IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 June 2024, accepted 12 July 2024, date of publication 15 July 2024, date of current version 24 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3428829

RESEARCH ARTICLE

ADHN: Sentiment Analysis of Reviews for MOOCs of Dilated Convolution Neural Network and Hierarchical Attention Network Based on ALBERT

LISHA YAO
School of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei, Anhui 230088, China
e-mail: jsjyaolisha@163.com

This work was supported in part by the Key Research Project of Natural Science in Universities of Anhui Province under Grant KJ2020A0782; in part by the Key Scientific Research Project of Anhui Provincial Research Preparation Plan, in 2023, under Grant 2023AH051806 and Grant 2023AH051807; in part by Anhui Xinhua College Quality Engineering Project under Grant 2020 sysxx01 and Grant 2020ylzyx02; in part by Anhui Xinhua University Level Research Project under Grant 2022zr003; and in part by the 2022 Anhui Province Quality Engineering Construction Project under Grant 2022xsxx089, Grant 2022sx060, Grant 2020jxtd120, Grant 2020jyxm0793, and Grant 2022jyxm659.

ABSTRACT Massive Open Online courses (MOOCs) are increasingly utilized by learners for knowledge acquisition and skill development. Accurate extraction of emotional information from MOOC course reviews plays a pivotal role in enhancing the quality of MOOC courses and fostering sustainable growth of MOOC platforms. Currently, sentiment analysis of MOOC course reviews predominantly focuses on general aspects, overlooking the hierarchical structure of text. Moreover, recurrent neural networks suffer from recursion limitations leading to reduced computational efficiency, while word embedding fails to address the issue of one-time polysemy. In this study, we propose ALBERT-DCNN-HAN (ADHN), an advanced text sentiment analysis model based on Dilated Convolution Neural Network and Hierarchical Attention network derived from ALBERT (A Lite BER) Network. The model primarily relies on the continuous updating of DCNN to compensate for the lack of hierarchical structure in deep neural networks, while also addressing the conflict between traditional word segmentation techniques and the trend towards emotion expression. Firstly, we employ the ALBERT model to generate ALBERT word vectors that integrate contextual features and dynamic semantics. ALBERT further incorporates contextual features from the sentence in which each word is located into its corresponding word vector, thereby generating distinct semantic vectors based on different meanings of polysemous words. Subsequently, these ALBERT word vectors are sampled and computed using DCNN to extract text features across multiple scales of context. Moreover, in order to capture both sentence-level and word-level characteristics comprehensively during text emotion expression, we fully consider hierarchy by integrating a hierarchical attention mechanism. Finally, Conditional random field (CRF) is employed for emotion prediction. Through analysis of information from empirical dataset derived from reviews of MOOC courses, our results demonstrate that this model effectively extracts valuable textual information and achieves an improved bias classification accuracy on review datasets compared with other neural network models.

INDEX TERMS MOOC, dilated convolution, HAN (hierarchical attention network), sentiment analysis, ALBERT (A Lite BER).

I. INTRODUCTION

With the swift progress of Internet technology, massive Open online Courses (MOOC) [1], [2], an online learning platform,

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

have garnered significant attention. While participating in these courses, learners frequently provide comments in the designated section. These reviews encompass not only assessments of course quality but also direct feedback on technical issues encountered with the MOOC platform. Given the substantial volume of comment data, manual methods for statistical analysis prove challenging. Employing sentiment analysis is deemed optimal to ascertain the emotional inclination within comment texts and effectively extract and mine a vast array of valuable information from them. This approach not only aids learners in selecting suitable courses but also assists platform administrators in identifying potential problems.

Emotional analysis of text is a subjective process that leverages natural language processing techniques to extract and analyze the emotional nuances within the text. Currently, there are three main approaches for emotion analysis: emotion dictionaries, machine learning, and deep learning [3], [4]. When using an emotion dictionary approach, it is crucial to select a high-quality and comprehensive emotional lexicon in order to accurately capture the emotional content of words. On the other hand, machine learning-based methods rely on manually created and extracted features for word-level emotional analysis. In recent years, with advancements in various learning technologies, deep learning has gained widespread adoption.

The research motivation of this paper lies in the following aspects:

(1) The traditional deep learning network model overlooks the hierarchical structure of text, while convolutional neural networks fail to effectively capture the sequential information between word sequences. Additionally, recurrent neural networks suffer from recursion, resulting in low computational efficiency.

(2) Furthermore, the implicit expression and blurred boundaries of emotional words pose challenges for polysemy and emotional analysis.

To address the aforementioned issues, this paper proposes a text sentiment analysis model that leverages ALBERT's dilated convolutional neural network and hierarchical attention mechanism. By utilizing ALBERT's word vectors, this model captures the depth of words and embeds contextual human dimensions to extract multi-scale context information through a dilated convolutional neural network. This compensates for the lack of hierarchical structure in deep neural networks and facilitates the extraction of multi-scale context information. Furthermore, we introduce a hierarchical attention mechanism to capture correlations and semantic information among words in the text, enabling more accurate modeling of long-distance dependencies between sentences while synthesizing sentence-level and word-level information. Experimental results demonstrate that our proposed model achieves promising performance in critical emotion analysis.

The main contributions of this paper are as follows:

(1) We propose a model based on ALBERT-DCNN, where the ALBERT embedded layer is utilized as input and combined with hierarchical attention mechanism for hierarchical text analysis. To enrich the embedded information, we introduce two modes, namely sentence and word.

(2) In order to further enhance the ability of extracting multi-scale features, we adopt parallel multi-void rate void convolution groups and expand the receptive fields in different ranges to enrich the diversity of features within the same layer.

(3) We employ a hierarchical attention network (HAN) to extract internal structural information from key words and sentences, capturing characteristics at different levels to capture contextual relevance effectively and thereby improving original text analysis capabilities.

The paper is structured as follows: Section I provides an introduction to the background, research motivation, and main contributions of sentiment analysis. Section II presents a comprehensive review of related work in the field of sentiment analysis. In Section III, we present the architecture and principles underlying our proposed model. Section IV details the experimental setup, results comparison, and analysis. Finally, Section V summarizes the key findings of this study and outlines future directions.

II. RELATED WORK

Currently, there exist three types of emotion analysis methods: emotion dictionary, machine learning, and deep learning. The approach based on an emotion dictionary necessitates the artificial construction of a high-quality and comprehensive emotion lexicon to accurately capture the emotional content of text; however, it overlooks contextual semantic information. The machine learning-based method requires feature extraction followed by classification using classifiers such as support vector machines [5], naive Bayes [6], or deep forest [7]. Nevertheless, these approaches rely on manual feature engineering.

In recent years, deep learning networks have been extensively employed for sentiment analysis due to their ability to automatically extract text features, which exhibit characteristics such as low labor cost, minimal domain knowledge requirements, and wide applicability. Attardi and Sartiano [8] utilized a convolutional neural network (CNN) for emotion analysis and achieved promising results on three-category emotion datasets. However, CNNs are limited in effectively capturing the sequential information of word sequences. To address this limitation, Bahdanau et al. [9] employed Recurrent Neural Networks (RNNs) for sentiment analysis as they can capture semantic features of the text; however, RNN's recursive nature leads to computational inefficiency and inadequate handling of long-distance textual information. Long Short-Term Memory (LSTM) models were introduced as an improvement to overcome these challenges. Zhang et al. [10] incorporated LSTM with attention mechanism to enhance model efficiency and

demonstrated that LSTM is capable of extracting semantic features more effectively than other approaches. Behera et al. [11] highlighted the effectiveness of deep convolutional networks in local feature selection while emphasizing that recursive networks like LSTM yield favorable results in sequence analysis tasks involving lengthy texts. By combining CNN and LSTM models, issues related to slow convergence and low recognition accuracy are resolved while enabling extraction of deep abstract features from the data at hand.Jiang et al. [12] proposed a fine-grained LSTM-CNN classification model incorporating attention mechanism wherein Bi-LSTM is used for contextual information retrieval between text contexts followed by addition of attention mechanism prior to CNN pooling stage thereby preserving lost information during pooling process leading to improved accuracy in text emotion classification tasks. Cui et al. [13] proposed a discrete self-attention mechanism that utilizes sparse substitution instead of the softmax algorithm to induce sparsity in sentiment analysis. Zhang et al. [14] employed convolution operation and attention mechanism to capture semantic features for extracting feature information from multi-word N-gram grammar, thereby accomplishing sentiment analysis tasks. Gan et al. [15] introduced the CNN-BiLSTM model with an attention mechanism, which incorporates a multi-channel extended joint structure capable of extracting both original context features and high-level multi-scale context features, resulting in significantly improved accuracy across multiple datasets; however, this model structure is complex as it requires multiple channels of CNN-BiLSTM for feature extraction. Wu et al. [16] proposed a word vector-based representation method and incorporated BiLSTM and an attention mechanism to effectively address issues related to inaccurate word segmentation and dependency on attention parameters. Huang et al. [17] put forward ERNIE2.0-BiLSTM-Attention model for implicit emotion analysis task, which can better capture the context semantics of implicit emotion text. Chen et al. [18] used memory network and hierarchical attention mechanism to obtain text representation. Li et al. [19] proposed a deep self-attention Bi-LSTM model to enhance the emotional information related to the object. In order to solve the problem of high computational complexity of LSTM threshold unit, Chen et al. [20] proposed an alternative scheme of LSTM-Gated Recurrent Unit,GRU). Compared with LSTM model, GRU model has a simpler structure and can greatly improve the training and reasoning speed of the model. Using bidirectional slice GRU to enhance the depth of semantic extraction shows the necessity of extracting depth information. The above models can't model the complete context well and ignore the hierarchical structure of the text, which leads to the lack of rich emotional features. Attention mechanism is only used to assign higher weights to important features after feature extraction, and the overall training speed of the model will also decrease due to the cyclic dependence mechanism of the recurrent neural network. Yang et al. [21] proposed a model for classifying emotions at the aspect level using graph neural networks. The model utilizes graph convolutional networks to enhance node representation, allowing it to learn global semantic and syntactic structures of sentences effectively. He et al. [22] introduced a trapezoidal structure bidirectional LSTM model that performs similarly to the standard structure but with fewer parameters. Zhang et al. [23] proposed an SA-Model that integrates multiple encoders to extract text features at different levels, thereby enhancing the accuracy and generalization capability of the model for analyzing poetic emotions. Rahman et al. [24] introduced a multi-layer classification approach for supervised machine learning on social media texts, while Mohamed et al.'s study [25] demonstrated the superior performance of LexDeep models over support vector machines in sentiment analysis tasks specific to Twitter datasets. Xu et al.'s work [26], based on BERT and hypergraph dual attention mechanism, significantly improved sentiment analysis accuracy for short online Chinese texts by dynamically extracting features and aggregating correlation information through dual graph attention mechanisms. Liu et al. method [27], which incorporates self-attention and dynamic word/sentence characteristics, effectively encodes comments for emotion analysis purposes.

Effectively expressing word vectors in sentiment analysis is of utmost importance. Commonly employed methods for word embedding learning include word2vec [28] and GloVe [29]. However, these models only provide a single vector representation for the same word in different contexts, which fails to effectively address the issue of word polysemy. Additionally, static word vectors are unable to capture polysemy adequately. The Bert (Bidirectional Encoder Representation from Transformers) model, based on Transformer architecture, exhibits strong language representation and feature extraction capabilities [30]. Nevertheless, this model suffers from poor reproducibility and slow convergence speed, necessitating substantial computational power. To enhance the slow convergence of the BERT model, Lan et al. [31] proposed ALBERT (A Lite BERT), a two-way coding feature representation model that pre-trains extensive text data and subsequently fine-tunes it according to downstream tasks to improve downstream prediction performance. The ALBERT pre-training language model serves as a bidirectional Trandformer encoder-based feature representation that can extract key features from text while minimizing parameter usage. In order to address the limitations of one-time polysemy and accurately express word features, this study employs the ALBERT model to obtain embedding vectors that combine contextual features with diverse semantics. These word vectors not only encompass semantic characteristics inherent in words but also integrate contextual attributes thereby compensating for deficiencies observed in traditional approaches towards word embedding methods. Moreover, owing to its lightweight nature, it is

well-suited for large-scale deployment with notable engineering application advantages.

In recent years, the field of MOOCs has witnessed numerous studies. Tao et al. [32] employed a sentiment analysis approach based on emotional and semantic features to comprehend the emotions exhibited by MOOC students through their participation and emotional behaviors. This study primarily relies on traditional sentiment analysis techniques and semantic feature extraction methods. Hew et al. [33], on the other hand, utilized a supervised machine learning method that combines sentiment analysis with gradient boosting trees (Gradient Boosting Trees) to predict students' satisfaction with MOOC courses. Although this method is rooted in conventional machine learning technologies and sentiment analysis, it does not incorporate deep learning models. Li et al. [34], however, applied natural language processing (NLP) for performing sentiment analysis on learners' comments while exploring key factors in MOOC teaching. Their study focuses on identifying characteristics of successful courses through student comments and places greater emphasis on pedagogy rather than utilizing deep learning techniques or traditional sentiment analysis approaches alone. This paper proposes the ADHN model which integrates multiple deep learning technologies including contextual semantic understanding from ALBERT, multi-scale feature extraction from DCNN, and hierarchical attention mechanism from HAN to address computational inefficiency issues associated with RNN as well as polysemy-related problems encountered in traditional word embedding methods during text sentiment analysis process. The proposed model demonstrates significant improvements thereby enhancing the quality and user experience of MOOC courses thus validating its effectiveness in practical applications.

III. ALBERT-BASED DILATED CONVOLUTION NEURAL NETWORK AND HIERARCHICAL ATTENTION MODEL

A. OVERALL STRUCTURE

The ALBERT-DCNN-HAN model, as illustrated in Figure 1, is proposed in this paper by integrating the Albert and Han models. It comprises several components: the ALBERT embedding layer, DCNN module, hierarchical attention module, and CRF module. Firstly, to address the limitation of traditional word embedding methods in expressing multiple meanings, this study generates character-level word vectors based on pre-training with the ALBERT model. Subsequently, the ALBERT word vector employs parallel dilated convolution groups with various dilated rates to extract text semantic features and aggregate multi-scale semantic dependencies. Simultaneously, the HAN module is utilized to consolidate both sentence-level and word-level information by attending to local contextual cues that highlight frequently occurring characters and words within a sentence. The features learned by the DCNN module (H) and the features learned by the HAN module (d) are concatenated. Subsequently, a Linear layer is employed to linearly combine the features from different feature maps, thereby extracting higher-level feature information that can be effectively utilized for enhanced classification. Finally, a CRF module is incorporated for comprehensive sentence output modeling.

B. BERT MODEL

The language model in natural language processing is primarily utilized to compute the probability of a sequence of linguistic elements $X_1, X_2, \cdots, X_n$. The calculation method is shown in Eq. (1).

$$p(S) = p(X_1, X_2, \cdots, X_n) = \prod_{i=1}^{n} p(X|X_1, X_2, \cdots, X_{i-1}) \tag{1}$$

The conventional unidirectional neural network language model fails to incorporate contextual information, while simultaneously lacking the ability to capture word ambiguity due to fixed word embeddings. The BERT model [35], [36], [37] and the ELMo model [38], [39], [40], [41] demonstrate proficient solutions to the aforementioned issues. In comparison with the ELMo model. The benefits of the GPT (Generative Pre-Training) [42] and ELMo [43] models are combined in the BERT model, which uses a bidirectional Transformer [43] as an encoder. In contrast to LSTM, it makes use of a superior transformer. On the other hand, the bidirectional language model enables BERT to obtain contextual information, which in turn makes word embedding richer in semantic information.

Figure 2 depicts the layout of the transformer coding unit. Transformer mainly uses the idea of attention mechanisms and obtains the internal relations of the sequence through self-attention. The calculation method is as shown in Eq. (2).

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})\mathbf{V} \tag{2}$$

The calculation method is demonstrated in Eq. (3)-(4) through the results obtained from the splicing of multi-head structures.

$$\begin{aligned} &MultiHead\ (Q, K, V) \\ &\quad = Concat(head_1, head_2, \cdots, head_h)W^O \end{aligned} \tag{3}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{4}$$

The Transformer coding unit incorporates residual connections and layer normalization to enhance the network's trainability, as depicted in Eq. (5)-(6).

$$LN\ (x_i) = \alpha \times \frac{x_i - \mu_L}{\sqrt{\sigma_L^2 + \varepsilon}} + \beta \tag{5}$$

$$FFN = \max\ (0, xW_1 + b_1)\ W_2 + b_2 \tag{6}$$

The input incorporates positional encoding to address the issue of attention mechanism overlooking temporal characteristics in time series data. The computational approach is illustrated in Eq. (7)-(8).

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{7}$$

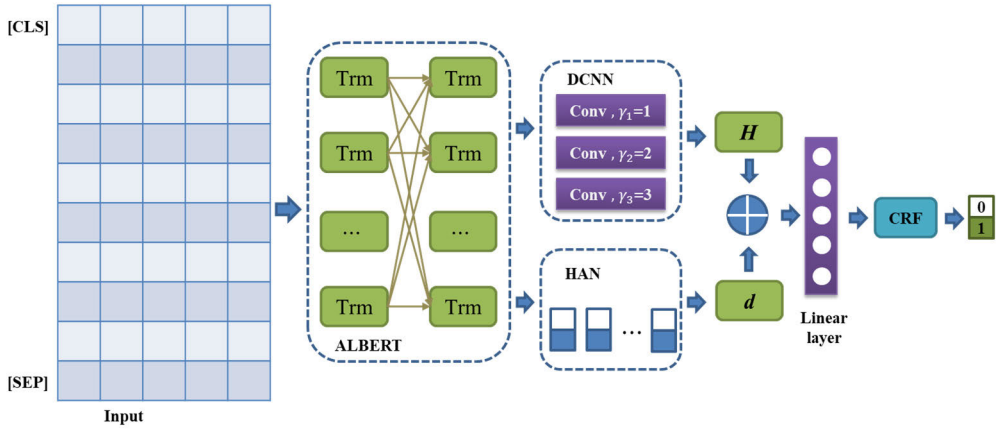

FIGURE 1. ALBERT-DCNN-HAN model structure.

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{model}}) \tag{8}$$

The position embedding and word embedding are concatenated and subsequently fed into the BERT model.

Although the model structure is only an improvement of GPT and ELMO, BERT innovatively put forward two tasks: ''Masked language model'' and ''next sentence prediction''. The masked language model randomly selects 15% of the of the words in the corpus to be replaced by the mask, and 80% of these selected words are replaced normally, 10% are replaced by another word, and 10% remain unchanged. The subsequent sentence prediction task involves selecting two sequentially connected sentences from the document as positive samples and randomly choosing sentences from different documents to follow the first sentence as negative samples during pre-training of the language model. These tasks capture information between words and sentences, respectively, and integrating them during training can enhance the global expressiveness of word embeddings.

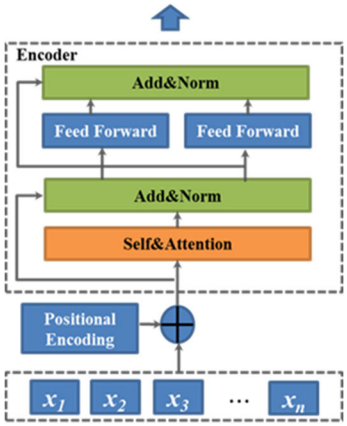

FIGURE 2. Transformer encoder.

C. IMPROVEMENT OF ALBERT

The ALBERT model incorporates two parameter reduction techniques to optimize memory consumption and enhance the training efficiency of BERT.

To eliminate unnecessary special characters such as punctuation marks and expressions in the dataset, regular expressions are employed for data cleansing and extraction of meaningful words. Subsequently, the text is segmented at the character level using a word divider, followed by removal of stop words. The characters are then encoded using a vocabulary, with non-existent characters being replaced by [UNK]. Additionally, preceding and succeeding the text, sentence vector [CLS] and clause marker [SEP] are added respectively. Finally, as one of the inputs to the ALBERT model, the character vector is transformed into a dynamic word vector incorporating contextual information.

The ALBERT model incorporates word embedding matrix decomposition and cross-layer parameter sharing strategy to optimize the number of parameters and enhance its semantic comprehension capability. Additionally, it replaces the original Next Sentence Prediction (NSP) task with Sentiment Order Prediction (SOP) task. Figure 3 illustrates the architecture of this model, where $X_1, X_2, \cdots, X_n$ represent words in a text sequence, and $E_1, E_2, \cdots, E_n$ denote the corresponding extracted textual feature vectors.

The ALBERT model addresses the issue of a large number of BERT parameters by utilizing bidirectional Transform to capture text characteristics in its encoder output. Building upon BERT, ALBERT introduces three enhancements.

1) WORD EMBEDDING VECTOR FACTORIZATION

By incorporating large word features into matrix decomposition, we obtain two relatively compact matrices, effectively separating the hidden layer. Through factorization, the model's complexity is reduced from $O(VH)$ to $O(VE+EH)$,

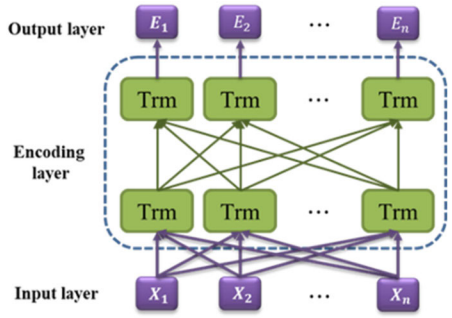

FIGURE 3. Structure of ALBERT model.

resulting in the following changes in complexity:

$$O(VH) \rightarrow O(VE + EH) \tag{9}$$

The length of the vocabulary is denoted by V in Eq. (9), while H represents the size of the hidden layer. E indicates the word embedding size, and when $E \ll H$, it significantly reduces the parameter quantity.

2) CROSS PARAMETER SHARING

To address the issue of excessive parameters, ALBERT employs a multi-layer parameter sharing strategy to effectively reduce the number of parameters. This includes parameter sharing in both the fully connected layer and attention layer. Such an approach ensures that the model network does not suffer from an increase in parameter count as its depth increases.

3) PARAGRAPH CONTINUOUS TASK

The ALBERT model enhanced the NSP (Next Sentiment Prediction) task in BERT and introduced the SOP (sentence-order prediction) task to mitigate the impact of topic recognition. This approach enables the selection of positive and negative samples within a single text, independent of sentence order.

The ALBERT pre-training model, functioning as a word embedding layer, incorporates the self-attention mechanism to effectively capture the interdependencies between words and generate dynamic word vectors with enhanced representational capacity. Notably, the initial vector in ALBERT's word vector sequence is denoted by [CLS], which can be leveraged for downstream classification tasks. Additionally, the clause vector [SEP] serves as a separator to demarcate distinct sentences.

D. DCNN MODULE

The second layer serves as the feature extraction layer in the DCNN model, aiming to capture the emotional information of textual content within a global context and mitigate potential recursion issues.

The CNN model demonstrates limited proficiency in learning from sequential dimensions and requires the stacking of multiple layers to effectively capture contextual information. In contrast, the proposed neural network with dilated convolution adeptly addresses this issue. Dilated convolution primarily differs from ordinary convolution in terms of its design for the convolution kernel. Furthermore, this type of dilated convolutional neural network is not significantly constrained by CNN, thereby facilitating feature map extraction and ensuring the preservation of feature integrity. Additionally, this network exhibits hierarchical recognition capabilities for text structure without compromising spatial dimensionality. During the process of convolution, dilated convolution expands the receptive field by selectively skipping elements.

In essence, the objective of the dilated convolutional network is to retain dense emotional analysis information while maximizing the receptive field. To achieve this, it is necessary to eliminate the pooling layer in the convolutional neural network to prevent loss of emotion analysis features during pooling. Additionally, increasing the receptive field requires transforming the convolution layer into a dilated convolution layer. The primary purpose of dilated convolution is to expand the receptive field through multiple consecutive dilations. The traditional convolution operation and its corresponding calculation formula are presented in Eq. (10).

$$z(x, y) = \sigma\left(\sum_{i,j} f(x + i, y + j) * g(i, j) + b\right) \tag{10}$$

$$z(x, y) = \sigma\left(\sum_{i,j} f(x + i * d, y + j * d) * g(i, j) + b\right) \tag{11}$$

The working principle of dilated convolution is illustrated in Figure 4. Dilated convolution employs a 3∗3 convolution kernel with an expansion degree of 0, as depicted in Figure 4(a). The operational procedure resembles that of general convolution; however, the pivotal component in dilated convolution neural networks lies within the convolution kernel. Upon completion of the operation, each grid effectively represents information from the original 3*3 grid, thereby resulting in a receptive field size of 3*3. The dilated convolution kernel depicted in Figure 4(b) has a size of 3*3, and the expansion degree d is set to 2. In Figure 4(b), despite having an expansion degree of 2, the actual spatial extent of the convolution kernel is highlighted in red. This convolution operation is performed on the feature map subsequent to the convolution operation shown in Figure 4(a). Consequently, compared to the convolution in Figure 4(a), there is a modification in the effective receptive field of the feature map resulting in a size change to 7*7. The convolution kernel in Figure 4(c) exhibits a size of 3*3 and an expansion degree (d) of 4, resulting in a receptive field of 15*15. When combined with the conventional three-layered 3*3 convolution kernel, the receptive field expands to reach dimensions of 7*7. This clearly demonstrates that the integration of dilated convolution kernels exponentially increases the receptive field value. Consequently, by employing this structure, dilated convolutions can effectively eliminate pooling processes

without sacrificing information loss while enhancing each convolution output's informational content.

Convolution operation primarily involves the identification and analysis of emotions in a physical context, aiming to extract relevant characteristics. If the convolution kernel effectively identifies emotional features during region analysis, the activation value of the feature can be computed using the aforementioned formula. Consequently, a larger Z value will activate in the new feature map Z, thereby accomplishing emotional feature identification. However, it is crucial to note that each layer of convolutional layers contains different convolution kernels for extracting distinct information during feature extraction and recognition processes. The local connectivity and weight sharing advantages inherent in convolutional layers make them more advanced than fully connected layers. Moreover, despite being composed of only a few parameters, convolution kernels possess remarkable sensitivity towards identifying pertinent information. Specifically regarding emotional details identification, multiple cooperation between convolution kernels is required to complete feature extraction in emotional analysis while significantly expanding receptive fields and enhancing recognition capabilities for fundamental features obtained through convolutions.

The DCNN module proposed in this study is a parallel multi-dilated rate dilated convolution group, which aims to investigate the semantic interdependence among words at varying distances and capture contextual information across different scales. E is the word vector matrix, $E \in R^{n\times d}$. Where n represents the number of words and d represents the dimension of the word vector. Dilated convolution can expand the basic convolution kernel to different sizes by different expansion rates. F is defined as the basic convolution kernel, and its receptive field is $h \times w$, where h represents the height of the convolution kernel and w represents the width of the convolution kernel. The expansion rate is $(\gamma, \emptyset)(\gamma, \emptyset \geq 1)$, and the receptive field of the expanded dilated convolution kernel F^{γ} is $D(h, \gamma) \times D(l, \emptyset)$. $D(h, \gamma) = (h-1) \times \gamma + 1, D(l, \emptyset) = (l-1) \times \emptyset + 1$. In this paper, the basic convolution kernel with $h = 3, w = 3$ is selected, and the dilated operation is performed on the first dimension of the basic convolution kernel ($\emptyset = 1$). The receptive field of F^{γ} is $(2\gamma+1)\times 3$.

The expression of the characteristic graph H^{γ} after dilated convolution can be formulated as Eq. (12).

$$H^{\gamma} = f(Conv(F^{\gamma}, E)) \tag{12}$$

Conv stands for convolution operation, $f()$ stands for activation function, and ReLu activation function is adopted here. Let the number of channels of F^{γ} be c^{γ}, then $H^{\gamma} \in R^{n\times d\times c^{\gamma}}$. In order to mitigate overfitting, the Dropout operation is employed subsequent to the dilated convolution operation, thereby reducing model parameters and mitigating the risk of overfitting. Eq. (13) can be reformulated as follows:

$$H^{\gamma} = Dropout\left(ReLu\left(Conv\left(F^{\gamma}, E\right)\right)\right) \tag{13}$$

The feature maps from each dilated convolution layer within the parallel multi-dilated rate dilated convolution group are integrated to effectively capture contextual semantic features across various scales. The fused text features $H \in R^{n\times d\times c}, c = \sum c^{\gamma}$.

The feature map after dilated convolution is preserved to match the input feature size by filling each dilated convolution layer with zeros. Additionally, employing an equal number of channels in each dilated convolution layer not only simplifies model parameter configuration but also facilitates parameter sharing among these layers. In this study, we enhance the expressive capacity of our model by aggregating multi-scale semantic features through parallel multi-void volumes while reducing the number of model parameters through parameter sharing.

The architecture diagram of the DCNN module proposed in this study is illustrated in Figure 5. The word vector matrix E, which has been processed by the ALBERT word embedding layer, is separately fed into 3×3 dilated convolutions with dilation rates of 1, 2, and 3. Subsequently, the dilated convolution operation is performed to generate three feature maps: H^1, H^2, H^3. Finally, these three feature maps are fused to obtain the ultimate feature map H.

E. HIERARCHICAL ATTENTION NETWORK (HAN)

For comprehensive emotional analysis of text, it is essential to consider contextual semantic relationships and pay attention to the significant impact of certain words on sentence emotion expression. To analyze text features from various angles and levels, this paper proposes a novel hierarchical attention model (HAN) that facilitates text analysis. From a sentence perspective, expanding attention mechanisms at both word and sentence levels can better capture long-distance dependencies between sentences and comprehend their informational content. The use of hierarchical attention mechanism not only improves performance but also dynamically focuses on sentences and words that aid in text analysis.

The HAN module is composed of four main layers: word coding layer, word attention layer, sentence coding layer, and sentence attention layer. Let L represent the number of sentences in the text, and T represent the number of words in each sentence. The structure of the HAN module is illustrated in Figure 6.

1) WORD CODING LAYER

Word-level coding is achieved through the use of a two-way Gated Recurrent Unit (GRU). GRU replaces the complex relationship between cell state and hidden state in LSTM with a linear relationship between hidden state and candidate hidden state, resulting in a simpler structure that is easier to calculate and train. The information transmission control structure of GRU is also relatively simple, consisting mainly of Update Gate and Reset Gate. The update gate determines

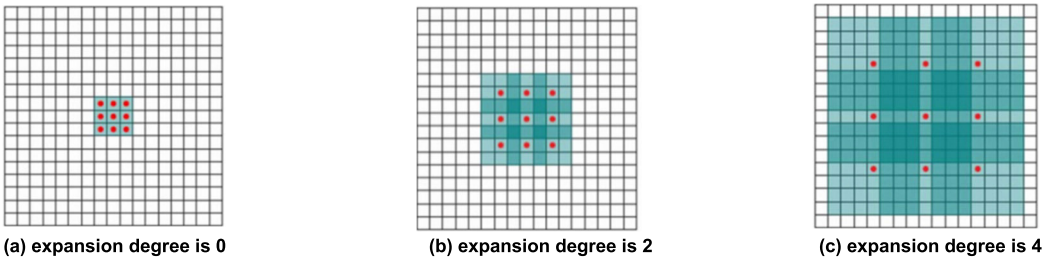

FIGURE 4. Transformer encoder.

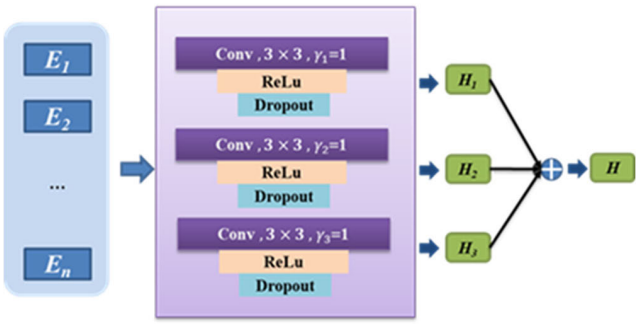

FIGURE 5. DCNN module structure diagram.

the ratio of previous moment's hidden state information to current moment's hidden state information as well as the ratio of current candidate hidden state information to current hidden state, while the reset gate determines the ratio of previous moment's hidden state information to current candidate hidden state. At time t, updating formulas for GRU's updated gate, candidate hidden states, reset gate and its ownhidden states are given by Eq. (14)-(17).

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{14}$$
$$z_t = \sigma\left(W_z x_t + U_z h_{t-1} + b_z\right) \tag{15}$$
$$\tilde{h}_t = tanh(W_h x_t + r_t * (U_h h_{t-1}) + b_h) \tag{16}$$
$$r_t = \sigma\left(W_r x_t + U_r h_{t-1} + b_r\right) \tag{17}$$

where h_t, $\tilde{h}_t$, h_{t-1} represent the hidden state at time t, the candidate hidden state and the hidden state at time $t-1$. x_t represents the input word vector, $t \in [1, T]$, and T represents the number of words in the sentence. z_t and r_t denote update gates and reset gates. *tanh* and σ represent *tanh* activation function and sigmoid activation function respectively. W_z, W_h, W_r, U_z, U_h, U_r represent the corresponding weights of GRU neurons. b_z, b_h, b_r represent the deviation of GRU neurons. * stands for element multiplication (Hadamard product).

The structure of the word coding layer is represented as follows:

$$\vec{h}_{it} = \vec{GRU}(x_{it}), t \in [1, T] \tag{18}$$
$$\overleftarrow{h}_{it} = \overleftarrow{GRU}(x_{it}), t \in [T, 1] \tag{19}$$

Among them, $\vec{h}_{it}$ and $\overleftarrow{h}_{it}$ respectively represent the forward hidden state and the reverse hidden state of the word w_{it}. The two words are merged to get the bidirectional hidden state $h_{it} = [\vec{h}_{it}, \overleftarrow{h}_{it}]$ of the word w_{it}.

2) WORD ATTENTION LEVEL

Firstly, the hidden layer vector u_{it} is obtained by applying a basic multi-layer perceptron network operation to the bidirectional hidden state h_{it} of the word w_{it} generated by the word coding layer. Subsequently, we introduce the softmax operation to derive the weight a_{it}, and finally obtain the sentence feature s_i through summation. The corresponding formula is expressed as Eq. (20)-(22).

$$u_{it} = tanh(W_w h_{it} + b_w) \tag{20}$$
$$a_{it} = softmax(u_w, u_{it}) \tag{21}$$
$$s_i = \sum_t a_{it} h_{it} \tag{22}$$

Among them, W_w, u_w and b_w represent the weight parameters and deviations associated with the word vector layer respectively.

3) SENTENCE CODING LAYER

The sentence feature s_i is encoded in a manner similar to that of the word coding layer, following the formulation expressed as Eq. (23)-(24).

$$\vec{h}_i = \vec{GRU}(s_i), i \in [1, L] \tag{23}$$
$$\overleftarrow{h}_i = \overleftarrow{GRU}(s_i), i \in [L, 1] \tag{24}$$

L stands for the number of sentences. Combining the two-way hidden states of the sentence s_i, we get $h_i = [\vec{h}_i, \overleftarrow{h}_i]$.

4) SENTENCE ATTENTION LEVEL

The structure of the sentence attention layer bears resemblance to that of the word attention layer. Emotional feature d is generated through the utilization of the sentence attention layer. The calculation formula can be represented as Eq. (25)-(27).

$$u_i = tanh\left(W_s h_i + b_s\right) \tag{25}$$
$$a_i = softmax\left(u_s, u_i\right) \tag{26}$$
$$d = \sum_i a_i h_i \tag{27}$$

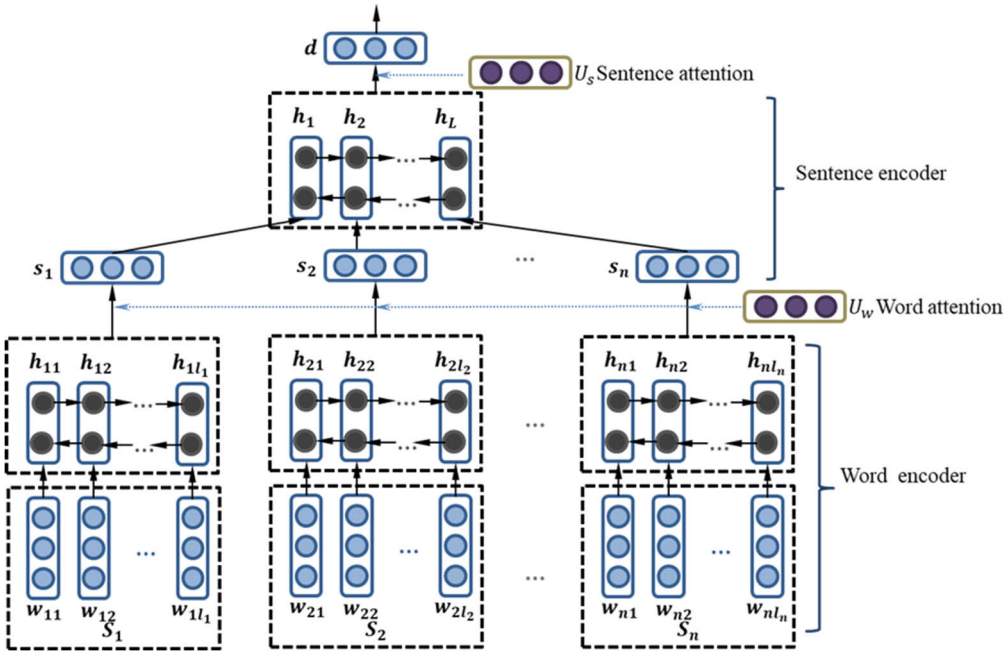

FIGURE 6. HAN module structure diagram.

Among them, W_s, u_s and b_s respectively denote the weight parameters and biases associated with the attention level of the sentence.

The purpose of the HAN module is to analyze the features extracted by the ALBERT-DCNN module in all directions. Additionally, the sentence attention layer enables more accurate capture of long-distance dependencies between sentences, effectively integrating fundamental information from both words and sentences. By leveraging a hierarchical attention mechanism, it facilitates comprehensive analysis at both word and sentence levels. The introduction of this hierarchical attention mechanism significantly enhances emotional analysis accuracy while dynamically attending to text segments that contribute to analysis. Furthermore, the HAN module assigns distinct attention weights to semantic coding, thereby discerning vector semantic coding importance and improving model accuracy.

F. CRF LAYER

The CRF is widely employed in natural language processing due to its ability to model sequential data and has numerous applications in this domain. On the other hand, the Deep Neural Network (DNN) excels at capturing long-range information and considers contextual details of input word (character) vector sequences; however, it overlooks tag dependencies despite their strong correlation. To address this limitation, we introduce CRF into our study for predicting optimal tags of text vectors.

Reconfigure the architecture by establishing a connection between the output H of the DCNN module and the output D of the HAN module, as depicted in Eq. (28).

$$M_i = tanh(W_M(H_i + MHead_i) + b_M) \tag{28}$$

M_i is the input of CRF layer. Where W_M and b_M respectively represent the weight matrix and the bias term.

The input sequence is $X = \{x_i\}_1^N$, and the score of the corresponding prediction sequence $S = \{s_i\}_1^N$ can be formalized as shown in Eq. (29).

$$s(X, S) = \sum_{i=1}^{N} \left(O_{i,s_i} + T_{s_{i-1},s_i}\right) \tag{29}$$

Among them, the vector of the ith column in the model input matrix is the vector M_i obtained from Eq. (28), and O_{i,s_i} represents the non-normalized probability that the input x_i is mapped to the tag s_i. A label transition matrix T, T_{s_{i-1},s_i} is introduced to represent the transition probability of two consecutive labels.

The specific scoring method can be expressed as shown in Eq. (30):

$$P\left(\frac{S}{X}\right) = \frac{e^{s(X,S)}}{\sum_{\hat{S}\in S_x} e^{s\left(X,\hat{S}\right)}} \tag{30}$$

Among them, $\hat{S}$ represents the real tag sequence, and S_x represents all the useful tag sets.

When conducting training, the output sequence can be obtained through the utilization of the likelihood function. The specific computational approach is illustrated in Eq. (31).

$$Log\left(P\left(\frac{S}{X}\right)\right) = s\left(X, S\right) - log\left(\sum_{\hat{S}\in S_x} e^{s\left(X,\hat{S}\right)}\right) \tag{31}$$

When determining the optimal label, it is assumed that the sequence S^* yielding the highest probability can be derived, as demonstrated in Eq. (32).

$$S^* = argmaxS\left(X, \hat{S}\right) \tag{32}$$

In the training phase, the function loss can be minimized through backpropagation, while during testing, the Viterbi algorithm enables us to obtain the label sequence with maximum probability.

IV. EXPERIMENTAL ANALYSIS

A. DATA SET AND EVALUATION CRITERIA

In order to validate the model, this paper collected a comprehensive dataset from China's MOOC website comprising 80,000 authentic comments written in Chinese characters using web crawling technology. These comments were specifically obtained for exceptional computer science courses and originated from diverse countries worldwide. The distribution of this dataset is presented in Table 1, while further details are provided in Table 2.

TABLE 1. Data set polarity statistics.

Data set	Positive proportion(%)	Negative proportion(%)	Number of samples
Training set	64.63	35.37	48000
Testing set	60.16	39.84	32000

TABLE 2. Demonstration of a sentiment polarity label.

Text	Label
This is one of the best machine learning, Python-like courses I've ever seen.	P
The downside is obvious. Lack of depth.	N

In order to validate the model, this paper employs Accuracy (Acc) and $MacroF_1$ as indicators for model verification, with Eq. (33) presented below.

$$\begin{cases} P = \dfrac{TP}{TP + FP} \\ R = \dfrac{TP}{TP + FN} \\ MacroP = \frac{1}{n}\sum_{i=1}^{n} P_i \\ MacroR = \frac{1}{n}\sum_{i=1}^{n} R_i \\ MacroF_1 = \frac{2 \times MacroP \times MacroR}{MacroP + MacroR} \\ Acc = \frac{T}{N} \end{cases} \tag{33}$$

where P represents precision, R represents recall, and n represents the number of classifications. TP denotes the count of correctly predicted positive samples. TN denotes the count of correctly predicted negative samples. FP denotes the count of falsely predicted positive samples among actual negatives. FN denotes the count of falsely predicted negative samples among actual positives. $MacroF_1(F_1)$ signifies the average value of F_1 for each category. T stands for true positive sample and N stands for total sample.

B. EXPERIMENTAL PLATFORM

The experimental platform and environment utilized in this study are presented in Table 3.

TABLE 3. Experimental platform.

Experimental environment	Specific information
Operating system	Microsoft Windows10
Processor	Intel Core i3-13100F 4.5GHZ
Graphics card	NVIDIA GeForce GTX1630
Memory	16GB
Development language	Python 3.7
Development environment	anaconda
Experimental environment	Specific information

C. EXPERIMENTAL PARAMETERS

The performance and recurrent processing ability of the model during training can be influenced by numerous parameters and factors. The model parameters are presented in Table 4. The word vector dimension employed in this study is set to 128. The maximum sequence length considered is 100. Dropout regularization technique is implemented to mitigate overfitting, with a dropout rate of 0.3, while the Adam optimizer is utilized for optimization purposes. The initial learning rate is configured as 0.0001, and a batch size of 64 samples per iteration is adopted throughout the training process. A total of 30 epochs are executed for convergence achievement. The ALBERT-DCNN module consists of three convolutional layers with a kernel size of 3×3 and 128 channels, employing the ReLU activation function. The dilation rates γ_1, γ_2 and γ_3 are respectively set to 1, 2, and 3. Additionally, the hidden layer within HAN has a dimensionality of 128. The Linear Layer neuron is set to 64.

TABLE 4. Parameter setting.

Parameters	Value
Word vector dimension(d)	128
Sequence length(n)	100
Dropout	0.3
Optimizer	Adam
Learning rate(baselr)	0.0001
Batch size	64
Epoch	30
Filter size	3×3
Number of convolution channels	128
Activation function	ReLu
Parameters	Value
Word vector dimension(d)	128

D. EMBEDDED LAYER ANALYSIS

In this paper, the results of two embedding layers are compared by experiments at word level and word segmentation level respectively. Text data is segmented according to word level, and word vectors are randomly initialized. The accuracy results of two embedding layers of the model are shown in Table 5.

TABLE 5. Model accuracy with different representation of embedded layer.

Embedding layer representation	Acc(%)
Character embedding	96.18
Word embedding	94.42

The present study compares the outcomes of two embedding layers through experiments conducted at both the word level and word segmentation level. Textual data is segmented based on individual words, with randomly initialized word vectors. The accuracy results for the two embedding layers employed in our model are presented in Table 5.

E. TEXT VECTOR REPRESENTATION ANALYSIS

To validate the efficacy of the ALBERT model in this study, diverse word vectors are employed to represent the model and distinct word vectors are trained as inputs for the embedding layer. The proposed model from this paper is utilized to analyze emotions on the identical dataset.

The ALBERT model, as demonstrated in Table 6, exhibits the highest Accuracy of 96.18% in this study. Unlike static word vectors such as Word2vec and GloVe that fail to effectively capture textual information for expressing polysemous words, ALBERT and ELMo offer dynamic representations of word vectors by fine-tuning their semantics based on downstream tasks and incorporating domain-specific knowledge, resulting in more feature-rich word vectors. However, while ELMo utilizes bidirectional LSTM to dynamically calculate semantic vectors within a contextual framework, ALBERT leverages the Transformer module with enhanced feature extraction capabilities, thereby yielding superior application performance compared to ELMo. The BERT model exhibits superior accuracy compared to ELMo, while also boasting the largest number of parameters. Consequently, the proposed model in this study achieves a harmonious equilibrium between parameter reduction and ensuring optimal performance for sentiment analysis.

TABLE 6. Comparative results of different word vector models.

Embedded learning method	Number of Parameters	Acc(%)
Word2vec	10M	92.36
GloVe	12M	93.67
ELMo	14M	94.25
BERT	110M	95.96
ALBERT	12M	96.18

F. LEARNING RATE ANALYSIS

In order to expedite the initial training phase and maintain a consistent learning rate during the later stages, thereby enhancing model convergence, we employ an exponential decay strategy for the learning rate as depicted in Eq. (34).

$$lr = max\left(baselr * 0.96^{globalstep}, 0.001\right) \tag{34}$$

where lr stands for learning rate. $baselr$ is the initial learning rate. $globalstep$ is the current iteration number.

To validate the rationality of the learning rate employed during model training, Figure 7 illustrates the impact of different learning rates on model performance, encompassing fixed learning rates (lr) of 0.01 and 0.05, as well as decaying learning rates with basic learning rates ($baselr$) of 0.1 and 0.05. As depicted in the figure, employing an exponentially decaying learning rate leads to faster convergence during training compared to a constant learning rate approach, resulting in higher final accuracy. Additionally, it is observed that smaller learning rates correspond to slower convergence speed. However, combining an appropriate decay strategy with a suitable initial learning rate not only ensures timely model convergence but also yields improved accuracy levels. This validates the selection of a reasonable decay strategy and corresponding value for the chosen learning rate in our experiment.

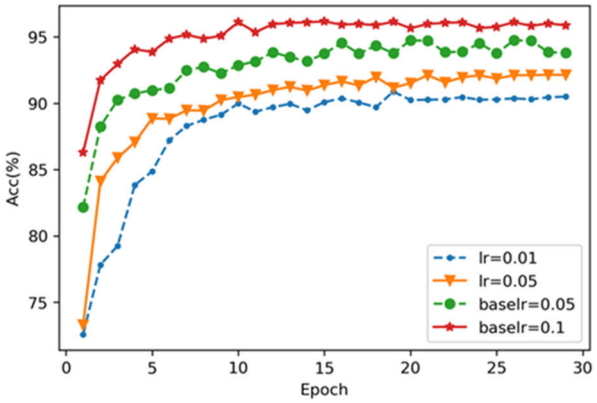

FIGURE 7. Influence of different learning rates on the model.

G. MODEL TRAINING ANALYSIS

In the process of model training, we also closely monitor the changes in accuracy and loss for both the training and test datasets, as depicted in Figure 8 and Figure 9 respectively. The x-axis represents the number of epochs trained on the entire dataset, while the y-axis represents accuracy and loss values. From Figure 8, it is evident that when training reaches Epoch = 16 on the complete dataset, our model achieves its highest accuracy on the test set. Although not reaching its absolute minimum value, at this point, the corresponding loss for the test set remains close to its minimum. However, as training progresses further, overfitting becomes apparent with a decline in test set accuracy accompanied by an upward trend in test set loss.

H. HIERARCHICAL ATTENTION ANALYSIS

In order to further validate the efficacy of the HAN module proposed in this study, three models are employed for comparative analysis.

① ALBERT-DCNN model: This model is obtained by removing the HAN module from the original architecture.

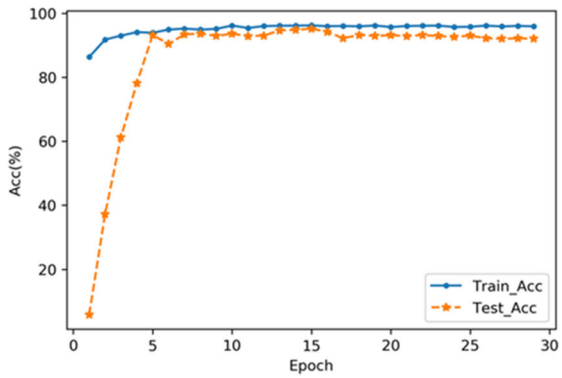

FIGURE 8. Acc variation curve of training set and test set.

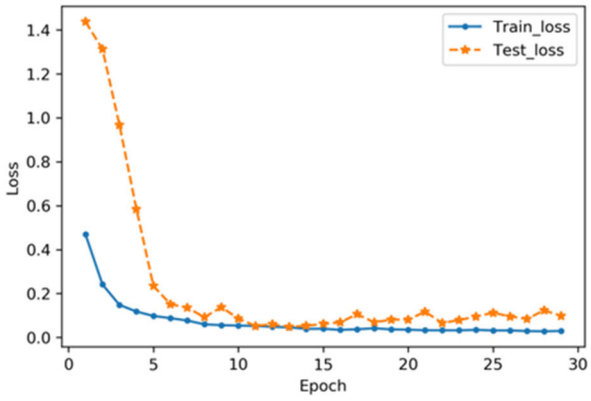

FIGURE 9. Loss variation curve of training set and test set.

② ALBERT-DCNN-ATT model: In this variant, the HAN module is substituted with a self-attention mechanism.

③ ALBERT-DCNN-HAN model: The proposed approach in this study involves incorporating an enhanced hierarchical attention network (HAN) into the ALBERT-DCNN framework.

The Acc comparison diagram of three models is presented in Figure 10. As depicted, this model exhibits the highest accuracy (Acc). The HAN module facilitates feature weight distribution, enabling the model to learn diverse levels of weighted features. This capability enhances the model's ability to accurately and swiftly capture higher-level weight information. Comparatively, when compared to the ALBERT-DCNN-HAN model without HAN mechanism, our proposed ALBERT-DCNN model achieves a significant improvement in Acc by 3.21%. This result underscores that incorporating the HAN mechanism enhances overall performance and plays a pivotal role in analyzing target-context interactions.

By comparing the ALBERT-DCNN-HAN model with the ALBERT-DCNN-ATT model, we can observe the superiority of the HAN module. The accuracy of the HAN-enhanced model is 1.16% higher than that of the ATT-enhanced model, indicating that incorporating hierarchical attention considers

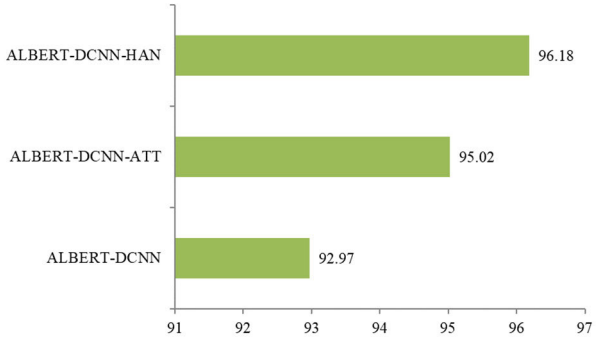

FIGURE 10. Text 2 character-level attention weight diagram.

both hierarchical text information and deep emotional features when extracting textual representations, thereby achieving superior performance compared to traditional self-attention mechanisms.

I. CLASSIFIER ANALYSIS

To validate the effectiveness and performance of the CRF classifier, we compare it with both the Softmax classifier and a CRF classifier based on this model. Figure 11 illustrates the changing trend of test accuracy for both classifiers. It is evident that compared to the Softmax classifier, the CRF classifier not only achieves higher classification accuracy but also exhibits faster convergence speed. This observation highlights that while the widely used Softmax classifier predicts labels solely based on maximum probabilities without considering label dependencies, our proposed CRF classifier takes into account such dependencies before and after each word in the classification process, leading to improved overall training efficiency.

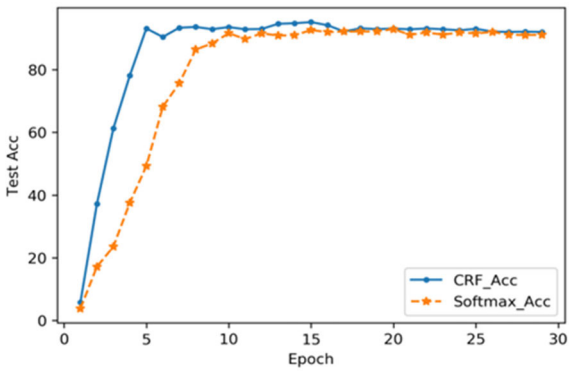

FIGURE 11. Performance comparison chart of Softmax classifier and CRF classifier.

J. CONTRAST EXPERIMENT

To ascertain the efficacy of the Albert-CNN-HAN model in sentiment analysis, this study conducted a comprehensive comparison with other prominent models including ERNIE2.0-BiLSTM-Attention [17], a graph neural network-based aspect-level emotion classification model [21], a BERT

and Hypergraph dual attention mechanism-based text emotion analysis model [26], GloVe-CNN-BiLSTM [29], and ALBERT pre-training model [31].

Meanwhile, in order to further validate the model's generalization capability proposed in this paper, it will be assessed on an additional dataset. The experimentation on an alternative dataset is conducted using Coursera's Course Reviews Dataset, which comprises 100K reviews and can be accessed at https://www.kaggle.com/septa97/100k-courseras-course-reviews-dataset. Each entry in this dataset represents a review for a specific course obtained from Coursera's website. The reviews were pre-labeled based on their corresponding ratings: very positive for a rating of 5, positive for 4, neutral for 3, negative for 2, and very negative for 1. This dataset encompasses feedback from 1835 distinct courses, comprising a total of 140320 reviews.

We conducted comparative experiments between the model proposed in this paper and the aforementioned models on both datasets within a consistent experimental environment. Performance assessment measures such as accuracy rate (Acc) and F_1 values were utilized to gauge their effectiveness, while presenting comparative outcomes in Table 7 and Figure 12-13.

TABLE 7. Model comparison results.

Model	China's MOOC		Coursera MOOCs review	
	Acc(%)	F_1 (%)	Acc(%)	F_1 (%)
Ref. [17]	94.23	94.17	91.67	89.02
Ref. [21]	94.79	94.06	91.52	88.59
Ref. [26]	95.83	94.59	92.27	89.16
Ref. [29]	93.38	91.87	90.01	88.27
Ref. [31]	92.76	91.86	88.92	87.54
Ours	**96.18**	**94.73**	**92.78**	**89.53**

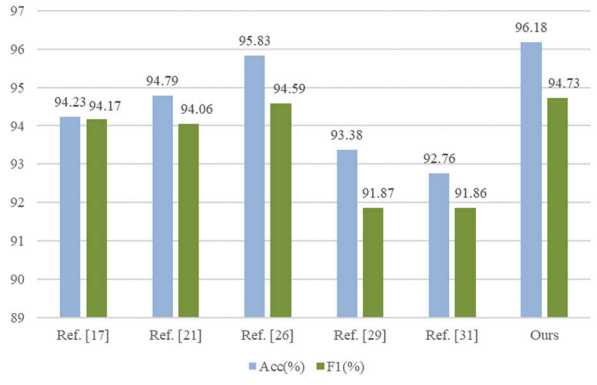

FIGURE 12. Compare the experimental results on China's MOOC.

The ERNIE2.0-BiLSTM-Attention model proposed by [17] exhibits superior proficiency in capturing the contextual semantics of implicit affective texts, thereby resulting in significant advancements in emotion analysis, as evident from Table 7 and Figure 12-13.

The model introduced by [21] utilizes a graph neural network to analyze and classify emotions across various

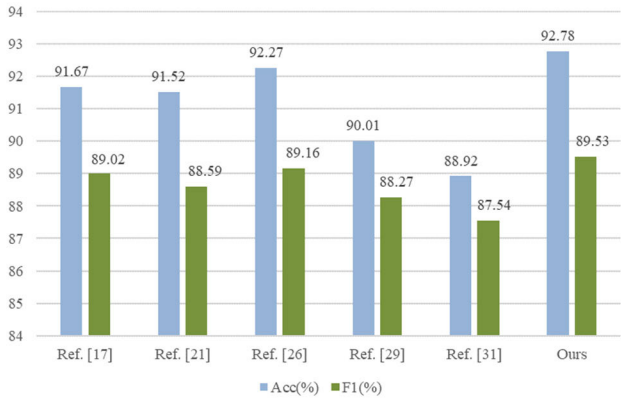

FIGURE 13. Compare the experimental results on Coursera MOOCs review.

aspects of the text, enabling more precise detection of subtle emotional fluctuations.

The approach presented in [26] leverages BERT, a pre-trained language representation model, as a feature extractor and integrates a hypergraph dual attention mechanism for context modeling, showcasing robust generalization capabilities when handling novel or unfamiliar samples.

Additionally, the approach proposed in [29] synergistically integrates GloVe word vectors, CNN, and BiLSTM to effectively harness their respective strengths: capturing intricate semantic relationships with GloVe word vectors, extracting localized features using CNN, and modeling enduring dependencies through BiLSTM. The ALBERT model introduced by [31] exhibits proficiency in extracting dynamic semantic features.

The proposed ALBERT-DCNN-HAN model in this study achieves the highest accuracy (Acc) and F_1 scores on both datasets. Specifically, on China's MOOC dataset, the Acc and F_1 scores reach 96.18% and 94.73%, respectively, surpassing the best method mentioned in [26] by a margin of 0.35% and 0.14%. Similarly, on the Coursera MOOCs review dataset, the Acc and F_1 scores achieve 92.78% and 89.53%, respectively, outperforming the best method described in [26] by a margin of 0.51% and 0.37%. Notably, the ALBERT model stands out as a lightweight and highly efficient pre-trained language representation model with fewer parameters compared to traditional BERT models, while still delivering exceptional performance. Consequently, sentiment analysis methods based on ALBERT can provide rapid and accurate results. By incorporating a dilated convolutional layer into our approach, we are able to capture information pertaining to the associations between different distances within the text and effectively process lengthy sequences of text. Moreover, by employing hierarchical attention mechanisms, our model can simultaneously focus on crucial features at various levels such as words and sentences, thereby facilitating a better understanding of textual structure and semantic information for conducting sentiment analysis across multiple levels. Through the integration of ALBERT, dilated convolutional neural network, and hierarchical attention network, we can

leverage their distinctive advantages to achieve enhanced performance in sentiment analysis tasks. The experimental results validate the effectiveness of the proposed model architecture.

V. CONCLUSION

This paper presents a comprehensive sentiment analysis of feedback on MOOC courses. To address the issue of large parameters and lengthy training time in BERT pre-training language models, we employ ALBERT pre-training model for word embedding training. In contrast to static word vector tools, ALBERT endows words with dynamic semantics that align with their context, thereby resolving polysemy concerns. This study establishes the ALBERT-DCNN-HAN model by integrating hole convolutional neural networks with a hierarchical attention mechanism. The hole convolutional neural network captures abstract relationships among characters or words, reducing the likelihood of recurrence problems. By leveraging word and sentence attentions, we effectively identify pertinent information for emotional tendency assessment. Furthermore, through incorporating a hierarchical attention network mechanism and analyzing at both word level and sentence level respectively, we fully exploit the characteristics and semantic combinations of words. The experimental results show that the model proposed in this paper achieves an Acc of 96.18% and an F_1 score of 94.73% on the China's MOOC dataset, and an Acc of 92.78% and an F_1 score of 89.53% on the Coursera MOOCs review dataset, which are higher than those of other models.

The findings of this research will contribute to a comprehensive understanding of students' course evaluation and feedback on online education platforms, thereby providing educational institutions and platforms with more robust data support. Additionally, sentiment analysis can facilitate timely identification and resolution of issues or dissatisfactions encountered by learners during the learning process, thus optimizing course design and teaching methodologies. Overall, this study holds significant implications for enhancing the quality of online education, fostering personalized learning experiences, and augmenting user satisfaction. In future investigations, we intend to delve deeper into sentiment analysis pertaining to MOOC course reviews while exploring factors influencing learners' sentiments in such evaluations.

DATA AVAILABILITY

The accompanying author can provide some of the models, data, or code created or utilized during the study upon request.

REFERENCES

[1] B. Wu, "Influence of MOOC learners discussion forum social interactions on online reviews of MOOC," *Educ. Inf. Technol.*, vol. 26, no. 3, pp. 3483–3496, May 2021.

[2] Y. Lin, S. Feng, F. Lin, J. Xiahou, and W. Zeng, "Multi-scale reinforced profile for personalized recommendation with deep neural networks in MOOCs," *Appl. Soft Comput.*, vol. 148, Nov. 2023, Art. no. 110905.

[3] H. T. Phan, N. T. Nguyen, and D. Hwang, "Convolutional attention neural network over graph structures for improving the performance of aspect-level sentiment analysis," *Inf. Sci.*, vol. 589, pp. 416–439, Apr. 2022.

[4] H. Zhang, Z. Chen, B. Chen, B. Hu, M. Li, C. Yang, and B. Jiang, "Complete quadruple extraction using a two-stage neural model for aspect-based sentiment analysis," *Neurocomputing*, vol. 492, pp. 452–463, Jul. 2022.

[5] C. C. Huang, L. M. Suolang, Z. Y. Lamu, and N. Qun, "Study and implementation of sentiment analysis on SVM-based Tibetan microblogging text," *Plateau Sci. Res.*, vol. 4, no. 1, pp. 92–96, 2020.

[6] Y. Su, Y. Zhang, B. Hu, and X. H. Tu, "Sentiment analysis research based on combination of naive Bayes and latent Dirichlet allocation," *J. Comput. Appl.*, vol. 36, no. 6, pp. 1613–1618, 2016.

[7] Z. H. Zhou and J. Feng, "Deep Forest: Towards an alternative to deep neural networks," *Nat. Sci. Rev.*, vol. 6, pp. 74–86, 2017.

[8] G. Attardi and D. Sartiano, "UniPI at SemEval-2016 task 4: Convolutional neural networks for sentiment classification," in *Proc. 10th Int. Workshop Semantic Eval.*, 2016, pp. 220–224.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to alignand translate," *Comput. Sci.*, vol. 24, no. 5, pp. 1345–1349, 2014.

[10] Z. L. Zhang, L. C. Li, X. Q. Zhu, and H. Y. Ma, "Aspect sentiment analysis combining ON-LSTM and self-attention mechanism," *J. Chin. Mini-Micro Comput. Syst.*, vol. 41, no. 9, pp. 1839–1844, 2020.

[11] R. K. Behera, M. Jena, S. K. Rath, and S. Misra, "Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data," *Inf. Process. Manage.*, vol. 58, no. 1, Jan. 2021, Art. no. 102435.

[12] M. Jiang, W. Zhang, M. Zhang, J. Wu, and T. Wen, "An LSTM-CNN attention approach for aspect-level sentiment classification," *J. Comput. Methods Sci. Eng.*, vol. 19, no. 4, pp. 859–868, Nov. 2019.

[13] B. Cui, Y. Li, M. Chen, and Z. Zhang, "Fine-tune BERT with sparse self-attention mechanism," in *Proc. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3548–3553.

[14] Y. Zhang, B. Xu, and T. Zhao, "Convolutional multi-head self-attention on memory for aspect sentiment classification," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 4, pp. 1038–1044, Jul. 2020.

[15] C. Gan, Q. Feng, and Z. Zhang, "Scalable multi-channel dilated CNN–BiLSTM model with attention mechanism for Chinese textual sentiment analysis," *Future Gener. Comput. Syst.*, vol. 118, pp. 297–309, May 2021.

[16] X. H. Wu, L. Chen, T. T. Wei, and T. T. Fan, "Sentiment analysis of Chinese short text based on self-attention and Bi-LSTM," *J. Chin. Infonnation Process.*, vol. 33, no. 6, pp. 100–107, 2019.

[17] S. C. Huang, D. H. Han, B. Y. Qiao, G. Wu, and G. R. Wang, "Implicit sentiment analysis method based on ERNIE2. 0-BiLSTM-attention," *J. Chin. Comput. Syst.*, vol. 42, no. 12, pp. 2485–2489, 2021.

[18] Y. Chen, T. Zhuang, and K. Guo, "Memory network with hierarchical multi-head attention for aspect-based sentiment analysis," *Int. J. Speech Technol.*, vol. 51, no. 7, pp. 4287–4304, Jul. 2021.

[19] L. Li, X. H. Wu, and J. Liu, "Sentiment analysis model of bi-LSTM with key opinion target recognition and deeper self-attention," *J. Chin. Mini-Micro Comput. Syst.*, vol. 42, no. 3, pp. 504–509, 2021.

[20] H. Chen, B. Y. Gao, L. N. Chen, and C. Yu, "Sentiment classification model combining attention mechanism and bidirectional slice GRU," *J. Chin. Mini-Micro Comput. Syst.*, vol. 41, no. 9, pp. 1793–1799, 2020.

[21] J. Yang, Y. Li, H. Zhang, J. Hu, and R. Bai, "Aspect-level sentiment analysis incorporating semantic and syntactic information," *J. Comput. Commun.*, vol. 12, no. 1, pp. 191–207, 2024.

[22] Z. He, "Text sentiment analysis based on multi-layer bi-directional LSTM with a trapezoidal structure," *Intell. Autom. Soft Comput.*, vol. 37, no. 1, pp. 639–654, 2023.

[23] L. Zhang, Y. Wu, Q. Chu, P. Li, G. Wang, W. Zhang, Y. Qiu, and Y. Li, "SA-model: Multi-feature fusion poetic sentiment analysis based on a hybrid word vector model," *Comput. Model. Eng. Sci.*, vol. 137, no. 1, pp. 631–645, 2023.

[24] H. Rahman, J. Tariq, M. A. Masood, A. F. Subahi, O. I. Khalaf, and Y. Alotaibi, "Multi-tier sentiment analysis of social media text using supervised machine learning," *Comput., Mater. Continua*, vol. 74, no. 3, pp. 5527–5543, 2023.

[25] A. Mohamed, Z. M. Zain, H. Shaiba, N. Alturki, G. Aldehim, S. Sakri, S. F. M. Yatin, and J. M. Zain, "LexDeep: Hybrid lexicon and deep learning sentiment analysis using Twitter for unemployment-related discussions during COVID-19," *Comput., Mater. Continua*, no. 4, pp. 1577–1601, 2023.

[26] G. X. Xu, L. Y. Liu, J. C. Wang, and Z. Chen, "Text sentiment analysis based on BERT and hypergraph with dual attention network," *Appl. Res. Comput.*, vol. 41, no. 3, pp. 786–793, 2024.

[27] Q. Liu, J. S. Zhu, L. L. Zhao, Y. C. Sha, S. D. Liu, and Y. M. Ji, "Emotional analysis approach based on dynamic word-sentence features and self attention," *J. Data Acquisition Process.*, vol. 39, no. 1, pp. 193–203, 2024.
[28] S. Hong, J. Kim, H.-G. Woo, Y.-C. Kim, and C. Lee, "Screening ideas in the early stages of technology development: A word2vec and convolutional neural network approach," *Technovation*, vol. 112, Apr. 2022, Art. no. 102407.
[29] L. Xiaoyan, R. C. Raga, and S. Xuemei, "GloVe-CNN-BiLSTM model for sentiment analysis on text reviews," *J. Sensors*, vol. 2022, pp. 1–12, Oct. 2022, doi: 10.1155/2022/7212366.
[30] S. Lamsiyah, A. E. Mahdaouy, S. E. A. Ouatik, and B. Espinasse, "Unsupervised extractive multi-document summarization method based on transfer learning from BERT multi-task fine-tuning," *J. Inf. Sci.*, vol. 49, no. 1, pp. 164–182, Feb. 2023.
[31] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," *Comput. Sci.*, vol. 1, Aug. 2020, Art. no. 11942v5.
[32] X. Tao, A. Shannon-Honson, P. Delaney, C. Dann, H. Xie, Y. Li, and S. O'Neill, "Towards an understanding of the engagement and emotional behaviour of MOOC students using sentiment and semantic features," *Comput. Educ., Artif. Intell.*, vol. 4, Jul. 2023, Art. no. 100116, doi: 10.1016/j.caeai.2022.100116.
[33] K. F. Hew, X. Hu, C. Qiao, and Y. Tang, "What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach," *Comput. Educ.*, vol. 145, Feb. 2020, Art. no. 103724.
[34] L. Li, J. Johnson, W. Aarhus, and D. Shah, "Key factors in MOOC pedagogy based on NLP sentiment analysis of learner reviews: What makes a hit," *Comput. Educ.*, vol. 176, Jan. 2022, Art. no. 104354.
[35] T. Wang, H. Shi, W. Liu, and X. Yan, "A joint FrameNet and element focusing sentence-BERT method of sentence similarity computation," *Expert Syst. Appl.*, vol. 200, Aug. 2022, Art. no. 117084.
[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[37] K. Ethayarajh, "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings," 2019, *arXiv:1909.00512*.
[38] Y. O. Zhao, J. Z. Zhang, Y. B. Li, and Y. K. wang, "Sentiment analysis based on hybrid model of ELMo and transformer," *J. Chin. Inf. Process.*, vol. 35, no. 3, pp. 115–124, 2021.
[39] E. Brugnera, L. Haney, C. Grimsley, M. Lu, S. Walk, A. Tosello-Trampont, I. G. Macara, H. D. Madhani, G. R. Fink, and K. S. Ravichandran, "Unconventional RAC-GEF activity is mediated through the Dock180-ELMO complex," *Nature cell Biol.*, vol. 4, no. 8, pp. 574–582, 2002.
[40] A. S. Kale, V. Pandya, F. Di Troia, and M. Stamp, "Malware classification with Word2 Vec, HMM2Vec, BERT, and ELMo," *J. Comput. Virol. Hacking Techn.*, vol. 19, no. 1, pp. 1–16, Apr. 2022.
[41] M. Affi and C. Latiri, "BE-BLC: BERT-ELMO-based deep neural network architecture for English named entity recognition task," *Proc. Comput. Sci.*, vol. 192, pp. 168–181, Jul. 2021.
[42] P. Li, J. Yu, J. Chen, and B. Guo, "HG-News: News headline generation based on a generative pre-training model," *IEEE Access*, vol. 9, pp. 110039–110046, 2021.
[43] Y. Zhang, T. Qian, and W. Tang, "Buildings-to-distribution-network integration considering power transformer loading capability and distribution network reconfiguration," *Energy*, vol. 244, Apr. 2022, Art. no. 123104.

LISHA YAO was born in Anhui, China, in 1986. She received the master's degree in applied computer technology from Anhui University, in 2011, and the Ph.D. degree in computer science from the National University of the Philippines, in 2020.

She has been a Teacher with Anhui Xinhua University, since 2011. She is an Associate Professor. She presided more than six scientific research projects, published more than 20 papers in domestic and foreign academic journals and international conferences, and obtained one national invention patent.