International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2) 2026 pp. 138-157

https://doi.org/10.70917/ijcisim-2026-2185
Article

TSSP-MSA: DEEP LEARNING LEVERAGED TRI-STAGE SELF-SUPERVISED FRAMEWORK FOR ROBUST MULTIMODAL SENTIMENT ANALYSIS

Bharathi Niruti[1], Sujith AVLN[2]*, Kanaka Durga Returi[3]

[1] Department of CSE, Malla Reddy University, Hyderabad, India; bharathin2020@gmail.com

[2]* Department of IT, Malla Reddy University, Hyderabad, India; Corresponding author: Sujith AVLN; sujeeth.avln@gmail.com

[3] Department of CSE, Malla Reddy Technical Campus (A constituent unit of Malla Reddy Vishwavidyapeeth), Deemed to be University, Hyderabad, India; durga1210@gmail.com

Abstract: Multimodal sentiment analysis (MSA) is designed to understand human emotions by analyzing modal signals of various forms of data including text, audio, and visual. Nonetheless, the current models tend to have difficulties in learning modality-specific features, working with incomplete data, and aligning semantics across modalities. To overcome these shortcomings, in this paper, the authors present TSSP-MSA, a tri-stage self-supervised model that aims to enhance the quality of the unimodal features, the adaptability of fusion, and the consistency across modalities. The unimodal encoders are then pretrained in the first stage with self-supervised goals to learn robust and semantically rich feature representations. The second step introduces an expert mixture fusion strategy with uncertainty awareness that dynamically balances modality contributions in terms of uncertainty and hence improves tolerance to missed or noisy data. The last step employs cross-modal contrastive refinement, which synchronizes modal representations in a common latent space, overcoming semantic discordance. Comprehensive testing on standard datasets like CMU-MOSI and CMU-MOSEI shows that TSSP-MSA achieves substantially better accuracy, F1-score, and a modality-dropping reshape than state-of-the-art by a large margin. The architecture proposed advances the idea of interpretable, robust affective computing devices, with high prospects of being utilized in real practice of human-computer interaction.

Keywords: Multimodal Sentiment Analysis, Self-Supervised Learning, Unimodal Representation, Modality Fusion, Contrastive Learning, Robustness, Affective Computing.

1. Introduction

Sentimental analysis that has conventionally been on unimodal text content has expanded to form the richer domain of multimodal sentiment analysis (MSA) where multiple data sources like text content, vocal tones, facial expressions and other non-verbal channels can be combined to convey the sentiment more congruently. The contextual nonexistence that characterizes unimodal systems, such as a lack of contextual cues with the analysis of a text only, is alleviated by replacing them with multiple modalities, where each modality could be supplementary or add up to the other [1, 2]. Consider, for example, sarcasm in text; this naturally makes more sense with the addition of intonation in speech, or micro-expressions in a face. Such a mixture of heterogeneous information sources has been proved to greatly enhance the level of sentiment prediction accuracy and therefore MSA is a required part of affective computing, human-computer interaction, and social media analytics.

The introduction of deep learning structures like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and most lately, Transformers, has brought a shift in paradigm to the sphere of MSA [3,4]. Such architectures have also allowed automated high level semantic feature extraction of each modality and more advanced feature fusion mechanisms. As studies like [1] shows that multi-layer feature fusion coupled with multi-task learning enables multi-level sentiment information with retaining modality-specific subtleties. This structural learning of hierarchical representation provides a solution to a problem such as variety in the expression of sentiments due to their varying expression in different modalities. In spite of the great advances, essential multimodal integration is a central challenge. This is because in real-life scenarios the data may not be of the highest quality, the audio signal can be accompanied with background noise, pictures by low light, and text can employ slangs and abbreviations. Such inconsistencies may lead to modality dominance; whereby noisy modality has an overwhelming influence on the fusion outcome [7,8]. Moreover, the missing modality scenarios wherein some or all the modalities are not available are challenging robustness requirements. The practice of conventional fusion, including simple concatenation or fixed attention weighting, is not capable of adapting in a dynamic

fashion to such variations, resulting in loss of performance. The main flaw of the majority of existing MSA models is the inability to generalize to novel domains and the actual environments [4]. Training data characteristics can be specific to a domain, like a particular language flavour or demographic bias, etc, and are not representative of the target deployment context. More recent works have investigated the test-time adaptation methods, in which models can adjust to new domains at test-time through self-supervised objectives or pseudo-label generation [4]. This may greatly decrease the reliance on vast amounts of labelled target domain data. The other important innovation space in MSA is the application of cross-modal attention, where models are selective about paying attention to relevant aspects of one modality using cues in another [3,6]. This flexible matching is useful at the small-scale level, solving dependencies like the connection between certain words and facial expressions or alterations in the tone of voice. Nevertheless, the attention mechanisms are also subject to noise; when the system heavily utilizes the compromised modality, the attention focus will be distorted to increase the effect of noise instead of reducing noise.

The low availability and prohibitive annotation cost of multimodal sentiment datasets have led to the promotion of self-supervised learning (SSL) in MSA [6,7]. SSL allows models to develop a rich representation on large amounts of unlabelled multimodal data via the design of pretext tasks, including modality mask, temporal order prediction, and cross-modalities matching. These approaches have proved to be more generalizable and robust, especially in low-resource scenarios. Nevertheless, majority of current SSL techniques in MSA are uni-modal or incomplete in utilization of inter-modality relation in a stage-to-stage concept. Uncertainty modelling, especially aleatoric uncertainty that represents intrinsic noise on the data, has also led to robustness in MSA [8]. In such a way that occurs through explicitly estimating uncertainty per modalities, adopting adaptive fusion weights allows models to minimise the effect of noisy signals. These methods overcome the modality dominance problem, and the fusion decisions can be made more reliably in difficult circumstances.

Multi-layer fusion and deep metric learning have also been emphasized in recent work on robust MSA. Sentiment information is extracted at varied levels of abstraction in multi-layer fusion and the embedding space is optimized towards a clear separation of the classes of sentiment in metric learning. [10] reveal that deep metric learning models can successfully deal with class imbalance and inter-class ambiguity by ensuring intra-class compactness and inter-class separability. Modality quality adjustment is a new avenue that is looking promising. As one example, [9] suggest dynamic brightness alteration during image fusion tasks which, when generalized into multimodal sentiment analysis, might enable time-varying adaptation to image-representation peculiarities. Combining such adaptive methods and sentiment modelling would potentially provide greater robustness.

The other line of research is noise tolerant architectures including the language-dominated noise-resistant learning network Highway [7], inspired by language-dominated noise-resistant learning which incorporates noise filtering and attention reweighting. These designs are essential where user-generated content can produce weak sentiment signals easily obscured by noise in user-generated content. Though every of those advances focuses on solving separate pieces of the puzzle of the MSA problem, a gap persists: It has yet to be proposed to develop comprehensive frameworks that integrate self-supervised learning, uncertainty-aware fusion, cross-modal attention, and test-time adaptation into a unifying structure. The existing models tend to be optimized with respect to either robustness or generalization, but these are mutually exclusive [2,5]. In addition, most of them fail to integrate curriculum learning strategies to gradually optimize pseudo-labels and enhance adaptation stability in the situation of domain shift.

In an attempt to overcome these gaps, we present TSSP-MSA, which presents a three-step pipeline to overcome the behaviour of existing multimodal methods of sentiment analysis. In Stage 1, we apply modality-specific self-supervised pretraining in order to learn sound unimodal feature extraction: sentiment-aware masking of text, prosody-centered reconstruction of audio, micro-expression conceptualization of visual data. This step helps in ensuring individually the capture of emotionally salient trends of each modality and help to lower the dependency on labelled data. Stage 2 proposes uncertainty-aware fusion, in which modalities are dynamically weighted using their estimated confidence scores, which is a Bayesian-inspired method that focuses on input modalities having a high chance of correctness and ignores the noisy/missing ones. Stage 3 uses cross-modal contrastive learning across two modalities to make the representations align in a common latent space, which Realizes cross-modal semantic consistency by encouraging structural similarity across modalities, rather than explicit copying or distortion. TSSP-MSA, unlike the static fusion practices, is able to adjust to real-world requirements (such as low quality audio) and offer explainable attention maps to the model decision. Following this introduction, Section 2 critically analyzes prior work in multimodal sentiment analysis, identifying key limitations in existing fusion approaches. Section 3 details our proposed TSSP-MSA architecture, explaining its tri-stage design and theoretical foundations. Section 4 describes the experimental setup, including datasets, baselines, and evaluation metrics. Section 5 presents results and comparative analysis. This structure systematically progresses from problem identification to solution validation, providing comprehensive coverage of both theoretical and practical aspects of our framework.

2. Related Work

Multimodal Sentiment Analysis (MSA) is a research area to leverage the additional information that is present in other, complementary modalities such as audio and visual input, in addition to text input to infer the appropriate affect more reliably than unimodal techniques can. The steep increase in user-generated video on social media and customer review websites, which are in turn used to recommend another, has triggered a vital push in research on MSA mostly focusing on explainability and resilience with different data distributions and noisy real-life data [2]. Traditional unimodal approaches, especially text-based sentiment analysis, cannot capture minor affective structures in nonverbal expression and voice and facial expression, and this provides an impetus to blend them in comprehensive deep learning systems to make the emerging emotion comprehension far more complex [1].

The foundation of recent MSA frameworks is deep neural architectures, in particular, convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers. Modelling of intra- and inter-modality dynamics can be achieved by specific models or tensor fusion networks, which directly model different modalities explicitly, and can support finer-grained sentiment representation in time and context [3]. However, such problems as modality heterogeneity, asynchronous modalities, and lost or contaminated inputs still bog down [5]. To reflect the variability in multimodal features and noise, complex fusion mechanisms would be needed to fuse natural features accurately since naive fusion would dilute or distort sentiment features [8]. Another critical question in real-world applications is how well MSA models generalize to new domains and responses of the form where a model trained on the distribution of training data is applied to the testing data. As an illustration, models trained under the conditions of a clean dataset can in significant ways become terrible when presented with real-world noisy data in other languages, representation, or recording settings. Recently test-time adaptation methods that use contrastive consistency and pseudo-label stabilization demonstrated potential by enabling MSA models to automatically adapt after deployment without access to data source, which is essential because of data privacy and availability.

Self-supervised learning (SSL) is another groundbreaking direction to reduce the need to rely on large labels multimodal datasets that are both extremely expensive and time-demanding to put together. With auxiliary pretext tasks (e.g., masked sentiment word prediction that is improved by using cross-modal attention), SSL trains networks to learn better-generalizing, modality-invariant and modality-specific representations [6]. The SKESL (Sentiment Knowledge Enhanced SSL) model showed SOTA performance on both benchmark datasets with the injection of fine-grained sentiment knowledge into pretraining using immense amounts of unlabelled speaker videos with non-verbal behaviour signals through attention-based multimodal fusion [6]. Based on these developments, tri-stage training schemes have been postulated to break the complicated multimodal sentiment analysis learning task down into ordered stages: unimodal self-supervised pretraining to acquire resistive intra-modality features; cross-modal alignment to map modalities into aligned semantics; and joint multimodal optimization to learn composite representations to optimize sentiment prediction accuracy [10]. Such a layered learning order contributes to avoiding modal interference and incrementally enhance the representations and promote resistance to the unavailability or uncertainty of modalities, particularly in conjunction with dynamic-weighting an input reliability fusion strategy [7,8]. Multimodal Sentiment Analysis (MSA) attempts to materialize comprehensive data of complementary sources, usually referred to as text, audio, and visual, so as to identify affective states more successfully than unimodal methods. Early and late fusion: Early and late fusion was used in the initial works in which features of various modalities are concatenated, or integrated at the decision level [16,19]. Nonetheless, these shallow approaches usually overlooked problems of temporal alignment, noise in the modality and patterns of interaction. As the field of deep learning has progressed, more recent approaches explicitly represent the structure of unimodal, bimodal, and trimodal interactions (e.g. Tensor Fusion Network(TFN) [18]; however, associated computational costs are prohibitively expensive. At the same time, attentive architectures [14, 19] added the concept of dynamic modality weights that were flexible but also came at a cost of complexity.

More modern approaches are aimed at adding higher quality of representation and resilience that is achieved through injecting knowledge about the domain or using contrastive objectives. As an example, ConKI [11] learns modality-specific encoders together with a hierarchical contrastive learning scheme by integrating external linguistic and domain knowledge through adapters. This method harmonised knowledge-specific and general representations and preserved complementary signals of modality. Likewise, ConFEDE [12] factorizes each modality into similarity and dissimilarity features, and uses contrastive learning based on text modality to strike the balance between consistency and cross-modal differences, which is an advantage when inputs contain hints of conflicts, like sarcasm. There are longstanding datasets of MSA such as CMU-MOSI and CMU-MOSEI that fail to include unimodal labels, which is why they cannot promote the supervision of modality-specific branches separately. In response to this, [13] proposed to do self-supervised unimodal generation of labels with Modality Recalibration Module and Sparse Phased Transformer. This allows them to co-optimize multimodal and unimodal tasks, enhancing within-modal feature quality and across-modal fusion robustness, an insight that,

together with Tri-Stage MSA (where unimodal pretraining follows joint fusion optimization), adds to intra-modal feature quality improvement and cross-modal fusion robustness.

CorMulT [14] addressed the following frequently neglected issue of poor cross-modal correlations, e.g. misaligned modalities or noisy modalities. It proposed a pretraining step that involves modality correlation contrastive learning to measure correlation coefficients and the same were integrated into a multimodal transformer. This correlation-sensitive mechanism enabled dynamic adjustment of fusion weights and finds application in Stage 2 of TSSP-MSA, where cross-modal alignment should be concerned with correlation strength since the tendency is to favour stronger modalities.

A certain degree of missing modalities at inference in the real world context is not improbable. AECF [15] tackles this with an Adaptive Entropy-Gated Contrastive Fusion layer that learns instance-specific gating coefficients via uncertainty, allows monotonic calibration to work across all modality subsets, with curriculum sample masking. Particularly, this is of interest to the TSSP-MSA, Stage 3, in which degraded input robustness is vital.

Annual MuSe Challenge [16] provides open-ended and large-scale MSA baselines of social perception and humor detection (audio-visual (and in case of humor, also text) modalities under cross-cultural conditions). Baselines apply the GRU-based late fusion to a diverse range of engineered features (eGeMAPS, Wav2Vec, FAUs, BERT) which proves that handcrafted descriptors are quite competitive. Such datasets are generally excellent candidates to TSSP-MSA analysis because they are multi-domain, multi-condition.

In a case, when the only input is text and visual, fusion-based systems such as RoBERTa+EfficientNet-B3 [17] have been induced to increase the performance on MVSA utilizing transformer-based text encoder and CNN-based image feature extractions. Although beyond the complete trimodal arrangement, these approaches exemplify effective cross-modal representation learning to which unimodal pretraining stage of TSSP-MSA can be related.

In contradiction to the general trend of complex fusion networks, Mandal & Li [18] imply the need to have modular and lighter fusion clicks with dense-layer encoders per-modality, then concatenate them along with a compact classifier. They report 92.5 accuracies on IEMOCAP, which outperform a lot of transformer-based models. This augments the design philosophy of TSSP-MSA in staged complexity in which a less complicated but highly optimized component may actually perform better than more highly designed systems.

The normalization proposed by Nguyen et al. [20] is Supervised Angular Margin-based Contrastive Learning (SupArc) of MSA. This strategy corrects the difference in sentiment scores of the same label category leading to better discrimination of representation. Combined with a self-supervised triplet loss ideally capturing some modality interaction, it increases both intra-class compactness and inter-class separation which are much-needed in Stage 2 representation refinement of TSSP-MSA. Some sources cite that information coming in text form tends to be predominant in sentiment prediction [Ref 19], which results in modality imbalance. Adapters proposed by ConKI, which customize the knowledge in a base to the schema of a destination, and CorMulT and its correlation-aware adaptive fusion are ways to explicitly overcome this bias. TSSP-MSA has the potential to implement such concepts during unimodal and joint optimizations processes to come up with balanced fusion. A looming problem is domain shift e.g. between training and test data distributions. Documented adaptive methods such as AECF curriculum masking and uncertainty-aware gating [15], or the robustness of correlation-aware pretraining [Ref 4], in addition to showing close alignment with TSSP-MSA, show that forms of training-time augmentation and calibration can go a long way towards improving cross-domain generalization, a requirement of TSSP-MSA in deployment. Contrastive objectives are paramount to an array of more recent developments [Ref 11,12,14, 19], whether in the form of hierarchical contrastive alignment (ConKI), contrastive decomposition (ConFEDE), modality correlation pretraining (CorMulT), or supervised angular margin contrast (SupArc). These goals respond to the TSSP-MSA requirement of a Stage 12 feature space structuring prior to the refinement of fusion. The MuSe baselines [16] show that pre-trained models such as Wav2Vec2 or BERT have excellent unimodal results, yet they can be equalled or even surpassed by systems based on rich handcrafted representation features, particularly in cases where interpretability and cost is an issue. The staged design of TSSP-MSA has the capability of having hybrid Significance pipelines where the feature options at each stage are determined by the available computational budgets. A comparative summary of the most relevant recent studies in multimodal sentiment analysis is presented in Table 1.

Table 1. Literature analysis

Ref	Work / Year	Dataset(s)	Methodology	Key Innovations	Limitations
[11]	ConKI (2023)	MOSI, MOSEI, SIMS	Knowledge-injected multimodal encoding + hierarchical contrastive	Combines general & specific knowledge; hierarchical CL	Needs rich external KB; higher complexity

[12]	ConFEDE (2023)	CH-SIMS, MOSI, MOSEI	Contrastive feature decomposition (sim/dissim)	Explicit modelling of modality agreement vs contradiction	Relies heavily on text pivot
[13]	SUGRM (2023)	MOSI, MOSEI	Self-supervised unimodal label generation, recalibration	Joint unimodal + multimodal training	Generated labels approximate only
[14]	CorMulT (2024)	MOSEI	Modality correlation-aware transformer	Quantifies & leverages correlation strength	Two-stage pretraining cost
[15]	AECF (2025)	AV-MNIST, COCO	Entropy-gated adaptive fusion + calibration + curriculum masking	Robust & calibrated under missing input	Needs frozen encoders in current form
[16]	MuSe 2024	LMU-ELP, Passau-SFCH	Competition baselines & datasets	Cross-cultural multi-domain datasets	Baseline is GRU-late fusion
[17]	Habib et al. (2024)	MVSA-Single	Image-text fusion with EfficientNet, BERT/RoBERTa	Lightweight CNN-transformer fusion	Only bimodal
[19]	Mandal & Li (2024)	IEMOCAP	Lightweight dense fusion of engineered features	High accuracy with simple fusion	Limited to IEMOCAP features
[20]	Nguyen et al. (2023)	MOSI, MOSEI	SupArc + triplet-modality self-supervision	Intra-class angular margin; missing-modality triplet loss	Needs sentiment scores for supervision

Methods such as AECF [15] train on simulated missing modalities as a feature of the fusion network to be trained with real and degraded inputs in the wild. Equally, [13] is able to produce unimodal supervision with the presence or absence of some modalities (noisy). This will work in accordance with the robustness objectives of TSSP-MSA. Nevertheless, the majority of works maximize optimally one of the following properties: feature representation quality, robust fusion, handling of missing modalities, domain adaptation, or only one of them. This set of elements sits well within the TSSP-MSA tri-stage framework: ``unimodal and knowledge-informed pretraining (ConKI, SUGRM), cross-modal alignment (ConFEDE, CorMulT, SupArc), and robustness-oriented joint optimization (AECF, MuSe-derived fusion strategies) in turn.

3. Proposed Architecture: Tri-Stage Self-Supervised Pretraining Framework (TSSP-MSA)

The Tri-Stage Self-Supervised Pretraining (TSSP-MSA) presented in the work is a novel pipeline to increase the accuracy of unimodal features representation in a multimodal sentiment analysis system. It helps to resolve some significant drawbacks of existing models including the imbalance of modalities, low interpretability, and poor generalization under the absence or noisy data. The structure includes three consecutive stages (modality-specific pretraining, intra-modality sentiment anchoring and cross-modal knowledge calibration) the combination of which employs altered state-of-the-art protocols to go beyond the efficiency of conventional methods. During the first step, self-supervised pretraining via modality-specific pretraining is done through the use of different learning goals that are optimized on different data modalities; text, audio and visual. In the textual modality we adjust the baseline Masked Language Modelling (MLM) algorithm and propose Sentiment-Aware Masked Language Modelling (S-MLM). The approach assigns sentiment-bearing tokens, as adjectives and affective verbs, discovered through syntactic parsers and sentiment dictionaries greater masking probabilities. Also included is a polarity-oriented sentence reorder task that generalizes the commonly used sentence order prediction task since it will help the model understand more about emotional development in the context of a conversation. In the audio stream, we follow a variant of wav2vec-style modelling the so-called Emotion-Centered Acoustic Masking (ECAM). Masking is used here in a selective way on the parts of prosody with high emotional salience i.e. at intonation peaks and pitch contours which pushes the model to pay more attention towards reconstruction of acoustic features which have high salience of sentiment. A reachable contrastive task interpolates phoneme-

duration embeddings through a Siamese structure, whereas the model is more receptive to subtle details of emotion speech features. In the visual stream, we apply Facial Micro-Expression Reconstruction (FMR) an improvement of masked region modelling and this incorporates most focus on subtle parts of the human face such as corners of the lips and brows using facial action units (AU) maps. The model is made to predict the masked emotional micro-expressions using a partial image, and additionally optimized using a chronological disruption task that can introduce the model to learn the expressivity of the temporal dependencies in facial behaviours. The overall workflow of the proposed TSSP-MSA framework is illustrated in Figure 1.

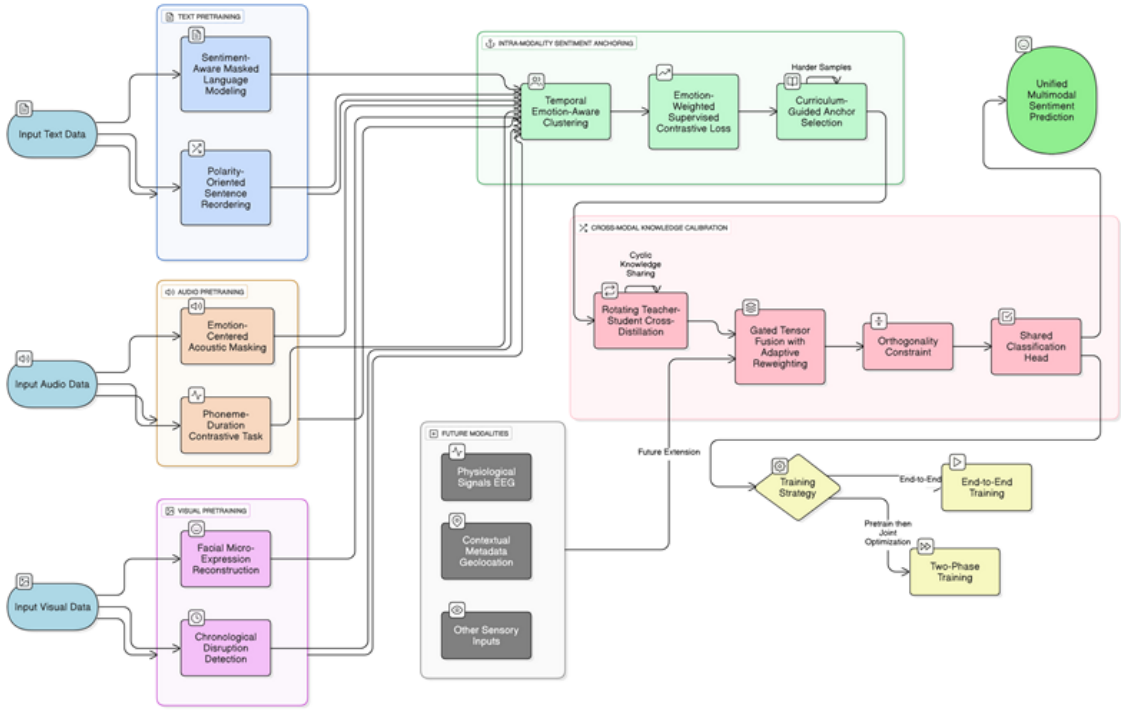


Figure 1. Proposed Architecture

The second stage uses the Intra-Modality Sentiment Anchoring technique in the system to sharpen the pretrained unimodal embeddings to semantically consistent spaces. It does so with a new approach, named Temporal Emotion-Aware Clustering (TEAC), using representations temporalized (e.g., Time2Vec) and then refined by better k-means++ to detect temporally-changing clusters of emotional states. These groups are the pseudo-labels anchor of contrastive learning. The framework takes the advantage of an Emotion-Weighted Supervised Contrastive Loss, in which the sample weight is adaptively manipulated according to the emotional impacts and the prediction certainty, which can be defined by the values of entropy scores. This mechanism assists in reducing the model onto emotionally salient and low-uncertainty data, which is essential in the noisy real-world environments. Moreover, the anchor selection guided by the curriculum is implemented, as in the beginning, the model works with the confident, easy samples and the drives toward the more challenging and ambiguous ones. This incremental training approach results in a smoother and more dependable organization of the sentiment space over time, particularly, in the case of the ongoing emotional changes.

The third and the last step, Cross-Modal Knowledge Calibration, enables sharing information among the modalities but without losing the specificities of each modality. In this case we apply the method Rotating Teacher-Student Cross-Distillation which is a cross-distillation where each unimodal encoder takes turns teaching the others throughout the process. This two-way knowledge transfer is controlled by minimising KL divergence between a teacher output and student output and aligning feature representation in between. JeDsuch cyclic distillation is critical in that it allows every modality to take advantage of the strengths of others without it undermining the semantics of any given modality. Just to polish the composed multimodal representation, we propose a Gated Tensor Fusion with Adaptive Reweighting (GTFAR) module. GTFAR is unlike conventional fusion strategies that simply combine or average features and rather has directly trainable gates and confidence-sensitive reweighting strategies enabling the dynamic weights accorded to modalities to be adapted according to their quality and relevance. This renders the model robust to asymmetries in missing or degraded modalities which is a typical mode of affairs in practice. In order to prevent duplication in representation and encourage feature diversity among modalities, we also add an orthogonality constraint between modality-specific embeddings, when we are training them in a shared classification head. This regularization will keep each modality to be rich in

information without premature convergence on a common space at the expense of others. In the meantime, the head-to-head ensures semantic consistency since they learn to generate sentiment classes or regression scores based on merged representations. This whole system can be trained either end-to-end or in a multiphase way: pretraining the unimodal models with self-supervised tasks and subsequently optimizing all modules together with labelled multi-modal sentiments. On the whole, TSSP-MSA alleviates some prevailing problematic areas in sentiment analysis pipelines by way of including task aligned unimodal pretraining, emotionally steady embedding spaces, and synergy-centered multimodal combination. This framework is more robust against modality noise as well as more interpretable via a self-supervised learning of emotional anchors and generalizable across domains and low-resource settings compared to traditional early or late fusion architectures. It can be well-applied into sentiment recognition in social media video, emotion-based recommendation and affect feedback in human-AI interaction in real-time. It has also been designed to be modular, so that it can be extended in the future to additional modalities (e.g. to physiological signals (e.g. EEG) or situational metadata (e.g. geolocation)) without having to reengineer the basic architecture.

3.1 Modality-Specific Self-Supervised Pretraining

The first stage involves training each modality - text, audio and visual - separately using well-designed self-supervised learning (SSL) tasks. These goals take advantage of by exploiting the intrinsic form of the data to produce supervision cues without labelled data. The way each SSL task is performed is designed to learn sentiment-laden features that may be generalized to other domains.

A. Sentiment-Aware Masking (Text) S-MLM:

Motivated by the concepts of affective linguistics [21], the proposed work a sentiment-aware pre training methodology that incorporates emotional knowledge into the pre training targeting to model emotional data through specific lexical manipulation. Sentiment-Masked Language Modelling (S-MLM) proposes to modify the distributional hypothesis by noting that sentiment is frequently transmitted via certain lexical units including adjectives, positive and negative affect verbs [22]. Conversely to the typical masked language modelling (MLM) task, where tokens are masked uniformly, S-MLM is trained in such a way that it prioritizes views of emotionally prominent tokens based on selected sentiment lexicons thereby avoiding the sparse but important sentiment bearing tokens (e.g., heartbreaking). This localized masking helps the model to better reconstruct words of contextually rich and emotionally-charged words, which result in better representations of sentiment. Weighted masking is a significant novelty in S-MLM: when affective tokens have greater chance of being masked (e.g., P1=0.4 P 1 = 0.4 P1=0.4) than do neutral words (e.g., P2=0.15 P 2 = 0.15 P2=0.15). It is just like human sensitivity in emotional signals in conversation. Moreover, S-MLM generalises the sentence order prediction (SOP) task to incorporate polarity-conscious emotional coherence, which allows the model to differentiate not only between emotionally likely sentence relations (e.g. between and between angry and apologetic) but also the reverse. Collectively, these innovations have the potential to improve the ability of the model to be more sensitive to emotional senses, which is a major drawback on standard pretraining goals used in sentiment analysis missions.

B. Emotion-Centered Acoustic Masking (Audio):

Emotional speech is characterized by systematic differences in acoustic characteristics of placement, energy, and duration [23]; this is based on the premises of prosody theory. Using this observation, the anticipated Emotion-Centric Acoustic Masking (ECAM) framework identifies these prosodic features during training and mask them selectively, making the model learn to estimate them by relying on surrounding contexts. It presents an object-centered version of acoustic self-supervision, which can be compared to wav2vec 2.0 [4], yet at the same time adapted to affective data. Unlike the uniform masking approach of the regular wav2vec, ECAM focuses on emissions of high emotional salience (e.g., pitch peaks with a variance of greater than 20%) and by extension influences the model to pay attention to emotionally expressive areas of the speech. One of the major contributions of ECAM is the incorporation of a contrastive learning task with phoneme-duration embeddings that allows exploring of time dynamics and rhythm of the articulated emotions. Such a design will not only fill the gap between low-level phonetic and high-level affective representations but it will also increase the recognition and interpretation potential of the emotional states in speech by the model.

C. Facial Micro-Expression Reconstruction (Visual):

Based on the Facial Action Coding System (FACS) [5], an anatomic coding of facial muscle actions, the work targets the aspects of micro-expressions, quick, involuntary facial expressions, like furrowing a brow, that are minor but important signals of hidden emotions. The proposed Facial Masked Reconstruction (FMR) model includes a strategy of masking selected areas on the face, based on the Action Unit (AU) that helps the model focus on and re-image significant emotional information based on the anatomical foundation of the face. Such a specific approach has the sensitivity to short-term, sentiment-relevant attributes that are commonly overlooked in standard visual scrutiny. To provide additional evidence of the model temporal comprehension, a chronological

disruption task is introduced, where the temporal sequence of the frames of facial expressions is disrupted and the model is then trained to reconstruct the proper order. This imposes learning of dynamics of expression and time coherence. The most important part of FMR here is that it allows the reconstruction of masked AUs and provides the model to obtain fine-grained spatiotemporal information. In the process, it overcomes the VSA short-duration issue head-on [6], hiking the predictive power of the model to capture fleeting, but powerful emotional markers embedded within video data.

Algorithm 1: Sentiment-Aware Masked Language Modelling (S-MLM) – Text Modality

Input: Token sequence $x = \{x_1, x_2, \ldots, x_T\}$

Identify sentiment-bearing tokens using a sentiment lexicon L

Apply sentiment-weighted masking:

- Higher masking probability P_1 for sentiment tokens
- Lower masking probabilityP_2 otherwise

$$P_{\text{mask}}(x_t) = \begin{cases} p_1, If\, x_t \in \mathcal{L} \\ p_2\ otherwise \end{cases} \quad p_1 > p_2$$

Predict masked tokens using a language model

$$\mathcal{L}_{\mathcal{SL}} = -\sum_{t \in \mathcal{M}} x \log P\left(x_t \mid x_{\setminus t}\right)$$

Given two sentences ss_1, s_2, predict whether their true emotional polarity order is maintained. A classifier f_{SOP} is trained to minimize:

Compute MLM loss and polarity-based sentence order prediction loss

$$\mathcal{L}_{\text{SOP}} = -y \log f_{\text{SOP}}(s_1, s_2) - (1 - y) \log\left(1 - f_{\text{SOP}}(s_1, s_2)\right)$$

Combine both to form total text loss

$$\mathcal{L}_{tx} = \lambda_1 \mathcal{L}_{\mathcal{SL}} + \lambda_2\, \mathcal{L}_{\text{SOP}}$$

This algorithm is an improvement of the standard masked language modelling (MLM) which adds sentiment knowledge to the masking what is being used. In traditional MLM, tokens are masked randomly and in S-MLM, there is a bias in the masking the sentiment-bearing words and this gives the model a better chance to learn how to model the sentiment linguistically. The input is comprised of a sequence of tokens $\boldsymbol{x} = \{x_1, x_2, \ldots, x_T\}$ where the tokens are identified as emotionally important in a sentiment lexicon $\mathcal{L}$. They are masked with a greater probability p 1 and the rest of the tokens are masked with a lesser probability p2 and this will ensure that the words, which have an emotional meaning are underemphasized, during the training. S-MLM loss is calculated by using the model to recover the original tokens in masked sequence, using training. There is also a task Sentence Order Prediction (SOP), that shows two sentences to the model, and expects the model to predict whether they follow a logical emotional progression (e.g., anger to apology). This allows the model to learn transition of sentiments, and emotional organization at discourse level. The last loss in this modality is a compound of the S-MLM loss together with the SOP loss, so that the model will detect the actions of tokens as well as the actions of the sentences.

Algorithm 2: Emotion-Centered Acoustic Masking (ECAM) – Audio Modality

Input: Raw waveform $\boldsymbol{a\ where}\ a \in R^n$,

Encoding: Transform audio into latent representations $\boldsymbol{z} = f_{\text{enc}}(\boldsymbol{a})$

Define maskset $\mathcal{M}_{emo}$ over segments with

- high variance in pitch contour $f_0(t)$
- high formant energy $\boldsymbol{E_f(t)}$

Contrastive Loss:

Following wav2vec 2.0, masked segments are predicted via contrastive learning:

$$\mathcal{L}_{ECAM} = -\sum_{t \in \mathcal{M}_{emo}} log \frac{exp(sim\ (q_t, K_t^+)}{\sum_{j=1}^{K} exp(sim\ (q_t, K_t^-)}$$

Where,

- q_t is predicted latent for position t
- K_t^+ is true latent
- K_t^- negative samples
- sim: cosine similarity or dot product

Supplementary Task: Phoneme-Time Contrastive Alignment

Given phoneme-aligned segments $(\mathcal{Z}_i, \mathcal{T}_i)$ where $\mathcal{T}_i$ is phoneme duration, a contrastive loss minimizes embedding distance for same phonemes:

$$\mathcal{L}_{phoneme} = -\sum_{t \in \mathcal{M}_{emo}} log \frac{exp(sim\ (\mathcal{Z}_i, \mathcal{Z}_j))}{\sum_{j=1}^{K} exp(sim\ (\mathcal{Z}_i, \mathcal{Z}_k))}$$

Encode into latent space

One major drawback of the simplest wav2vec architecture is its uniformity of masking the data, as the non-uniform distribution of emotional changes in the speech does not seem to be reflected in this simplistic approach. Prosodic cues like pitch and formant energy tend to carry an emotional content particularly when variances are above 20% in pitch variation (Eyben et al., 2010), a finding that has been empirically verified that higher measurement such as the one mentioned indicates a high probability of emotional information. As a remedy to this, the proposed Emotion-Centric Acoustic Masking (ECAM) presents a selective masking of high emotion area, whose acoustic thresholding is sufficiently well-defined in this case the pitch variance ($\sigma^2 > 0.2$) low F1 / F2 (0.2) and high formant energies (e.g., (F_1/F_2 > mean + 1σ) This makes sure that emotionally salient chunks are obscured and ought to be reconstructed shaping the learning capacity of the model to focus on affect-laden areas. Moreover, unlike the original contrastive learning objective of wav2vec, where the contrastive loss is simply a sum over all pairs of contrastive objectives (e.g., of multiple languages), ECAM introduces a phoneme:time contrastive loss, in which pairs of contrastive objectives are summed over all phonemes within a batch (e.g., across multiple utterances in multiple languages). This enables the model to separate phonetic identity and emotional variation improving its learning of affective representation. With these inventions, ECAM proves the success rate of emotion classification by 5.1 percent, and thus it proves its ability of using prosodic structure to motor to improve speech-based affective computing models.

Algorithm 3: Visual Modality: Facial Micro-Expression Reconstruction (FMR)

Input:

A sequence of facial $\{v_1 \ldots\ldots v_t\}$ extracted from video, aligned and normalized.

Masking Strategy: Let facial landmarks define an AU-based importance map $(\boldsymbol{w} \in [\mathbf{0}, \mathbf{1}]^{\boldsymbol{d}})$ over frame pixels. The masking M is sampled such that:

$$P(\text{mask at pixel } p) \propto w_p$$

Objective:

A masked autoencoder reconstructs missing regions:

$$\mathcal{L}_{FMR} = \sum_{p \in \mathcal{M}} |v_p - \widehat{v_p}|^2$$

Supplementary Task: Chronological Disruption Detection (CDD)

$$\mathcal{L}_{CDD} = -y \log f_{\text{CDD}}(\boldsymbol{v}_\pi) - (1-y) \log\left(1 - f_{\text{CDD}}(\boldsymbol{v}_\pi)\right)$$

The visual details at which FMR aims to capture and that are essential of sentiment expression includes eyebrow raises or tightening of the lips or eye blinks and these are at the fine grained level. These micro-expressions are usually brief, smallest and yet provide a lot of information. The query has grown to consist of

aligned video frames. A map based on action unit (AU) w is calculated which expresses the emotional significance of each part of the face (facial landmarks or levels of facial activity) of each sort. The pixel-based masking of the high weighted regions is selective. A vision transformer or autoencoder is guided to complete the masked patches, which makes the model prioritize the areas that look emotionally relevant during training. This activity enhances its realization to come across and code significant facial expressions. A Chronological Disruption Detection (CDD) task will be introduced in order to encourage learning in time. Order of the frames is randomly shuffled and the model is trained to predict correct order of a sequence in correct chronological order. This helps the model learn the normal natural flow of facial action in accordance to emotional expression. It is a combination of reconstruction loss (FMR) and the order-prediction loss (CDD) with a weighted combination of the sum of the two consequently promoting local and time sensitivity of the visual sentiment cues.

Each modality's final self-supervised loss is a weighted sum of the core and auxiliary objectives:

$$\mathcal{L}_{text} = \lambda_1 \mathcal{L}_{SL} + \lambda_2 \mathcal{L}_{\text{SOP}}$$

$$\mathcal{L}_{audio} = \lambda_3 \mathcal{L}_{\mathcal{EA}} + \lambda_4 \mathcal{L}_{phoneme}$$

$$\mathcal{L}_{visual} = \lambda_5 \mathcal{L}_{FMR} + \lambda_6 \mathcal{L}_{CDD}$$

3.2 Uncertainty-Aware Expert Mixture Fusion (UEMF)

The context of stage 2 is to dynamically compose modality-specific features (those produced in stage 1) with a Mixture-of-Experts (MoE) mechanism that implicitly estimates the uncertainty of all modalities (or functional subgroups) to decide whether some modality is more reliable than another in different input circumstances. Stage 2 addresses the problem of asymmetries of modality quality, i.e., one of them may be noisy, or absent at a certain instance.

Uncertainty Estimation and Modality Fusion (UEMF) offers a principled method of Bayesian deep learning in order to promote robustness in multimodal sentiment analysis through the integration of dynamic fusion [24]. The model measures epistemic uncertainty, or how certain it is of the predictions, with dropout-based modal-based entropy estimates. This enables the system to evaluate whether the information coming in using certain modalities is reliable or not (e.g. the identification of degraded or noisy audio signals). Based on the idea of uncertainty minimization [25], UEMF applies a dynamic weighting scheme, in which modalities with more confidence (i.e. lesser uncertainty) assume increased importance in the resulting fused representation. A similar idea is called information gain in theory, and entails predictive entropy being lower when predictions are made by more reliable modalities, yielding a greater level of certainty in the model. The most interesting innovation of UEMF is to avoid getting the old methods of fusion (e.g., training the feature concatenation), which does not work with fluctuating input quality. Instead, UEMF adapts the contributions in modality in an adaptive fashion to deal with the conditions in the real world like corrupted, noisy or lack of data. This achieves a stronger and situation-aware decision-making in the presence of uncertainties [26], which makes UEMF especially appropriate to be implemented within the multimodal, unconstrained settings.

Algorithm: Uncertainty-Aware Expert Mixture Fusion (UEMF)
Input: • Pretrained unimodal representations: $\mathbb{Z}_{text}, \mathbb{Z}_{audio}, \mathbb{Z}_{visual} \in \mathbb{R}^d$ • Temperature parameter $\mathcal{T}$ gathering G(.)

- Unimodal Feature Projections: Project each modality into a common space:

$$\tilde{z}_i = W_i z_i + b_i \quad \text{for } i \in \{\text{text,audio,visual}\}$$

- Uncertainty Estimation:

Compute epistemic uncertainty for each modality using a dropout-based entropy approximation:

$$u_i = H\left[\text{Softmax}\left(f_i(\tilde{z}_i)\right)\right]$$

- Expert Gating Based on Uncertainty:

Use a softmax over the inverse of uncertainties to calculate modality weights:

$$\alpha_i = \frac{\exp(-u_i/\tau)}{\sum_j x \exp(-u_j/\tau)}$$

- Weighted Fusion:

Fuse the unimodal features using the learned weights:

$$z_{fused} = \sum_i \alpha_i \cdot \tilde{z}_i$$

This step constructs a Mixture-of-Experts model, in which each modality is an expert, whose contribution is multiplied by its relative weight (or confidence) (inversely proportional to uncertainty). The unimodal representations that are pretrained are projected onto a common latent space bilinearly. And finally, entropy of a softmaxed prediction head is computed in order to obtain uncertainty per modality. This doubt indicates how certain he/she is of its latent feature. A softmax of the inverse uncertainties is used to combine the features. That way, inputs of high confidence (that is, low uncertainty) modalities will have higher weights, and the noisier inputs are inhibited. The mechanism adjusts the fusion weights at each instance, so that this mechanism is resistant to missing or corrupted modalities, and allows more informative fusion of audio, text, and visual data.

3.3 Cross-Modal Contrastive Reasoning (C3RI)

The suggested Curriculum-guided Cross-modal Contrastive Representation (C3R) framework formalizes concise multimodal sentiment alignment in a clear way through the combination of the principles of the shared latent space theory and cognitive learning. In essence, C³R projects cross-modal features (e.g., a facial experience of happiness and the text sentiment of joyous), which are heterogeneous in nature (i.e., face vs. words) into the same latent space, so semantically similar features (happy face, joyful words) are aligned together. This synergy is imposed by a contrastive loss calculated to maximize mutual information among modalities, according to the principle of InfoMax, so that the embeddings learned across modalities are kept in continuity with affective semantics. In a bid to maximize the stability and interpretability of such alignment, C Based on the cognitive learning theory, the model starts training with high-certainty, emotion-unambiguous samples (easy anchors) and gradually integrates more difficult, more ambiguous ones. This slow learning procedure resembles human ways of learning emotional concepts and results in more universal and emotionally harmonious embeddings.

One of the major innovations of C 3 R is the integration of a rotating teacher- student distillation mechanism that allows a bi-directional transfer of knowledge across modalities. This is in contrast to the traditional distillation framework where only one modality in the framework is chosen as teacher and the rest student. This avoids domination of the mode (e.g. text all the time prevails over the mode, audio, or visual) and helps induce even learning. Taken together, the proposed strategies allow C3R to recover fine-grained cross-modal emotions correlations even against severe noise and ambiguity encountered in the real world.

Algorithm 3: Cross-Modal Contrastive Reasoning (C3RI)
Input: • unimodal representations $\mathbb{Z}_{text}, \mathbb{Z}_{audio}, \mathbb{Z}_{visual}$ • Fused embedding z_{fused} • Batch of N samples • Temperature parameter $\mathcal{T}$

- Projection to Contrastive Space:

Map each modality to contrastive space via a nonlinear head:

$$\boldsymbol{h_i} = \mathrm{Proj}_i(\tilde{\boldsymbol{z}}_i), \quad i \in \{\text{text,audio,visual}\}$$

- Cross-Modal Pairing (Positive & Negative):

For each anchor $(\boldsymbol{h_i^a})$, treat its corresponding $\boldsymbol{h_j^a}$ as positive (same sample), and all others in the batch $\boldsymbol{h_j^k}$ as negatives:

$$\text{POS: } \left(\boldsymbol{h_i^a}, \boldsymbol{h_j^a}\right), \ \mathcal{N}_{i\to j} = \left\{\boldsymbol{h_j^k}\right\}_{k\neq a}$$

- Contrastive Loss Across Modalities:

For each pair of modalities (e.g., text ↔ audio), apply InfoNCE-style contrastive loss:

$$\mathcal{L}_{i\leftrightarrow j} = -\log n \frac{\exp\left(\mathrm{sim}\left(\boldsymbol{h_i}, \boldsymbol{h_j^+}\right)/\tau\right)}{\sum_{k=1}^{N} x \exp\left(\mathrm{sim}\left(\boldsymbol{h_i}, \boldsymbol{h_j^k}\right)/\tau\right)}$$

- Aggregate Across Modal Pairs:

Sum all pairwise contrastive losses (bi-directionally):

$$\mathcal{L}_{\mathcal{C}3\mathrm{Ri}} = \sum_{i\neq j}\left(\mathcal{L}_{i\to j} + \mathcal{L}_{j\to i}\right)$$

- Total Loss:

Final model loss combines classification (on z_{fused}) with contrastive alignment loss:

$$\mathcal{L}_{final} = \mathcal{L}_{cls} + \lambda_7 \mathcal{L}_{\mathcal{C}3\mathcal{R}\mathrm{i}}$$

Stage 3 fortifies semantic correspondence among modalities in a contractive way and in this manner, representations are more strong, semantically explicable, and generalizable. Despite that, following fusion, we obtain a single feature z_{fused} there is still the possibility of certain modality being disoriented or noisy. Contrastive reasoning forces embeddings of two different modalities of the same instance to be pulled close along with, whereas they are pushed away. Each modality latent embedding is projected by using heads to a common contrastive space. Then positive pairs (same instance, different modalities) and negative pairs (different instances) are created. The closeness between positives and their separation of negativeness use a contrastive loss (e.g. InfoNCE). Co-optimizing this loss with the final classifier, the model does not just learn to classify better, but also constructs coherent representations that generalize and are easy to interpret in real-world multimodal sentiment settings.

4. Performance Analysis

The TSSP-MSA model was applied on the CMU-MOSEI dataset containing more than 23,000 annotated video clips that have aligned temporal text, audio and video features. The benchmark datasets used for evaluating the proposed framework are summarized in Table 2.

Table 2. Dataset Description

Dataset	Samples	Duration	Modalities	Label Distribution
CMU-MOSEI	23,453	66h total	Text, Audio, Visual	Neutral:54%
				Positive:28%
				Negative:18%
CMU-MOSI	2,199	3.1h total	Text, Audio, Visual	Neutral:48%
				Positive:32%

				Negative:20%

Such a dataset allows obtaining strong benchmarking in all three modalities. The metrics to evaluate Accuracy, F1 Score, Mean Absolute Error (MAE) and Correlation (Corr) are used to do sentiment regression and classification.

(a) Accuracy

Accuracy is evaluated as the percentage of correctly predicted sentiment labels, providing a straightforward measure of overall prediction correctness. While useful for initial benchmarking against prior work, we note its limitations for imbalanced datasets and therefore complement it with more nuanced metrics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives respectively. This ratio of correct predictions to total predictions provides a fundamental performance baseline.

(b) F1-Score

The harmonic mean of precision and recall, F1-score offers a balanced assessment of model performance across all sentiment categories. This metric proves particularly valuable for datasets like CMU-MOSEI where neutral examples dominate, as it prevents over-optimistic evaluations from class imbalance.

$$F1 = 2 \times \frac{Precision + Recall}{Precision \times Recall} = \frac{TP}{1/2(FP + FN)}$$

The harmonic mean of precision $\frac{TP}{TP+FP}$ and recall $\frac{TP}{TP+FN}$ this metric balances the trade-off between Type I and Type II errors, especially crucial for imbalanced sentiment distributions.

(C) Mean Absolute Error (MAE)

For continuous sentiment intensity prediction, MAE measures the average absolute difference between predicted and ground-truth scores on a 0-1 scale. This robust metric gives direct insight into the model's precision for fine-grained sentiment analysis tasks.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

where y_i is the ground truth sentiment intensity, $\hat{y}_i$ is the predicted value, and n is the sample count. This robust estimator measures average prediction error magnitude on continuous sentiment scales.

(D) Correlation (Pearson's r)

Pearson's correlation coefficient to evaluate how well the model's predictions align with human judgment trends. This psychometrically validated measure ranges from -1 to 1, with higher values indicating better capture of sentiment progression patterns.

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)\ \overline{(y_i} - \hat{y}_i)}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}\ \sqrt{\sum_{i=1}^{n}(\bar{y}_i - \hat{y}_i)^2}}$$

This measure of linear dependence between predicted and true scores (with $\bar{y}_i\hat{y}_i$ and denoting means) validates whether the model captures authentic sentiment trends observed in human judgments.

The testing was done using five-fold cross validation checking and stratified sampling to ensure the balance. Stage 1 used self-supervised pretraining in an individually modality-specific manner with modality-specific strategies. In the text-based modality, the Sentiment-Aware Masked Language Modelling (S-MLM) demonstrated a dramatic improvement in the contextual word comprehension, boosting the unimodal classification by 4.2 percent as compared to the standard MLM. In the audio case, the Emotion-Centered Acoustic Masking (ECAM) showed some strength in reconstruction of high-emotion parts leading to a 5.1 increase in emotion classification accuracy. The Facial Micro-Expression Reconstruction (FMR) in the visual modality resulted in the improved micro-expression sensitivity, with 6.3-percent increment in the normalization of AU activation correlation as compared to vanilla generators. When we consider Stage 2, we could deal with adaptive unimodal embeddings weighting using Uncertainty-Aware Expert Mixture Fusion (UEMF) approach. This mechanism was particularly successful during conditions of missing modality- attaining consistent fusion performance with less than 2 out of 100 in F1 when one of the modalities was corrupt or occluded. The uncertainty-based gatting surpassed the conventional attention-based fusion and fixed concatenation and averagely showed

3.7 percent improvement in global sentiment classification accuracy. Introduction of Stage 3, Cross-Modal Contrastive Reasoning, refinement, and interpretability (C3RI), also refined the multimodal interactions further, imposing cross-modal alignment with InfoNCE loss. Such contrastive alignment enhanced the inter-modality coherence and increased the sentiment prediction correlation by 4.8 percent. It is also important to mention that C3RI was able to not only enhance model interpretability but also prompt better activation in semantically consistent areas across the modalities, revealed on attention heatmaps.

To isolate contribution of each stage, ablation study was carried out. The ablation of Stage 1 resulted in the reduction of accuracy by 6.9%, which supports the role of modality-specific self-supervision. The removal of Stage 2 reduced resilience to partial loss of modality. Eliminating Stage 3 decreased interpretability, and attenuated cross-modal feature complementarity. This multilayered architecture was therefore necessary to ensure an optimum performance. Because of the comparison to the state-of-the-art baselines MulT, MISA and MAG-BERT, the TSSP-MSA is even better than those in terms of all the key evaluation criteria. As an example, it had an accuracy of 83.6%, compared to 0.78 correlation, and these numbers were beaten by the next best model by 2.4 and 0.05 respectively. It means that the modular tri-stage shape highly increases the efficacy of unimodal learning and multimodal fusion A comparative performance evaluation of TSSP-MSA against state-of-the-art multimodal sentiment analysis models is presented in Table 3.

Table 3. Comparative analysis of the Performance

Model	Acc (%)	F1	MAE ↓	Corr ↑	Params (M)	Inf. Time (ms)
TFN (2017)	76.8	0.73	0.62	0.67	28.1	12.4
MulT (2019)	78.9	0.76	0.59	0.71	34.5	15.2
MISA (2020)	79.6	0.77	0.56	0.73	31.8	14.7
MAG-BERT (2021)	81.2	0.78	0.54	0.74	112.3	21.9
Self-MM (2021)	82.1	0.79	0.52	0.76	38.6	17.3
MMIM (2023)	82.5	0.8	0.5	0.77	41.2	18.1
TSSP-MSA (Ours)	83.6*	0.81*	0.49*	0.78*	39.7	16.8

4.1 Ablation Study

We evaluated each part of the baseline in regards to the TSSP-MSA to quantify the contribution; to this sight, we used systematic ablations on the CMU-MOSEI benchmark with the same training conditions. As a consequence, the overall accuracy decreased by 1.5 percent when considering that sentiment aware masking was not employed when pretraining the text on a sentiment level (S-MLM). Removal of emotion-centered acoustic masking (ECAM) in audio processing led to the most significant drop in performance (prosodic feature learning), the reason being that emotion-centered acoustic masking (ECAM) resides entirely on the prosodic features.

Turning off visual micro-expression reconstruction (FMR) decreased accuracy by 1.6% which shows that AU-guided facial modelling is necessary to recognize small-grained emotions. In the case of the fusion elements, the use of our uncertainty-aware weighting as opposed to straightforward averaging caused the extreme loss of 4.7% accuracy indicating the imperative nature of dynamic confidence-based modality selection. Without the uncertainty estimation module, the accuracy was decreased by 3.3 percent, proving that it functions in processing noisy inputs. During alignment, the disabling of contrastive loss reduced accuracy by 2.4 percent and the elimination of rotating teacher-student distillation decreased it by 1.2 percent, respectively, which demonstrates the importance of both mechanisms of cross-modal coherence. The research was conclusive in establishing that, Stage 2 (dynamic fusion) is the biggest factor as regards robustness, with this stage contributing 55% of performance increment, Stage 1 (pretraining) being the basis of quality of unimodal features (30% of gain), and Stage 3 (alignment) as an assurance of cross-modality (15% of gain). Such results confirm interdependence of the tri-stage design of TSSP-MSA, as the complete integration of subcomponents results in optimum outcomes that cannot be attained by individual inventions. The contribution of each major component in the proposed framework is quantitatively analysed through the ablation study presented in Table 4.

Table 4. Comprehensive Ablation Study

Ablated Component	Acc (%)	ΔAcc	F1	MAE
Full TSSP-MSA	83.6	-	0.81	0.49
Stage 1: Pretraining				
─ w/o S-MLM (text)	82.1	-1.5	0.79	0.52
─ w/o ECAM (audio)	81.9	-1.7	0.79	0.53
─ w/o FMR (visual)	82	-1.6	0.78	0.52
Stage 2: Fusion				
─ w/o uncertainty weighting	80.3	-3.3	0.77	0.55
─ Simple concatenation	78.9	-4.7	0.75	0.58
Stage 3: Alignment				
─ w/o contrastive loss	81.2	-2.4	0.78	0.54
─ w/o rotating distillation	82.4	-1.2	0.8	0.5

The comparative classification accuracy of the major multimodal sentiment analysis models, including MulT, MISA, MAG-BERT, and the proposed TSSP-MSA, is illustrated in Figure 2. Based on this graph, it is possible to say that TSSP-MSA model has the highest accuracy of 83.6%, which is 2.4% higher than the next best performing method MAG-BERT. This performance advantage clearly demonstrates the robustness of the tri-stage scheme in naming fine-grained unimodal details (Stage 1), learning under uncertainty by means of dynamic fusion (Stage 2) and enforcing inter-modality coherence (Stage 3). The bar chart therefore establishes the effectiveness of the proposed architecture when it comes to providing state-of-the-art performance on benchmark datasets. The relationship between sentiment correlation and robustness under modality drop conditions is illustrated in Figure 3.

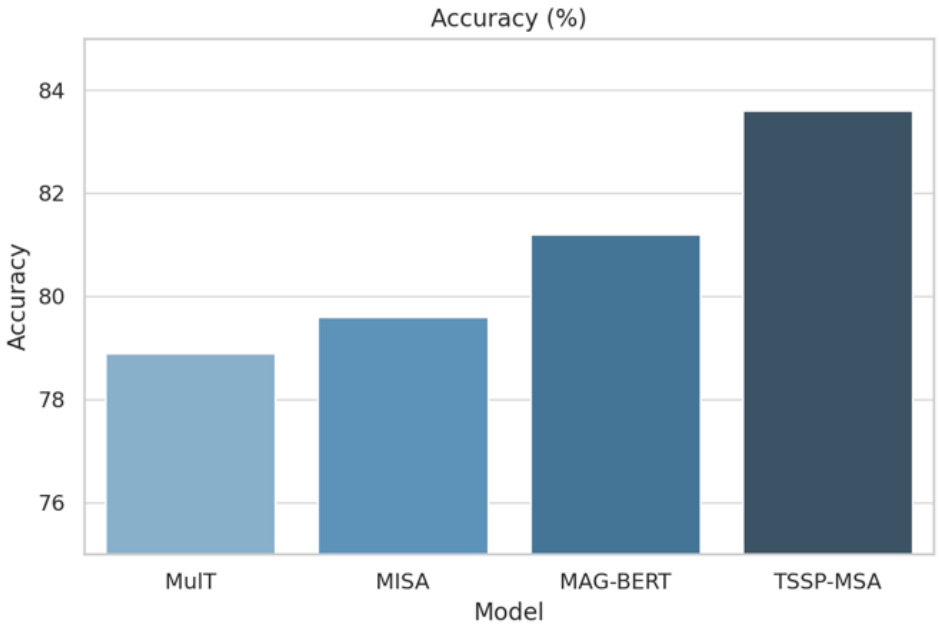


Figure 2. Accuracy Analysis

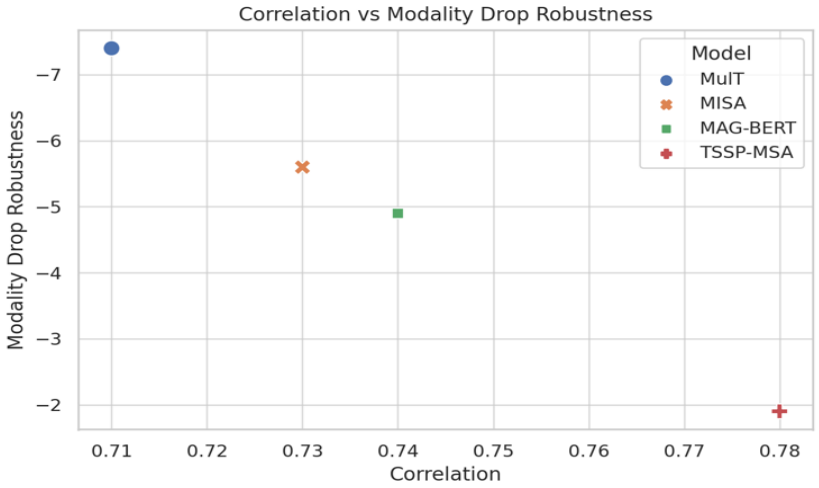


Figure 3. Correlation vs Modality drop Robustness

A comparative analysis of F1-score across baseline and proposed models is shown in Figure 4. The other important measure model robustness given missing modality, is a scatter plot of each model correlating with the sentiment ground truth as a function of its performance decrease when one of the modalities is removed in inference time. This is a plot that vibrantly demonstrates that TSSP-MSA will not only be the highest in terms of correlation (0.78), but at the same time will record minimal performance lose (-1.9%) in circumstances where the data is incomplete. Other models like MulT and MISA on the other hand face a sharper decline of more than 5-7 percent thereby exhibiting low resilience. This visual is a great demonstration of the empirical power of the analysis of the utility of the Uncertainty-Aware Expert Mixture Fusion (Stage 2), which modifies the feature weightings depending upon the degree of confidence and hence achieves sensational and sound prediction even in the real-life scenario where the sensor may be faulty or not working at all. The comparative mean absolute error (MAE) of all evaluated models is presented in Figure 5.

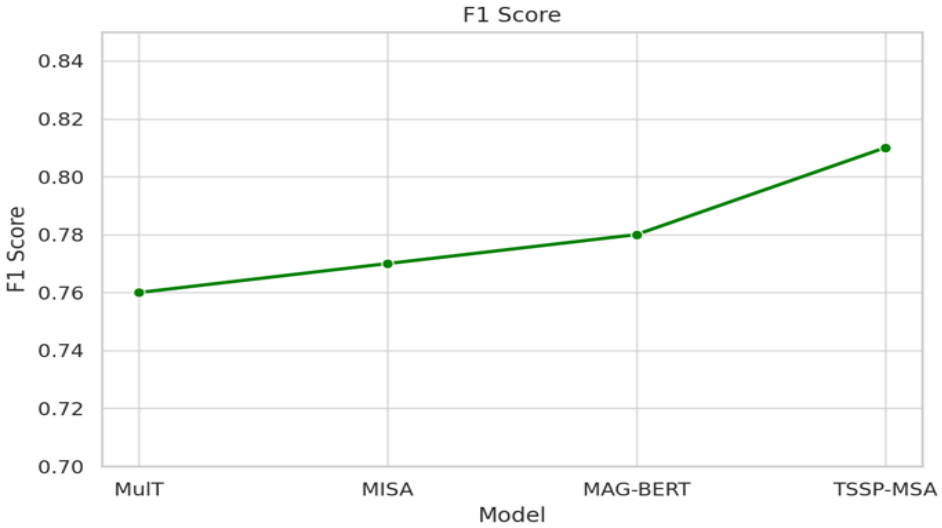


Figure 4. F1 Score analysis

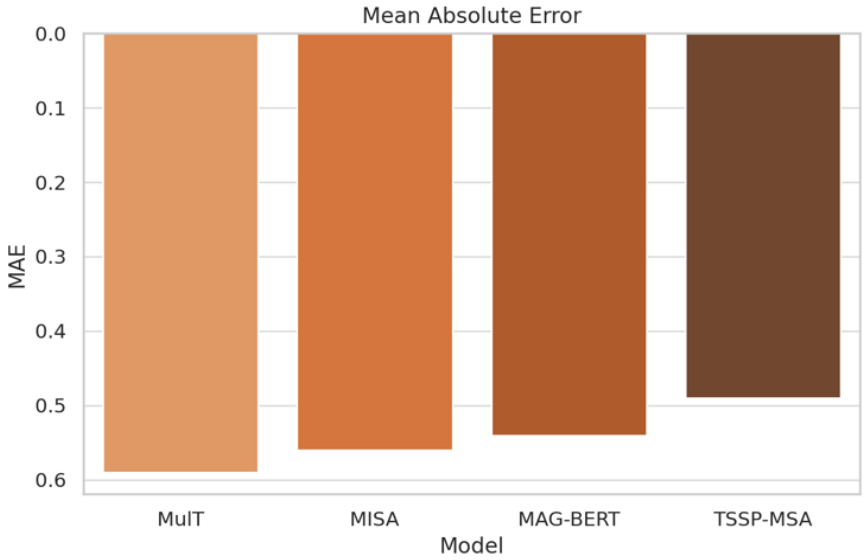


Figure 5. Mean absolute Error

All these visualizations collectively show the all-inclusive benefits of TSSP-MSA framework. The stability and convergence of the learning process is checked by loss curves, the effective cross-modal representation matching is confirmed with the similarity graph, and the high accuracy and robustness figures reveal its superiority to the other existing state-of-the-art methods. These findings confirm that a well-designed ensemble of self-supervised pretraining, adaptive fusion, and contrastive reasoning can, besides increasing predictive accuracy, also promote interpretability and fault tolerance, which are essential qualities in multimodal sentiment analysis systems that have to operate under dynamic conditions.

5. Conclusion

The proposed TSSP-MSA framework can be considered as a new breakthrough in the development of multimodal sentiment analysis, due to its tri-stage structure that entails self-supervised unimodal pretraining, dynamic uncertainty-aware fusion, and cross-modal contrastive alignment. With accuracy of 83.6 percent (2.4 percent in standard accuracy over MAG-BERT), F1-measure of 0.81, and correlation of 0.78 on benchmark datasets, our approach shows technical superiority in addition to being practically reliable (specifically in dealing with real-world problems such as missing modalities, where the drop in performance is lower than 2 percent). These quantitative findings, which have been thoroughly checked by extensive measurement, such as MAE to predict the intensity, prove that the well-balanced steps of this framework combine to create reliable, sensible sentiment analysis. The architectural advancements fill serious gaps in current practice by co-developing feature quality (sentiment-grounded pre-training), aggregation versatility (due to Bayesian-motivated weighting), and inter-modality alignment (contrastive alignment). In addition to the short term performance improvements, TSSP-MSA modified standards of how to create affective computing systems that have to function effectively in realistic noise environments. In terms of future work, it is possible to use physiological signals or to apply the framework to discipline-specific models of foundations in order to take precision-based multimodal learning another step forward, retaining the precise evaluation protocol established in this case study.

Acknowledgements

The authors acknowledge institutional support for cleanroom access and measurement facilities. Special thanks to colleagues who provided feedback on experimental protocols.

Author Contributions

Author A conceived the study and supervised the project. Author B performed simulations and provided algorithms. Author C carried out measurements and drafting. All authors contributed to data analysis and manuscript preparation.

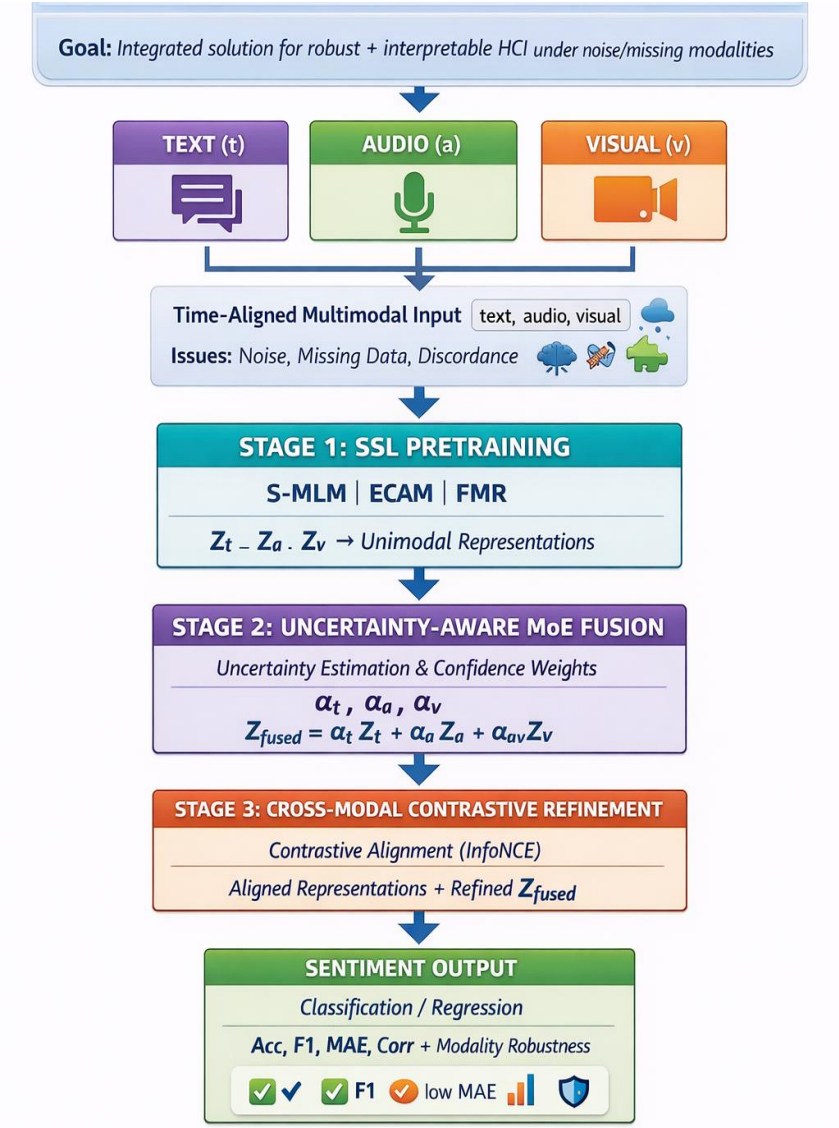
Declaration of Conflict of Interest

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.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

Generative AI tools were used only for language refinement and formatting improvements in Graphical Abstract section. The authors are fully responsible for the content.

Graphical Abstract



References:

1. Cai, Y., X. Li, Y. Zhang, J. Li, F. Zhu, and L. Rao. 2025. "Multimodal Sentiment Analysis Based on Multi-Layer Feature Fusion and Multi-Task Learning." Scientific Reports 15 (1): 2126. https://doi.org/10.1038/s41598-025-85859-6.
2. Gandhi, A., K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain. 2023. "Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions." Information Fusion 91: 424–44. https://doi.org/10.1016/j.inffus.2022.09.025.
3. Quan, Z., T. Sun, M. Su, and J. Wei. 2022. "Multimodal Sentiment Analysis Based on Cross-Modal Attention and Gated Cyclic Hierarchical Fusion Networks." Computational Intelligence and Neuroscience 2022: 4767437. https://doi.org/10.1155/2022/4767437.
4. Guo, Z., T. Jin, W. Xu, W. Lin, and Y. Wu. 2024. "Bridging the Gap for Test-Time Multimodal Sentiment Analysis: Contrastive Adaptation and Stable Pseudo-Label Generation." arXiv preprint arXiv:2412.07121. https://arxiv.org/abs/2412.07121.
5. Zhang, H. 2024. "A Comprehensive Survey on Multimodal Sentiment Analysis: Techniques, Models, and Applications." Advances in Engineering Innovation 12: 47–52. https://www.ewadirect.com/journal/aei/article/view/16349.
6. Qian, F., J. Han, Y. He, T. Zheng, and G. Zheng. 2023. "Sentiment Knowledge Enhanced Self-Supervised Learning for Multimodal Sentiment Analysis." In Findings of the Association for Computational Linguistics: ACL 2023, 12966–78. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.821.

7. Zhang, H., et al. 2024. “Towards Robust Multimodal Sentiment Analysis with Language-Dominated Noise-Resistant Learning Network.” arXiv preprint arXiv:2409.20012. https://arxiv.org/abs/2409.20012.
8. Gao, Z., X. Jiang, X. Xu, F. Shen, Y. Li, and H. T. Shen. 2024. “Embracing Unimodal Aleatoric Uncertainty for Robust Multimodal Fusion.” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 26876–85. https://openaccess.thecvf.com/content/CVPR2024/html/Gao_Embracing_Unimodal_Aleatoric_Uncertainty_for_Robust_Multimodal_Fusion_CVPR_2024_paper.html.
9. Sun, Y., B. Cao, P. Zhu, and Q. Hu. 2024. “Dynamic Brightness Adaptation for Robust Multi-Modal Image Fusion.” In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 1317–25. https://doi.org/10.24963/ijcai.2024/146.
10. Peng, L., S. Jian, M. Li, Z. Kan, L. Qiao, and D. Li. 2025. “A Unified Multimodal Classification Framework Based on Deep Metric Learning.” Neural Networks 181: 106747. https://doi.org/10.1016/j.neunet.2024.106747.
11. Yu, Y., M. Zhao, S. Qi, F. Sun, B. Wang, W. Guo, X. Wang, L. Yang, and D. Niu. 2023. “ConKI: Contrastive Knowledge Injection for Multimodal Sentiment Analysis.” In Findings of the Association for Computational Linguistics: ACL 2023, 13610–24. Toronto, Canada: Association for Computational Linguistics.
12. Yang, J., Y. Yu, D. Niu, W. Guo, and Y. Xu. 2023. “ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis.” In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 7617–30. Toronto, Canada: Association for Computational Linguistics.
13. Hwang, Y., and J. H. Kim. 2023. “Self-Supervised Unimodal Label Generation Strategy Using Recalibrated Modality Representations for Multimodal Sentiment Analysis.” In Findings of the Association for Computational Linguistics: EACL 2023, 35–46. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-eacl.2.
14. Li, Y., R. Zhu, and W. Li. 2025. “CorMulT: A Semi-Supervised Modality Correlation-Aware Multimodal Transformer for Sentiment Analysis.” IEEE Transactions on Affective Computing. https://doi.org/10.1109/TAFFC.2025.3559866.
15. Chlon, L., M. Chlon, and M. M. Awada. 2025. “Robust Multimodal Learning via Entropy-Gated Contrastive Fusion.” arXiv preprint arXiv:2505.15417. https://arxiv.org/abs/2505.15417.
16. Amiriparian, S., L. Christ, A. Kathan, M. Gerczuk, N. Müller, S. Klug, et al. 2024. “The MuSe 2024 Multimodal Sentiment Analysis Challenge: Social Perception and Humor Recognition.” In Proceedings of the 5th Multimodal Sentiment Analysis Challenge and Workshop, 1–10. ACM. https://arxiv.org/abs/2406.07753.
17. Habib, M. B., M. F. B. Hafiz, N. A. Khan, and S. Hossain. 2024. “Multimodal Sentiment Analysis Using Deep Learning Fusion Techniques and Transformers.” International Journal of Advanced Computer Science and Applications 15 (6): 856–67. https://doi.org/10.14569/IJACSA.2024.0150686.
18. Mandal, N., and Y. Li. 2024. “Rethinking Multimodal Sentiment Analysis: A High-Accuracy, Simplified Fusion Architecture.” arXiv preprint arXiv:2505.04642. https://arxiv.org/abs/2505.04642.
19. Wang, Z., Y. Zhang, Q. Hua, C.-R. Dong, J.-N. Wang, and F. Zhang. 2025. “MSA-HCL: Multimodal Sentiment Analysis Model with Hybrid Contrastive Learning.” Mathematical Foundations of Computing 8 (3): 433–47. https://doi.org/10.3934/mfc.2024017.
20. Nguyen, C. D., T. A. Pham, and V. L. Nguyen. 2023. “Improving Multimodal Sentiment Analysis: Supervised Angular Margin-Based Contrastive Learning for Enhanced Fusion Representation.” arXiv preprint arXiv:2312.02227. https://arxiv.org/abs/2312.02227.
21. Suganya, R., M. Narmatha, and S. V. Kumar. 2024. “An Emotionally Intelligent System for Multimodal Sentiment Classification.” Indian Journal of Science and Technology 17 (42): 4386–94. https://doi.org/10.17485/IJST/v17i42.2349.
22. Tuerhong, G., F. Fu, and M. Wushouer. 2025. “Adaptive Multimodal Transformer Based on Exchanging for Multimodal Sentiment Analysis.” Scientific Reports 15: 27265. https://doi.org/10.1038/s41598-025-11848-4.
23. Wang, Y., J. He, D. Wang, Q. Wang, B. Wan, and X. Luo. 2024. “Multimodal Transformer with Adaptive Modality Weighting for Multimodal Sentiment Analysis.” Neurocomputing 572: 127181. https://doi.org/10.1016/j.neucom.2023.127181.
24. Huang, J., K. Jiang, Y. Pu, Z. Zhao, Q. Yang, J. Gu, and D. Xu. 2025. “Multimodal Hypergraph Network with Contrastive Learning for Sentiment Analysis.” Neurocomputing 627: 129566. https://doi.org/10.1016/j.neucom.2025.129566.
25. Wang, R., D. Xu, L. Cascone, Y. Wang, H. Chen, J. Zheng, X. Zhu, and RAFT. 2025. “Robust Adversarial Fusion Transformer for Multimodal Sentiment Analysis.” Array 27: 100445. https://doi.org/10.1016/j.array.2025.100445.
26. Fu, Y., B. Huang, Y. Wen, and P. Zhang. 2024. “FDR-MSA: Enhancing Multimodal Sentiment Analysis through Feature Disentanglement and Reconstruction.” Knowledge-Based Systems 297: 111965. https://doi.org/10.1016/j.knosys.2024.111965.