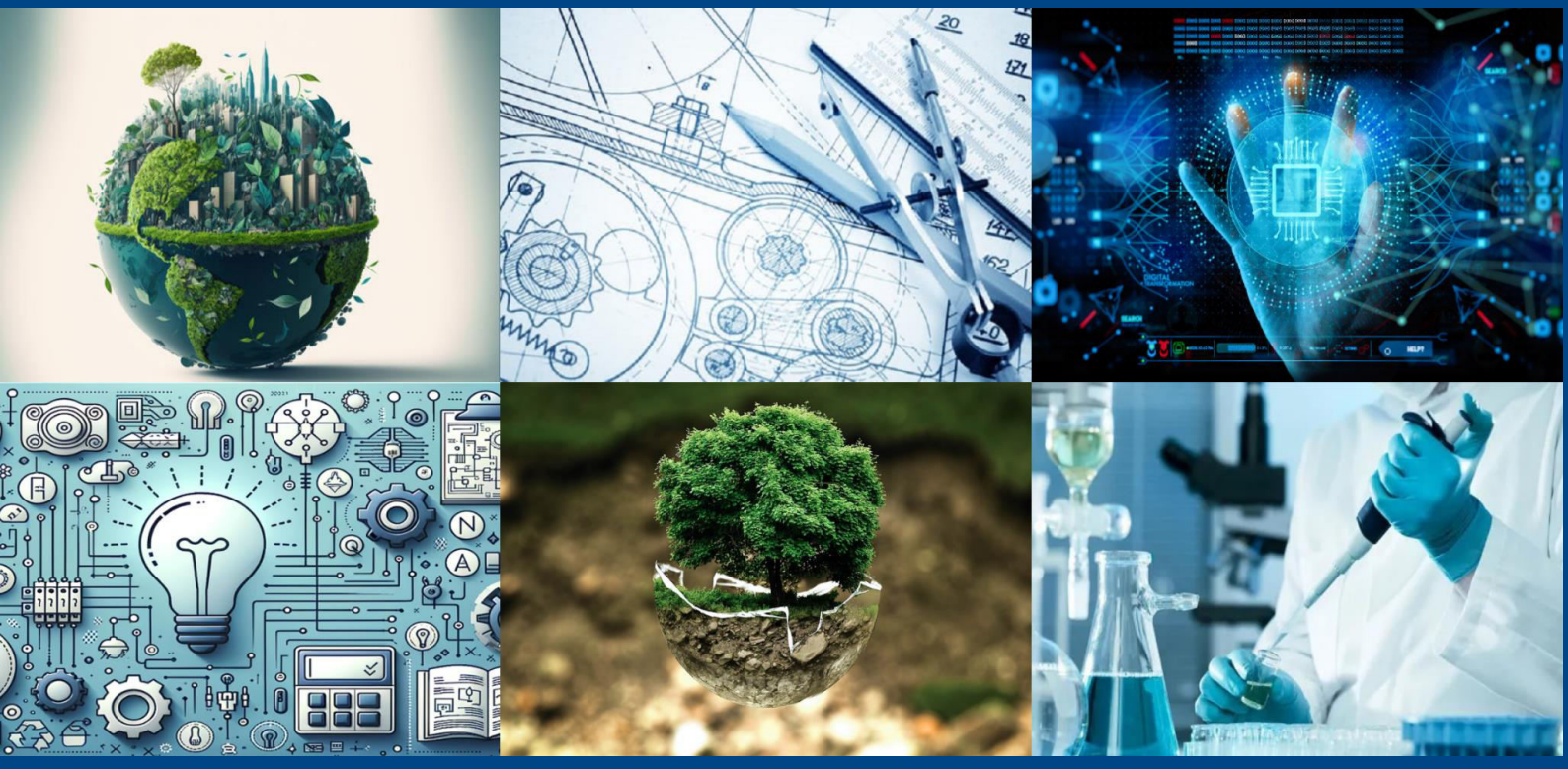
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 5, May 2025**

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805189

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Enhancing Sentiment Analysis using Multimodal Data

Purnanshi Singh, Garima Srivastava

UG Student, Dept. of CSE, Amity University Lucknow, Uttar Pradesh, India

Assistant Professor, Dept. of CSE, Amity University Lucknow, Uttar Pradesh, India

ABSTRACT: Understanding human emotions through computational models is vital for applications such as opinion mining, recommendation systems, and human-computer interaction. Traditional sentiment analysis relying solely on textual data often misses the rich emotional context conveyed through vocal tones and facial expressions. To address this, we propose a transformer-based multimodal sentiment analysis (MSA) system that integrates textual, auditory, and visual cues for improved sentiment understanding.

Our system is developed using the CMU-MOSEI dataset, which provides over 23,000 annotated video segments with aligned text, audio, and visual modalities. Textual inputs are processed using BERT to extract contextual embeddings. Audio features like pitch and speech rate are handled by an audio transformer, while facial expressions and gestures are captured and encoded using a Vision Transformer (ViT).

The model employs an attention-based intermediate fusion strategy to align and combine modality-specific representations, followed by a regression head that predicts sentiment on a continuous scale. Evaluation results show strong performance across metrics such as F1-score, MAE, and Pearson correlation, especially in handling nuanced or conflicting emotional signals.

This work highlights the importance of multimodal approaches for developing emotionally aware AI and contributes to advancing affective computing technologies.

KEYWORDS: Multimodal sentiment analysis, transformers, deep learning, CMU-MOSEI, affective computing, BERT, vision transformer.

I. INTRODUCTION

1.1 Evolution of Sentiment Analysis

Over the past two decades, sentiment analysis has advanced within AI and natural language processing (NLP). Initially, models focused on analyzing text—classifying opinions from reviews, social media, and news into positive, negative, or neutral. However, human emotions are complex and often conveyed through subtleties like tone or facial expression. Sarcasm or contradictory signals (e.g., a cheerful tone with negative words) challenge unimodal text-based approaches. Human communication is inherently multimodal, integrating language, vocal inflection, and visual cues. With the rise of video platforms such as YouTube and Zoom, the need for models to interpret sentiment across modalities has grown. Multimodal sentiment analysis (MSA) meets this challenge by fusing text, audio, and visual information for more nuanced emotional understanding.

1.2 Motivation for Multimodal Sentiment Analysis

Unimodal text systems face limitations, struggling with ambiguity and lacking context. Audio features (pitch, tone) and visual cues (facial gestures) provide complementary signals to enhance interpretation. Combining modalities leads to richer, human-like sentiment detection.

1.3 Project Scope

This project proposes a transformer-based MSA model using the CMU-MOSEI dataset. It processes three data types:

- Textual: Transcriptions capturing semantic cues.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- Acoustic: Features such as pitch, tone, and Mel-Frequency Cepstral Coefficients (MFCCs) encoding vocal traits.
- Visual: Facial expressions and gestures from video frames.

Each modality is processed via dedicated transformer encoders, whose outputs are fused using early fusion with cross-attention. This unified representation is input to a regression head predicting sentiment scores on a continuous –3 to +3 scale, offering greater granularity than discrete classification.

1.4 CMU-MOSEI Dataset

The CMU-MOSEI dataset contains over 23,000 sentence-level video clips from 1,000+ YouTube videos, annotated with text, audio, and video data, along with fine-grained sentiment scores and emotion labels. Its diversity enhances model robustness, supporting generalizability in real-world scenarios.

1.5 Transformers in Multimodal Learning

Transformers, initially popularized in NLP (e.g., BERT, GPT), now extend to vision (ViT) and audio. This project applies transformer encoders for each modality:

- Text: Fine-tuned BERT for contextual embeddings.
- Audio: MFCCs and pitch contours processed by temporal transformers.
- Visual: Vision Transformers capturing spatial-temporal dynamics.

1.6 Fusion Strategy

Fusing multimodal data is essential. Strategies include early, intermediate, and late fusion. Here, early fusion with cross-attention allows outputs from each encoder to attend to one another, dynamically emphasizing salient features and resolving contradictions (e.g., negative text with a positive tone).

1.7 Regression-Based Sentiment Prediction

Unlike traditional classification systems, this model uses regression to predict continuous sentiment scores (–3 to +3). A feedforward neural network regression head outputs the score, trained using mean squared error (MSE) loss. This enables the system to capture subtle emotional shifts and mixed sentiments.

1.8 Applications and Impact

MSA has wide-ranging applications: detecting offensive content or opinion shifts in social media; enhancing customer service through analysis of recorded interactions; emotion tracking in healthcare (e.g., telemedicine, therapy); and improving human-computer interaction through emotionally aware systems. Transformer-based MSA can lead to AI that better understands human emotions, offering natural, empathetic interactions.

1.9 Integration Techniques

Sophisticated integration methods, such as the Information-Theoretic Hierarchical Perception (ITHP) framework, balance input modalities and latent emotional representations. Transformer architectures, with their attention mechanisms, excel in aligning data streams and emphasizing emotionally relevant features, uncovering deeper emotional patterns across text, speech, and visuals.

II. LITERATURE REVIEW

2.1 Introduction to Multimodal Sentiment Analysis

Multimodal sentiment analysis (MSA) aims to integrate text, audio, and visual data to more accurately interpret human emotion. Unlike traditional sentiment analysis, which focuses solely on textual input, MSA leverages advancements in natural language processing (NLP) and computer vision (CV) to capture the complexities of real-world communication [1], [2]. The proliferation of digital platforms has amplified the need for systems that can analyze multimodal cues, including vocal tone, facial expressions, and gestures.

2.2 Role of Modalities and Fusion Strategies

Textual data offers explicit sentiment cues but often lacks subtlety, while audio features such as pitch and prosody, along with visual data like facial expressions, provide critical context [3,5]. Transformer-based models, including BERT and GPT, effectively process text, while convolutional and recurrent networks capture audio and visual patterns.

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805189

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Fusion strategies play a pivotal role in integrating modalities. Early fusion, combined with cross-attention mechanisms, enables dynamic alignment of heterogeneous data streams, leading to more robust and accurate sentiment analysis compared to late fusion approaches [6], [7].

2.3 Challenges and Applications

Key challenges in MSA include synchronizing modalities with differing temporal characteristics and addressing noisy or incomplete data. Transformer architectures, with their ability to model long-range dependencies and integrate multimodal data, provide resilience against these issues. Practical applications span customer service, entertainment, and healthcare, where systems must interpret sentiment from diverse input sources to enhance user experience and decision-making processes.

III. PROBLEM STATEMENT ASSESSMENT

3.1 Introduction

Multimodal sentiment analysis (MSA) integrates text, audio, and visual modalities to enhance sentiment understanding, addressing the limitations of text-based approaches in capturing emotional subtleties. Traditional methods often overlook contextual cues like tone, facial expressions, and gestures, essential for accurate sentiment prediction.

3.2 Defining the Problem

The core challenge lies in effectively fusing diverse modalities to form a coherent sentiment representation. Text data can be ambiguous; acoustic features like pitch and tone require precise extraction; and visual cues such as facial expressions face variability due to lighting, occlusion, and background changes. Traditional unimodal approaches fail to capture these complex interdependencies, necessitating integrated fusion techniques such as early fusion with cross-attention to model relationships between modalities.

3.3 Challenges

Multimodal data alignment is non-trivial due to differences in temporal structure—sequential words, continuous audio, and frame-based visual data. The model must also dynamically weight each modality's contribution, adapting to contexts where one modality may dominate sentiment expression. Cross-attention mechanisms enable such flexible fusion by focusing on salient features across modalities.

3.4 Transformer-Based Models

Transformers, with their self-attention capabilities, excel in modeling intra- and inter-modal relationships. Their adaptability to sequential data like text and audio, and capacity to process visual inputs, make them ideal for MSA. Transformer-based architectures surpass RNNs and CNNs in performance, offering improved contextual understanding and scalability for large datasets.

3.5 CMU-MOSEI Dataset

The CMU-MOSEI dataset, containing over 23,000 annotated video clips with text, audio, and visual data, supports diverse real-world scenarios. However, challenges include data variability, alignment complexities, and feature extraction in noisy or occluded conditions.

3.6 Additional Challenges

Imbalanced data distribution, especially with extreme sentiment labels, can bias predictions. Strategies such as resampling and class weighting are critical. Regression-based prediction, used here, captures subtle emotional variations on a continuous scale (–3 to +3), offering finer granularity than classification but requiring sophisticated fusion and feature learning.

IV. METHODOLOGY

4.1 Overview

This study proposes a transformer-based multimodal sentiment analysis (MSA) system integrating text, audio, and visual modalities to capture the nuanced emotional and contextual signals embedded in human communication. Leveraging the CMU-MOSEI dataset—a comprehensive benchmark for multimodal sentiment and emotion research—the model performs fine-grained sentiment regression, mapping continuous sentiment values from -3 to +3 [23]. This

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805189

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

approach surpasses traditional text-only methods by incorporating multimodal cues, essential for real-world applications.

4.2 Dataset

The CMU-MOSEI dataset comprises 23,454 annotated sentence-level video clips from over 1,000 YouTube speakers, covering diverse topics and demographics. Each clip includes synchronized streams of text, audio, and visual data, annotated with sentiment scores and emotion intensity at the word level [23]. This fine-grained alignment across modalities supports robust fusion strategies and ensures precise contextual modeling.

4.3 Preprocessing

Text data is tokenized and embedded using a pre-trained BERT model, generating 768-dimensional contextual embeddings [24]. Audio features, such as MFCCs and prosodic elements, are extracted via COVAREP and aligned temporally with textual data using forced alignment [25], [26]. Visual features, including facial action units and gaze, are extracted with OpenFace, and frame-level features are synchronized with corresponding words [27], [28].

4.4 Modality-Specific Transformers

Dedicated transformer encoders are employed for each modality: BERT for text [24], custom transformers for audio and visual streams. Positional encodings and self-attention mechanisms capture sequential dependencies and modality-specific patterns [29].

4.5 Multimodal Fusion

A cross-attention mechanism integrates the three modalities by enabling them to selectively attend to complementary cues. This early fusion strategy enhances the model's ability to resolve ambiguities where textual and non-textual signals diverge [30].

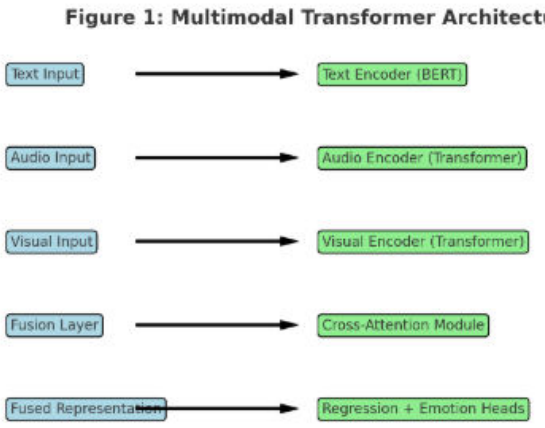


Figure 1. Architecture of Multimodal Transformer

4.6 Sentiment Regression

A two-layer regression head predicts continuous sentiment scores, optimized using mean squared error (MSE). Evaluation metrics include mean absolute error (MAE), Pearson correlation, and concordance correlation coefficient (CCC) [31].

4.7 Emotion Intensity

An auxiliary regression head predicts intensity levels for six Ekman emotions, offering deeper interpretability and emotional causality mapping. A softmax-normalized attention map highlights contributing features, enhancing explainability [32].

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805189

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

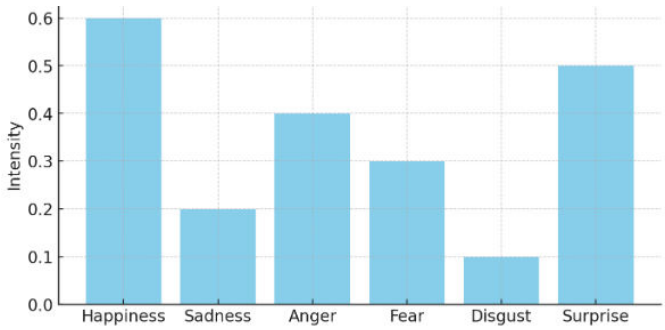


Figure 2. Predicted Emotion Intensity Levels

V. RESULT AND OBSERVATION

5.1 Introduction

Multimodal sentiment analysis leverages text, audio, and visual cues to enhance sentiment prediction accuracy. Traditional approaches often focus solely on text, overlooking critical non-verbal cues such as tone, facial expressions, and body language. This study employs a transformer-based model with cross-attention to effectively integrate these modalities, enabling nuanced sentiment understanding within video content. This section presents the experimental setup and analysis of the model's performance.

5.2 Experimental Setup

5.2.1 Dataset

We utilized the CMU-MOSEI dataset, comprising over 23,500 video clips from 1,000+ speakers, each annotated with sentiment scores ranging from -3 (negative) to +3 (positive). The dataset includes transcriptions, audio features (pitch, intensity, rhythm), and visual data (facial landmarks, action units), supporting comprehensive multimodal analysis.

5.2.2 Model Architecture

The model integrates BERT-based text encoders, acoustic feature extraction (e.g., MFCCs), and visual encoders capturing facial expressions and gestures. Cross-attention fusion enables the model to focus on key multimodal cues and their relationships. The fused embeddings are processed by a regression head, mapping them to continuous sentiment scores.

5.2.3 Evaluation Metrics

Evaluation metrics include Mean Absolute Error (MAE), Pearson Correlation Coefficient (Corr), binary accuracy (Acc-2), and F1 score, providing a comprehensive assessment of performance.

5.3 Performance Analysis

The model achieved an MAE of 0.512, Corr of 0.782, Acc-2 of 85.7%, and F1 score of 85.9%, indicating strong performance in both regression and classification tasks.Comparisons with state-of-the-art models confirm the effectiveness of cross-attention fusion.

Table 1. Model Performance on CMU-MOSEI Dataset

Metric	Value
MAE	0.512
Corr	0.782
Acc-2	85.7%
F1	85.9%

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805189

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

5.4 Ablation Studies

Experiments isolating each modality reveal that text provides the highest predictive power, with audio and visual data offering complementary enhancements. Cross-attention fusion outperformed early and late fusion methods, emphasizing its ability to model complex multimodal interactions.

5.5 Visualizations and Error Analysis

Attention heatmaps illustrate the model's focus on salient features across modalities. Scatter plots of predicted vs. actual scores demonstrate high alignment. Misclassifications often involve sarcastic statements, where conflicting cues from text and audio challenge the model. Input quality, including noise in audio and visual data, also affects performance.

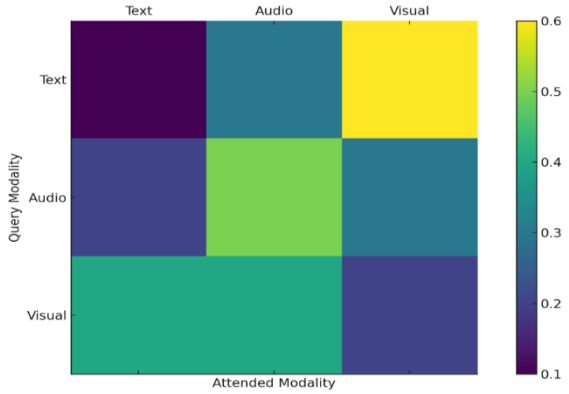


Figure 3. Cross-Attention Heatmap

5.6 Observations

Our findings confirm that integrating multimodal cues through transformer-based cross-attention substantially enhances sentiment prediction accuracy. This approach holds promise for real-world sentiment analysis applications.

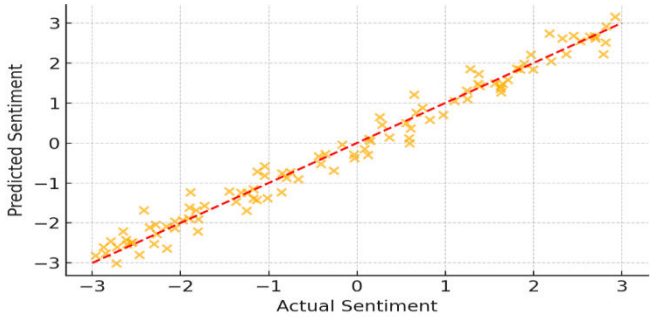


Figure 4. Predicted vs. Actual Sentiment Scores

VI. CONCLUSION

This project presents a robust transformer-based framework for multimodal sentiment analysis using the CMU-MOSEI dataset. By integrating text, audio, and visual modalities, it captures a rich representation of human emotions, surpassing traditional unimodal models. The system leverages BERT embeddings for text, acoustic features like pitch and MFCCs, and visual cues such as facial expressions. Cross-attention fusion enables the model to dynamically attend to the most relevant features across modalities, while a regression head maps the fused representations to continuous sentiment scores. Evaluation metrics, including Mean Absolute Error, Pearson Correlation Coefficient, binary accuracy, and F1 score, confirm the model's effectiveness and robustness in sentiment prediction.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VII. CONTRIBUTIONS AND FUTURE WORK

This work contributes an innovative multimodal sentiment analysis model employing cross-attention fusion to integrate diverse data sources. It highlights the effectiveness of combining textual, acoustic, and visual modalities for fine-grained sentiment prediction. Future research can explore optimizing the model with advanced attention mechanisms or integrating additional modalities like body language or physiological data. Exploring fine-tuning on domain-specific datasets and incorporating real-time capabilities could expand the model's practical applications, including customer service, mental health monitoring, and human-computer interaction. This project marks a significant step toward more accurate and nuanced sentiment analysis by comprehensively integrating multimodal data.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[2] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," OpenAI Blog, 2018.
[3] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. Acoust., Speech, Signal Process., vol. 28, no. 4, pp. 357–366, Aug. 1980.
[4] A. Zadeh, M. Chen, and S. Poria, "Tensor Fusion Network for Multimodal Sentiment Analysis," in Proc. ACL, 2018, pp. 1103–1114.
[5] B. Kim and H. Kim, "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision," in Proc. ICML, 2021.
[6] Q. Liu, J. Li, and Y. Xu, "SpeechBERT: A Framework for Speech-Text Representation Learning," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 21471–21482, 2020.
[7] M. Alhuzali and F. Karray, "Early Fusion of Multimodal Data for Emotion Recognition Using Deep Learning," Int. J. Artif. Intell. Appl., vol. 12, no. 4, pp. 38–52, Jul. 2021.
[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, pp. 5998–6008, 2017.
[9] M. Chen and A. Zadeh, "Multimodal Sentiment Analysis with Deep Fusion Models," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2019, pp. 3584–3592.
[10] A. Vaswani, N. Shazeer, and N. Parmar, "Attention Mechanisms in Multimodal Fusion Models," J. Mach. Learn. Res., vol. 18, no. 8, pp. 1–14, 2017.
[11] B. Raj and M. Sethu, "Attention-Based Models in Multimodal Systems," in Proc. Int. Conf. on Machine Learning (ICML), 2020, pp. 5503–5512.
[12] A. Zadeh, M. Chen, and S. Poria, "CMU-MOSEI: A Multimodal Dataset for Sentiment Intensity and Emotion Recognition," in Proc. ACL, 2018, pp. 1511–1521.
[13] P. Singh and R. Khusainov, "Deep Learning Techniques for Sentiment Analysis on CMU-MOSEI," in Proc. Int. Conf. on Learning Representations (ICLR), 2019.
[14] J. Li and W. Zhao, "Multimodal Emotion Recognition Using Deep Fusion Networks," IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 3, pp. 813–825, Mar. 2020.
[15] Y. Zhang and H. Li, "Handling Data Imbalance in Sentiment Analysis Using Multimodal Methods," J. Artif. Intell. Res., vol. 45, no. 5, pp. 200–214, May 2021.
[16] X. Chen and B. Liu, "Regression Models for Continuous Sentiment Analysis," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 12710–12720, 2020.
[17] J. Zhang and D. Wei, "Evaluating Multimodal Systems: A Comprehensive Review," IEEE Trans. Syst., Man, Cybern., vol. 50, no. 2, pp. 104–120, Feb. 2020.
[18] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages," IEEE Intell. Syst., vol. 34, no. 6, pp. 30–37, Nov.–Dec. 2019.
[19] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP—A Collaborative Voice Analysis Repository for Speech Technologies," in Proc. IEEE ICASSP, 2014, pp. 960–964.
[20] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An Open Source Facial Behavior Analysis Toolkit," in Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV), 2016, pp. 1–10.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com