-attention mechanism for semantic encoding, because the core of self-attentive is to augment the semantic representation of the target word with other words in the text, so that the information of the context can be better utilized and the semantics of the sentence can be preserved. The specific calculation is shown in Equation 7:

$$MHSA = MultiHead(X, X, X) \quad (7)$$

In this paper, words are semantically encoded by MHSA, $s_i = \{s_1, s_2, \dots, s_l\} \in R^{d_s \times l}$, and d_s denotes the dimension of MHSA. The formula is then calculated as follows:

$$s_i = MultiHead(R, R, R) \quad (8)$$

Sentence Encoder

Given the sentence vector s_i , the document vector, in a similar way, is obtained to the word encoding. The sentences using bidirectional GRU are encoded as follows:

$$\begin{cases} \vec{h}_i = \overrightarrow{GRU}(s_i), i \in [1, L] \\ \overleftarrow{h}_i = \overleftarrow{GRU}(s_i), i \in [L, 1] \\ h_i = [\vec{h}_i, \overleftarrow{h}_i] \end{cases} \quad (9)$$

where h_i is the output of Bi-GRU for sentence encoding, by concatenating the sentence forward state and backward state to obtain information about sentence i .

Sentence Attention

Considering that different sentences in a document contribute differently to the document, and that the importance level is not fixed but determined by the contextual environment, an attention mechanism is introduced at the sentence level with a sentence-level context vector u_s to learn this importance:

$$u_i = \tanh(W_s h_i + b_s) \quad (10)$$

$$\alpha_i = \frac{\exp(u_i^T u_s)}{\sum_i \exp(u_i^T u_s)} \quad (11)$$

$$d = \sum_i \alpha_i h_i \quad (12)$$

where d is the document vector that summarizes all the information of the sentences in the document. u_s is the sentence-level context vector that is randomly initialized during training and then learned to be updated.

Finally, the authors transfer the document representation to the softmax classifier and obtain the probability distribution of sentiment polarity as follow:

$$y = \text{softmax}(Wd + b) \quad (13)$$

The model in this paper uses the sum of categorical cross-entropy as the loss function, defined in Equation 14:

$$L = -\sum_i \sum_j y_i^j \log \hat{y}_i^j + \lambda \|\theta\|^2 \quad (14)$$

where i is the subscript of the i -th sample and j is the subscript of the j -th sentiment category; y is the true distribution of sentence sentiment polarity and \hat{y} is the predicted distribution of sentence sentiment polarity. $\lambda \|\theta\|^2$ is the L2 regular term and θ is the parameter set of the model. The Adam algorithm (Kingma & Ba, 2014) is used to train the proposed model.

EXPERIMENT

Datasets and Experiment Parameters

The authors' model (HHNN-MHA) is evaluated for document-level sentiment classification on publicly available datasets, including Yelp 2013, Yelp 2014, Yelp 2015, and Amazon Food reviews. Details of each dataset are presented in Table 1. The authors split the documents into sentences and labeled each sentence, and used 80% of the data for training. Detailed statistical information of these datasets is shown in Table 1. *Yelp reviews* are from the Yelp dataset challenge (Tang et al., 2015) in 2013, 2014, and 2015. The ratings are divided into five levels from 1 to 5 (higher is better). These review ratings with manual tags are considered as the gold standard sentiment tags. Thus, it is not necessary to manually annotate the sentiment tags of the documents. *Amazon Food* reviews (McAuley & Leskovec, 2013) were obtained from reviews of food on Amazon.com. Similar to Yelp reviews, the ratings are range from 1 to 5.

In this paper, the word2vec pre-trained models from Glove (Pennington et al., 2014) are used to initialize the documents in the experiments, whose word vector dimension is chosen as $d = 50$. All words that are not in the word vector dictionary are initialized as zero vectors, while the biases are all set to zero. Meanwhile, the model in this paper is implemented using the deep learning framework Pytorch 0.4, and Adam is used as the optimizer of the model. The corresponding learning rate is set to 0.001; the batch size is set to 64 (Yelp 2013, Yelp 2014) or 256 (Yelp 2015, Amazon Food reviews); the dropout is set to 0.5. Testing accuracy is used as the evaluation criterion in this experiment.

Baseline Methods

In this section, the authors present the comparative models for document-level sentiment classification to evaluate the working performance of the proposed model in this paper, as described below:

- **BiLSTM (Hochreiter & Schmidhuber, 1997):** The LSTM is a variant of the RNN. In order to overcome the directionality of the RNN, this method uses the bidirectional Long Short-Term Memory framework BiLSTM to control the selection of information by introducing a memory unit.
- **BiGRU (Cho et al., 2014):** This is another RNN variant with a simpler structure. This method encodes and decodes documents using a bidirectional gated recurrent unit framework.
- **BiGRU-Attention:** An attention model is added to BiGRU to assign different weights to different words in the document, and then a vector representation of the document is obtained by weighting the word vectors with these different weights.
- **TextCNN (Kim, 2014):** This is a standard convolutional neural network for sentiment classification, proposed by Kim et al.
- **TextHCNN (Zheng et al., 2019):** TextHCNN is a hierarchical neural network derived from TextCNN. Separate convolutional operations are performed.

Table 1. Statistics of the dataset: aw/d and mw/d denote the average and maximum number of words in each document, respectively, and as/d and ms/d denote the average and maximum number of sentences in each document

Datasets	Classes	Data Size	aw/d	mw/d	as/d	ms/d
Yelp 2013	5	63181	178	1476	8.9	151
Yelp 2014	5	63743	182	1599	9.2	155
Yelp 2015	5	1569264	151.9	1199	8.97	151
Amazon Food	5	239400	42.028	5668	2.959	222

- **HAN (Yang et al., 2016):** This is a hierarchical attention network with a word encoding layer and a sentence encoding layer that uses attention mechanisms to obtain sentence and document representations at the word level and the sentence level, respectively.
- **Bi-layers multi-head attention network:** This is a modification of the HHNN-MHA model, which uses a multi-headed attention mechanism at the word level and sentence level.

The experimental results of HHNN-MHA and other comparative models are shown in Table 2. These models are split into two parts. The first part is a non-layered network, which mainly includes some classic neural networks. The second part lists the hierarchical neural networks, i.e., TextHCNN, HAN, two-layer multi-headed attention networks, and the model of this paper, HHNN-MHA. The authors re-implemented the other comparative models on their dataset for the document-level sentiment classification task.

RESULT AND ANALYSIS

From Table 2, it can be seen that BiGRU-Attention is much more accurate than BiGRU (from 3.43% to 5.05%) on the Yelp2013 and Yelp2014 datasets due to the addition of the attention mechanism. The two models are essentially equal in Yelp2015 and performed slightly weaker in Amazon Food, probably because the attention mechanism could obtain more practical information and improve accuracy when dealing with a broader range of categories. TextCNN does not perform as well as LSTM and GRU in all datasets, probably because the RNN has more advantages than the CNN structure when processing long-distance text. On the other hand, GRU outperforms LSTM on the Yelp2015 dataset and the Amazon-Food dataset, probably because these two datasets have more words compared to other datasets (see Table 1). Thus, when the documents are too long, LSTM is not powerful enough to capture the long dependencies of the documents better.

It can be observed that the hierarchical network TextHCNN shows higher accuracy compared to TextCNN, with varying degrees of improvement on most datasets. Furthermore, the hierarchical networks, i.e. TextHCNN, HAN, Bi-layers-MHA, and HHNN-MHA show higher accuracy on each dataset than the non-hierarchical networks, indicating that hierarchical structures outperforms the non-hierarchical networks in terms of document classification. Besides, compared to HAN and Bilayers-MHA, which use the same method at word and sentence levels, HHNN-MHA achieves the

Table 2. Average accuracy of different datasets

Models	Accuracy on Test Set			
	Yelp2013	Yelp2014	Yelp2015	Amazon Food
<i>Non-hierarchical Models</i>				
BiLSTM	54.11	54.67	55.59	47.14
BiGRU	50.99	54.02	57.6	48.15
BiGRU+attention	56.04	57.45	57.23	45.76
TextCNN	40.62	41.76	44.28	43
<i>Hierarchical Models</i>				
TextHCNN	40.97	41.51	44.7	43.43
HAN	58.83	58.35	63.46	54.96
Bi-layers MHA	57.89	58.15	61.51	50.01
HHNN-MHA	59.36	58.36	63.61	53.28

best results on all three datasets, with an average accuracy rate of 0.34% and 1.76% higher, respectively. This demonstrates the rationality of designing different attention mechanisms for the word-level and sentence-level in the authors' model. The accuracy of the HAN model on the Amazon Food is slightly higher than that of HHNN-MHA. A possible explanation is that more specialized words exist in this category, which are more straightforward to identifying instead of analyzing the structure and the relation of the given document.

HHNN-MHA achieves better results on most datasets. This comes from several superiorities. First, the authors used a hierarchical structure to model long texts, making full use of the structural knowledge of the documents. Second, considering the long-distance dependency between sentences and the local features within sentences, the outputs of the CNN and BiGRU are fused to obtain sentence representations through a gating mechanism. Third, different attention mechanisms are applied to the word encoding level and sentence encoding level, and a multi-head attention mechanism is introduced at the word level to extract more symbolic semantic representations.

SENSITIVITY ANALYSIS

Ablation Study

To investigate the effects of the multi-head attention mechanism and the CNN on their model, the authors set up four controlled experiments for ablation studies based on the HHNN-MHA:

- **-MHA -CNN:** The multi-head attention and the CNN from the model are removed. In this case, the authors' model structure becomes a BiGRU-Attention model.
- **-MHA +CNN:** The multi-head attention mechanism is removed from the model while the CNN is kept. In this case, the authors use the output of the fused CNN and BiGRU of the gating mechanism in the first layer as the sentence representation vector.
- **+MHA -CNN:** The CNN is removed from the model and the multi-head attention is retained. In this case, the output of BiGRU is transferred directly to the multi-head attention mechanism.
- **+MHA +CNN:** All components are preserved.

The results are shown in Table 3. It can found that a significant drop in model accuracy from 0.91% to 7.52% after removing the CNN and multi-head attention mechanism. However, the model still outperforms most non-hierarchical models and a hierarchical model TextHCNN.

Adding CNN to the model improves the accuracy by 1.53%, 0.51%, 5.27%, and 6.53% on Yelp2013, Yelp2014, Yelp2015, and Amazon Food, respectively. This is because the CNNs can capture local features from sentence representations. Similarly, it can be noted that the model's accuracy is also improved considerably by adding a multi-head attention mechanism, and by adding all components, the model outperforms all hierarchical models except HAN on the Amazon Food dataset.

Table 3. Ablation studies of HHNN-MHA on document classification

Models	Accuracy on Test Set			
	Yelp2013	Yelp2014	Yelp2015	Amazon Food
-MHA -CNN	56.04	57.45	57.23	45.76
-MHA +CNN	57.57	57.96	62.5	52.13
+MHA -CNN	58.41	58.34	63.04	52.72
+MHA +CNN	59.33	58.36	63.61	53.28

Visualization of Attention Mechanism

Through the authors' experiments, it is found that the attention mechanism can improve the performance of the model. Therefore, the authors visualize the attention of words to find out which words contribute more to the sentiment information of the sentence. As shown in Figure 5, darker colored words imply greater attention weights. For example, in the sentence "We had excellent food with large portions," the word "excellent" is darker than the other words, proving that it contributes more sentiment information in this sentence. In this way, the outcome of the document can be predicted.

CONCLUSION

In this paper, the authors proposed a hierarchical hybrid neural network HHNN-MHA model based on a multi-head attention mechanism to address the problems in existing document classification studies. Such as, the RNNs do not focus sufficiently on local features and may lose sentiment information words at a distance when modeling long-dependent sentences, as well as the inability to perform parallel processing of input data. In the first layer, the authors used CNN-BiGRU and a multi-head attention mechanism for encoding and then fed the sentence vector to BiGRU in the second layer. For the purpose of sentence processing, a different attention mechanism was used in the second layer to obtain the final document representation.

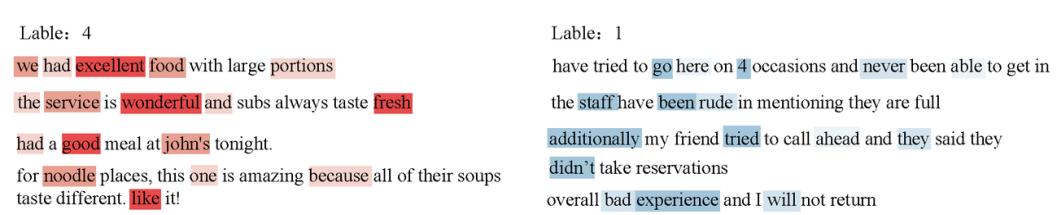
The experimental results on four datasets, Yelp2013, Yelp2014, Yelp2015, and Amazon Food, show that the HHNN-MHA model proposed in this paper significantly improves the results compared to the deep learning-based model, with an accuracy of 59.33%, 58.36%, 63.61%, and 53.28%, respectively.

Although the current model achieves a satisfying performance, there is still much research work that needs to be done in the future. First, the model can be further tuned to ensure performance, and its complex structure is still a problem, which leads to slower convergence. Second, how to improve the multi-head attention mechanism with a large number of parameters to avoid premature over fitting on small datasets will also be the focus of further research.

ACKNOWLEDGMENT

The acknowledgment of the article also needs to be revised. The revised acknowledgment is as follows: This work was supported by the National Statistical Science Research Project of China under Grant No. 2016LY98, the Characteristic Innovation Projects of Guangdong Colleges and Universities (Nos. 2018KTSCX049), the Science and Technology Plan Project of Guangzhou under Grant Nos. 202102080258 and 201903010013. This work was supported by Humanity and Social Science Foundation of the Ministry of Education of China (21A13022003), Zhejiang Provincial Natural Science Fund (LY19F030010), Zhejiang Provincial Social Science Fund (20NDJC216YB), Zhejiang Educational Science Fund (GH2021642).

Figure 5. Visualization of the attention mechanism



CONFLICT OF INTEREST

The authors of this publication declare there is no conflict of interest.

FUNDING AGENCY

The Open Access Processing Charge for this article was covered in full by the authors of this article.

REFERENCES

- Ali, M., Khalid, S., & Aslam, M. H. (2017). Pattern based comprehensive urdu stemmer and short text classification. *IEEE Access: Practical Innovations, Open Solutions*, 6, 7374–7389. doi:10.1109/ACCESS.2017.2787798
- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
- Cai, T., Shen, M., Peng, H., Jiang, L., & Dai, Q. (2019, October). Improving transformer with sequential context representations for abstractive text summarization. In *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 512-524). Springer. doi:10.1007/978-3-030-32233-5_40
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078. doi:10.3115/v1/D14-1179
- Collobert, R., & Weston, J. (2008, July). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167). doi:10.1145/1390156.1390177
- Du, J., Gui, L., He, Y., Xu, R., & Wang, X. (2019). Convolution-based neural attention with applications to sentiment classification. *IEEE Access: Practical Innovations, Open Solutions*, 7, 27983–27992. doi:10.1109/ACCESS.2019.2900335
- Gao, S., Ramanathan, A., & Tourassi, G. (2018, July). Hierarchical convolutional attention networks for text classification. In *Proceedings of The Third Workshop on Representation Learning for NLP* (pp. 11-23). doi:10.18653/v1/W18-3002
- Guan, Q., Ye, S., Yao, G., Zhang, H., Wei, L., Song, G., & He, K. (2009, July). Research and design of internet public opinion analysis system. In *2009 IITA International Conference on Services Science, Management and Engineering* (pp. 173-177). IEEE. doi:10.1109/SSME.2009.62
- Guggilla, C., Miller, T., & Gurevych, I. (2016, December). CNN-and LSTM-based claim classification in online user comments. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers* (pp. 2740-2751). Academic Press.
- Habimana, O., Li, Y., Li, R., Gu, X., & Yu, G. (2020). Sentiment analysis using deep learning approaches: An overview. *Science China. Information Sciences*, 63(1), 1–36. doi:10.1007/s11432-018-9941-6
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735 PMID:9377276
- Hu, Z., Hu, J., Ding, W., & Zheng, X. (2015, October). Review sentiment analysis based on deep learning. In *2015 IEEE 12th International Conference on e-Business Engineering* (pp. 87-94). IEEE. doi:10.1109/ICEBE.2015.24
- Huang, Y., Chen, J., Zheng, S., Xue, Y., & Hu, X. (2021). Hierarchical multi-attention networks for document classification. *International Journal of Machine Learning and Cybernetics*, 12(6), 1639–1647. doi:10.1007/s13042-020-01260-x
- Johnson, R., & Zhang, T. (2017, July). Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 562-570). doi:10.18653/v1/P17-1052
- Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188. doi:10.3115/v1/P14-1062
- Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
- Li, H., Xue, Y., Zhao, H., Hu, X., & Peng, S. (2019, October). Co-attention networks for aspect-level sentiment analysis. In *CCF International Conference on Natural Language Processing and Chinese Computing* (pp. 200-209). Springer.

- Liang, B., Liu, Q., Xu, J., Zhou, Q., & Zhang, P. (2017). Aspect-based sentiment analysis based on multi-attention CNN. *Journal of Computer Research and Development*, 54(8), 1724.
- Liu, F., Zheng, L., & Zheng, J. (2020). HieNN-DWE: A hierarchical neural network with dynamic word embeddings for document level sentiment classification. *Neurocomputing*, 403, 21–32. doi:10.1016/j.neucom.2020.04.084
- Liu, W., & Wang, T. (2010). Index-based Online Text Classification for SMS Spam Filtering. *Journal of Computing*, 5(6), 844–851. doi:10.4304/jcp.5.6.844-851
- McAuley, J., & Leskovec, J. (2013, October). Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems* (pp. 165-172). doi:10.1145/2507157.2507163
- Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543). doi:10.3115/v1/D14-1162
- Sarioglu, E. S. (2014). *Effective classification of clinical reports: natural language processing-based and topic modeling-based approaches* (Doctoral dissertation). The George Washington University.
- Tang, D., Qin, B., & Liu, T. (2015, July). Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) (pp. 1014-1023). doi:10.3115/v1/P15-1098
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wang, X., Jiang, W., & Luo, Z. (2016, December). Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 2428-2437). Academic Press.
- Wang, X., Liu, Y., Sun, C. J., Wang, B., & Wang, X. (2015, July). Predicting polarities of tweets by composing word embeddings with long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) (pp. 1343-1353). doi:10.3115/v1/P15-1130
- Xiao, L., Hu, X., Chen, Y., Xue, Y., Chen, B., Gu, D., & Tang, B. (2020). Multi-head self-attention based gated graph convolutional networks for aspect-based sentiment classification. *Multimedia Tools and Applications*, 1–20. doi:10.1007/s11042-020-10107-0
- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1480-1489). Academic Press.
- Zhao, H., Xue, Y., Gu, D., Chen, J., & Xiao, L. (2021, May). Modeling Inter-aspect Relationship with Conjunction for Aspect-Based Sentiment Analysis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 756-767). Springer. doi:10.1007/978-3-030-75765-6_60
- Zheng, J., Cai, F., Chen, W., Feng, C., & Chen, H. (2019). Hierarchical neural representation for document classification. *Cognitive Computation*, 11(2), 317–327. doi:10.1007/s12559-018-9621-6

Weihao Huang is a graduate student in the School of Electronics and Information Engineering of South China Normal University. He is currently studying for a master's degree. His main research fields include natural language processing, text sentiment analysis and deep neural networks.

Jiaojiao Chen obtained her master's degree from the School of Electronic and Information Engineering, South China Normal University, Foshan, China. Her research interest mainly includes natural language processing, deep learning, and text classification.

Qianhua Cai received her master's degree in the school of computer science of South China Normal University. After graduation, she joined the School of Electronics and Information Engineering of South China Normal University. Research interests: data mining, text data analysis.

Xuejie Liu is currently a lecturer at the School of Electronics and Information Engineering, South China Normal University, Foshan, China. She received her Bachelor Degree in Telecommunication Engineering and International Economy and Trade from Ludong University in 2011, the Master of Science in Acoustics from South China University of Technology in 2014, and PhD degree from University of New South Wales, Canberra, Australia in 2019. She is now working on brain-computer interface and electroencephalographic signal processing.

Yu-Dong Zhang worked as a postdoc from 2010 to 2012 with Columbia University, USA; and as an assistant research scientist from 2012 to 2013 with Research Foundation of Mental Hygiene (RFMH), USA. He served as a Full Professor from 2013 to 2017 with Nanjing Normal University. Now he serves as Professor with School of Informatics, University of Leicester, UK. His research interests include deep learning and medical image analysis. He is the Fellow of IET (FIET), and Senior Members of IEEE, IES, and ACM. He was the 2019 recipient of "Web of Science Highly Cited Researcher". He is included in "Top Scientist" in Guide2Research. He is the author of over 300 peer-reviewed articles, including more than 50 ESI Highly Cited Papers, and 4 ESI Hot Papers. His citation reached 16192 in Google Scholar (h-index 71), and 9444 in Web of Science (h-index 54). He has conducted many successful industrial projects and academic grants from NIH, Royal Society, GCRF, EPSRC, MRC, British Council, and NSFC.

Xiaohui Hu received the BS degree from Beijing Institute of Technology in 1993 and the MS degree from Hunan University in 1999. She is currently an associate professor of School of Electronics and Information Engineering. Her research interests include natural language processing, data mining and pattern recognizing.