International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.2 (2025) pp. 2092-2104
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

A Transfer Learning-Based Text-Centric Model for Multimodal Sentiment Analysis

Shaowei YI[1], Suhaila Zainudin[2*]

[1]The National University of Malaysia, Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Selangor, Malaysia
Email: p131872@siswa.ukm.edu.my-**ORCID:** 0009-0003-9021-4511

[2] The National University of Malaysia, Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Selangor, Malaysia
* **Corresponding Author Email:** suhaila.zainudin@ukm.edu.my - **ORCID:** 0000-0003-2352-5312

Article Info:

DOI: 10.22399/ijcesen.1548
Received : 25 January 2025
Accepted : 05 April 2025

Keywords :

Multimodal Sentiment Analysis,
Transfer Learning,
Text-Centric Model,
Information Fusion.

Abstract:

Multimodal sentiment analysis (MMSA) is a research method that extracts effective information from heterogeneous modal information. Then, MMSA processes the multimodal data and performs sentiment analysis. Along with big data and machine learning development, multimodal sentiment analysis has become a hot research direction in multimodal learning and natural language processing. Although various feature extraction methods and information fusion methods have been continuously proposed, challenges exist in MMSA research. First, in terms of feature extraction, pre-trained models trained with many data sets can obtain higher quality features, but research on how to use these feature extraction methods to extract the best features is still needed. Currently, the more popular feature fusion methods do not focus on the interaction between multiple modal information and the retention of basic information. To overcome these problems a multimodal sentiment analysis model utilizes text features as core modal features, using video and audio modal features as auxiliary modal features, multimodal feature modality attention mechanism to extract the intrinsic connection between different modalities. The attention mechanism uses the features of video modality and audio modality as the focus and then enhances the text modality with the fusion of video modality and modality. To improve the quality of extracted features, this method chooses the transfer learning training method and uses the pre-trained model for processing. This research uses the CMU-MOSI dataset to test the proposed method. Experimental results show that the performance of the proposed model in emotion score prediction and emotion classification tasks exceeds traditional methods and baseline methods.

1. Introduction

Sentiment analysis research initially focused on unimodal text sentiment analysis, identifying sentiment tendencies by mining the text's vocabulary, syntax, and context features [1]. Traditional methods mainly rely on machine learning techniques, such as support vector machines (SVM) and naive Bayes classifiers, which combine sentiment dictionaries or features such as TF-IDF for classification [1]. With the introduction of deep learning, models such as recurrent neural networks (RNN) and convolutional neural networks (CNN) have demonstrated strong performance in sentiment analysis tasks, especially in capturing contextual semantics and long-term dependencies [2]. In recent years, many large computing companies have greatly improved the accuracy of text sentiment analysis by using pre-trained language models and increasing the corpus [3].

For example, a text may have sarcasm or ambiguity, but voice intonation or facial expressions can compensate for this deficiency. Therefore, Multimodal sentiment analysis has emerged, improving the robustness and accuracy of emotion recognition by integrating multiple modalities (such as text, voice, and images) [4]. Voice data contains rich emotional clues, such as intonation, volume, and rhythm; images, especially facial expressions, can intuitively reflect emotional

changes. Combining multimodal data can improve the model's emotional understanding ability and get closer to the human judgment mechanism of emotions under multi-sensory input.

At present, the research on multimodal sentiment analysis mainly focuses on two directions: 1) how to effectively integrate data from different modalities [4,5], and 2) how to deal with the inconsistency and missingness of multimodal data [6-28]. Early studies mostly used feature-level fusion methods, that is, directly connecting or weighting the features of text, speech, and images [6]. However, it is difficult to capture the interactive relationship between modalities fully. In recent years, deep learning-based attention mechanisms and graph neural networks have been widely used to model information interaction between modalities, significantly improving the performance of multimodal sentiment analysis.

Recently, deep learning methods have brought new ideas to information fusion. Attention mechanisms and graph neural networks (GNNs) based on deep learning have been widely used in modeling inter-modal interactions. For example, the attention mechanism can dynamically adjust the contribution weights of different modalities to sentiment analysis, thereby improving the fusion effect [7]; GNN captures the high-order interaction relationship of multimodal data by constructing an inter-modal relationship graph, effectively solving the problem of insufficient modeling of modal correlation in traditional methods [8]. In addition, methods such as cross-modal alignment and co-learning representation have also become research hotspots. Cross-modal alignment technology aims to map data of different modalities to a shared semantic space to reduce the heterogeneity problem between modalities [9]; co-learning representation learning uses a joint optimization strategy to complement each modality's information and improve the overall performance of sentiment analysis.

Early text sentiment analysis methods mostly combined sentiment lexicons with traditional machine learning algorithms. In recent years, language models based on deep learning (such as BERT) have greatly improved the accuracy of text sentiment analysis [10]. The advantage of the text modality is the strong ability to parse explicit expressions of emotions. Its accuracy in emotion recognition is usually higher than other modalities, especially when processing structured data.

Visual modality sentiment analysis focuses on emotional signals in images or videos, especially non-verbal features such as facial expressions and body language. Models based on convolutional neural networks (CNNs) perform well in facial expression recognition tasks and can automatically extract local features related to emotions in images [11]. Visual modality has a unique advantage in providing intuitive emotional clues, but its accuracy may be limited by factors such as lighting [29], occlusion [30], individual differences [31], and cultural background [32]. Although unimodal sentiment analysis has made great progress, it isn't easy to fully reflect real emotions by relying solely on information from a single modality [33]. There are still some challenges in recognizing implicit emotions, such as sarcasm and puns [34].

Auditory modality sentiment analysis mainly analyzes speech signals and captures emotional information through intonation, speech rate, pitch, and energy [12]. The auditory modality can capture dynamic and subtle emotional changes and is suitable for identifying complex emotional states. However, its accuracy is easily affected by background noise, recording quality, and language differences, resulting in its lack of robustness in practical applications.

In comparisons of single-modality sentiment analysis, text modality often performs the best, especially when deep learning techniques are used. It shows strong sentiment recognition capabilities [13]. However, in practical applications, there is no absolute advantage or disadvantage between the modalities, and the applicability of different modalities depends on specific task requirements and data environments. Various modal data also provide important research impetus for developing multimodal sentiment analysis.

Based on previous research, this paper proposes a multimodal emotion analysis model with text modality as the central modality and visual and auditory modalities as auxiliary modalities because text modality as the basic modality can provide more detailed basic emotion information, and auditory modality and visual modality are used as auxiliary information. This model simulates how humans perceive video emotions: first, understand the central content by understanding the text information and then judge the text by observing multimodal information such as actions, sounds, and expressions, such as whether it contains sarcasm or hidden emotions. The information fusion method belongs to feature-level fusion but is not a simple model splicing. Instead, the feature vectors of the three modalities are unified in dimension, and then the visual and auditory features are used to assist in optimizing the text features. In terms of optimization, the attention mechanism should be used to explore the potential relationship between the visual modality, the auditory modality, and the text features. For the dataset, the study selected the CMU-MOSI dataset,

which is a public multimodal dataset video file commonly used in sentiment analysis experiments. It includes many video clips collected on YOUTUBE and annotated with sentiment labels.

A multimodal sentiment analysis model was introduced, focusing primarily on the text modality while incorporating an improved attention mechanism to integrate visual and auditory features into textual features. For data extraction, transfer learning was employed to obtain more accurate modality features. Compared to simple baseline methods, this model demonstrates a more effective and reasonable improvement in the accuracy of multimodal sentiment analysis on the CMU-MOSI dataset.

In the following part of this paper, the second part reviews current research on multimodal sentiment analysis and transfer learning, the third part presents a detailed experimental model framework, and the fourth part presents the experimental results and data analysis. The fifth part summarizes the results of this paper and provides prospects.

2. Literature Review

2.1 Multimodal Sentiment Analysis

Multimodal sentiment analysis combines information from multiple modalities (such as speech, text, and visual signals) to achieve more comprehensive and accurate emotion recognition in complex situations. Single-modal sentiment analysis may perform well in specific scenarios. Still, its performance is often disappointing when the emotional expression is complex, the modal information is insufficient, or the noise interference is serious. Therefore, multimodal sentiment analysis has gradually become a research hotspot in sentiment analysis because the data it uses can complement each other's missing information and fully use the data's effectiveness [5].

Early multimodal sentiment analysis research usually adopts feature- or decision-level fusion to model information from different modalities at different times jointly. For example, in extracting audio features of speech, word vector features of text, and expression features of vision, feature splicing is directly performed, and classification is performed using models such as support vector machines or random forests. However, this method faces the problems of heterogeneous feature expressions and time series alignment between modalities, and its effect is limited. Alternatively, a single-modal analysis is performed first to obtain the classification result directly. Then, the prediction results of each modality are integrated using a weighted average, voting mechanism, or maximum strategy. Finally, the sentiment analysis result is output. These methods are relatively easy to implement but require manual setting of weights or rules, and their performance is very limited in complex emotion expression scenarios [4,5].

Deep learning methods have significantly improved the performance of multimodal sentiment analysis. Models based on deep neural networks can achieve automatic learning and effective fusion of modal features. For example, RNN and LSTM are widely used to process time series modalities (such as speech and video), while CNN can extract local sentiment features in image modalities [14]. In addition, the introduction of the attention mechanism enables the model to dynamically model the model according to the context of emotional expression, further enhancing the feature extraction effect [15].

The rise of deep learning has significantly changed the way feature extraction is extracted, from manually designed features to data-driven automated feature learning. This method is based on the ability of large models to train neural networks to extract emotion-related features from raw data, which not only improves the expressiveness of features but also effectively solves the problem of different data being difficult to model uniformly.

Although multimodal sentiment analysis has made significant progress, it still faces challenges, such as excessive modal information categories (such as the lack of some modalities), differences in cross-cultural emotional expressions, and the interpretability of emotion recognition systems and real-time issues. Future research directions include stronger modeling of modality missing robustness, more efficient cross-modal alignment methods, and domain-specific emotion dynamics modeling.

2.2 Feature Extraction and Transfer Learning

Feature extraction is the core of sentiment analysis. It extracts effective features from multimodal data to provide usable and effective information for sentiment recognition models. Good feature extraction methods can reduce noise and interference in data [16]. With the development of technology, feature extraction has transformed traditional manual feature extraction into automatic feature extraction driven by deep learning.

The rise of deep learning has significantly changed the way feature extraction is extracted, from manually designed features to data-driven automated feature learning. This method uses deep neural networks to automatically extract emotion-related features from raw data, improving the expressiveness of the features and effectively

solving the problem of unified modeling of heterogeneous data.

Convolutional neural networks have a strong performance in image processing, so they are often used as the core technology for visual modality feature extraction. CNN can extract local features through multi-layer convolution operations and gradually form high-level semantic representations. For example, models such as AlexNet and VGGNet are used for facial expression emotion analysis, automatically extracting the edge, texture, and semantic information of the expression area and converting it into a visual feature vector [17]. In addition, researchers have further improved the CNN structure to adapt to emotion recognition tasks, such as extracting facial key points and emotion feature points through multi-task joint learning [18].

Recurrent neural networks and related models (LSTM, GRU) are widely used for feature extraction in language and visual modalities because they can construct time-dependent sequence data, as the context of language and human expressions are time-dependent. RNNs can capture the dynamic characteristics of data that change over time and are an important tool for audio sentiment analysis. For example, by inputting audio waveforms or mel-spectrograms, LSTM models can learn the temporal dependencies of emotional signals [19]. In the video modality, researchers applied RNNs to model continuous expression changes, significantly improving the accuracy of time series emotion recognition [20].

The Transformer architecture and its variants (BERT, ViT) have demonstrated powerful feature extraction capabilities in sentiment analysis tasks. Due to its excellent encoding method and model design, the Transformer can capture long-term correlations between different modalities. For example, the BERT model excels in processing text information. It learns general language representations through pre-training and then adjusts them to achieve efficient and general sentiment classification [3]. The ViT model in the visual modality breaks away from the limitations of traditional CNNs by directly processing image patches, showing great potential in feature extraction [21].

Although deep learning-driven automatic feature extraction has achieved great success in sentiment analysis, it still faces some challenges: the features extracted by deep learning are generally high-dimensional and lack intuitive interpretability; there are also differences between modalities, and data alignment and weight distribution are also issues that need to be addressed when facing feature fusion. Future research directions mainly include more accurate and effective small-sample learning methods, enhancing the interpretability of extracted features, and finding a suitable multimodal general framework [35,36].

Transfer learning is one of the important research directions in machine learning in recent years. It effectively alleviates the bottleneck problem of insufficient training data by transferring knowledge learned in one field (source domain) to another related field (target domain) with limited data. Transfer learning is widely used in sentiment analysis, especially multimodal sentiment analysis. Transfer learning provides a new solution for challenges such as modality differences, task generalization, and data scarcity.

Multimodal sentiment analysis needs to deal with heterogeneity and alignment issues between modalities. Transfer learning has made breakthroughs in the following aspects: Using rich data from one modality to transfer knowledge to another modality. For example, large-scale text sentiment data can guide sentiment analysis in audio or video modalities [22].

By pre-training multimodal models (such as CLIP and FLAVA), learning joint modality representations, and then transferring to sentiment analysis tasks, the collaborative modeling capabilities of different modalities are improved [23]. Given the differences in domain distribution in multimodal data (such as emotional expressions from various cultural backgrounds), domain adaptation methods are used to reduce the differences in modal feature distribution between the source and target domains [24].

3. Material and Methods

This paper proposes a multimodal sentiment analysis model, then experiments and verification on the multimodal sentiment dataset CMU-MOSI. The methodology will describe the model framework, data collection and preprocessing steps, and the final performance evaluation.

3.1 Overall Experimental Framework Design

Figure 1 is the text-centric multimodal sentiment analysis model proposed in this study. The general framework shows how to extract modal features from a multimodal video file, then align the features based on the time step, fuse the information of different modal data, and finally perform a weighted fusion of the two enhanced feature vectors to predict the sentiment score. The cross-modal text enhancement module adopts a multimodal attention model, which mainly includes

a multi-head attention mechanism, a weight fusion module, and a regression prediction layer. Improved multi-head attention mechanism: The text modality is enhanced by audio and video modality, respectively, to generate audio-enhanced text features and video-enhanced text features.

The learnable weight parameters are introduced to fuse the two enhanced text features to generate the final multimodal representation. The multimodal representation is mapped to a sentiment intensity score (-3 to +3) through a fully connected layer.

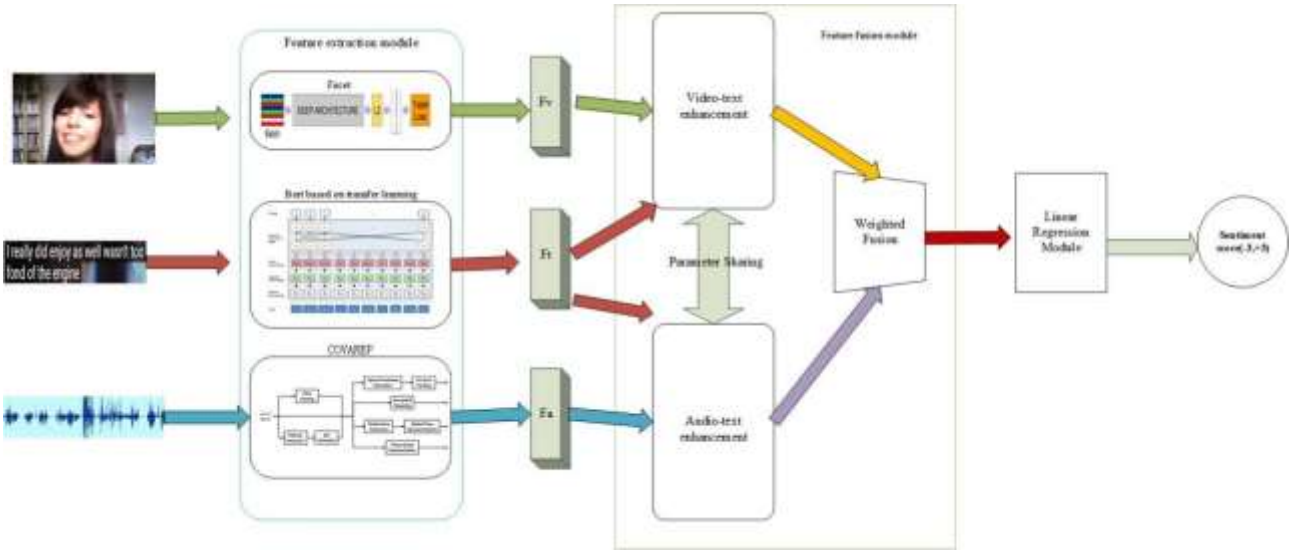

Figure 1. *Experimental Model Framework*

This study chooses text as the central modality because text, as the main carrier of emotional expression, contains directly interpretable emotional information and relatively stable semantics. Text modality information contains vocabulary, grammatical structure, and contextual information of emotional expression, directly reflects the emotional state, and is less affected by noise. Therefore, it is often regarded as the core modality of emotional analysis tasks. Some researchers have shown through experiments that subjectivity, emotional vocabulary, and contextual clues in text modality play a key role in emotional classification tasks [4]. The CMU-MOSI dataset used in this paper is a multimodal sentiment dataset. Its original data is short video clips from YouTube, containing three modal information (text, video, and audio). Among the three features, text features contain stable emotional information, which can help us quickly understand the emotional content expressed by the data and obtain key information.

In addition, the model uses a multi-head attention mechanism as an inter-modal fusion tool to increase the interaction between modalities and the ability to fuse different features. The multi-head attention mechanism can effectively learn the interactive information between different modalities and dynamically focus on key features, thereby improving the model's robustness to noise and time delay. Some studies [7] have shown that the attention mechanism can focus on the correlation between modalities, making the experimental results more accurate. In the experiment, the attention mechanism can help the model focus on the changing features of key time steps in different modalities, such as sentiment words reflected in text, fundamental frequency changes in audio, or environmental changes in video, thereby capturing the sentiment correlation between these features. In this way, the multi-head attention mechanism effectively improves the prediction accuracy of multimodal sentiment analysis tasks.

In general, choosing text as the central feature can ensure that the model obtains stable and clear emotional information from text features, and using a multi-attention mechanism can effectively solve the noise and delay problems in multimodal emotional tasks and fully explore the modal interaction between text, audio, and video to achieve the effect of improving model accuracy. This design method is consistent with the research findings of the field literature and in line with the characteristics of the data used.

3.2 Dataset Introduction and Data Preprocessing by Transfer Learning.

This study selected the CMU-MOSI dataset, a standard public basic dataset in multimodal sentiment analysis. The CMU-MOSI dataset collects 93 YouTube videos, which contain 2199 video clips, each of which includes the following information:

Sentiment label: A continuous value in the range of [-3, +3], reflecting the intensity of sentiment.

Three modal messages: text modality, audio modality, video modality.

The length of the modal data for each video clip is inconsistent, and the specific number of time steps depends on the length of the video clip.

To conduct experiments efficiently and obtain better experimental preprocessing effects, this study loaded and preprocessed data based on CMU-Multimodal SDK [25].

CMU-Multimodal SDK is an open-source toolkit for multimodal learning developed by Carnegie Mellon University, specifically designed for processing multimodal data (text, audio, and video). It provides efficient tools and methods to help researchers quickly load, process, and analyze multimodal data. It is particularly suitable for multimodal sentiment analysis, behavior recognition, and emotion detection tasks.

CMU-Multimodal SDK integrates multiple well-known multimodal datasets (CMU-MOSI, CMU-MOSEI, IEMOCAP), simplifying the data preprocessing process. CMU-Multimodal SDK has an efficient feature alignment tool that supports time series alignment. It can synchronize audio, video, text, and other modal data to the same timeline, automatically handle asynchronous problems between modalities, and ensure data consistency during information fusion.

In the feature extraction phase, a text feature extraction method based on transfer learning was proposed to extract sentiment features from the CMU-MOSI dataset for sentiment regression tasks. It also adopted the uptown/bert-base-multilingual-uncased-sentiment model [37,38], leveraging its strong capabilities in sentiment analysis tasks and applying it to new tasks to obtain efficient feature representations of text.

Transfer learning is a technique that leverages learned knowledge and applies it to different but related tasks. This study uses a fine-tuned pre-trained language model trained on multilingual sentiment analysis tasks and can predict sentiment scores based on text content. The model is based on the BERT architecture and has strong language understanding capabilities, which can capture complex semantic and sentiment information in text.

In the text feature extraction phase, performed the following steps:

Data preprocessing: First, the text data in the CMU-MOSI dataset was cleaned and preprocessed to ensure that the text fits the input format of the model.

Text input and feature extraction: The processed text is input into the pre-trained model. The model generates contextual representations of text features based on its sentiment analysis capabilities, especially extracting the hidden layer representation [CLS] token as the sentiment feature of the text. This feature vector encodes the context and semantics of the input text.

Advantages of transfer learning: Using the pre-trained knowledge of the model, feature vectors with high sentiment expression capabilities can be obtained without additional fine-tuning. This enables us to quickly and efficiently provide reliable features for subsequent sentiment score regression tasks.

Use the Facet tool for feature extraction in the video feature extraction process. It extracts AU1: inner eyebrow lift, AU6: cheek lift, AU12: mouth corner pull (smile), Facial posture: describes the rotation and tilt angle of the head, including pitch, yaw, and roll, and then converts it into a related feature vector through its encoding method.

Use the COVAREP tool to extract audio features. These features mainly include pitch, which indicates the high and low changes in the sound. Energy: Indicates the strength or loudness of the audio signal. Normalized Amplitude Quotient (NAQ) describes the characteristics of the glottal source. Mel-frequency cepstral coefficients (MFCCs): Used to capture the acoustic characteristics of audio. Peak Slope: Describes the changes in the peak in the spectrum. Energy Slope: Indicates the trend of energy changes over time.

If data from different modalities cannot be combined simultaneously, the resulting fusion may have a counterproductive effect. For example, a certain video frame shows a smiling face, but the corresponding text modality is not in this time step, and the text modality shows anger. This may cause the fusion of the model to have a counterproductive effect. Since the time steps of different modalities are inconsistent, to ensure that multimodal features can be effectively aligned and fused on the time axis, this experiment uses the text modality as the reference modality and uses the alignment function provided by CMU-Multimodal SDK to align the audio and video modal features in time, and also aligns the labels to the text modality. After the alignment is completed, the three modal features of each video clip are guaranteed to be consistent in the time step dimension.

To improve the model's generalization ability, standardize the feature-length sequences of all modalities to a fixed N, which can discard data beyond the feature sequence length or fill in the gaps within the feature-length to ensure that input is of the same length.

The data is divided into 10 subsets; one is used as the test set, eight is used as the training set, and the remaining is used as the validation set. All modal features and labels are converted into PyTorch

tensors to facilitate subsequent model training and verification.

3.3 Model and Algorithm Design

This study proposes a text-centric sentiment analysis model based on a multimodal attention mechanism, which aims to use text as stable and interpretable basic information, audio and visual modal features as auxiliary information, and improve the accuracy of sentiment intensity prediction through multimodal feature fusion. The model uses a multi-head attention mechanism to enhance text features separately and then generates the final multimodal feature representation through weighted fusion for regression prediction of sentiment intensity. The design drawings of the two modules are similar; Figure 2 shows the structure diagram of the audio enhancement module. The structure diagram of the video enhancement module is like that of the audio module.

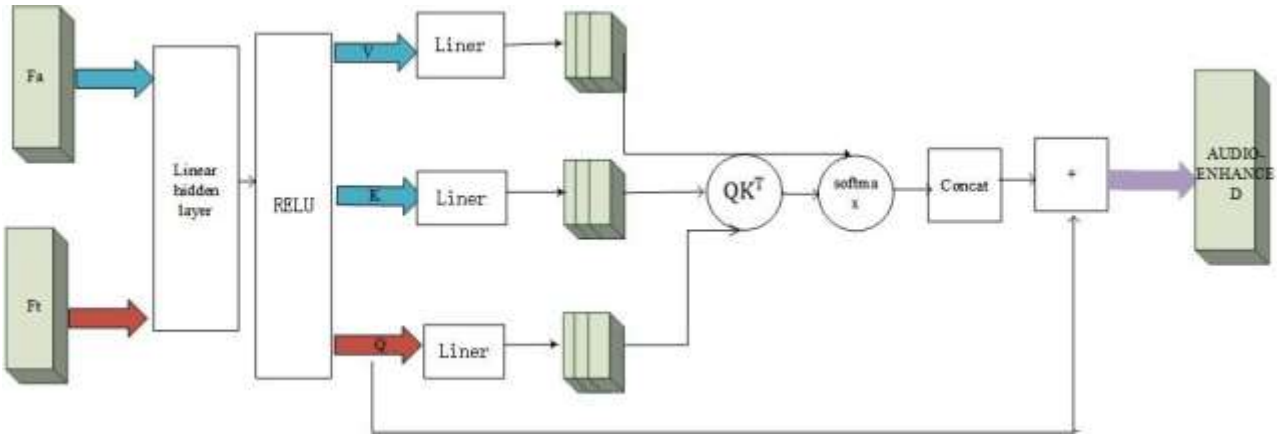

Figure 2. *Audio enhancement module structure diagram*

The input of the model is three modal features:

- The text modality dimension is $d_{\_text} = 768$.
- The audio modality dimension is $d_{\_audio} = 74$.
- The video modality dimension is $d_{\_video} = 47$.

These three features are mapped to a unified hidden layer dimension through a linear transformation module $d_{hidden} = 128$. After this linear mapping, the information of the three modes can be learned and calculated in the same feature space. It is mapped through a learnable weight matrix, and each batch of experiments will be adjusted according to the loss function. After this operation, the eigenvector of each mode is transformed into ($N*d_{hidden}$). It is convenient for us to carry out subsequent attention mechanism calculations.

The model designs two modality-enhanced text modules based on the multi-head attention mechanism, which are explained as follows:

The multi-head attention mechanism requires three inputs: Q (query), K (key), and V (value). The three input features are linearly transformed through a learnable weight matrix in equation 1:

$$MultiheadAttention = Concat(head_1, head_2, \ldots, head)W^O \quad (1)$$

The calculation formula for each$head_i = attention(QW_i^Q, KW_i^K, VW_i^V)$,W^O is the learnable weight matrix of each head in equation 2.

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

The characteristic dimension of the matrix QK^T is (N*d_{hidden}) *(d_{hidden}*N) = (N*N), the size of the entire output matrix is （N*N） (N*d_{hidden}) = (N*d_{hidden}), It has the same size as input Q.

- Audio Enhanced Text Module: Use text features (F_{text}) as Query, audio features (F_{audio}) as Key and Value and generate audio-enhanced text features through the attention mechanism in equation 3:
- $F_{audio-enhance} = MultiheadAttention(Q = F_{text}, K = F_{audio}, V = F_{audio}) + F_{text}$ (3)
- Video Enhanced Text Module: Use text features (F_{text}) as Query and video features (F_{video}) as Key and Value. Generate video-enhanced text features through attention mechanism in equation 4 :

$$F_{video-enhance} = MultiheadAttention(Q = F_{text}, K = F_{video}, V = F_{video}) + F_{text} \quad (4)$$

The next module fuses the two enhanced feature vectors obtained and outputs the sentiment prediction score. The structure is shown in the figure 3.

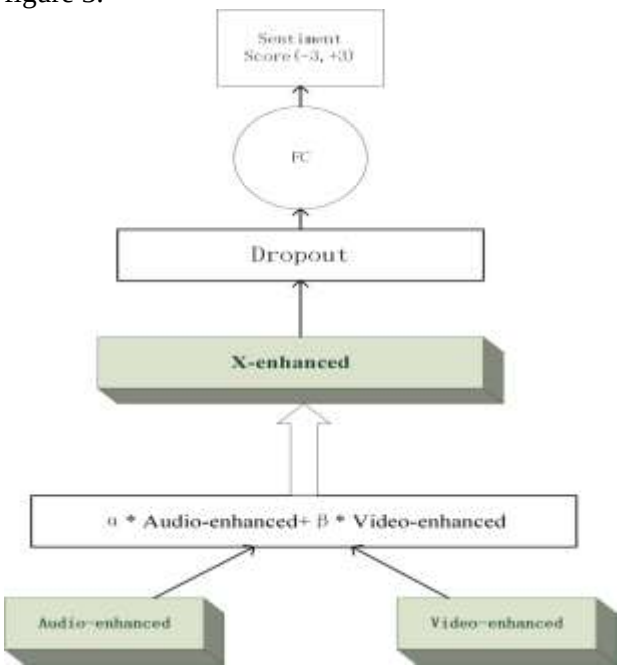

Figure 3. *Weighted fusion output module*

The model designs a weighted fusion module to combine the two text features after audio enhancement and video enhancement and make the three modalities fully play their roles. The fusion method is as follows:

Introduce two learnable weight parameters α and β, these two learnable weight parameters α and β express which enhancement modality contributes more to the fused multimodal representation. Initially, the contribution of the two enhancement modalities to the text modality is to be the same, that is, ($\alpha = \beta = 0.5$). The parameters will be dynamically adjusted through the loss function after each training. Use weights to perform weighted summation of enhanced features to generate the final multimodal representation in equation 5 (*X_enhanced*):

$$X_{enhanced} = \alpha \cdot F_{audio-enhance} + \beta \cdot F_{video-enhance} \quad (5)$$

Among them, α and β are weight parameters optimized by the model through training, which dynamically adjust the contribution of audio and video modalities to emotion prediction.

The fused multimodal representation is input into the fully connected layer for regression prediction (y_{pred}) of the sentiment intensity score in equation 6:

$$Y_{pred} = FC(Dropout(X_{enhance})) \quad (6)$$

$$\square_{\square\square\square} = FC(Dropout(\square_{\square\square\square\square\square\square\square})) \quad (6)$$

W_{out} is a weight matrix for the linear layer. $b_{\_out}$ is the bias term, which is a scalar. y_{pred} is a scalar representing the predicted sentiment intensity score (-3, +3).

Training loss: By comparing the obtained y_{pred} with the label y_{true}, MSE is the loss function for the model to perform forward propagation to optimize the parameters of the entire model in equation 7.

$$lossMSE = \frac{1}{N}\sum_{i=1}^{N}(y_{pred}^{(i)} - y_{true}^{(i)})^2 \quad (7)$$

3.4 Experiment

The model proposed is designed based on the sentiment score prediction task. The model's parameters will also be tuned according to the loss function under this task. Because it is a specific score prediction, MAE and Pearson are better standards for measuring the model's performance.

In order to verify the effect of the model in the nonlinear regression task, the sentiment score labels of the data set were converted into classification labels through manual classification. Then, they modified the relevant neuron functions used by the model to let the model output the prediction of sentiment classification. Designed two classification experiments. The first one is that the model performs a binary classification task. Mark the original labels belonging to [-3, 0) as negative and the original labels belonging to [0, +3] as positive. The other experiment is a three-classification task. Based on the binary classification, a new neutral classification is refined for the experiment. The original labels belonging to [-3, -1) are marked as negative, [-1, 1] are marked as neutral, and the original labels (1, +3] are marked as positive for the sentiment classification task, mainly calculated the F1-score and ACCURACY2 of the binary classification task as evaluation indicators. As an extension of the binary classification task, only ACCURACY3 was recorded as the evaluation indicator for the three classification tasks.

3.5 Comparative Experiments and Ablation Experiments

Conduct comparative experiments on some baseline methods to prove that the proposed model has better results in multimodal sentiment analysis and use evaluation indicators to confirm the model's performance. The ablation experiment removes the proposed module or replaces it with a general

module to demonstrate the necessity and contribution of the proposed method in the model.
This paper selects the following baseline models for comparative experiments:

- TFN [5]: Tensor Fusion Network fuses data between multiple modalities by converting features into tensors, promoting dynamic learning between modalities.
- MulT [26]: A Transformer-based multimodal sentiment analysis model that focuses on the interaction of multimodal information across different time steps, potentially transferring information from one modality to another.
- MFM [6]: The interaction relationship between multiple modalities and the independent information unique to each modality is discovered by factorizing the modality-specific and modality-shared features.
- MISA [27]: It learns modality-invariant and modality-specific representations to effectively fuse multimodal information, extract various modality-specific information, extract shared representations between multimodalities, and form new multimodal representations.
- MMLATCH [39]: Mining multimodal information interactions through dynamic and long-term dependency modeling. It emphasizes the dynamic interaction between modalities and the integration of time series features.

By comparing the results with these baseline methods, evaluate the strengths and weaknesses of the model through various indicators.
Ablation experiments are set up based on the multimodal information and attention mechanism that was used:

- Single-modal input model:

The input features were reduced to single-modal features to compare the performance of various modalities in independent states. This approach allows for an analysis of the effects and influences of features from different modalities in various aspects.

- Remove the modality enhancement module:

The feature enhancement module was removed, and direct concatenation was used to replace the feature fusion strategy. This approach helps verify whether the proposed feature fusion module is effective.

- Change the central modality :

The experimental results were compared by making either video or audio the central modality in the center enhancement module, with the other modalities serving as auxiliary modalities. This comparison was conducted to verify that the method proposed, with text as the central modality, yields the best performance.

3.6 Experimental environment

This experiment was conducted on RTX 3060, using python3.8 and pytorch1.9.0. The feature-length N used was 40, the number of heads K of the multi-head attention mechanism was 4, the training cycle was 50, the batch size was 16, and the initial model fusion parameters α and β of the Adam optimizer were set to 0.5 and 0.5 (assuming that the audio and video modalities have the same impact on the text modality, their range is set to (0, 1)). The dropout rate was set to 0.3. The early stopping parameter was set to 5, which was used together to avoid overfitting caused by the model being too close to the data.

4. Results and Discussions

The comparison between experiments and the baseline method in the sentiment score regression task is shown in Figure 4. The comparison between experiment and the baseline method on the CMU-MOSI dataset is shown in the figure. MAE represents the absolute error of the model on this dataset. MAE is lower than other models, proving that the model can output more accurate sentiment scores in the sentiment score prediction task. The highest Pearson indicates that the model is more relevant to the sentiment label score in actual prediction; the model is more in line with the experimental data.

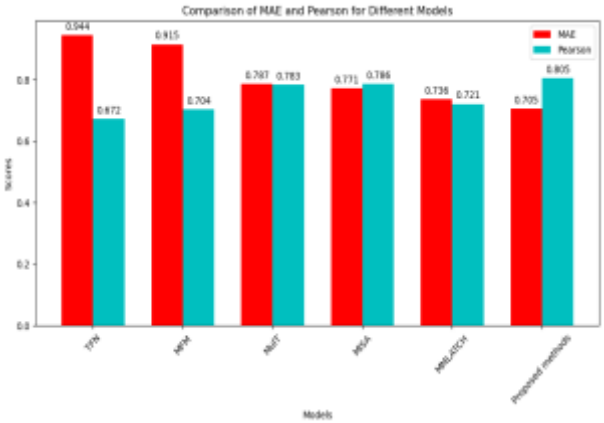

Figure 4. *Comparison of MAE and Pearson for Different Models*

The accuracy of the sentiment classification task is shown in the figure 5. On the CMU-MOSI dataset, the model achieved the best results in sentiment classification tasks because the BERT model fine-tuned on the movie ratings dataset used to extract text features can extract more accurate feature vectors on more similar sentiment classification

tasks, and it is better at sentiment classification tasks.

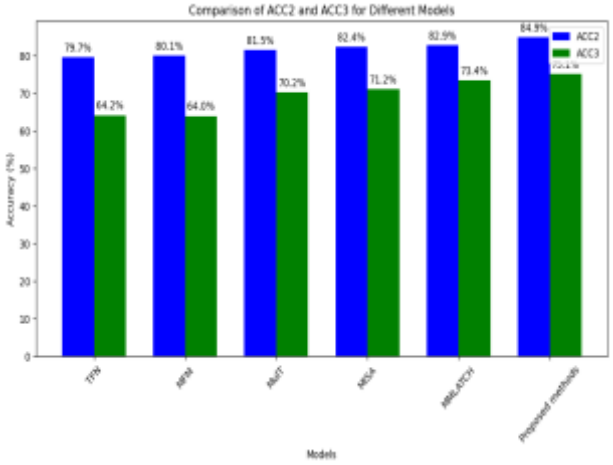

Figure 5. *Comparison of ACC2 and ACC3 for Different Models*

The experimental results of the f1 score based on the two-classification are shown in Figure 6. The Model can also achieve the best results in the f1 score. It proves the model can more accurately classify positive and negative classes in binary classification tasks.

From the overall results, in the binary classification experimental task, compared with the best-performing baseline model, the ACCURACY of the model increased by 2%, and the F1 score also increased by 2.4%. In the three-classification experimental task, ACCURACY3 increased by 1.7%. The model also performed well in the regression task of sentiment score prediction. The MAE index decreased by 2.1%, and the correlation coefficient index increased by 1.9% compared with the highest model. These show that the model has better experimental performance than the current commonly used methods.

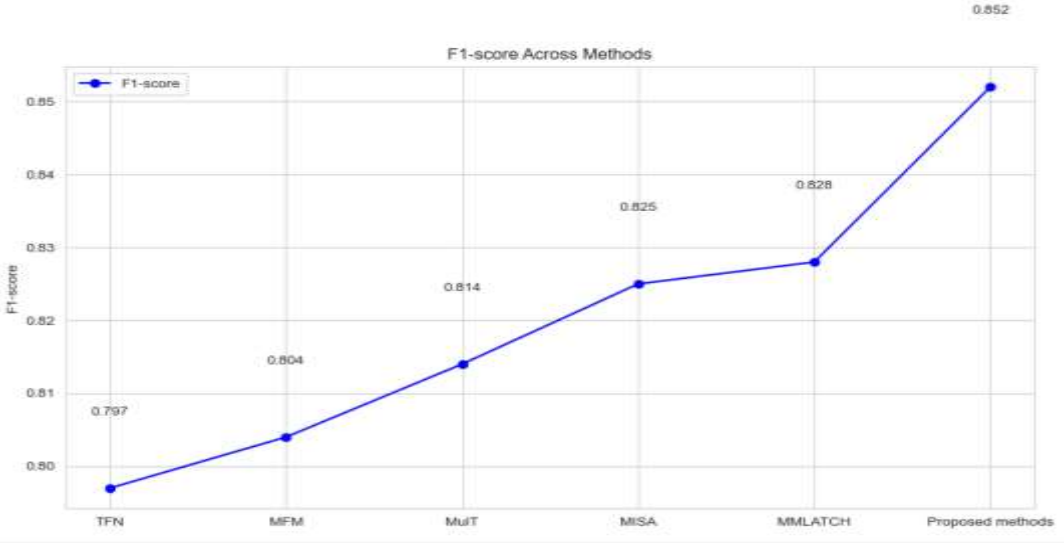

Figure 6. *Comparison F1-score for Different Models(2-classification)*

Table 1 shows the experimental results of the ablation experiments. The experimental results show that under unimodal model input, the text modality performs best in all indicators compared with the visual and auditory modalities, which shows that the text modality is more suitable for multimodal sentiment analysis. The performance in the task is better than that of other modalities, which proves the rationality of the experimental method of using text modality as the central modality in this paper. In terms of the replacement of the central modality, the experimental models centered on the visual modality and the auditory modality have improved compared with the single modal input, but compared with the text-centered model, ACCURACY2 They are 2.8% and 1.9% lower, respectively, and the F1 scores are 3.3% and 1.7% lower respectively. In the sentiment score prediction task, the MAE is 10.3% and 5.1% higher, respectively, and the Pearson correlation coefficient is 3.7% and 2.5% lower, respectively, proving that the performance of the text-centric model is significantly higher than that of the video-centric model-centric and audio-centric models. Compared with the direct concatenation model, the three modality enhancement models using the improved attention mechanism performed better in all performance indicators. When using direct concatenation features, the model's accuracy is even lower than using a single text modality feature. Only by choosing an effective feature fusion method can the operating efficiency of the model be effectively improved. Transfer Learning is studied in literature and reported [40-46].

4. Conclusions

This study proposes a sentiment analysis model that combines multimodal feature fusion with transfer

Table 1. *Experimental Results of Single-Modal Input, Changing the Central Modality, And Feature Splicing Methods*

Model	Evaluation Metrics				
	MAE	**PEARSON**	**ACC2**	**F1(2)**	**ACC3**
Text (BERT）	0.793	0.769	82.5%	0.822	64.3%
Audio	0.971	0.667	77.0%	0.770	61.1%
Visual	0.944	0.723	78.7%	0.788	62.3%
Audio-centric	0.808	0.768	82.1%	0.819	68.9%
Video-centric	0.756	0.780	83.0%	0.835	70.2%
Direct splicing	0.858	0.733	80.2%	0.795	66.4%
Proposed methods	0.705	0.805	84.9%	0.852	75.1%

learning. A better multimodal fusion feature vector is obtained by setting the text modality as the central modality. This study chooses to use transfer learning. When receiving text features, the BERT model that has been fine-tuned in other datasets and sentiment classification tasks is used to obtain a better text feature vector, and the CMU-Multimodal SDK tool is used for data preprocessing to ensure the alignment of modal information. The study shows that text modality features contain core emotional information in multimodal features, and other modal features play a more auxiliary role. Introducing a multi-head attention module in the fusion between features enhances the expression ability of fused multimodal features. The method proposed in this paper demonstrates model performance in sentiment regression and classification tasks through various experimental designs on the CMU-MOSI dataset. Although the experimental method proposed in this work has achieved good results in multimodal sentiment analysis tasks, there are still some shortcomings: in the CMU-MOSI dataset, the number of neutral class samples is small, the boundary division of sentiment classification is relatively vague, and the generalization ability of the model may be affected. The future goal is to combine multimodal sentiment analysis with the currently popular generative AI and explore more strategies to enhance the correlation between modalities through generative AI so as to explore the development path of future multimodal sentiment analysis agents.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors acknowledge the Fundamental Research Grant Scheme research was supported by the Fundamental Research Grant Scheme (FRGS), grant number FRGS/1/2022/ICT02/UKM/02/7, funded by the Ministry of Higher Education (MOHE) Malaysia.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *arxiv preprint cs/0205070*, 2002.

[2] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1746–1751.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter Association for Computational Linguistics: Human Language Technologies, vol. 1*, Long and Short Papers, 2019, pp. 4171-4186.

[4] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annual Meeting of the Association for Computational Linguistics*, vol. 1, Vancouver, Canada, 2017, pp. 873–883.

[5] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arxiv preprint arxiv:1707.07250*, 2017.

[6] Y. H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, (2018). Learning factorized multimodal representations, *arxiv preprint arxiv:1806.06176*,

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., (2017). Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30.

[8] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, (2020). A comprehensive survey on graph neural networks, *IEEE Trans. Neural Networks Learn. Syst.*, 32(1);4-24.

[9] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, (2017). Tensor fusion network for

multimodal sentiment analysis, *arxiv preprint arxiv:1707.07250*.

[10] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, (2019). Bert: Pre-training of deep bidirectional transformers for language understanding, in *Proc. 2019 Conf. North American Chapter Association for Computational Linguistics: Human Language Technologies, vol. 1*, Long and Short Papers, pp. 4171-4186.

[11] K. He, X. Zhang, S. Ren, and J. Sun, (2016) Deep residual learning for image recognition, in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. 770-778.

[12] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge, *Speech Communication*, 53(9-10);1062-1087.

[13] P. P. Liang, A. Zadeh, and L.-P. Morency, (2018). Multimodal local-global ranking fusion for emotion recognition, in *Proc. 20th ACM Int. Conf. Multimodal Interaction*, pp. 472-476.

[14] L.-P. Morency, R. Mihalcea, and P. Doshi, (2011). Towards multimodal sentiment analysis: Harvesting opinions from the web, in *Proc. 13th Int. Conf. Multimodal Interfaces*, pp. 169-176.

[15] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, in *Proc. 56th Annual Meeting of the Association for Computational Linguistics*, vol. 1, Melbourne, Australia, pp. 2236-2246.

[16] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, (2017). End-to-end multimodal emotion recognition using deep neural networks, *IEEE J. Select. Topics Signal Process.*, 11(8);1301-1309.

[17] K. Simonyan and A. Zisserman, (2014). Very deep convolutional networks for large-scale image recognition, *arxiv preprint arxiv:1409.1556*,

[18] B. Hasani and M. H. Mahoor, (2017) Facial expression recognition using enhanced deep 3D convolutional neural networks, in *Proc. IEEE Conf. Computer Vision Pattern Recognition Workshops*, pp. 30-40.

[19] A. Satt, S. Rozenberg, and R. Hoory, (2017) Efficient emotion recognition from speech using deep learning on spectrograms, in *Interspeech*, pp. 1089-1093.

[20] X. Li, W. Zheng, Y. Zong, H. Chang, and C. Lu, (2021). Attention-based spatio-temporal graphic LSTM for EEG emotion recognition, in *2021 Int. Joint Conf. Neural Networks (IJCNN)*, 2021, pp. 1-8.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al., (2020). An image is worth 16x16 words: Transformers for image recognition at scale, *arxiv preprint arxiv:2010.11929*,

[22] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, (2017) Adversarial discriminative domain adaptation, in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. 7167-7176.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., (2021) Learning transferable visual models from natural language supervision, in *Int. Conf. Machine Learning*, pp. 8748-8763.

[24] F. Zhuang, et al., (2020). A comprehensive survey on transfer learning, *Proc. IEEE*, 109(1);43-76.

[25] A. A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annual Meeting of the Association for Computational Linguistics*, vol. 1, Melbourne, Australia, 2018, pp. 2236-2246.

[26] Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, (2019) Multimodal transformer for unaligned multimodal language sequences, in *Proc. Assoc. Comput. Linguistics*, p. 6558.

[27] D. Hazarika, R. Zimmermann, and S. Poria, (2020) MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis, in *Proc. 28th ACM Int. Conf. Multimedia*, 1122–1131.

[28] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, (2018). Multimodal sentiment analysis: Addressing key issues and setting up the baselines, *IEEE Intell. Syst.*, 33(6);17-25

[29] Z. Shao, C. Wang, and X. Li, (2015) Facial expression recognition under varying lighting conditions, *IEEE Trans. Affective Comput.*, 6(2);161-172.

[30] N. P., (2024). Occlusion-aware facial expression recognition: A deep learning approach," *Multimedia Tools Appl.*, 83(11);32895-32921,

[31] P. R. Sackett, F. Lievens, C. H. Van Iddekinge, and N. R. Kuncel, (2017). Individual differences and their measurement: A review of 100 years of research, *J. Appl. Psychol.*, 102(3);254,

[32] D. Matsumoto and H. S. Hwang, (2012). Culture and emotion: The integration of biological and cultural contributions, *J. Cross-Cultural Psychol.*, 43 (1);91-118

[33] Z. Madhoushi, A. R. Hamdan, and S. Zainudin, (2015) Sentiment analysis techniques in recent works, in *2015 Science and Information Conf. (SAI)*, pp. 288-291.

[34] Z. Madhoushi, A. R. Hamdan, and S. Zainudin, (2019). Aspect-based sentiment analysis methods in recent years, *Asia-Pacific J. Inf. Technol. Multimedia*, 8(01);79-96,

[35] Z. Madhoushi, A. R. Hamdan, and S. Zainudin, (2023). Semi-supervised model for aspect sentiment detection, *Information*, 14(5);293, 2023.

[36] E. Z. U. A. N. A. Sukawai and N. A. Z. L. I. A. Omar, (2020). Corpus development for Malay sentiment analysis using semi-supervised approach, *Asia-Pacific J. Inf. Technol. Multimedia*, 9(01);94-109,

[37] M. A. Latiffi and M. R. Yaakub, (2018) Sentiment analysis: An enhancement of ontological-based using hybrid machine learning techniques," *Asian J. Inf. Technol.*, 7;61-69.

[38] NLP Town, (2023). bert-base-multilingual-uncased-sentiment (Revision edd66ab), Hugging Face, [Online]. Available: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment. [Accessed: Mar. 6, 2025].

[39]G. Paraskevopoulos, E. Georgiou, and A. Potamianos, (2022). Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis," in *ICASSP 2022–2022 IEEE Int. Conf. Acoustics, Speech Signal Process.*, Virtual and Singapore, 4573–4577.

[40]S. Ranjana, & A. Meenakshi. (2025). Breast Cancer Detection using Convolutional Autoencoder with Hybrid Deep Learning Model. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.1225

[41]D. Naga Jyothi, & Uma N. Dulhare. (2025). Understanding and Analysing Causal Relations through Modelling using Causal Machine Learning. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.1018

[42]Olola, T. M., & Olatunde, T. I. (2025). Artificial Intelligence in Financial and Supply Chain Optimization: Predictive Analytics for Business Growth and Market Stability in The USA. *International Journal of Applied Sciences and Radiation Research*, 2(1). https://doi.org/10.22399/ijasrar.18

[43]Johnsymol Joy, & Mercy Paul Selvan. (2025). An efficient hybrid Deep Learning-Machine Learning method for diagnosing neurodegenerative disorders. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.701

[44]Ibeh, C. V., & Adegbola, A. (2025). AI and Machine Learning for Sustainable Energy: Predictive Modelling, Optimization and Socioeconomic Impact In The USA. *International Journal of Applied Sciences and Radiation Research* , 2(1). https://doi.org/10.22399/ijasrar.19

[45]Sivananda Hanumanthu, & Gaddikoppula Anil Kumar. (2025). Deep Learning Models with Transfer Learning and Ensemble for Enhancing Cybersecurity in IoT Use Cases. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.1037

[46]Hafez, I. Y., & El-Mageed, A. A. A. (2025). Enhancing Digital Finance Security: AI-Based Approaches for Credit Card and Cryptocurrency Fraud Detection. *International Journal of Applied Sciences and Radiation Research*, 2(1). https://doi.org/10.22399/ijasrar.21