Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2023 SCPE. Volume 24, Issues 4, pp. 895–907, DOI 10.12694/scpe.v24i4.2454

SPEECH EMOTION ANALYSIS OF SHORT ENGLISH READINGS BASED ON THE CAM-SPAT MODEL

QINGYUAN LI*

Abstract. With the development of technology, voice sentiment analysis has also undergone rapid development, and its application fields are constantly expanding. Multimodal models have become a key focus of researchers due to their ability to better predict emotions. In order to help English learners improve their oral English proficiency, a deep learning based emotional analysis model for English short text reading is proposed, and this model is used to analyze emotions in English reading. Additionally, a cross-modal attention mechanism based on a prediction-assisted task was developed to identify emotions in English reading aloud in state and a two-layer attention-based bi-directional long- and short-term memory network was created to classify emotions in English reading aloud. The results of the research revealed that the classification model's mean F1 value was 98.54%, the detection model's mean F1 value was 85.13%, and the speech emotion analysis model's mean F1 value was 73.55, which was not significantly different from the mean of the professionals' ratings. The significance of the study lies in providing English learners with a method and pathway to improve their oral English proficiency.

Key words: read aloud speech; emotion; CAM-SPAT; deep learning; feature extraction

1. Introduction.

One of the most essential aspects of human existence is emotion, and the key to emotion research is gathering data on these traits and utilising models to assess them [1]. Unimodal and multimodal models of sentiment analysis are two broad categories. Single modal models have the disadvantage of being unable to dynamically, multi-dimensionally, and multi-dimensionally analyze human emotions. Therefore, multimodal models have gradually become the focus of emotional analysis, and most of its current research is mainly focused on finding a better multimodal fusion method [2]. With the development of technology, the application area of speech sentiment analysis is gradually expanding [3]. In addition, based on advances in technologies such as artificial neural networks and deep learning (DL), more and more researchers have started to combine sentiment analysis with DL [4]. Researchers such as S Dahmani used different neural architectures to synthesize emotional speech and studied the performance of neural networks in learning visual and auditory modal features under different emotions [5]. Experts such as F Jia have constructed an emotional lexicon for texts and proposed an emotional orientation analysis model to analyze the emotional characteristics of participants in online public opinion [6]. However, the combination of deep learning methods also has certain shortcomings, such as difficulties in personalised feature extraction, difficulties in cross-modal interaction and limitations in distribution modelling [7]. Based on these problems, the study innovatively proposes a DL-based model for sentiment analysis of English short-text read-aloud speech. The model consists of a two-layer Dual Attention-based Bidirectional Long Short-term Memory Networks (DABLSTM) model and a Cross-modal Attention Mechanism (AM) with Sentiment Prediction Auxiliary Task (CAM-SPAT) model. There are two innovative points in the research. The first point is the combination of dual attention mechanisms and bidirectional long-term and short-term memory networks. The second point is the combination of emotion prediction assistance tasks and cross modal attention mechanisms. The research aims to help English learners better understand the specific situation of their oral English proficiency and improve their oral English proficiency.The study is divided into four parts: the speech emotion (SE) analysis portion is the first, the SE feature extraction portion is the second, the SE analysis model portion is the third, and the study conclusion portion is the fourth.

2. Related Works.

Emotion is a necessary component of human existence, and as science and technology advance, so does study on voice sentiment analysis. Researchers like C Park. developed a prospective profile analysis employing exploratory cross-sections to examine the surface and deep performances of nurses in order

[*] School of Foreign Languages, Nanchang Institute of Technology, Nanchang, 330044, China (liqingyuan1001@163.com)

to study the management of nurses' emotional work. The study's findings demonstrated that nurses with highly regulated traits and surface acts were more likely to cause emotional weariness [8]. To study and mine the emotional tendencies of web writings, Haichao Sun et al. experts proposed a BPSO-based approach of integrated learning text emotion categorization in stochastic subspaces. By chunking the web texts with various granularities of sentiment, the approach categorises web texts at various levels of detail. The experimental findings demonstrated that the study's suggested algorithm had a greater level of classification accuracy [9]. Researchers I Gupta et al. incorporated an enhanced negation computation based on Twitter sentiment analysis and trained the classifier using various sorts of data attributes in order to analyse sentiment on social media. Study findings demonstrated that the support vector machine classifier performed better than other classifiers [10]. To advance the study of sentiment analysis on social concerns, academics like Y Mehmood developed an improved lexicon-based methodology. The approach combines the use of verbs and multilayer lexical dependencies with General Inquirer. The approach has good results and can be accurate up to 83%, according to experimental results [11]. Experts like J Barnes have suggested a multi-task strategy to produce precise forecasts about emotion. The approach uses a cascaded, hierarchical neural network with sentiment analysis and the addition of negative data. According to experimental findings, the proposed technique can greatly enhance sentiment analysis [12].

To address the issue of efficiently producing training data for sentiment analysis, R Ghasemi and colleagues suggested a cross-linguistic DL framework. They also used cross-linguistic embedding to change the sentiment analysis model into a migration learning model. The model can classify text using a wide range of deep architectures. The study's findings demonstrate the model's definite superiority, with a total improvement in model performance of 16% [13]. In order to cut down on the time and labour needed for data annotation, R Alahmary and other specialists suggested a semi-automatic method based on plain Bayesian, and the method was utilised to annotate new data sets. According to the experimental findings, the basic Bayesian classifier had an accuracy rate of up to 82.9%, which can significantly reduce the amount of time and labour needed for data annotation while simultaneously accelerating the process [14]. Experts like Zhang Hua have suggested a two-stage neural network approach to analyse sentiment in texts and spoken words. In order to extract categories and polarities of viewpoint terms in texts or utterances, the model contains of modules like BiLSTM and positional coding [15]. For the study of picture sentiment, Liang Yun and other researchers have suggested a chain-centre loss. This loss function regulates both local and global spatial distributions and was built on the foundation of central loss and triple loss. The study's findings demonstrated that this loss function enhances the chain-centre loss performance and enables the building of a prospective space of sentiment relevance [16]. GG Kim et al. The performance of the model before and after profanity data removal was simulated for the purpose to examine the influence of profanity data on sentiment categorization using comment data from the web as source data. The study examined whether the profanity data decreased the model's accuracy by comparing its performance before and after data elimination. According to the experimental findings, including profanity data as noisy data caused the model's accuracy to drop by 1.8% [17]. For the purpose to evaluate the recovery after a disaster, experts like D Contreras proposed a method for analysing sentiment analysis based on web data. The method used supervised classification and expert rules to define the polarity of the network data. The study's findings revealed that the method's total accuracy was 56.8% [18].

In conclusion, both domestically and internationally, there is a lot of study on SE analysis, and the algorithms and techniques used are varied. These research do, however, also have several drawbacks, including challenges with cross-modal interaction, limitations with distribution modelling, and challenges with extracting customised information. A novel sentiment analysis technique for reading short English texts out loud is developed based on these issues. The DABLSTM model and the CAM-SPAT model combine to create a model that can classify and identify the emotions expressed in spoken English in brief paragraphs.

3. DL-based SE Analysis Model Design.

A DL-based SE analysis model was designed for the analysis of emotion in short English speech read aloud. The model extracts feature information from speech through a DABLSTMl, and analyses multimodal emotion through a cross-modal attention (CMAM) model based on a prediction-assisted task.

3.1. Design of DL-based SE Feature Extraction Method.

Low-level features and deep-level features are the two primary divisions of read aloud speech characteristics [19]. Different features have different

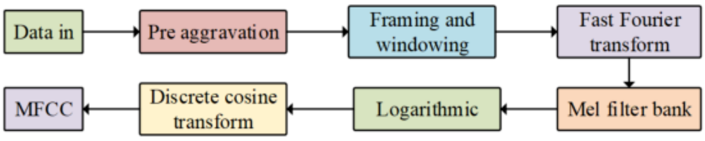


Fig. 3.1: MFCC' s extraction process

extraction methods. Before extracting the acoustic features of the low-level features, the low-level features will be pre-processed with the acoustic signal. There are three main pre-processing methods for acoustic signals, namely pre-emphasis, framing and windowing. Pre-emphasis is mainly used in the high frequency part of the speech signal, which is shown in equation 3.1.

$$y(t) = \chi(t) - \alpha\chi(t-1) \tag{3.1}$$

In equation 3.1,$\chi(t)$represents the signal before the pre-emphasis filter, $y(t)$ represents the resultant signal, t is the moment,α represents the pre-emphasis factor, and takes values in the range [0.9,1.0]. After the continuous acoustic signal has been processed by pre-emphasis, the pre-processing of the acoustic signal requires a framing operation. By dividing the continuous speech signal into several short-time speech segments, the framing operation makes it possible to obtain a relatively stable short-time signal, which can also be used for the extraction of specific information content, as shown in equation 3.2.

$$w(t) = 0.54 - 0.46cos\frac{2\pi}{N-1} \tag{3.2}$$

In equation 3.2, N is the length of the Hamming window, and n takes values in the range of [0, N-1]. For the extraction of acoustic features, the main purpose is to obtain the Mel Frequency Cepstrum Coefficient (MFCC) and the logarithmic Mel spectrogram, which is revealed in Fig 3.1 In Fig 3.1, the first step in extracting the MFCC is to pre-emphasise the speech signal, performing the operations of framing and adding windows. The second is to perform a fast Fourier change, and the third is a Meier filter bank. The fourth step is to take the logarithm and obtain a logarithmic Meier spectrogram. The fast Fourier change converts the time domain signal into a spectrum and the corresponding spectrum is generated for each frame as shown in equation 3.3.

$$X_v = \sum_{k=0}^{v-1} X_k e^{\frac{-2\Pi jkv}{v}} \tag{3.3}$$

In equation 3.3, V represents the number of frame segments, X represents the data value of k, and $e^{\frac{-2\Pi jkv}{v}}$ is the Fourier transform factor. The Meier spectrogram is a spectrogram transformed from frequency to Meier scale, and the transformation is shown in equation 3.4.

$$h = 2395log_{10}\left(1 + \frac{f}{700}\right) \tag{3.4}$$

In equation 3.4, f represents the actual linear frequency and h represents the Mel frequency. The logarithmic Meier spectrum can be calculated by taking the logarithm on the Meier spectrum, and the equation is shown in equation 3.5.

$$X_n^{`} = 10log_{10}\left(\frac{X_n}{ref}\right) \tag{3.5}$$

In equation 3.6,$X^{`}$ is x_n scaled with respect to ref . After the logarithm has been processed, the MFCC features can be obtained by the discrete cosine transform. The discrete cosine transform is shown in equation 3.6.

$$C_i(j) = \sum_{m=0}^{M-1} X_j^{`}(m)cos\left(\frac{\pi j(m-0.5)}{M}\right) \tag{3.6}$$

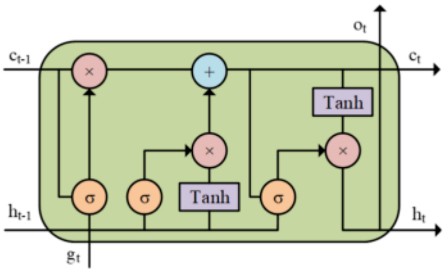


Fig. 3.2: The specific structure of LSTM

In equation 3.6, M is the number of MFCCs, j represents the number of variations from 1 to MFCC, and j takes values in the range [0,L], where L is the number of MFCC features, $C_i(j)$ represents the jth MFCC coefficient of the first frame, and $X_j^{'}(m)$ represents the logarithmic energy of the mth Meier filter. MFCC features can only represent the static characteristics of the speech signal. Equation 3.7 illustrates how to calculate the time derivative.

$$\Delta C_m(i) = \frac{\sum_{\tau=-1}^{M} \tau C_m(i+\tau)}{\sum_{\tau=-1}^{M} \tau^2} \tag{3.7}$$

In equation 3.7, $C_m(i)$ represents the static coefficients of frame i , the first-order (FO) difference feature is calculated based on the preceding and following M frames, the value of M is usually set to 2, and τ represents the time difference of the DABLSTM derivatives. The study employs a linear interpolation technique, which is computed as indicated in equation 3.8, to fill in the gaps in the data.

$$y = y_\circ \left(1 - \frac{x-x_\circ}{x_1-x_\circ}\right) + y_1 \left(1 - \frac{x-x}{x_1-x_\circ} \quad \right) \tag{3.8}$$

In equation 3.8, (x, y) represents the coordinates of the interpolation point. $(x_\circ, y_\circ)$ & (x_1, y_1) represent the coordinates of the fixed point. For the extraction of deep features of speech signals, the study mainly used Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) network and AM in DL [20, 21].

In Fig 3.23.2, the LSTM can provide some relief to the long-time dependency problem [22]. The LSTM has three steps: deciding what data should be deleted, deciding what new data should be saved, and deciding what output should be made of the result. The expression is shown in equation 3.9.

$$\delta_t = sigmoid(W_C \left[h_{t-1}, g_t\right] + b_c) \tag{3.9}$$

In equation 3.9, W_c is the weight,b_c represents the bias, h_{t-1} is the hidden state at moment $t-1$ and g_t represents the input at the current moment. The confirmation of the new message is calculated as shown in equation 3.10.

$$\begin{cases} \widetilde{c} = tanh(W_\delta \left[h_{t-1,} g_t\right] + b_\delta) \\ d_t = sigmoid \left[h_{t-i}, g_t\right] W_d + b_d \\ c_t = \delta_t * c_{t-1} + d_t * \widetilde{c} \end{cases} \tag{3.10}$$

In equation (10), d_t represents the control coefficient of the input gate, W_d and W_δ are the weights, b_d and b_δ represent the bias, $\widetilde{c_t}$ is the state information at the current moment, d_t and $\widetilde{c_t}$ are multiplied together to represent whether the current information should be retained, and c_t is the long-time hidden state. The output is calculated as shown in equation (11).

$$\begin{cases} o_t = sigmoid(W_\circ \left[h_{t-1}, g_t\right] + b_\circ) \\ \eta_t = o_t * tanh(c_t) \end{cases} \tag{3.11}$$

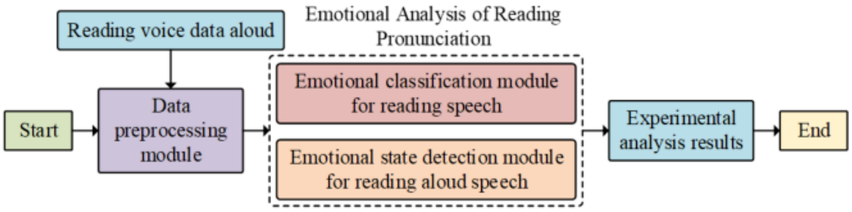

Reading voice data aloud
Emotional Analysis of Reading Pronunciation
Start
Data preprocessing module
Emotional classification module for reading speech
Emotional state detection module for reading aloud speech
Experimental analysis results
End

Fig. 3.3: Structure diagram of SE analysis model

In equation (11), o_t represents the control coefficient of the output gate, η_t is the output result under c_t condition, $W_\circ$ is the weight and $b_\circ$ is the bias. The Bidirectional LSTM (BLSTM) can process sequential data in both directions, and its output sequence is shown in equation 3.12.

$$\varphi_t = \mathbf{W}_{\rightarrow h}\varphi \rightarrow_h t\,(-\mathbf{W}_{\leftarrow h}\varphi \leftarrow_h t - \mathbf{b}_\varphi) \tag{3.12}$$

In equation 3.12, φ is the output sequence, $\rightarrow_t$ represents the LSTM hidden state of the forward processing input, and $\leftarrow_t$ represents the LSTM hidden state of the reverse processing input. The Gated Recurrent Unit (GRU) is a variant of the LSTM, which is shown in equation 3.13.

$$\begin{cases} h_t = z_t \Theta h_t + (1 - z_t)\Theta h_{t-1} \\ h_t = tanh(W_h g_t + U_h b_t) r_t \Theta h_{t-1} \\ z_t = \sigma(W_z g_t + U_z h_{t-1} + b_z) \\ t_t = \sigma(b_r + W_r g_t + U_r h_{t-1}) \end{cases} \tag{3.13}$$

In equation 3.13, U represents the weight matrix of h_{t-1} , the sigmoid activation function is $sigma$, r_t represents the vector of reset gates, z_t represents the vector of update gates, $tanh$ is the hyperbolic tangent function, the element multiplication is represented by Θ , and h_t is the candidate activation. The use of the AM requires the determination of the attention weights, which are calculated as shown in equation 3.14.

$$\partial_l = \frac{exp(f(g_l))}{\sum_u exp(f(g_u))} \tag{3.14}$$

In equation 3.14, $f(g)$ is the scoring function, g_l and g_u represent each input vector, and ∂ represents the attention weights. The output of attention is calculated as shown in equation 3.15.

$$attentive_g = \sum \partial_l g_l \tag{3.15}$$

3.2. DL-based Model Construction for SE Classification and State Detection. Traditional research on emotion recognition is mainly based on discrete emotion description models and dimensional emotion description models. In order to analyse the emotion in the speech of short English texts read aloud, the study designed a DL-based SE analysis model. The model not only allows for real-time monitoring of speech signals, but also allows for the analysis of the emotions of English learners when they read aloud short English texts. The structure is shown in Fig 3.3. The SE analysis model for reading aloud short English texts is broken down into three key parts, as seen in Fig 3.3. The pre-processing of the data comes first, followed by the development of the SE categorization and state detection modules, and finally the experiment. For data preprocessing, the conventional approach is to forcibly align words with different modalities, which can lead to insufficient interaction between modalities. Therefore, in terms of data preprocessing, the study adopted a method of not forcing

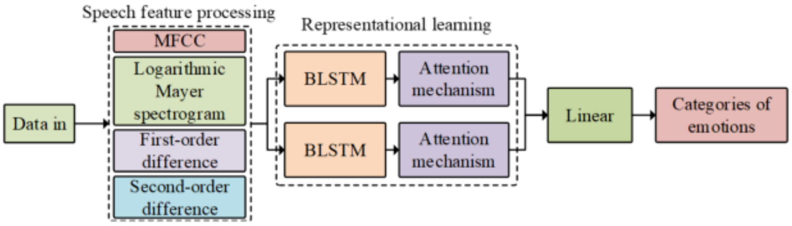


Fig. 3.4: The specific structure of the DABLSTM model

word alignment, while also using feature extraction tools to obtain feature vectors from different modalities. For text modality, research uses Transformer based Bidirectional Encoder Representation from Transformers (BERT) word embedding to represent modal features. The acoustic modality uses the COVAREP tool to represent modal features, while the visual model uses the Facet framework to represent modal features. For the construction of the SE classification module, a DABLSTM model was designed, and the specific structure of this model is shown in Fig 3.4.

The processing of linguistic feature data and representation learning are the two fundamental components of the DABLSTM model, as depicted in Figure 3.4. The linguistic feature processing includes MFCC features, log-Meier spectrograms. Representation learning mainly includes BLSTM and AMs. Traditional neural network models have insufficient emotional information acquisition in language emotion classification, while BLSTM can not only avoid the problem of vanishing gradient in recurrent neural networks, but also effectively capture future and past information in feature sequences. For log-Meier spectrograms, the study mainly uses linear interpolation for data pre-processing. The operation of linear interpolation is completed in the time series direction of the number Mel spectrum. When the audio duration is less than 10 seconds, a shape that is consistent with 10 seconds of audio can be obtained. When the audio duration is greater than 10 seconds, it is necessary to shorten the sample length by discarding some samples. For deep features, the study mainly uses CNN for extraction. In addition, the temporal information in MFCC and the local and full feature information in deep features can be extracted by DABLSTM model. For the construction of the SE state detection module, the study designed a CMAM with Sentiment Prediction Auxiliary Task (CAM-SPAT) model which is shown in Fig. 3.5.

In Fig 3.5, the multimodal feature processing module mainly involves BERT word embedding, COVAREP tool, and Facet framework. The representation learning module focuses on the encoding of sentiment feature vectors in different modalities using a Bidirectional Gated Recurrent Unit (BiGRU). BiGRU can not only delete invalid information from different modal feature data, but also efficiently obtain effective information from feature data. The learned representations are fed into the weighted CMAM module and the emotion prediction assistance task module. The weighted CMAM module is mainly concerned with the calculation of spatial correlations across multiple attention and time steps, and linear processing of the output data. The mood prediction auxiliary task module involves linear processing of the data across modalities and modification of the loss function, which is finally combined with the processing results of the weighted CMAM module as the output of the model.

4. Analysis of the DL-based SE Analysis Model results. The study conducted experiments on the DABLSTM model on the IEMOCAP dataset and validated it by metrics such as unweighted accuracy, weighted accuracy, loss function and confusion matrix. In addition, the study conducted experiments on the CAM-SPAT model on the CMU-MOSEI and CMU-MOSI datasets. Also it validated the performance of the model by metrics such as F1 value and 7-class classification accuracy.

4.1. Analysis of the Outcomes of the DL-based SE Classification Model. In the validation of the DABLSTM model, for the purpose to investigate the effects of the number of layers of parallel structure and input features on the performance of the model, the study was carried out in a single-layer Attention-based

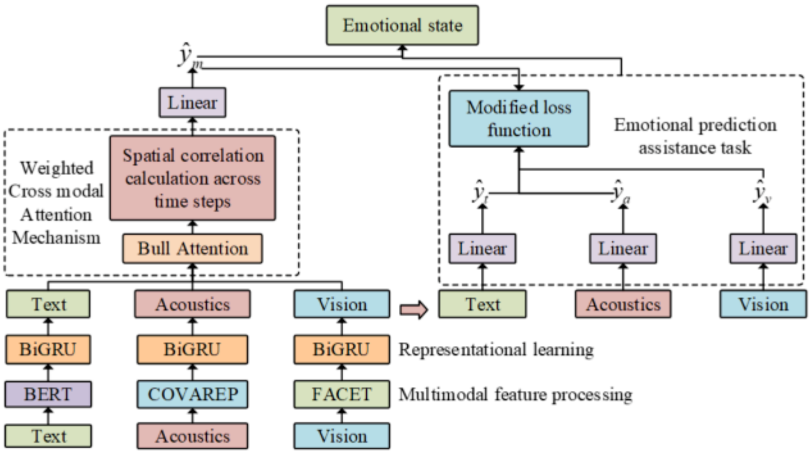


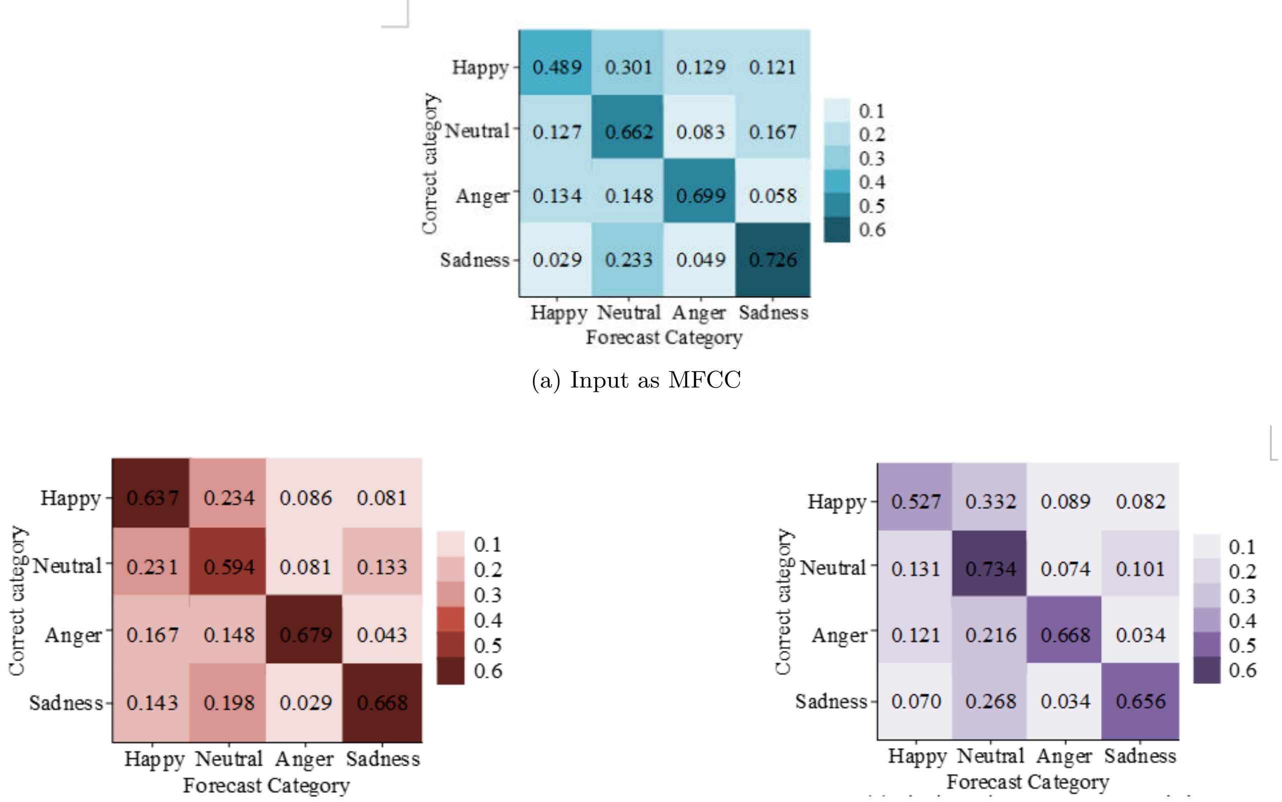
Fig. 3.5: The CAM-SPAT model's specific structure

Bidirectional Long Short-term Memory Networks model (ABLSTM). With various input characteristics and compared confusion matrices for various features, Fig ?? illustrates the comparison of the confusion matrices for various features.

It can be learned from Fig 4.1a that when the input feature is MFCC, the recognition accuracy of happy category is 48.9%, neutral type is 66.2%, angry category is 69.% and sad category is 72.6%. Meanwhile, the Unweighted Accuracy (UA) of ABLSTM was 62.7%. From Fig 4.1b, when the input features were MFCC with DABLSTM difference, the recognition accuracy was 63.7% for the happy category, 59.4% for the neutral type, 67.9% for the angry category and 66.8% for the sad category. the UA of ABLSTM was 63.7%. In Fig 4.1c, when the input features are MFCC with second-order differences, the recognition accuracy is 52.7% for the happy category, 73.4% for the neutral type, 66.8% for the angry category and 65.6% for the sad category. the UA of ABLSTM is 64.3%. It can be seen that the highest UA value was achieved when the input features were MFCC with second-order differences. A comparison of the model performance under different layer structures is shown in Fig 4.2.

In Fig 4.2a, when there are in layer 2, the recognition rates for the happy, neutral, furious, and sad categories are 68.4%, 68.6%, 73.1%, 77.7%, and 73.1% respectively. the UA of DALSTM is 71.2%. In Fig 4.2b that when there are in layer 3, the recognition accuracy of the happy category is 63.2%, the recognition accuracy of the neutral type is 73.3%, the recognition accuracy of the angry category is 71.0%, and the recognition accuracy of the sad category is 60.7%. The UA of the model at this point was 66.4%. From Fig 4.2c, the recognition accuracy of the joyful category is 62.1%, the neutral type is 66.7%, the angry category is 71.6%, and the recognition accuracy of the sad category is 76.8%. The UA of the model at this point was 68.6%. This shows that the DALSTM model with a two-layer structure has the highest UC and the best results. The study compared it with other models and the comparison is shown in Fig 4.3.

Fig 4.3a illustrates that the 3D-CRNNs model's WA on the training set is 62.98% and its UA is 61.93%. The WA and UA of the ABLSTM-FCNs + DNN model are respectively 60.7% and 61.1%. The WA of the ABLSTM-AFCN model is 69.% and the UA is 68%. the WA of the DALSTM model is 71.98% and UA was 71.29%. From Fig 4.3b, the WA of the 3D-CRNNs model on the validation set is 62.76% and the UA is 61.81%. the WA of the ABLSTM-FCNs + DNN model is 61.82% and the UA is 61.%. the WA of the ABLSTM-AFCN model is 69.8% and the UA is 68.6%. the WA of the DALSTM model is The DALSTM model has a WA of 72.72% and a UA of 72.21%. This shows that the DALSTM model outperforms the comparison model. The study compared and analysed the F1 values and loss functions of the different models, and the comparison results are shown in Fig 4.4. From Fig 4.4a, the maximum value (MaxV) of F1 for the 3D-CRNNs model



(a) Input as MFCC

(b) The input is MFCC accompanied by first-order difference

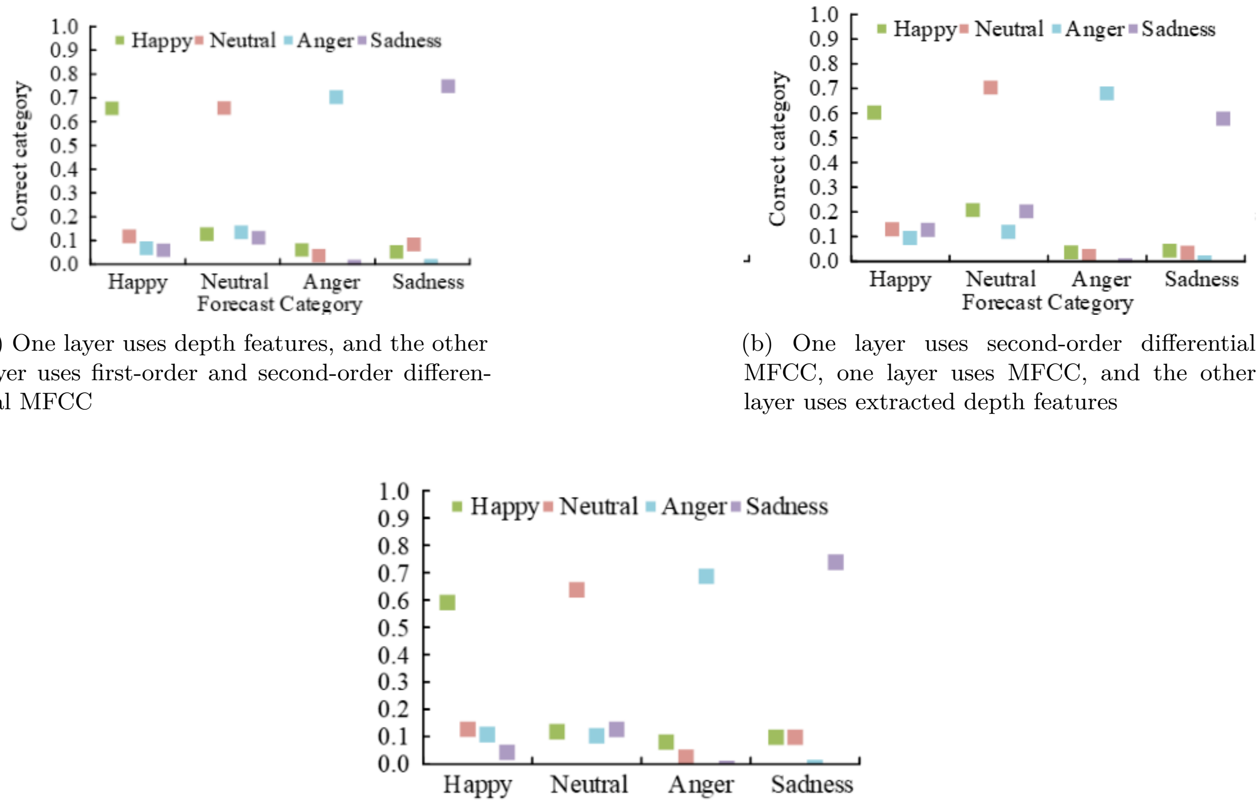
(c) The input is MFCC accompanied by second-order difference

Fig. 4.1: Comparison of confusion matrix corresponding to different features

is 97%, the minimum value (MinV) is 95.7% and the mean value (MeV) is 96.48%. the MaxV of F1 for the ABLSTM-FCNs + DNN model is 96.5%, the MinV is 95.2% and the MeV is 95.9%. the MaxV of F1 for the ABLSTM-AFCN model is 97.5%, the MinV is 96.6% and the MeV is 97.04%. the MaxV of F1 for the DALSTM model is 99.20%, the MinV is 97.7% and the MeV is 98.54%. The MaxV of F1 for the DALSTM model is 99.20%, the MinV is 97.% and the AV is 98.54%. The MaxV of the loss function of the 3D-CRNNs model is 1.87 and the MinV is 0.71.

4.2. Analysis of the Results of the DL-based SE Condition Detection Model. The CMU-MOSI dataset and the CMU-MOSEI dataset were utilised in the study's validation of the CAM-SPAT model in order to evaluate the various models and to observe the regression and classification outcomes both with and without the CAM-SPAT model. The Binary Classification Accuracy (Acc-2), 7-class Classification Accuracy (Acc-7) and F1-value metrics were used to more clearly compare the regression and classification outcomes with and without the CAM-SPAT model. A comparison of the regression and classification results with and without the CAM-SPAT model for the different datasets is shown in Fig4.5.

On the CMU-MOSI dataset, the MaxV of Acc-2 is 85.69%, the MinV is 51.87%, and the MeV is 75.87%, as shown in Fig 4.5a. The MaxV of Acc-7 is 47.94%, the MinV is 16.89% and the MeV is 39.03%. the MaxV of F1 value is 85.61%, the MinV is 44.20% The MaxV for F1 was 85.61%, the MinV was 44.20% and the AV was 74.21%. The MaxVs for all three metrics occurred when the CAM-SPAT model was used and the input features were text acoustic and visual, the MinVs for Acc-2 and F1 occurred when the CAM-SPAT model was not used and the input features were acoustic and visual, and the MinVs for Acc-7 occurred when the CAM-SPAT model

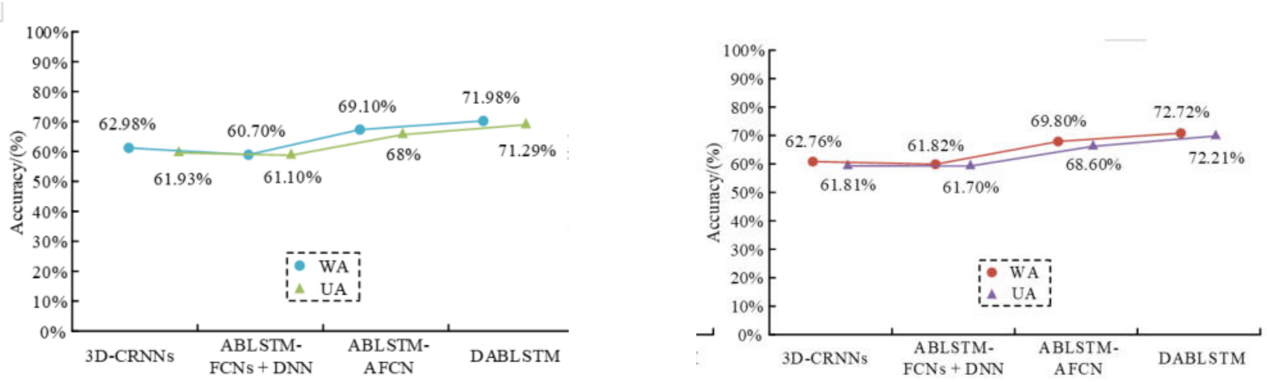


(a) One layer uses depth features, and the other layer uses first-order and second-order differential MFCC

(b) One layer uses second-order differential MFCC, one layer uses MFCC, and the other layer uses extracted depth features

(c) One layer uses second-order differential MFCC, one layer uses first-order differential MFCC, and the other layer uses MFCC

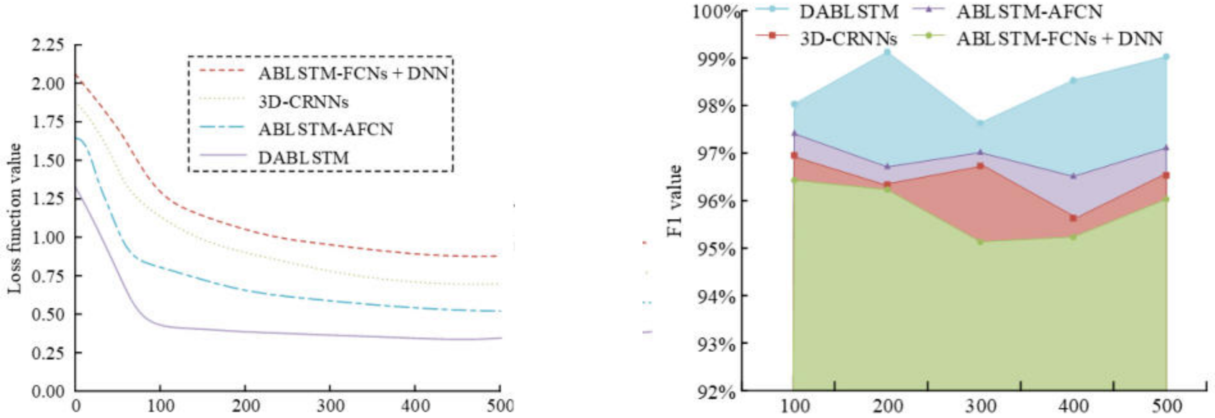
Fig. 4.2: Comparison of the performance of the lower model with different layers



(a) Comparison of the models on the training set

(b) Comparison of the models on the validation set

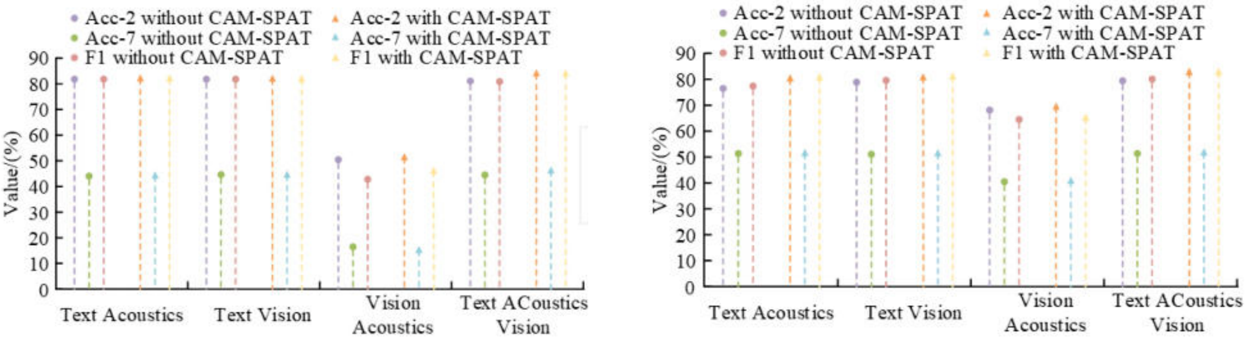
Fig. 4.3: Comparison of DALSTM model and other models on UA and WA



(a) Comparison of loss functions of different models

(b) Comparison of F1 values between different models

Fig. 4.4: Comparison of F1 values and loss functions of different models



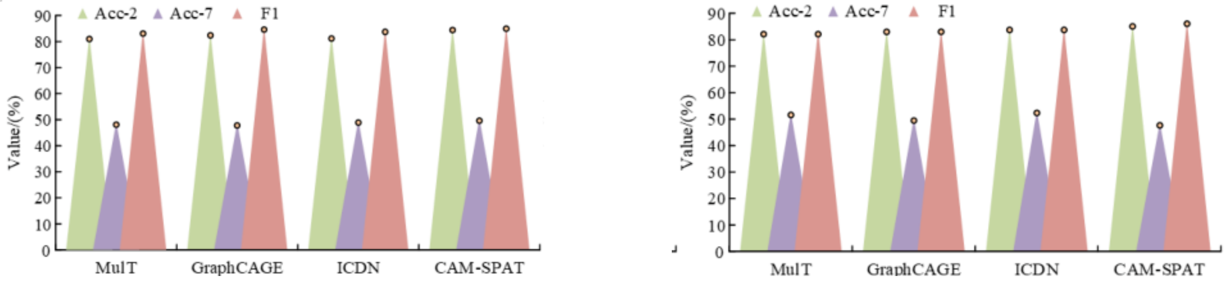
(a) CMU-MOSI Dataset

(b) CMU-MOSEI Dataset

Fig. 4.5: Comparison of regression and classification results using and not using CAM-SPAT models under different datasets

was used and the input features were acoustic and visual. From Fig 4.5b, on the CMU-MOSEI dataset, the MaxV of Acc-2 is 84.82%, the MinV is 69.9% and the MeV is 78.93%. the MaxV of Acc-7 is 53.82%, the MinV is 42.3% and the MeV is 50.64%. The MaxV of F1 value is 84.84%, the MinV is 66.29% and the MeV the MaxV for F1 was 84.84%, the MinV was 66.29% and the AV was 78.34%. The MaxVs for all three metrics occurred when the CAM-SPAT model was used and the input features were text acoustic and visual, while the MaxVs for all three metrics occurred when the CAM-SPAT model was not used and the input features were acoustic and visual. The study compares the CAM-SPAT model with other models. The models compared are Multimodal Transformer (MulT), Graph Capsule Aggregation (GraphCAGE) and Integrating Consistency and Difference Network (ICDN). The comparison results are shown in Fig 4.6.

From Fig 4.6 4.6a, the MaxV of Acc-2 is 85.46%, the MinV is 82.01% and the MeV is 83.28% on the CMU-MOSI dataset. the MaxV of Acc-7 is 50.66%, the MinV is 48.82% and the MeV is 49.63%. the MaxV of F1 value is 85.99%, the MinV is 84.12% and the The MaxV of F1 was 85.99%, the MinV was 84.12% and



(a) CMU-MOSI Dataset

(b) CMU-MOSEI Dataset

Fig. 4.6: Comparison results of different models under different datasets

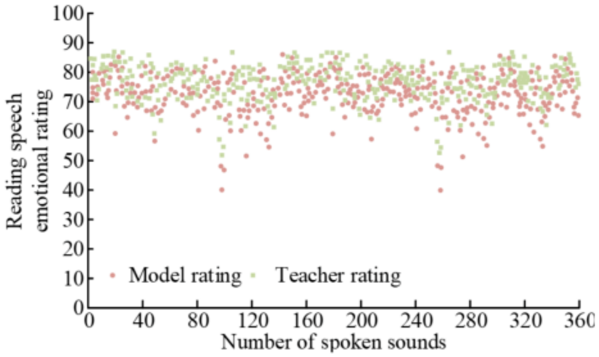


Fig. 4.7: Comparison of ratings between models and professionals under the same voice data

the average value (AV) was 85.13 The MaxVs of all three indicators were on the CAM-SPAT model. As can be seen through Fig 4.6b, on the CMU-MOSEI dataset, the MaxV of Acc-2 was 84.99%, the MinV was 82.08% and the MeV was 83.43%. the MaxV of Acc-7 was 52.66%, the MinV was 48.02% and the MeV was 50.61%. the MaxV of F1 value was 85.98%, the MinV was 82.75 with a MeV of 83.85%. The MaxVs for all indicators, except Acc-7, were on the CAM-SPAT model. This shows that the CAM-SPAT model has a higher accuracy rate. The study compared the scores of the model with the scores of the professionals for the same speech data, and the results of the comparison are shown in Fig 4.7.

In Fig 4.7, the majority of the data for both the model scores and the professional scores overlap with one another. The values for both the model ratings and the professional ratings are essentially in the range of 70 to 90. The MeV of the model scores was 73.55 and the MeV of the professional scores was 77.02. This shows that the model proposed by the study for analysing the sentiment of reading aloud short English texts has a high degree of reliability.

5. Conclusion. In order to enable English learners to improve their oral English proficiency, a deep learning based emotional analysis model for English short text reading pronunciation has been proposed, which includes the DALSTM model and the CAM-SPAT model. According to the study's findings, the DALSTM model had a training set WA and UA of 71.98% and 71.29%, respectively. On the validation set, the DALSTM model had a WA of 72.72% and UA of 72.21%. The DALSTM model's F1 values had a MaxV of 99.20%, a MinV of 97.7%, and a MeV of 98.54%. From this, it can be seen that the performance of the DALSTM model is superior to that of the comparative model. The DALSTM model fared better than the contrast model. The CAM-SPAT model produced the highest values for Acc-2, Acc-7, and F1 values on the CMU-MOSEI dataset, with respective values of 85.46%, 50.66%, and 85.99%. The greatest values of Acc-2 and F1 on the CMU-MOSEI dataset were all discovered on the CAM-SPAT model with 84.99% and 85.98%, respectively. This demonstrates the superiority of the CAM-SPAT model over the comparison model. Additionally, the sentiment analysis model's mean rating was 73.55, which was not significantly different from the mean rating of professionals. This demonstrates the great dependability of the sentiment analysis approach put forward in the study. The study's proposed sentiment analysis model has high reliability and superiority, but it also has some drawbacks, including the selection of a relatively small number of modalities and the omission of modal information like posture, which can be improved in future research.

6. Discussion. The research mainly designed an emotion analysis model for English short article reading speech, including an emotion classification model and an emotion state detection model. There are three main contributions to the research on the sentiment classification model DALSTM. The first one is to solve the problem of inconsistent audio signal length through linear interpolation. The second point is to verify that the double-layer structure improves the performance of the model by comparing the confusion matrix and model recognition accuracy. The third point is to verify the superiority of the performance of the DALSTM model. There are also three main contributions to the research on the emotional state detection model CAM-SPAT. The first one is to use a method of non mandatory word alignment and use different methods to represent features of different modalities. The second is to validate the effectiveness of the CAM-SPAT model through different classification indicators and comparison algorithms, and to verify the feasibility of the CAM-SPAT model by comparing professional ratings and model ratings. The third point is the combination of emotion prediction assistance tasks and cross modal attention mechanisms.

REFERENCES

[1] Mukhopadhyay, M., Pal, S., Nayyar, A., Pramanik, P., Dasgupta, N. & Choudhury, P. Facial Emotion Detection to Assess Learner's State of Mind in an Online Learning System. *5th International Conference On Intelligent Information Technology (ICIIT 2020)* . pp. 107-115 (2020)

[2] Sun, Y. & Lin, C. Design of Multidimensional Classifiers using Fuzzy Brain Emotional Learning Model and Particle Swarm Optimization Algorithm, Acta Polytechnica Hungarica, vol 18. *No.* **4** pp. 25-45 (2021)

[3] Hajhmida, M. & Oueslati, O. Predicting mobile application breakout using sentiment analysis of Facebook posts. *Journal Of Information Science.* **47**, 502-516 (2021)

[4] Aouani, H. & Ayed, Y. Speech Emotion Recognition with deep learning, Procedia Computer Science. (2020)

[5] Dahmani, S., Colotte, V., Girard, V. & Ouni, S. Learning emotions latent representation with CVAE for text-driven expressive audiovisual speech synthesis, Neural Networks, vol 141. *No.* **11** pp. 315-329 (2021)

[6] Jia, F. & Chen, C. Emotional characteristics and time series analysis of Internet public opinion participants based on emotional feature words, International Journal of Advanced Robotic Systems, vol 17. *No.* **1** pp. 105-108 (2020)

[7] Lu, X. Deep Learning Based Emotion Recognition and Visualization of Figural Representation. *Frontiers In Psychology.* **12** (2022)

[8] Park, C., Cho, H., Lee, D. & Jeon, H. Latent profile analysis on Korean nurses: Emotional labour strategies and well-being. *Journal Of Advanced Nursing.* **78**, 1632-1641 (2021)

[9] Sun, H., Wang, G. & Xia, S. Text tendency analysis based on multi-granularity emotional chunks and integrated learning, Neural computing & applications, vol 33. *No.* **14** pp. 8119-8129 (2021)

[10] Gupta, I. & Joshi, N. Feature-Based Twitter Sentiment Analysis With Improved Negation Handling, IEEE Transactions on Computational Social Systems, vol 8. *No.* **4** pp. 917-927 (2021)

[11] Mehmood, Y. & Balakrishnan, V. An enhanced lexicon-based approach for sentiment analysis: a case study on illegal immigration, Online Information Review, vol 44. *No.* **5** pp. 1097-1117 (2020)

[12] Barnes, J., Velldal, E. & Vrelid, L. Improving sentiment analysis with multi-task learning of negation, Natural Language Engineering, vol 27. *No.* **2** pp. 249-269 (2020)

[13] Ghasemi, R., Asli, S. & Momtazi, S. Deep Persian sentiment analysis: Cross-lingual training for low-resource languages. *Journal Of Information Science.* **48**, 449-462 (2020)

[14] Alahmary, R. & Al-Dossari, H. A semiautomatic annotation approach for sentiment analysis. *Journal Of Information Science.* **49**, 398-410 (2023)

[15] Hua, Z., Chen, Z., Bi, C., Biao, H., Mian, L., Cheng, Y. & Bo, J. Complete quadruple extraction using a two-stage neural model for aspect-based sentiment analysis, Neurocomputing, vol 492. *No.* **1** pp. 452-463 (2022)

[16] Yun, L., Maeda, K., Ogawa, T. & Haseyama, M. Chain centre loss: A psychology inspired loss function for image sentiment analysis, Neurocomputing, vol 495. *No.* **7** pp. 118-128 (2022)

[17] Kim, C., Hwang, Y., Kamyod, C. & Study, A. of Profanity Effect in Sentiment Analysis on Natural Language Processing Using ANN. *Journal Of Web Engineering.* **21**, 751-766 (2022)

[18] Contreras, D., Wilkinson, S., Balan, N. & James, P. Assessing post-disaster recovery using sentiment analysis: The case of L'Aquila, Earthquake Spectra, vol 38. *No.* **1** pp. 81-108 (2022)

[19] Kobayashi, M., Hamada, Y. & Akagi, M. Acoustic features correlated to perceived urgency in evacuation announcements, Speech Communication. (2022)

[20] Wang, X., Cheng, M., Eaton, J., Hsieh, C. & Felix, S. Fake node attacks on graph convolutional networks. *Journal Of Computational And Cognitive Engineering.* **1**, 165-173 (2022)

[21] Chen, Z. Research on internet security situation awareness prediction technology based on improved RBF neural network algorithm. *Journal Of Computational And Cognitive Engineering.* **1**, 103-108 (2022)

[22] Hiriyannaiah, S., Siddesh, G., Kiran, M. & Srinivasa, K. A comparative study and analysis of LSTM deep neural networks for heartbeats classification, Health and Technology, vol 11. *No.* **3** pp. 663-671 (2021)

Edited by: Rajni Mohana
Special issue on: Sentiment Analysis and Affective computing in Multimedia Data on Social Network
Received: Jul 27, 2023
Accepted: Sep 26, 2023