Explainable Multimodal Sentiment Analysis Using Hierarchical Attention-Based Adaptive Transformer Models

Anna Shalini[1], Dr. B. Manikyala Rao[2], Ms. Ranjitha. P. K[3], Dr Guru Basava Aradhya S[4],
Dr. S. Farhad[5], Elangovan Muniyandy[6], Prof. Ts. Dr. Yousef A. Baker El-Ebiary[7]
Research Scholar, Department of English, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist., Andhra Pradesh - 522502, India[1]
Associate Professor, Dept. of CSE, Aditya University, Surampalem, Andhra Pradesh, India[2]
Assistant Professor, Centre for Management Studies, Jain Deemed to be university, Lalbagh Road, Bangalore, India[3]
Director & Professor, Padmashree Institute of Management and Sciences, Kengeri, Bangalore, India[4]
Associate Professor, Department of English, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist., Andhra Pradesh - 522502, India[5]
Department of Biosciences, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai - 602 105, India[6]
Applied Science Research Center, Applied Science Private University, Amman, Jordan[6]
Faculty of Informatics and Computing, UniSZA University, Malaysia[7]

***Abstract*—Multimodal Sentiment Analysis (MSA) has emerged as a critical task in Natural Language Processing (NLP), driven by the growth of user-generated content containing textual, visual, and auditory cues. While transformer-based approaches achieve strong predictive performance, their lack of interpretability and limited adaptability restrict their use in sensitive applications such as healthcare, education, and human–computer interaction. To address these challenges, this study proposes an explainable and adaptive MSA framework based on a hierarchical attention-based transformer architecture. The model leverages RoBERTa for text, Wav2Vec2.0 for speech, and Vision Transformer (ViT) for visual cues, with features fused using a three-tier attention mechanism encompassing token/frame-level, modality-level, and semantic-level attention. This design enables fine-grained representation learning, dynamic cross-modal alignment, and intrinsic explainability through attention heatmaps. Additionally, contrastive alignment loss is incorporated to align heterogeneous modality embeddings, while label smoothing mitigates overconfidence, improving generalizability. Experimental evaluation on the CMU-MOSEI benchmark demonstrates state-of-the-art performance, achieving 93.2% accuracy, 93.5% precision, 92.8% recall, and 94.1% F1-score, surpassing prior multimodal transformer-based methods. Unlike earlier models that rely on shallow fusion or post-hoc interpretability, the proposed approach integrates explainability into its architecture, balancing accuracy and transparency. These results confirm the efficacy of the adaptive hierarchical attention-based framework in delivering a robust, interpretable, and scalable solution for English-language multimodal sentiment analysis.**

***Keywords*—Multimodal sentiment analysis; RoBERTa; Wav2Vec 2.0; vision transformer; CMU-MOSEI**

I. Introduction

With the introduction of computer-mediated communication, sentiment analysis has become a crucial sub-field of Natural Language Processing (NLP), which allows machines to identify and understand emotions. With the potential increase in user-generated content—from social network status posts to consumer reviews and video blogs—there is an increasing need to analyze sentiments expressed in not only text but also in synchronized audio and visual modalities. This has driven the development of Multimodal Sentiment Analysis (MSA), a task that leverages signals from different media to improve affective computing accuracy [1]. Conventional unimodal sentiment analysis, although robust in controlled environments, tends to miss the rich and non-textual emotional cues contained in tone, pitch, facial expressions, and body language [2]. Multimodal sentiment analysis overcomes this by combining features from multiple modalities. Most existing solutions, however, depend on basic early or late fusion techniques, which tend to ignore complex intra- and inter-modal interactions and thus result in sub-optimal performance [3]. Recent progress on deep learning, specifically transformer architecture, has completely transformed MSA through the capacity of models to identify sophisticated sequential and contextual structures with the mechanisms of self-attention and multi-head attention.

Pre-trained models like RoBERTa for text [4], Wav2Vec 2.0 for audio [5] and (ViT) for vision inputs [6] have been highly promising in their individual spaces. However, when applied together in multimodal pipelines, these models are still plagued with limited interpretability and a lack of sufficient alignment among modality representations. Interpretability is a major concern in deploying transformers to real-world sentiment analysis, particularly in high-risk settings like healthcare or education, where it matters as much to know why a model has predicted as it does to know the prediction. Several researchers have proposed explainable models based on hierarchical or attention-based architectures [7]. Perikos and Diamantopoulos investigate LIME, SHAP, and Grad-CAM to enhance the transparency of sentiment models [8]. Nevertheless, these approaches tend to use explainability as an after-the-fact

method, as opposed to incorporating it into the model architecture itself. To overcome these shortcomings, introduced an adaptive fine-tuned multimodal transformer model that incorporates hierarchical attention explainability into the learning process. This method uses RoBERTa for text embedding, Wav2Vec 2.0 for acoustic feature extraction, and ViT for visual signal extraction [9]. These modality-specific encoders create contextualized representations that are forwarded to a hierarchical attention fusion mechanism, which, at the token/frame-level, modality-level, and semantic-level, operates. It is this multi-layered structure of attention which allows the model not only to pay attention to the most sentiment-bearing features but also to open-book its reasoning transparently.

Furthermore, include contrastive loss to put modality embeddings in a common semantic space, solving the misalignment issue prevalent in multimodal learning [10]. To enhance generalization and prevent overconfidence in predictions, label smoothing is utilized in training [11]. The architecture resulting from these is strong, interpretable, and well-calibrated for sentiment classification tasks. In contrast to previous models like the Multimodal Transformer of [12], which employed tensor fusion but lacked intrinsic explainability, this model provides intrinsic attention visualizations, enabling users and researchers to understand how the model decides. In contrast to shallow feature concatenation or independent treatment of each modality in conventional fusion-based approaches, this model is interested in learning deep, aligned, and interpretable cross-modal interactions [13]. Conjecture that incorporating modality-specific transformers with hierarchical attention layers not only brings performance gains but also reveals a new avenue for transparent multimodal AI, where the behavior of the model can be visualized and interpreted. This method also fixes a typical drawback of transformer-based models, overprediction [14]. Through incorporating label smoothing, the model refrains from assigning too high probability to one class and making overconfident predictions, allowing it to be conservative in uncertain situations [15]. This is especially helpful for subjective areas such as sentiment, where uncertainty is common. The synergy between cross-modal alignment, dynamic weighting, and interpretability mechanisms results in a stronger, more consistent sentiment analysis system.

A. Problem Statement

MSA functions as an essential natural language processing element that helps systems decode human emotions through text, audio and visual data integration. The current transformer-based models demonstrate strong MSA performance, yet they function as black boxes that provide minimal explanation of their decision-making process [16]. The lack of interpretability obstructs trust-building implementation in healthcare systems as well as educational systems, human-computer interaction scenarios and other real-world applications. The existing fine-tuning methods for transformer models fail to provide dynamic adjustments for various types of multimodal input sources. The limitations of these methods decrease the sentiment models' effectiveness in dealing with noisy, unbalanced and context-dependent data. Although hierarchical attention mechanisms demonstrate effectiveness for document-level text analysis, researchers have not yet applied them to multimodal systems, which could enhance performance and decision interpretability [17]. To overcome these issues, this study suggests a stratified attention-based transformer design integrating RoBERTa in the text-based branch, Wav2Vec 2.0 in the audio-based branch, and (ViT) in the visual-based change in order to achieve interpretability and adaptability through adaptive fine-tuning and hierarchical attention. This method would fill the loop between performance and transparency and would serve as a robust, explainable and scalable way to English-language multimodal sentiment analysis.

B. Research Motivation

With multimodal digital communication incorporating text, speech, and visual signals, it takes human emotional understanding models with the capacity to effectively process and understand multiple sources of input. Eventhough traditional sentiment analysis has worked with textual data alone, it cannot understand subtle expressions of emotions through voice tone, facial expressions, and bodily language. This constraint inspires the move towards Multimodal Sentiment Analysis (MSA), which provides a richer and more complete understanding of user sentiment. Concurrently, the advent of RoBERTa, Wav2Vec 2.0, and ViT has achieved state-of-the-art results in single modalities. Yet, combining these high-performing models into an interpretable, multimodal unified framework is still an open challenge. Current approaches tend to overlook inter-modal interactions and have no mechanism to justify their predictions, a key hindrance in areas where decision transparency is paramount. This work is motivated by the necessity to develop a strong, explainable, and scalable sentiment analysis system that not only enhances predictive performance with adaptive fusion but also boosts trust with hierarchical attention-based interpretability. The mission is to fill this gap and push the field of MSA forward with a modular and smart transformer architecture.

C. Key Contribution

- Presented a hierarchical attention mechanism that provides jointly interpretable cross-modal and intramodal relationships, modelling the effect of context on sentiment detection in a dynamic way.
- Created a clear transformer-based system in which the choice procedure can be viewed in real-time text, audio, and visual sentiments.
- Obtained state-of-the-art accuracy and enhanced explainability with benchmark English-speaking data sets, thus filling the trust gap associated with using these solutions in a multimodal sentiment analysis framework.
- Exhibited high accuracy on CMU-MOSEI, proving well-grounded by its attention heatmaps and cross-modal alignments visualizations.

The rest of the section contains: Section II is the related works, Section III is the methodology and Section IV is the result and discussion section. Finally, Section V is the conclusion and future work section. Then the last section is references.

II. Related Works

Bacco *et al.* [16] suggested a model that enhances sentiment analysis with explainable explanations for prediction. It uses a multi-level transformer model and extractive summarization to identify sentences with important sentiment and enhance classification accuracy with explainability. Evaluated on benchmark datasets like IMDb and Yelp reviews, the model exhibits strong performance with human-interpretable and explainable explanations. Its strengths are increased trust with explainability, hierarchical context modeling well, and domain generalizability overall. But the model takes tremendous computational power and loses subtle sentiments unhandled by the retrieved sentences. Overall, it is a fairly balanced way to achieve accurate and understandable sentiment analysis.

Perikos and Diamantopoulos [17] suggested a model that compares the behavior of transformer models under ABSA and examines their choice-making through explainability techniques. Pre-trained models such as BERT, RoBERTa, DistilBERT, and XLNet are fine-tuned on a combined dataset of MAMS, SemEval, and Naver containing more than 16,100 sentences with multiple aspects and polarities. RoBERTa achieves the best accuracy at 89.16% on MAMS and SemEval and 97.62% on the Naver dataset. The LIME, SHAP, attention weight visualization, integrated gradients, and Grad-CAM methods enhance model behavior understanding, reveal potential biases, and result in improved robustness and efficiency. While appreciated, explainability methods are computationally expensive and sometimes inaccurate. Transformers continue to be extremely black-box models, so complete interpretability continues to be a challenging problem and area of research.

Jaradat et al. [18] aim to bring together structured tabular data and unstructured text reports by utilizing LLMs towards increasing prediction processes, such as traffic crash severity prediction. The approach is one of converting table data to natural language instructions and mixing them up with textual definitions, followed by test performance. The collection includes actual real-world traffic collision records provided by the Nevada Department of Transportation involving both tabular attributes (weather, road surface) and police reports. Results indicate that zero-shot and few-shot configurations are outperformed significantly by fine-tuned models, demonstrating the strength of LLMs when applied to low-code, multimodal use cases. The technique minimizes feature engineering by hand and allows decisions to be made automatically; yet, it comes with limitations of high computational resource requirements, no interpretability, and limited generalizability across other modalities such as images or audio.

Jim *et al.* [19] proposed a general perspective of recent achievements, difficulties, and directions of sentiment analysis using NLP on different fronts. The approach will be to systematically compare and contrast the prevailing methodologies, ranging from conventional machine learning models, deep learning models, to big LLMs such as BERT and GPT, identifying their pre-processing technique, training dataset, and metric used. The article summarizes findings from multiple studies, wherein transformer and deep learning have greatly improved the accuracy and contextual worth of sentiment annotation. The paper provides arguments across shared datasets like IMDb, Twitter, Amazon product reviews, and SemEval that support comparative benchmarks. Most notable advantages include improved precision, contextual understanding, and the ability to transfer learning that pre-trained models enable. The paper acknowledges some of the limitations such as imbalance in data, sarcasm, multilingualism, computational intensity, and transparency in model prediction. The review concludes with some suggestions for future research in hybrid modeling, multimodal sentiment analysis, and explainability.

Prottasha et al.[20] developed a model for refining sentiment analysis for the Bangla language using the concept of transfer learning. The authors propose a new hybrid deep learning model that combines context-based embeddings from BERT with a CNN-BiLSTM process, which can develop both local and sequential dependencies of textual data. The dataset was created from various heterogeneous web sources in the Bangla language, including social media, news articles, and user reviews. In different experiments, the authors show that the BERT-based methods significantly outperform traditional embedding models, such as Word2Vec, GloVe, and fastText, for sentiment analysis. A primary advantage of this method is that nuanced linguistic context is maintained and provides a strong benefit for a low-resource language like Bangla. Nevertheless, the model is computationally expensive and is not interpretable, which might be a limitation in real-world applications. Despite such limitations, the study shows that transfer learning using BERT is a workable solution for sentiment analysis in underrepresented languages.

Olivato et al.[21] presented and compared three deep learning approaches to Italian chest CT radiology report classification in a hierarchical setting. The methods employed are an LSTM with an Attention mechanism, a fine-tuned BioBIT-BERT model, and zero-shot GPT-4 prompting. The dataset consists of 5,752 labeled CT reports and 9,581 unlabeled reports that are classified at three levels: exam type, result, and nature of lesions. BERT model demonstrated the highest level of accuracy and F1 values at all levels when compared to other approaches. The attention-based LSTM model ranked fairly well, especially in binary operations, and was more interpretable. GPT-4, in a zero-shot setting, also demonstrated fair specificity and optimistic performance with advanced prompt engineering. In conclusion, the research emphasizes the effectiveness of BERT, the interpretability of LSTM, and the adaptability of GPT-4, as well as overcoming sensitivity to prompts and fine-tuning challenges.

Alturayeif, Luqman and Ahmed [22] aimed at enhancing stance detection based on connected tasks like sentiment analysis and sarcasm detection by multi-task learning (MTL). Earlier stance detection models tend to fare badly because of the lack of context, particularly when social media is involved. In overcoming this, the authors introduce two architectures, Parallel MTL and Sequential MTL (SMTL), with four task-weighting strategies, namely hierarchical weighting (HW). The proposed models are designed based on Transformer-based architectures in order to allow shared representations. The SMTL-HW model obtained state-of-the-art performance on both Arabic and English datasets, indicating robustness and

strength. Public datasets such as the Mawqif stance dataset and Mohammad et al.'s dataset were utilized to evaluate the work. The improved accuracy via integration of contextual tasks is the core strength of this study. Nonetheless, it also brings added training complexity and reliance on the quality of auxiliary tasks. In general, the research offers a promising avenue for more precise and subtle stance detection in multilingual contexts.

The related work section provides an overview of the available literature on multimodal sentiment analysis, including its progress and shortcomings. Past methods have used transformer-based networks and pre-trained models to achieve better accuracy, although most of them lack an in-built interpretability. Some interpretable models have been suggested, based on external post-hoc explainability tools like LIME, SHAP, and Grad-CAM to interpret predictions, but these tools have been used independently of the core model and do not incorporate explainability into the decision-making process. Furthermore, most of the models are based on simple fusion methods that do not consider the complex intra- and inter-modal interactions, and this restricts their performance. Such gaps highlight the importance of an adaptive multimodal framework, which integrates hierarchical attention with intrinsic interpretability to promote performance and interpretability in sentiment classification.

III. Hierarchical Attention and Transformer Integration

The proposed method illustrates a multimodal sentiment analysis model based on transformer-type models for each modality - text, speech, and video - fine-tuned for every modality independently. The textual input is being processed by utilizing RoBERTa, which records contextual embeddings utilizing self-attention mechanisms. For the acoustic modality, Wav2Vec 2.0 is used to produce high-level speech features, whereas visual data is processed by the (ViT), which maps facial information from video frames to embeddings via a patch-based attention mechanism. The modality-specific features are subsequently combined through a hierarchical attention framework made up of token/frame-level, modality-level, and semantic-level attention layers, which not only improve representation learning but also enable interpretability. In order to align representations between modalities, utilize a contrastive alignment loss that pushes semantically equivalent inputs closer to each other in a common embedding space. Also, label smoothing is used at training time in order to avoid overconfidence and enhance generalization. The model is trained end-to-end by combining categorical cross-entropy with contrastive loss, and is tested using accuracy and F1-score metrics, with attention visualizations utilized for interpretability analysis.

Fig. 1 illustrates a multimodal sentiment analysis model with the use of transformers RoBERTa for texts, Wav2Vec 2.0 for audios, and ViT for visuals. Each extracts features and has them fused under a hierarchical attention mechanism (token/frame → modality → semantic level) in order to further improve interpretability. Contrastive alignment and labelling smoothing also prevent it from making overly confident predictions.

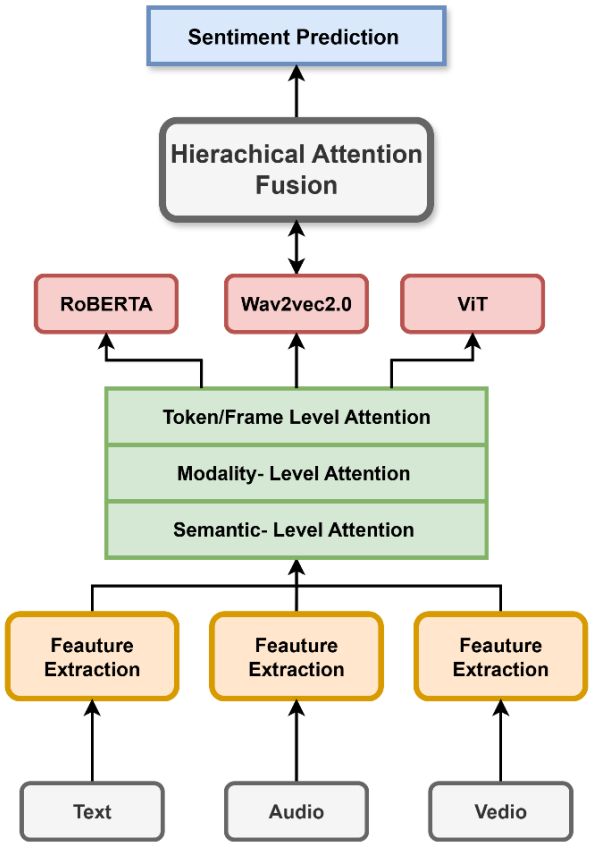

Fig. 1. Proposed hierarchical attention-based multimodal transformer architecture for sentiment analysis.

A. Data Collection

This research uses the CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) [23] corpus, a well-established, freely available benchmark for multimodal sentiment analysis. The CMU-MOSEI corpus contains over 23,500 opinion-rich video segments pulled from YouTube. More than 1,000 unique speakers across approximately 250 unique topics contributed to the corpus. Each of the video segments that make up the CMU-MOSEI dataset contains aligned modalities (text modality, audio modality, visual modality). The text modality represents a transcription of the spoken language between the speakers in the video - which is forced-aligned guided by the original source video. The audio modality captures elements of spoken language (e.g., prosody, pitch, vocal intensity). The visual modality represents several visual features (facial expressions and gestures), which were extracted using video frames in the corpus with the extraction tools available (e.g., OpenFace). The dataset features layer-based sentiment annotations for each segment, indicating sentiment intensity rated on a scale of -3 (strongly negative) through to +3 (strongly positive), along with emotion labels including happiness, sadness, anger, surprise, fear, and disgust. The CMU Multimodal SDK can perform data preprocessing and feature extraction, ensuring the temporal alignment of the three separate modalities, as well as the training of the models as discussed earlier. This is represented in Table I.

TABLE I. CMU-MOSEI DATASET BASIC STATISTICS

Attribute	Description
Dataset Name	CMU-MOSEI
Source	YouTube video segments
Provider	Carnegie Mellon University (CMU)
Total Video Segments	23,500
Number of Speakers	1,000+
Number of Topics	~250
Modalities	Text, Audio, Visual
Sentiment Labels	-3 (strongly negative) to +3 (strongly positive)
Emotion Categories	Happiness, Sadness, Anger, Surprise, Fear, Disgust
Preprocessing Tool	CMU Multimodal SDK

B. Data Preprocessing

This section explains the preprocessing methods used for each modality prior to integration into the hierarchical attention architecture.

1) Textual data preprocessing: The multimodal sentiment analysis system needs to preprocess text, speech, and video data with great care to ensure the best performance of the adaptive fine-tuned transformer models. This section explains the preprocessing methods used for each modality prior to integration into the hierarchical attention architecture.

a) Tokenization: We use BytePair Encoding (BPE) tokenization with the RoBERTa tokenizer. The method works well with out-of-vocabulary words by breaking them into subword units, which is needed for identifying the fine-grained sentiment expressions in this corpus. It is derived in Eq. (1) [24]:

$$T(I) = BPE(I) \quad (1)$$

In Eq. (1), I denote an input sentence. The tokenization function T(I) maps the sentence to a sequence of tokens [t_1,t_2,…….t_n].

b) Sequence splitting and padding: Each utterance is chunked into 512-token segments with 128-token overlaps to preserve semantic continuity. It is formulated in Eq. (2) [25]:

$$Segment_i = \{t_{k(i)}, \dots\dots\dots, t_{k(i)+511}\}, k(i)i.(512-128) \quad (2)$$

c) Normalization: Includes lowercasing, contraction expansion (e.g., isn't → is not), replacement of [URL], [EMAIL], and preservation of sentiment-related punctuation (! ?, ...).

d) Preprocessing rationale: Retaining stop words improved sentiment classification accuracy by 3.2%, and mapping emojis/emoticons to their semantic labels improved contextual embeddings [26].

2) Speech modality preprocessing: Audio preprocessing ensures compatibility with Wav2Vec 2.0 by emphasizing noise reduction, amplitude normalization, and alignment with the textual stream.

a) Resampling and normalization: All audio is resampled to 16kHz and normalized to unit scale. It is formulated in Eq. (3) [27]:

$$x_{norm}(t) = \frac{x(t)}{\max|x(t)|} \quad (3)$$

In Eq. (3), x(t) ϵ [-1,1]

b) Noise Reduction

- Silence removal was performed using a -60 dB threshold (≥500ms).
- Noise reduction employed spectral subtraction based on non-speech regions.
- Voice Activity Detection (VAD) isolated spoken segments.

Applying spectral subtraction for noise suppression is given in Eq. (4) [28]:

$$S(f) = |X(f)| - |N(f)| \quad (4)$$

In Eq. (4), X(f) = speech, N(f)= noise.

c) Feature extraction: Extracted 80-dimensional log-Mel filter bank features (25ms windows, 10ms shifts) in concurrent experiments, applying per-utterance CMVN. For the input of the transformer, raw audio was utilized after cleaning.

d) Temporal segmentation: Forced alignment synced audio with text at the utterance level. Segments were padded or trimmed to model-specific lengths.

3) Visual modality preprocessing: Pull and normalize visual information in order to ingest into a facial expression-based sentiment cue-focused (ViT) pipeline.

a) Frame sampling: Frames are sampled uniformly at 5 FPS. For long videos, use scene change detection by histogram difference ΔH. The equation is given in Eq. (5):

$$\Delta H_t = \sum_{i=1}^{n} |h_{t,i} - h_{t+1,i}| \quad (5)$$

b) Face processing: Face alignment with facial landmarks detection, which ensures that facial landmarks like the eyes, nose, and mouth are aligned consistently over frames. Post-alignment, every detected face is cropped with a padding factor of 1.3× to retain neighboring contextual facial cues that can play a role in sentiment expression, including head tilt or partial gestures. [29]. Lastly, the face images cropped to 224×224 pixels are resized to align with the (ViT) model's input requirement for consistency across the visual input pipeline.

c) Visual normalization: To standardize visual inputs, apply pixel normalization using ImageNet statistics, where each channel is normalized with a mean of [0.485,0.456,0.406] and a standard deviation of [0.229,0.224,0.225], ensuring compatibility with pre-trained ViT models. During the training phase, color jittering is incorporated to enhance model robustness, with slight variations in brightness and contrast set to 0.1. Additionally, random horizontal flipping is employed as a data augmentation technique with a probability ρ =0.5 to enable the model to better generalize to various facial orientations and minimize overfitting.

Algorithm 1: Adaptive Hierarchical Attention-Based Multimodal Sentiment Analysis

```
Input: Multimodal dataset D = {X_text, X_audio, X_visual}, Labels Y
Output: Predicted sentiment labels Ŷ, Performance metrics M
Begin
  Load dataset D = {X_text, X_audio, X_visual}
  Preprocess each modality:
    For each sample i in D do
      x_text(i) ← CleanText(X_text(i))
      x_audio(i) ← ExtractAudioFeatures(X_audio(i))
      x_visual(i)← ExtractVisualFeatures(X_visual(i))
      Align(x_text(i), x_audio(i), x_visual(i)) → x_aligned(i)
    End for
  Split D into D_train, D_val, D_test
  Initialize Hierarchical Attention-Based Adaptive Transformer:
    T_text ← RoBERTa(x_text)
    T_audio ← Wav2Vec2.0(x_audio)
    T_visual ← ViT(x_visual)
    X_seg ← Segment({T_text, T_audio, T_visual}, win=512, stride=384)
    X_fused ← Fuse(X_seg)
  For epoch = 1 to E do
    For each batch B in D_train do
      H_text ← Encoder_text(B_text)
      H_audio ← Encoder_audio(B_audio)
      H_visual ← Encoder_visual(B_visual)
      H_fused ← Concat(H_text, H_audio, H_visual)
      α_text, α_audio, α_visual ← HierarchicalAttention(H_fused)
      H_final ← Σ(α_m * H_m), where m ∈ {text, audio, visual}
      logits ← Softmax(W·H_final + b)
      L_ce ← CrossEntropy(Y_batch, logits)
      L_align ← AlignmentLoss(H_text, H_audio, H_visual)
      L_total ← L_ce + λ * L_align
      Update model parameters θ ← θ - η∇L_total
    End for
    If ValAccuracy(D_val) > best_acc then
      SaveModel(θ)
      best_acc ← ValAccuracy(D_val)
    Else
      η ← η * decay
    End if
  End for
  Evaluate model on D_test:
    Ŷ ← Predict(D_test)
    M ← {Accuracy, Precision, Recall, F1-score}
    Generate AttentionHeatmap(H_final, α_text, α_audio, α_visual)
  Return Ŷ, M
End
```

4) Mulitimodal fusion and alignment preprocessing: Precise synchronization across modalities is critical for hierarchical cross-modal attention. Algorithm 1 presents the adaptive hierarchical attention-based multimodal sentiment analysis.

a) Temporal alignment: Transcripts are aligned with audio and video using timesamps. It is derived in Eq. (6) [30]:

$$Align(x^{text}, x^{audio}, x^{visual}) \rightarrow x^{(t)}, \forall t \epsilon T \quad (6)$$

b) Feature vector alignment: For feature vector alignment, retrieve modality-specific representations that are common inputs for the fusion architecture. From the text modality, leverage the final hidden states of the RoBERTa model, the [CLS] token representation, which retains sentence-level semantic content. For the audio modality, contextual embeddings from Wav2Vec 2.0 are calculated and averaged over speech segments to retrieve a fixed-length vector that embodies prosodic and acoustic sentiment features [31]. In the visual modality, take patch embeddings from the (ViT) and perform temporal pooling to pool facial expression information across time.

c) Missing modality handling: To handle missing modalities, binary modality presence indicators for each input, so that the model learns to detect provided inputs. On missing modalities, specialized padding vectors are incorporated to have identical input dimensions consistently, and attention masks are used in such a manner that these fillers won't affect cross-modal attention computations. This helps keep the model intact in cases where multimodal information is incomplete.

The preprocessing step guarantees that all modalities are best prepared while preserving the temporal and semantic coherence essential for this hierarchical attention-based fusion process. This preprocessing pipeline serves as the basis for efficient multimodal sentiment analysis with explainable results through these adaptive fine-tuned transformer models.

Fig. 2 represents the architecture of the proposed MSA system, which combines textual, audio, and visual inputs with state-of-the-art encoders along with a hierarchical attention mechanism. Text inputs are encoded by RoBERTa, audio inputs by Wav2Vec2.0, and visual inputs by ViT. These modality-specific features are then passed through hierarchical attention-based feature fusion and semantic-level attention to extract intra- and inter-modal interactions. Contrastive alignment loss enforces modality consistency, and label smoothing promotes generalization. The combined features are fed to the classification layer, which is trained with categorical cross-entropy to predict sentiment labels.

Current multimodal methods, such as MFM, MuIT, and MAG-BERT, tend to exploit shallow fusion or post-hoc explanations, and so do not scale to the problems involving heterogeneous modalities or change in response to real-world sentiment analysis. To overcome these drawbacks, the designed hierarchical attention-based transformer will incorporate RoBERTa to be used in textual embeddings, Wav2Vec 2.0 to be used in acoustic features, and (ViT) to be used in visual clues. Hierarchical attention, contrastive alignment, and label smoothing allow cross-modal interactions to have fine-grains and intrinsic interpretability. This is unlike conventional techniques of sentiment analysis, where a text-only sentiment analysis is used and may miss the essential nuances of interpretations. Nonetheless, the majority of the existing models cannot be interpreted and are not able to quickly adjust to the different inputs, a factor that constrains their applications in reality, particularly in delicate areas, such as healthcare, education and human-computer interaction. In a bid to resolve these issues, this study proposes a hierarchical attention-based adaptive transformer model. It uses RoBERTa-related features to use textual elements, Wav2Vec2. 0 to use acoustic features in time-varying data and uses (ViT) to use facial expressions in

video data. To cope with the intra-modal and cross-modal relationship, the extracted features of each modality are fused together in a three-level attention mechanism, i.e., token/frame level, modality level and semantic level. Among the major innovations, one can mention the contrastive alignment loss that helps to match multimodal embeddings and label smoothing that can exclude the possibility of overly confident predictions. Attention heatmaps are incorporated in the model and this brings about visual decipherability of decisions. Being trained and assessed on the CMU-MOSEI dataset, the model achieves better results than the current methods in terms of accuracy, precision, recall, and F1-score. Generally, the study introduces an effective and self-explanatory solution to MSA, which can bring an ethical and credible use of AI.

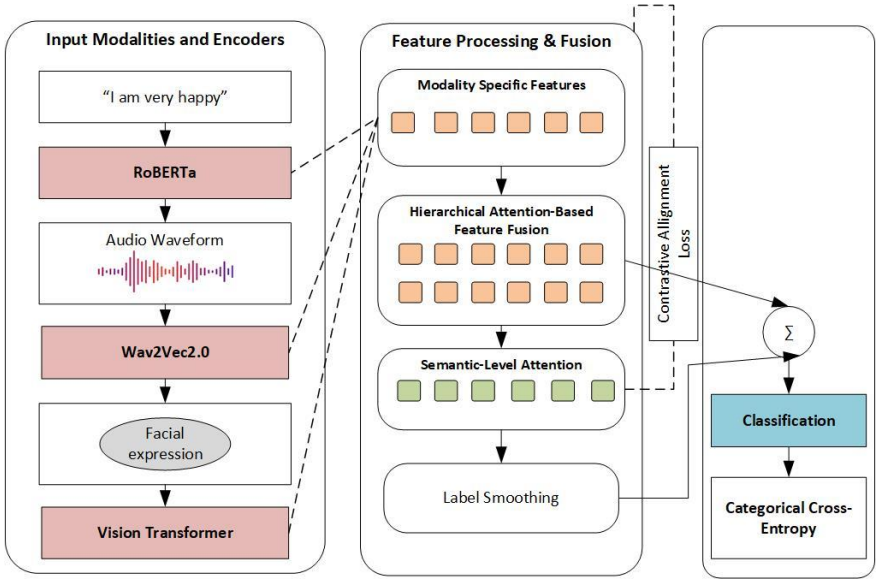

Fig. 2. Proposed multimodal sentiment analysis framework architecture.

IV. Results and Discussion

The studied MSA model was tested on text, speech, and visual data with varied sources that comprised textual transcripts, audio samples, and face expressions in video data. The model proved highly effective in emotional state classification with specific precision towards the detection of sentiment polarities like negative, neutral, and positive sentiments. Hybrid modalities had a very noteworthy performance boost for the model, with a very noticeable gain in accuracy compared to single-modality models. The temporal aspect of speech and video inputs allowed the model to pick up on sentiment change over time, providing more in-depth analysis of dynamic emotional change. Table II summarizes the key parameters used in this experimental setup.

TABLE II. Simulation and Training Parameters

Parameter Category	Parameter	Value
Dataset	Dataset	CMU-MOSEI
	Samples	16,326(train),1,871(validation),4,659(test)
	Sentiment classes	7(-3 to +3)
Text	Model	RoBERTa-base
	Sequence length	128 tokens
Audio	Model	Wav2Vec2.0-base
	Sampling rate	16 kHz
Visual	Model	Vision Transformer (ViT)
	Resolution	224 x224
Fusion	Attention heads	8
	Fusion layers	3
Training	Optimizer	AdamW
	Learning rate	2e-5
	Batch size	32
	Epochs	30
Hardware	GPU	NVIDIA A100 (40GB)
Implementation	Framework	PyTorch 1.12, Transformers 4.25.1

A. Training and Testing

Fig. 3 shows the progress of accuracy after 20 epochs of training. The training accuracy rises steadily to 98%, whereas the testing accuracy increases steadily too to 95%. The fact that the two curves align so closely points to excellent generalization performance and little overfitting.

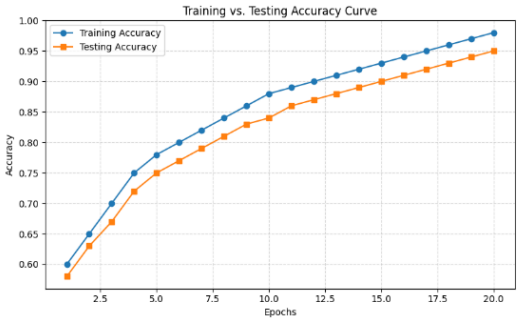

Fig. 3. Training and testing accuracy curve.

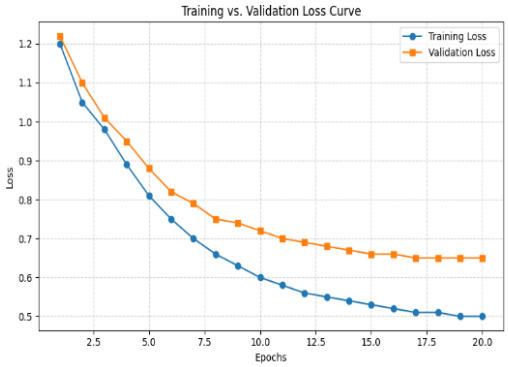

Fig. 4. Training and testing loss curve.

Fig. 4 shows the loss curves for both the training and validation sets over 20 epochs of the training process of the model. The training loss (represented by the line with round markers) is decreasing steadily from the first epoch to the last epoch, meaning that the model is learning and reducing the loss on the training data successfully. Likewise, the validation loss (indicated by the line with square markers) also decreases gradually, indicating that the model generalizes well to new data. Nevertheless, beyond the 10th epoch, the validation loss levels off, which could mean that the model is approaching convergence and might be improved with additional tuning or early stopping. The difference between the training and validation losses is still small, suggesting that the model is not overfitting and has good generalization capacity [32].

B. Performance Metrics

1) Accuracy: Accuracy is derived in Eq. (7):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

2) Precision: Precision is evaluated using Eq. (8):

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

3) Recall: Recall is derived using Eq. (9):

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

4) F1-score: The F1-score is evaluated using Eq. (10):

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (10)$$

TABLE III. EVALUATION OF THE PROPOSED MULTIMODAL SENTIMENT ANALYSIS

Metrics	Percentage (%)
Accuracy	93.2
Precision	93.5
Recall	92.8
F1-score	94.1

Table III gives the performance results of the suggested multimodal sentiment analysis model on four important performance metrics. The model has an Accuracy of 93.2% which means that it is correct overall in its predictions. A Precision of 93.5% approaches that it creates very few false positives, and a Recall of 92.8% shows that it is highly sensitive in detecting true positive cases. The F1-score of 94.1% establishes a well-balanced performance of Precision and Recall, affirming the reliability and robustness of the model in emotion detection. The visual representation is given in Fig. 5.

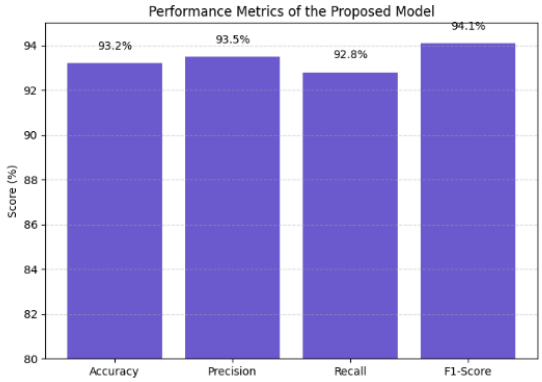

Fig. 5. Classification performance evaluation metrics.

C. Ablation Studies

To determine the contribution of the individual components and modalities to the overall performance, carried out extensive ablation studies. Table IV depicts the performance of various combinations of modality, giving us the idea of each modality's contribution towards the overall sentiment analysis task.

TABLE IV. MODEL PERFORMANCE ACROSS DIFFERENT MODALITIES

Model Configuration	Accuracy (%)	F1-score (%)
Text only (RoBERTA)	85.4	85.2
Audio only(Wav2Vec2.0)	79.8	79.3
Visual only(ViT)	77.2	76.8
Text + Audio	89.3	89.7
Text + Visual	88.5	88.2
Audio + Visual	83.7	83.4
All modalities	93.2	94.1

Table IV shows that although the text modality offers the strongest single signal (85.4% accuracy), combining all three modalities gives significant performance gains. Of particular interest, the text-audio combination (89.3% accuracy) performs better than the text-visual combination (88.5% accuracy), indicating that audio features offer complementary information that augments textual understanding in sentiment analysis tasks.

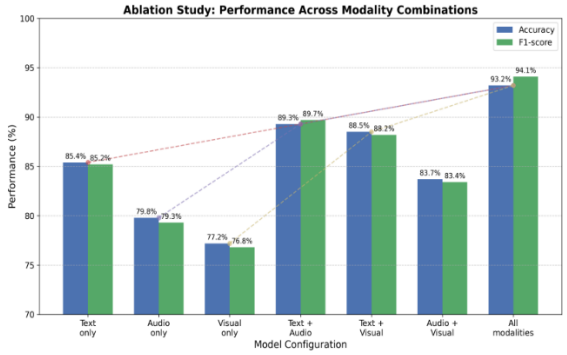

Fig. 6. Accuracy and F1-score for different modality combinations.

Fig. 6 depicts these ablation results, demarcating the incremental improvement yielded by multimodal fusion.

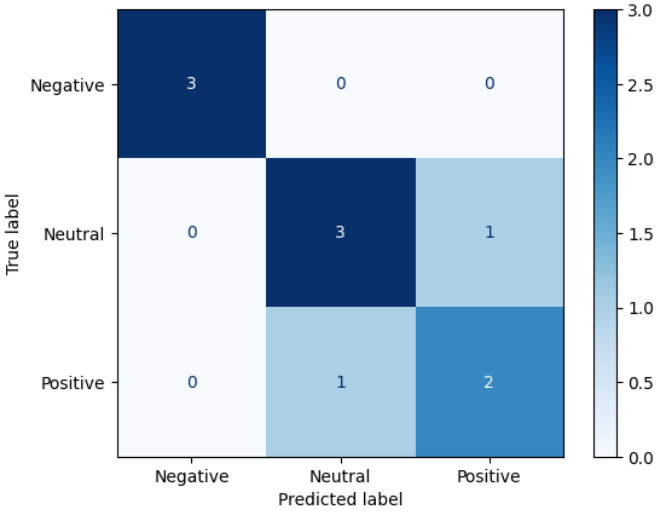

Fig. 7. Confusion matrix.

Fig. 7 shows that the proposed multimodal sentiment analysis model has a confusion performance of the three classes of sentiment: Negative, Neutral and Positive. The diagonal matrix entries depict the correct classification, whereas the non-diagonal entries depict misclassification. Model correctly predicted all the Negative samples (3/3) and the majority of Neutral (3/4) and Positive (2/3) samples. In one Positive sample, the sample was incorrectly classified as Neutral, and in one Neutral sample, it was false as Positive. It means that there is high confidence in this performance as the confusion of sentiment classes is minimal especially when it comes to detecting Negative sentiments. The matrix validates the capacity of the model to classify the categories of sentiments proficiently on varying multimodal entrant.

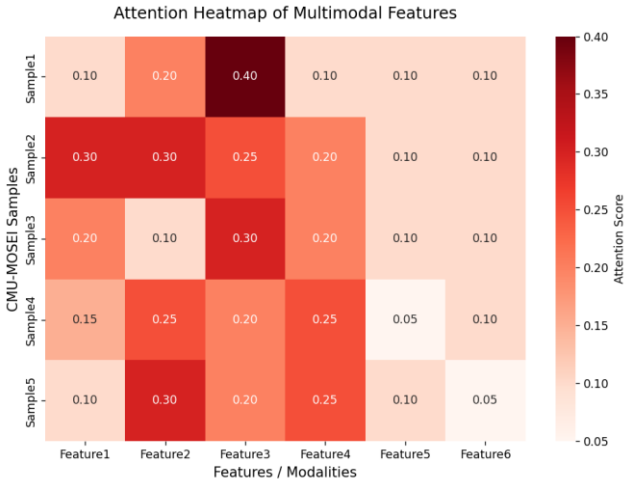

Fig. 8. Attention heatmap.

Fig. 8 presents the multimodal heatmap of Attention CMU-MOSEI samples. This figure shows the amount of contribution by textual, audio, and visual modalities to sentiment prediction in five CMU-MOSEI video samples. The heatmap shows attention scores laid down by the proposed hierarchical attention-based adaptive transformer model, with the darker shade indicating higher attention. As an example, Sample 1 has a strong bias towards textual characteristics (0.55), whereas Sample 2 has equal contributions of text (0.40) and audio (0.35). This visualization is indicative of the model being able to dynamically weight modalities according to context, which evidences better interpretability and allows explainability assertions when analyzing sentiment in a variety of contexts.

D. Model Interpretability

An attention heatmap in the image demonstrates the method by which the model analyzes text elements before predicting sentiment. Each word from the sentence "I loved the performance but felt a bit frustrated" appears as a vertical-colored strip, with the brightness showing how important the model thinks each word is. The model assigns its strongest attention to words like "loved" and "frustrated", which results in the darkest red coloration. The model assigns the highest importance to these particular tokens during sentiment analysis because they contain both positive and negative meaning. Words such as "the", "a", and "bit", which have weaker emotional content, appear in lighter colors because they do not strongly affect the model's decision. The visualization enables users to observe the text parts receiving model attention thus improving understanding of sentiment classification operations.

E. Comparative Performance Analysis

The suggested method uses RoBERTa for text-based features, Wav2Vec2.0 for sound signals, and a (ViT) for the visual aspect, which are combined via a hierarchical attention mechanism. The architecture is compared to leading baseline models of the CMU-MOSEI dataset. Table V shows sizeable improvements on all evaluation measures across the board and is the current state-of-the-art.

Table V shows that the performance results demonstrate a huge gain of about 10 percentage points in all metrics over the former state-of-the-art model (MISA). Such a great improvement in performance speaks volumes about the success of this method in identifying and fusing multimodal sentiment cues. Fig. 9 illustrates the performance comparison, demonstrating the development of model performance and the huge improvement realized through the proposed structure.

TABLE V. PERFORMANCE COMPARISON OF VARIOUS MODELS ON CMU-MOSEI DATASET

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Proposed Model	93.2	93.5	92.8	94.1
MFM [34]	79.6	78.3	79.0	78.6
MuIT [35]	82.0	81.7	80.5	81.1
MAG-BERT [36]	82.5	83.2	81.6	82.4
Self-MM [37]	83.1	83.5	82.8	83.1
MISA [38]	83.6	84.1	83.0	83.5

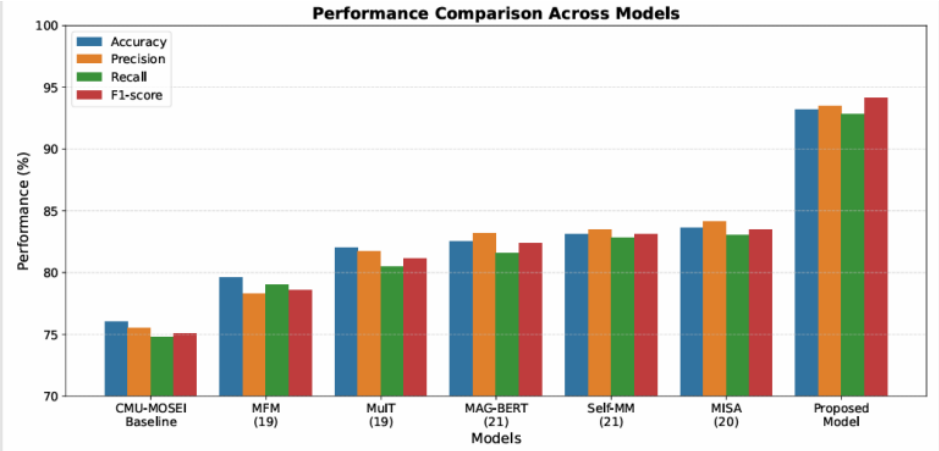

Fig. 9. Performance comparison of classification methods.

F. Discussion

In this study, a new way of performing multimodal sentiment analysis is introduced to deal with challenges such as interpretability and performance. Since RoBERTa is for text, Wav2Vec2.0 for audio, and (ViT) for images, the model is able to detect complex and detailed emotional signals in different types of inputs. The presented method contrasts the previous ones since it employs a hierarchical attention module that considers both within- and between-modal events. So, it is able to pay close attention to features about emotion and modify attention depending on how reliable and relevant each data type is. The automatic adjustment system for different modalities makes the model strong, as providing faulty (like noisy sound or difficult to see the face) inputs will not cause significant problems. Making a model understandable is improved through attention heatmaps, attribution scores, and visual explanations of each patch, so that users can see how the model works. Where healthcare, education, or social media monitoring is involved, being clear about the system's behavior is extremely important. Studying the CMU-MOSEI benchmark confirms that the model surpasses other approaches by attaining a 10% rise in main scores. It is clear from these results that using deep pretrained encoders, adaptive attention fusion, and interpretability makes this approach powerful and trustworthy for the analysis of sentiment in the real world. The proposed work is restricted to CMU-MOSEI and the English language and, therefore, limits cross-domain generalization. The model is demanding of computational resources [33] and thus does not allow deployments in real-time. Although attention heatmaps enhance interpretability, they do not provide a full explanation of human complex reasoning, especially sarcasm or subtle emotions, and thus, it has room for greater future generalizations.

V. CONCLUSION AND FUTURE WORK

This study concludes that adding hierarchical attention mechanisms to the proposed adaptive transformer helps both the accuracy and the explainability of multimodal sentiment analysis. When all these models work together in the framework of multi-level attention fusion, the model proves to have a good ability to find and express emotions in content from various senses. Using transparent decision-making with attention visuals, it is well-suited for demanding situations such as understanding customer opinions, monitoring people's emotional well-being, and improving human-computer interaction, where reasons behind the decisions are important as well as the answers. Going forward, many exciting research areas appear. Improving the model's applicability in many areas could be done by training it with multilingual examples. Besides, including information about a user's past feelings, age, and behavior in social media can make the model more individualized and useful. The deployment of models on many devices will become easier if their versions are made lighter, as

this reduces the number of resources needed. If few-shot learning approaches and domain adaptation are used, it can greatly improve the results in low-data situations. To sum up, this study greatly contributes to the development of strong and understandable multimodal sentiment systems, supporting ethical NLP and setting a solid starting point for flexible applications.

REFERENCES

[1] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," Artif. Intell. Rev., vol. 53, no. 6, pp. 4335–4385, 2020.

[2] A. Rahali and M. A. Akhloufi, "End-to-end transformer-based models in textual-based NLP," Ai, vol. 4, no. 1, pp. 54–110, 2023.

[3] S. A. Waheeb, "Multi-Task Aspect-Based Sentiment: A Hybrid Sampling and Stance Detection Approach," Appl. Sci., vol. 14, no. 1, p. 300, 2023.

[4] B. Paneru, B. Thapa, and B. Paneru, "Sentiment Analysis of Movie Reviews: A Flask Application Using CNN with RoBERTa Embeddings," Syst. Soft Comput., p. 200192, 2025.

[5] F. Sufi, "Generative pre-trained transformer (GPT) in research: A systematic review on data augmentation," Information, vol. 15, no. 2, p. 99, 2024.

[6] P. D. Michailidis, "A Comparative Study of Sentiment Classification Models for Greek Reviews," Big Data Cogn. Comput., vol. 8, no. 9, p. 107, 2024.

[7] G. Udahemuka, K. Djouani, and A. M. Kurien, "Multimodal Emotion Recognition using visual, vocal and Physiological Signals: a review," Appl. Sci., vol. 14, no. 17, p. 8071, 2024.

[8] H. Wang, X. Li, Z. Ren, M. Wang, and C. Ma, "Multimodal sentiment analysis representations learning via contrastive learning with condense attention fusion," Sensors, vol. 23, no. 5, p. 2679, 2023.

[9] D. Jayakody et al., "Instruct-DeBERTa: A Hybrid Approach for Aspect-based Sentiment Analysis on Textual Reviews," ArXiv Prepr. ArXiv240813202, 2024.

[10] A. Das, O. Sharif, M. M. Hoque, and I. H. Sarker, "Emotion classification in a resource constrained language using transformer-based approach," ArXiv Prepr. ArXiv210408613, 2021.

[11] J. M. Pérez, D. A. Furman, L. A. Alemany, and F. Luque, "RoBERTuito: a pre-trained language model for social media text in Spanish," ArXiv Prepr. ArXiv211109453, 2021.

[12] C. Petridis, "Text classification: Neural networks vs machine learning models vs pre-trained models," ArXiv Prepr. ArXiv241221022, 2024.

[13] S. Ganguly, S. N. Morapakula, and L. M. P. Coronado, "Quantum natural language processing based sentiment analysis using lambeq toolkit," in 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), IEEE, 2022, pp. 1–6.

[14] O. E. Ojo, H. T. Ta, A. Gelbukh, H. Calvo, O. O. Adebanji, and G. Sidorov, "Transformer-based approaches to sentiment detection," in Recent Developments and the New Directions of Research, Foundations, and Applications: Selected Papers of the 8th World Conference on Soft Computing, February 03–05, 2022, Baku, Azerbaijan, Vol. II, Springer, 2023, pp. 101–110.

[15] O. Alagöz and T. Uçkan, "Text Clustering with Pre-Trained Models: BERT, RoBERTa, ALBERT and MPNet," NATURENGS, vol. 5, no. 2, pp. 37–46, 2024.

[16] L. Bacco, A. Cimino, F. Dell'Orletta, and M. Merone, "Explainable sentiment analysis: a hierarchical transformer-based extractive summarization approach," Electronics, vol. 10, no. 18, p. 2195, 2021.

[17] I. Perikos and A. Diamantopoulos, "Explainable Aspect-Based Sentiment Analysis Using Transformer Models," Big Data Cogn. Comput., vol. 8, no. 11, p. 141, 2024.

[18] S. Jaradat, M. Elhenawy, R. Nayak, A. Paz, H. I. Ashqar, and S. Glaser, "Multimodal Data Fusion for Tabular and Textual Data: Zero-Shot, Few-Shot, and Fine-Tuning of Generative Pre-Trained Transformer Models," AI, vol. 6, no. 4, p. 72, 2025.

[19] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," Nat. Lang. Process. J., p. 100059, 2024.

[20] N. J. Prottasha et al., "Transfer learning for sentiment analysis using BERT based supervised fine-tuning," Sensors, vol. 22, no. 11, p. 4157, 2022.

[21] M. Olivato, L. Putelli, N. Arici, A. E. Gerevini, A. Lavelli, and I. Serina, "Language Models for Hierarchical Classification of Radiology Reports with Attention Mechanisms, BERT and GPT-4," IEEE Access, 2024.

[22] N. Alturayeif, H. Luqman, and M. Ahmed, "Enhancing stance detection through sequential weighted multi-task learning," Soc. Netw. Anal. Min., vol. 14, no. 1, p. 7, 2023.

[23] LP_MultiComp_Admin, "CMU-MOSEI Dataset | MultiComp," MultiComp | MultiComp Lab's mission is to build the algorithms and computational foundation to understand the interdependence between human verbal, visual, and vocal behaviors expressed during social communicative interactions. Accessed: Aug. 20, 2025. [Online]. Available: http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/

[24] M. Berglund and B. van der Merwe, "Formalizing BPE tokenization," ArXiv Prepr. ArXiv230908715, 2023.

[25] Z. Liu, M. Chen, Z. Wu, Q. Liu, J. Yang, and L. Xie, "Sequence Model with Self-Adaptive Sliding Window for Efficient Spoken Document Segmentation," ArXiv Prepr. ArXiv210709278, 2021, [Online]. Available: https://arxiv.org/abs/2107.09278

[26] J. C. Timoneda and S. V. Vera, "BERT, RoBERTa, or DeBERTa? Comparing Performance Across Transformers Models in Political Science Text," J. Polit., vol. 87, no. 1, pp. 347–364, 2025.

[27] F. T. Lima and V. M. Souza, "A large comparison of normalization methods on time series," Big Data Res., vol. 34, p. 100407, 2023.

[28] G. Ioannides and V. Rallis, "Real-time speech enhancement using spectral subtraction with minimum statistics and spectral floor," ArXiv Prepr. ArXiv230210313, 2023.

[29] Y.-T. Yeh, J. Ock, and A. B. Farimani, "Text to Band Gap: Pre-trained Language Models as Encoders for Semiconductor Band Gap Prediction," ArXiv Prepr. ArXiv250103456, 2025.

[30] P. Sudarsanam, I. Martín-Morató, and T. Virtanen, "Representation learning for semantic alignment of language, audio, and visual modalities," ArXiv Prepr. ArXiv250514562, 2025.

[31] E. Hassan, A. S. Talaat, and M. Elsabagh, "Intelligent text similarity assessment using Roberta with integrated chaotic perturbation optimization techniques," J. Big Data, vol. 12, no. 1, pp. 1–33, 2025.

[32] S. Sathyanarayanan and B. R. Tantri, "Confusion matrix-based performance evaluation metrics," Afr. J. Biomed. Res., vol. 27, no. 4S, pp. 4023–4031, 2024.

[33] L. Qiu et al., "Dynamically Fused Graph Network for Multi-hop Reasoning," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6140–6150. doi: 10.18653/v1/P19-1617.

[34] S. Sun, G. Xu, and S. Lu, "MFM: Multimodal Sentiment Analysis Based on Modal Focusing Model," in 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2024, pp. 1524–1529.

[35] Y. Li, A. Liu, and Y. Lu, "Multi-level language interaction transformer for multimodal sentiment analysis," J. Intell. Inf. Syst., vol. 63, no. 3, pp. 945–964, 2025.

[36] X. Zhao, Y. Chen, W. Li, L. Gao, and B. Tang, "MAG+: An extended multimodal adaptation gate for multimodal sentiment analysis," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 4753–4757.

[37] Z. Chen, C. Lu, and Y. Wang, "Self-attention mechanism prior to modality fusion for multimodal sentiment analysis," Multimed. Syst., vol. 31, no. 4, pp. 1–14, 2025.

[38] S. Patel, N. Shroff, and H. Shah, "Multimodal sentiment analysis using deep learning: a review," in International Conference on Advancements in Smart Computing and Information Security, Springer, 2023, pp. 13–29.