Journal of Kerbala University, Vol. 23, Issue 1, March , 2026

Available at https://www.iasj.net/iasj

Iraqi Academic Scientific Journals

Journal homepage:
https://journals.uokerbala.edu.iq/index.php/UOKJ

Research Article

A Survey on Multimodal Sentiment Analysis in Short Videos: Datasets, Architectures, Fusion Strategies, and Challenges

[1,]Zahraa Saleeh idan [2,] Hiba Jabbar Aleqabie

[1,]st Department of Computer Science, Faculty of Education University of Kufa, Najaf, Iraq

[2,]nd Artificial Intelligence Engineering Department, College of Information Technology Engineering, Al-Zahraa University for Women.

Article Info

Article history:
Received 6-11-2025
Received in revised form 15-12-2025
Accepted 22-1-2026
Available online 31 -3-2026

Keywords: Cross-Modal Attention, Deep Learning, Emotion Recognition, Multimodal Fusion, Multimodal Sentiment Analysis

Abstract:

The rapid expansion of short-form video platforms has increased the demand for intelligent systems capable of understanding human emotions expressed across multiple modalities. Multimodal sentiment analysis (MSA) addresses this challenge by integrating textual, acoustic, and visual information within unified deep learning frameworks. This survey provides a structured and analytical overview of recent multimodal sentiment analysis approaches, emphasizing early, late, and hybrid fusion strategies. Representative architectures, including transformer-based models, convolutional neural networks, and recurrent neural networks, are comparatively examined, with particular focus on cross-modal attention mechanisms that enhance interaction modeling and predictive performance. Benchmark datasets such as CMU-MOSI, CMU-MOSEI, MELD, IEMOCAP, CH-SIMS, and TikTok-10M are analyzed to highlight their importance for training and evaluating multimodal systems across diverse linguistic and cultural contexts. Comparative findings suggest that hybrid transformer-based architectures generally outperform unimodal approaches, especially in detecting subtle emotional expressions such as sarcasm and mixed affect. Evaluation practices are reviewed across both classification metrics (Accuracy, Precision, Recall, F1-score) and regression metrics (MAE, RMSE), emphasizing metric selection under class imbalance. Despite progress, challenges persist, including limited annotated datasets.

Corresponding Author E-mail: hiba.jabbar@alzahraa.edu.iq

Peer review under responsibility of Iraqi Academic Scientific Journal and University of Kerbala.

1. Introduction

The intensive growth of the social media and mobile internet technologies has turned short-form video into a leading tool of daily digital consumption. Unlike traditional text-based communication, short videos combine visual, acoustic and textual messages and create multimodal content that is rich and complex. The videos encode overlaid affective messages using facial expressions, body language, voice intonation, background music, and verbal/written speech [1] .However, the actual decoding and processing of emotional information of such multimodal video data is a daunting task. Affective cues are often subtle, context-specific and disproportionate to constituent modalities. Pictorial messages can conflict with the textual mood, the intonation in the voice can be used to express sarcasm, and the changing of the scenes can break the time continuity. To this end, the heterogeneity, diversity, as well as the synchronization complexities of multimodal signals, make inference of sentiment reliable .[4] Further, differences in linguistic behavior, characteristics of speakers, and cultural differences increase the challenge of achieving strong multimodal sentiment recognition [2][3]. Although sentiment analysis has been widely used in social media analytics, public-opinion monitoring, crisis management, marketing, and financial prediction [3], modern-day communication is becoming more and more video-based than text-based, and analytical frameworks that can integrate two or more modalities simultaneously are needed .[4]As a result, deep-learning models, such as transformer-based models and convolutional neural networks, have taken the center stage in the multimodal sentiment analysis because they can learn both hierarchical and cross-modal representation [5]. The survey gives an in-depth overview of multimodal sentiment analysis using deep-learning in the context of short-form videos, discusses benchmark datasets, architecture paradigms, combination approaches, evaluation metrics, current challenges, and future research directions.

2.Background and theoretical foundation

Sentiment analysis is a strategic niche of the more general concept of affective computing, which hopes to endow artificial intelligence systems with the abilities to interpret, sense, and respond to the emotional conditions of human beings (as described in [6]). Previous sentiment analysis methodologies mostly relied on sentiment lexicons and traditional machine-learning paradigms and textual content was the main source of inferring opinion polarity and emotion trends [7]. Although such stratagems have admirable efficacy in organized textual milieus, they are still limited in their nature in capturing the entire spectrum of human affective expression. Human emotions, as such, are dynamic, multi-modal and complex. The communicative purpose is often expressed not only by the lexical material but also by the inflections of prosodics, facial muscles, gestures, and the visual presentation of the context. A text-based analysis is poorly suited to capture such phenomena as sarcasm, irony, prosodic modulation, and non-verbal cues, which have a substantive effect on sentiment reading [8]. Therefore, the use of linguistic data is only sufficient in a narrow and actually unrealistic scope in the case of social-media or video-based situations .The dramatic growth of multimodal data has motivated the development of multimodal sentiment analysis (MSA) which combines heterogeneous modalities such as textual, acoustic, and visual streams into integrative deep-learning frameworks [8]. With training in synergistic and complementary representations in these modalities, multimodal systems are able to capture subtle affective patterns which unimodal arrangements may miss. These models are optimized regularly using gradient-based learning algorithms to minimize predictive error with the use of back-propagation. Modern multimodal architectures are used on a regular basis to employ transformer-based models, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to enable end-to-end sentiment classification [9][10]. Every constituent modality injects different affective

information: textual frameworks like BERT-based architectures encode semantic and contextual essence within language [11]; acoustic models disaggregate pitch, timbre, rhythm, and prosodic contours that reflect affective states and play a central role in human-computer interaction situations, e.g., voice-assistant systems [12]; visual models attempt to interpret facial expressions and body gestures, which are key elements of non-verbal discourse, but also are still difficult to interpret by automated systems .[13] Although methodological advances have been made, the smooth integration of textual, acoustic and visual modalities has remained a complex issue. The heterogeneity of feature spaces, modality time-desynchronization, and modality-specific noise are all working together to make reliable integration difficult. In a bid to reduce these barriers, researchers have come up with a spectrum of fusion solutions, which include early fusion (feature-level fusion), late fusion (decision-level fusion), and hybrid fusion (intermediate cross-modal interaction) [14]. Hybrid architectures, especially those with attention mechanisms and transformer layers, attempt to find a reasonable balance between representational depth and interpretability and enhance robustness against incomplete or missing modality data .Together multimodal techniques provide a more comprehensive and human-grounded sentiment perception system compared to unimodal and text-based systems. These models improve the ability to detect even the slightest affective states, such as sarcasm, ambivalent affect, and context-dependent ambiguity, through the use of both verbal and non-verbal signals [15].

3. Survey Methodology

To ensure a structured and comprehensive review, this survey follows a structured literature selection process covering studies published between 2015 and 2025. Relevant works were retrieved from major academic databases using predefined keywords related to multimodal sentiment analysis and short-form video. After screening and applying inclusion criteria—focusing on deep learning-based multimodal approaches evaluated on benchmark datasets—82 studies were selected for analysis. The selected literature was categorized based on architectural design and fusion strategy to enable systematic comparison.

4 .Datasets:

The availability of multimodal datasets is a key pillar in developing sentiment and emotion analysis models for short videos, enabling researchers to train models and test their ability to understand the interaction between text, audio, and visual content In this context, several datasets have emerged in recent years that specialize in multimodal sentiment analysis (Multimodal Sentiment/Emotion Analysis), varying in the nature of their content, language, and number of samples in them, and the type of media they contain (text, audio, image/video). Below is a review of the most prominent of these groups used in recent studies related to short video content[30][31][32][33][34] CMU-MOSI Dataset[27] .

CMU-MOSI is one of the first multimodal sentiment analysis datasets, developed at Carnegie Mellon University in 2016.This dataset focuses on sentiment intensity analysis and includes approximately 2,199 short video clips taken from YouTube reviews. The dataset contains three media: text (transcriptions), audio, and images, allowing for the study of the relationship between spoken language, tone of voice, and facial expressions. Annotation is performed at the sentence level, with each segment assigned a value on a scale from -3 (very negative) to +3 (very positive).

CMU-MOSI is a key benchmark for testing deep models in accurate sentiment analysis of short videos [25]. CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) is an expanded version of the MOSI, published in 2018 by Zadeh et al. of Carnegie Mellon University. It includes over 23,000 short video clips taken from over 1,000 YouTube videos, covering a wide range of topics and speakers. It contains three media: text, audio, and video, with instructions related to both emotion intensity and emotion rating. Each clip is rated

on a scale from -3 to +3, and emotions are classified into six categories: happiness, sadness, anger, disgust, surprise, and fear. It is widely used in fine-grained emotion recognition research and the study of the generalization of models across speakers in short videos.

MELD Dataset[35] the MELD (Multimodal Emotion Lines Dataset), released in 2019 by Poria et al. of Declare Lab, is one of the most popular databases for emotional dialogue analysis. It consists of approximately 13,000 sentences within 1,400 dialogues taken from the TV series Friends. It includes text, audio, and video media. Each sentence carries two categories: sentiment (positive, negative, neutral) and emotion (anger, joy, sadness, surprise, fear, disgust, neutral). MELD is widely used in studies of sentiment analysis of short dialogues and interactions between speakers.

OMG-Emotion Dataset[36] The OMG-Emotion (One-Minute Gradual Emotion) dataset was created in 2018 by Barros et al. as part of the IJCNN Emotion Recognition Challenge. It contains approximately 420 YouTube videos ranging in length from one to two minutes, showing gradual changes in emotion over time. The three media include text (when available), audio, and video. It includes continuous labels for emotional dimensions such as arousal and valence. It is designed to evaluate models capable of capturing the temporal dynamics of emotions, especially in short videos.

CH-SIMS Dataset [37] The CH-SIMS (Chinese Multimodal Sentiment Analysis) dataset was introduced in 2020 by Yu et al. at Tsinghua University, and an extended version (v2.0) was released in 2022 .The original version contains 2,281 videos, while the second version contains over 14,000 Chinese-language videos. It covers three media (text, audio, and image) and provides single- and multi-modal labels for sentiment intensity on a scale from -1 to +1 .This corpus is important for studying multi-modal sentiment analysis in Chinese and for comparing the consistency of different media.

MOUD Dataset [38] the MOUD (Multimodal Opinion Utterances Dataset) was developed in 2013 by Perez-Rosas et al. and focuses on sentiment analysis in Spanish .It includes 80 video clips containing approximately 498 sentences taken from YouTube reviews .Text, audio, and video media are available, and sentences are categorized into three categories: positive, negative, and neutral .It is one of the first non-English datasets used in cross-cultural sentiment analysis research.

IEMOCAP Dataset [39] the IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset was created in 2008 at the University of Southern California (USC) SAIL Lab. It includes 12 hours of dialogue from over 10 professional actors, along with audio, text, and video recordings, as well as Motion Capture data. For facial and body movements. The clips are categorized into six emotions: anger, happiness, sadness, frustration, enthusiasm, and neutral, in addition to continuous values (valence, arousal, and dominance). It is considered one of the oldest references in the field of multimodal emotion recognition.

TikTok-10M Dataset[40] the TikTok-10M dataset is one of the most recent datasets (published between 2023 and 2024 on Hugging Face). It contains over 10 million short videos from the TikTok platform, including audio, video, and accompanying text such as comments and hashtags. It is not issued by a single academic university, but rather is published on the Hugging Face Dataset Hub platform by an entity known as The Data Company (TDC). This open-source research company collects multimedia social data for artificial intelligence research purposes. This dataset is known as openly curated social media multimodal dataset for research use. Most research papers working on multimedia analysis of short video cite this dataset as a large-scale multimedia dataset for pre-training or fine-tuning short video models. This description is based on the official information provided on dataset page The-Daya-Company, 2024, on the Hugging Face platform.

Short-form video platforms like TikTok and YouTube present unique challenges for

multimedia sentiment analysis. These videos are typically short (15-60 seconds), with rapid scene changes and limited contextual information. Furthermore, multimedia signals may be inconsistent, as audio, visual, and textual signals do not consistently convey the same emotions, making integration and accurate analysis more difficult.

VGGSound Dataset[41] the VGGSound dataset was released in 2020 by the Visual Geometry Group at the University of Oxford. It contains over 200,000 10-second audio-visual clips, categorized into 310 audio categories (e.g., laughter, applause, dog barking). Although it is not specifically designed for emotion, it is commonly used to pre-train models for extracting audio and visual features before applying them to sentiment analysis tasks.

AVE Dataset[42] the AVE (Audio-Visual Event Dataset) was created in 2019 at Tsinghua University in China. It contains 4,143 10-second video clips, each depicting a visually and auditorily distinguishable event (such as "a baby crying" or "a car horn").It incorporates audio and video simultaneously and is used to evaluate audio-visual event fusion models. Although it is not intended for sentiment analysis, it is used to train multimedia models for short videos Table 1 represents a summary of all the datasets.

Table 1: Summary of the datasets

No.	Dataset Name	Year	Modalities	Language	Samples / Videos
1	CMU-MOSI	2016	Text, Audio, Visual	English	2,199 video segments
2	CMU-MOSEI	2018	Text, Audio, Visual	English	23,453 video segments from 1,000+ YouTube videos
3	MELD	2019	Text, Audio, Visual	English	13,000 utterances from 1,400 dialogues
4	OMG-Emotion	2018	Text (when available), Audio, Visual	English	420 YouTube videos (1–2 minutes each)
5	CH-SIMS (v1/v2)	2020 / 2022	Text, Audio, Visual	Chinese	v1: 2,281 videos / v2: 14,000+ videos
6	MOUD	2013	Text, Audio, Visual	Spanish	498 utterances from 80 videos
7	IEMOCAP	2008	Text, Audio, Visual, Motion Capture	English	12 hours of dialogue (≈10 actors)
8	TikTok-10M	2023–2024	Text, Audio, Visual	Multilingual (mainly English & mixed dialects)	10 million short videos
9	VGGSound	2020	Audio, Visual	English / Global	200,000 10-sec clips (310 sound classes)
10	AVE	2019	Audio, Visual	English / Global	4,143 10-sec video clips

5. Related Works

Recent advances in multimodal sentiment analysis (MSA) for short-form video have shifted from simple feature concatenation techniques to sophisticated transformer-based frameworks capable of modeling complex cross-modal interactions.

To provide structural clarity, representative works are grouped below according to architectural type rather than chronological order, following common categorizations adopted in recent multimodal sentiment analysis surveys [31][32].

Table 2: Classification of Representative Works by Architecture Type

Model / Study	Architecture Type	Fusion Strategy	Core Mechanism	Dataset
Poria et al. [16]	Late Fusion	Decision-Level	Parallel decision fusion	MOSI
Kumar & Vepa [21]	Hybrid Fusion	Intermediate (Attention-Based)	Self-attention, Cross-attention, Learnable Gating	CMU-MOSI / CMU-MOSEI
Yang et al. [17]	Hybrid (Graph + Transformer)	Intermediate	Graph learning + Transformer encoder	CH-SIMS / MOSI / MOSEI
Gulanbaier et al. [18]	Transformer-Based Hybrid	Intermediate	Adaptive multimodal Transformer with modality exchange	MOSI / CMU-MOSEI / CH-SIMS
Dellbrouck et al. [19]	Transformer-Based Joint Encoding	Early/Intermediate	Transformer-Based Joint Encoding (TBJE)	CMU-MOSEI
Lee et al. [20]	Hybrid Fusion	Early & Late Fusion	Multi-head attention for cross-modal interaction	Multimodal emotion datasets
Qiu et al. (HAEMSA) [22]	Hybrid (Expert-Based)	Intermediate	Evolutionary optimization + cross-modal knowledge transfer + multi-task learning	CMU-MOSI / CMU-MOSEI / IEMOCAP
Xie et al. [23]	Graph-Based Hybrid	Intermediate	Domain generalization + Graph Neural Networks	IEMOCAP / CMU-MOSEI / CMU-MOSI
Qiu et al. (MULG) [24]	Graph-Attention Hybrid	Intermediate	Cross-modal attention + learned graph units	CMU-MOSI / CMU-MOSEI / IEMOCAP
Zadeh et al. (DFG) [25]	Dynamic Graph Fusion	Hybrid	Dynamic Fusion Graph (interpretable fusion)	MOSI / MOSEI
Cai et al. [26]	Hybrid Fusion	Intermediate	Unimodal Feature Extraction Network + Multitask Fusion Network	Multimodal sentiment datasets
Gajjar et al. [27]	Transformer-Based Early Fusion	Early Fusion	BERT-based encoders with early multimodal fusion	Multimodal datasets
Chen et al. (VCAN) [28]	Hybrid (Auxiliary Network)	Intermediate	Video-based Cross-modal Auxiliary Network + ensemble learning	Multimodal sentiment datasets

Liu et al. [29]	Variational Hybrid Model	Intermediate	Information bottleneck + modality consistency/specificity decomposition	MOSI / MOSEI

Table 2 provides a comparison of three major paradigms of architecture, including early fusion, late fusion, and hybrid or frameworks of transformers (see the original work expressed in references [31][32]). The early fusion approach requires models to merge directly modality-specific feature spaces, whereas models based on late fusion keep the modality networks apart and later combine their eventual predictive performances [25][35]. Modern studies place hybrid architectures in the focus of the investigation, in large part due to the integration of intermediate cross-modal attention processes and shared-private representation learning units (see [17][28][43]). Specifically, hybrids based on transformers are the best at learning inter-modal relationships using multi-headed attention mechanisms (as discussed in [28][43][44]) whereas graph-based and variational models are dedicated to improving robustness, interpretability, and domain generalizability (see [23][29]). Despite the fact that Table 3 shows a categorical taxonomy of representative models, Section 7 focuses on a quantitative analysis with the focus on performance metrics and experimental conditions**.**

6. Deep Learning Architectures

Multimodal sentiment analysis (MSA) systems are built on deep learning architecture. The deep models used, unlike traditional machine learning models like Support Vector Machines (SVM) or Naive Bayes classifiers, use modality-specific neural networks to process textual, acoustic and visual data individually and then combine them later. This structure eases the hierarchical representation learning and enables more powerful feature extraction before fusion.[37] Current multimodal systems usually use transformer-based encoders, convolutional neural networks (CNNs), and recurrent neural networks (RNNs) so that they can perform end-to-end sentiment classification and emotional polarity detection [37].

6.1 Early Fusion

Early fusion, which is also known as feature-level fusion, merges feature that are modality-specific before the classification phase. Textual, auditory, and visual streams provide features, which are then concatenated and input to a common neural network and in such a way, the model learns cross-modal interactions at the earliest stages of processing. The strategy is especially beneficial in cases when affective signals are spread in heterogeneous modalities. As an illustration, the text can have a positive overture with the voice intonation or facial expression showing sarcasm or negativity. Through learning these feature spaces jointly, early fusion enables the model to learn these modality-based inconsistencies [25] [27]. However, early fusion is susceptible to the issues of modality heterogeneity and temporal heterogeneity, which can spread noise and decrease the robustness when inter-modal correlations are weak [25].

6.2 Late Fusion

Decision-level fusion, also known as late fusion, processes each of the modalities independently using special models and then combines their predictions at the resulting decision point [35]. Such a methodological decision gives the system the quality of flexibility and resistance to the loss or corruption of separate modalities. As an example, the Gated Mechanism of Attention based Multimodal Sentiment Analysis (GBS) uses the transformer-based text encoders, a CNNBiLSTM pipeline of acoustic modelling, and convolutional neural network models, including VGG, to extract visual features. Instead of combining intermediate representation of features, the ultimate classification results are combined through

confidence-weighted schemes [35]. Though, the late fusion removes the sensitivity of the system to noise and imbalance of modalities, it does not explicitly model cross-modal interactions, and thus may underutilize temporal and contextual relations across modalities. Hybrid fusion strategies attempt to strike a compromise between the strengths of early and late fusion by combining modalities at intermediate levels with preserving modality-specific representations [19]. Such systems often incorporate cross-modal transformer layers that enable learning of both common and private cross-modal representations [28]. Additional mechanisms, including contrastive feature decomposition (e.g., ConFEDE), are also included to reduce cross-modal interference and support discriminative ability [17]. The fused representations are further optimized in some pipelines by Graph Neural Networks (GNNs) to encode temporal dependencies, or by Attention-Weighted Pooling (AWP) to emphasize salient multimodal features before classification. Hybrid models made using transformers, such as MISA and MAG-BERT, have demonstrated better results in cross-modal dependency modelling, especially on short-form and conversational video data [43][44].

6.3 Hybrid Fusion

Hybrid fusion aims to balance the strengths of early and late fusion by integrating modalities in intermediate layers while preserving the representations specific to each modality.

These architectures often incorporate cross-modal transformer layer to learn shared and private representations across modalities [28], such as contrastive feature decomposition (e.g.,ConFEDE), to reduce cross-modal interference and enhance discriminative learning [17].In some frameworks, the fused representations are further refined by using Graph Neural Networks(GNNs) to model temporal dependencies or Attention Weighted Pooling(AWP) to identify prominent multimodal signals before classification.

Transformer-based hybrid models as Multimodal Interaction and shared-private Architecture (MISA) and Multimodal Adaptive Gate-Bidirectional Encoder Representations form Transformer(MAG-BERT) have demonstrated strong performance in modeling cross-modal dependencies, particularly on short and conversational video datasets datasets.[43][44].

6.4 Unified Multimodal Pipeline Framework

In order to provide a single conceptual framework to multimodal sentiment analysis, Figure 1 presents a generalized pipeline of integrating textual, acoustic and visual modalities in one architecture. Every modality is first encoded by a specific encoder to obtain high-level representations. Such representations are then fused through a fusion module- early, late or hybrid- based on the architectural design. The resulting merged representation is then provided to a classifier that predicts sentiment polarity or strength of emotion. This coherent system predicts the parallel processing architecture of multimodal systems and emphasizes the critical role played by fusion processes in shaping overall performance.

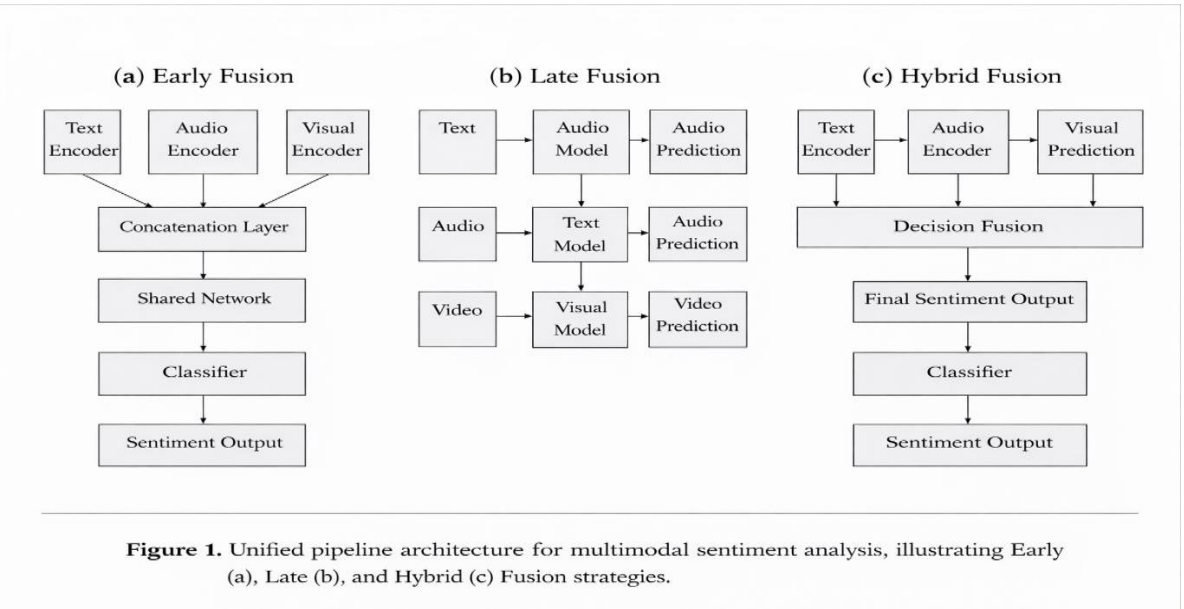

Figure 1. Unified pipeline architecture for multimodal sentiment analysis, illustrating Early (a), Late (b), and Hybrid (c) Fusion strategies.

Figure1 illustrates a generalized multimodal sentiment analysis across
multiple media (text, audio, and video). Each modality is processed by a dedicated encoder to extract high-level features, which are then combined using early, late, or hybrid fusion strategies. Finally, the fused representation is passed to a classifier to predict the sentiment polarity or intensity of the emotions.

6.5 Critical Comparative Analysis

Although hybrid fusion models based on transformers are often found to perform better than early or late fusion methods, one must remember that the performance difference depends on a variety of factors, such as the nature of datasets, task formulation, and evaluation procedures [43][44] .As an example, the models that achieve very high accuracy, such as over 90 percent are commonly evaluated in binary sentiment classification systems, which inherently are not as demanding as multi-class or regression-based sentiment prediction regimes.[45] Conversely, more difficult methods like CMU-MOSEI add multi-speaker variability, noisy modalities, and continuous sentiment scales, which result in more realistic performance results with a relatively lower price.[31] Scale of datasets also has a great impact. Massive corpora like MOSEI enable transformer-based models to harness multi-head attention to the fullest to learn cross-modal alignments and representations [28]. On smaller datasets, however, less complex fusion strategies can still be competitive due to reduced over-fitting risk and smaller training samples, e.g. MOSI or MOUD. The other important factor is modality dominance. Textual features tend to dominate overall performance in many benchmark datasets, but acoustic and visual streams provide complementary, although sometimes noisy, information. This observation raises critical concerns as to whether the reported gains can be attributed to multimodal integration or just to good unimodal backbones like BERT or RoBERTa.[31] Lastly, there should be consideration of computational complexity and scalability. Models based on hybrid transformers, even at state-of-the-art, require a large number of computational resources to train and implement, which may be a limiting factor in real-time short-video contexts.[34] Therefore, the fusion strategy should not be solely blamed as an origin of performance differences. Instead, they are a manifestation of a joint influence of dataset difficulty, label granularity, modality quality, evaluation setting, and

backbone architecture. An extensive quantitative comparison of representative models on a variety of datasets, evaluation procedures, and fusion strategies can be found in Section 7, which continues to shed light on the differences in performance and the context

7. Quantitative Comparison of Representative Models

Building upon the architectural grouping presented in Section 5, Table 2 provides a quantitative comparison of representative models in terms of reported performance, dataset characteristics, and fusion strategy.

Table 3: Quantitative Comparison of Representative Multimodal Sentiment Analysis Models

Model	Architecture Type	Fusion Strategy	Dataset	Reported Performance
DFG (Zadeh et al.)	Graph-based + LSTM	Dynamic Fusion	CMU-MOSEI	Competitive performance (ACL 2018)
MISA	Transformer-based	Hybrid	MOSI, MOSEI	Strong cross-modal learning
MAG-BERT	Transformer-based	Hybrid (Attention-guided)	MOSI, MOSEI	Significant F1 improvement
GBS	CNN + BiLSTM	Late Fusion	MOSI	83.9% Accuracy
Adaptive Multimodal Transformer	Transformer	Hybrid	MOSI, MOSEI	89.18% (MOSI)
Gajjar et al.	Transformer	Early Fusion	MOSEI	97.87% Accuracy
Lee et al.	Multi-head Attention	Early vs Late	Custom Dataset	72.39% Accuracy

When the comparative data provided in Table 3 is examined, some salient observations are found. To begin with, hybrid fusion models that are based on transformers tend to show strong performance on benchmark corpora including CMU-MOSEI and MOSI, which is supported by modern studies that highlight the effectiveness of attention-based architecture in cross-modal interaction representation [43][44]Secondly, models achieving remarkably high accuracy are typically tested in a binary sentiment classification model, which is less complicated by definition than a multi-class or fine-grained regression task. As an example, the performance on MOSI binary classification reported is likely to shadow the performance on multi-class MOSEI settings.[25] Thirdly, the size and the nonhomogeneous nature of the data have a significant impact on performance results. Big corpora like CMU-MOSEI have enough diversity to allow transformer-based models to take advantage of multi-head attention mechanisms. Smaller datasets in contrast can limit the generalization ability of deep architectures and increase the likelihood of overfitting. Therefore, performance differences cannot be interpreted as a sign of architectural excellence exclusively, instead, they are the joining impacts of the complexity of the dataset, formulation of tasks, fusion strategy, backbone strength, and the evaluation protocol.

8.Explainability in Multimodal Sentiment Analysis

With the ever-growing complexity of the multimodal sentiment analysis systems, the demand of interpretability has become the critical issue of modern studies. Contemporary transformer-based architectures are often black box models and thus fail to provide insight into the exact modalities or features that have the most significant effect on the ultimate prediction. As a result, a set of explainable artificial intelligence (XAI) methods have been utilized to increase the transparency of multimodal structures. The methods of attention-visualization can be used in the field of textual modality to identify the influential lexical items that determine sentiment classification. Visual analysis has also been helped by methods that include Grad-CAM that highlights the parts of the image that make predictions of emotions. In the case of multimodal fusion models, SHAP (SHapley Additive exPlanations) has been essential in the measurement of relative significance of each of the modalities in the final decision. The importance of explainability is particularly acute on the short-video platforms, where biased predictions can have a concrete impact on the content visibility and perception by the user. Open multimodal systems do not only promote trust but also make the process of debugging simpler, permit fairness analysis, and enable responsible deployment. Although such developments have occurred, little exploration of explainability in multimodal sentiment analysis has been done. This is especially the case with cross-modal attention processes, where modal interaction is complex and therefore offers significant interpretive difficulties.

9. Ethical Considerations and Dataset Bias

The emerging use of multimodal sentiment analysis on short-video applications like Tik Tok and YouTube Shorts is associated with a multiplicity of ethical issues. Massive data mining of social media usually comes with privacy concerns, as user-created material is generally subject to processing without the knowledge of downstream users . Another serious problem is the bias of algorithms. Many benchmark datasets, such as CMU-MOSI and CMU-MOSEI, are mostly English-based and reflect specific cultural conditions. Models that are only trained on these corpora run the risk of having reduced performance when applied to other languages, accents or other demographics. Modality imbalance might also lead to bias. As an example, speech recognition can be found to have different degrees of performance between different accents, and face analysis models can exhibit dissimilar accuracy between ethnicities. Such differences may lead to unfair predictions of the sentiment and the unintentional intensification of the established stereotypes. In addition to that, the introduction of automated sentiment classification systems to short video platforms can influence the visibility of content, moderation policies and recommendation algorithms and thus affect user exposure and engagement patterns. In turn, the following research should anticipate equity-conscious training, creation of more varied and representative databases, openness in model implementation, and a conscientious attitude towards introducing multimodal AI systems.

10 .Short-Form Video Specific Challenges

Compared to the classic long-form video sentiment analysis, short-form video platforms like TikTok and YouTube Shorts are complicated by a variety of unique technical and contextual problems. Such platforms require a re-conceptualization of traditional affective modelling frameworks because of the rapidity of these platforms, as well as, their purpose-driven brevity. To start with, the length of short videos is usually 15-60 seconds, which condenses

the contextual information and dramatically impairs the provision of time-related information that is critical in the process of recording subtle emotional shifts. Second, the high frequency of scene transitions and dynamic editing techniques that are characteristic of short videos tend to create sudden modality switches and make it difficult to synchronize the textual transcripts, acoustic cues, and visual frames in time. Third, deliberate modality incongruence is a stylistic tool; an example is background music intentionally incongruent with spoken speech, or visual displays that are incongruent with narration, in order to create comic or sarcastic effects, which provokes cross-modal attention processes. Fourth, expressive exaggeration, digital filters or audio overlays, which are frequently preempted by short-form content, may alter the natural emotional cues on which corpus-based models are based. Thus, it is unsurprising that multimodal models trained on long videos of conversations are unlikely to effectively generalise to short-form content unless specifically designed and applied with architecture adjustments and optimised temporal modelling methods.

11. Evaluation Metrics in Multimodal Sentiment Analysis

Evaluation metrics have taken an essential place in modern studies of multimodal sentiment analysis, acting as the test of how effective suggested systems are. However, the rational choice of these measures is inevitably connected with the specific formulation of the problem in question, inherent features of the data set, and distributional properties of the classes considered[45] . In the case of the classification problem being reduced to binary discernment of sentiment polarity, the canonical repertoire includes Accuracy, Precision, Recall, and F1-score. Despite providing an aggregate measure of model correctness, Accuracy becomes highly unhelpful in the presence of extreme class imbalance, a situation that is common in sentiment corpora. In this regard, Precision and Recall are often combined as a composite measure, the F1-score, which is often used as a more accurate measure of performance in these skewed settings [45]. Conversely, triadic sentiment classification cases (which include positive, neutral and negative) require a more sophisticated measure. Macro-F1, specifically, presupposes a dominant role since each of the classes has an equal weight and, thus, the distortion caused by the unbalance of the classes is reduced. This quality makes Macro-F1 particularly relevant to datasets with high differences in the frequency of classes, like CMU-MOSEI and MELD [47]. Sentiment prediction efforts that are regression-oriented utilize a unique set of quantitative evaluations. The Mean Absolute Error (MAE) [46] and Root Mean Square Error (RMSE) [50] provide direct approximations of the difference between the intensity forecasts of sentiment as provided by a model and the ground-truth values. Signal-to-Noise Ratio (SNR) is used to measure audio quality and robustness, while Pearson Correlation Coefficient (r) and Concordance Correlation Coefficient (CCC) are used to measure correlation and agreement between predictions and actual values [46][47][48][49]

Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) are commonly used to evaluate image or video quality in the visual modality [50][51]. Another perceptual metric that is in line with human visual perception is Gradient Magnitude Similarity Deviation [52].In addition to these, correlation measures, such as the Pearson correlation coefficient (r) and the Concordance Correlation Coefficient (CCC), have been used to evaluate the concordance of predicted and actual dimensions of emotion, thus providing a complementary view of model fidelity [49].[48] It is the responsibility of scholars to be keen on scrutiny when comparing

reported performances in different studies. Comparison theses should take into consideration differences in assessment schemes, partitioning of data sets, and the fineness of classification schemes. In other cases, what may seem to be performance improvements may actually be a result of simplifying the task by design, e.g. of reducing a multi-class problem to a binary schema.

Therefore, the selection and definition of evaluation measures should be closely related to the empirical properties of the data and the overall research goals and not be determined based on the actual accuracy rates only. Scholars should use methodology protection mechanisms such as class-weighted loss functions, focal loss schemes, and macro-averaged metrics, to obtain an equitable and fair evaluation of model performance across the range of sentiment classes in datasets with strong imbalance in classes.[52][53]

12. Challenges and Limitations

Although the progress in the field of multimodal sentiment analysis has been notable, there are still a number of vexing issues that have dogged the maturity of this field. To start with, lack of information and subjectivity of annotation still makes emergent models weak. The quality of multimodal corpora requires the delivery of textual, acoustic, and visual labels in a synchronous manner, which is both expensive and time-consuming. In addition, sentiment annotations are often a reflection of annotator bias and cultural subtext, thus providing variation between datasets.

Second, the accurate time-warping of different modalities is one of the main technical challenges. Not only are textual transcripts, audio wave forms, and visual frames all working on different timeframes, synchronizing them to the point of accuracy is a tedious task, particularly in the context of short videos that contain rapid scene change-ups and condensed contextual details. Third, there is a structural complication in the imbalance of modalities. Textual cues have an overrepresented predictive performance in many benchmark suites, with audio and visual streams playing a relatively small role. Such an imbalance raises the question of whether there is actual exploitation of cross-modal interaction because the models can be based largely on sound textual backbones instead of comprehensive multimodal integration. Fourthly, there is a daunting limit of computational complexity. Transformer-based architectures, despite being powerful, require significant resources to train and execute inference, thus constrained in use in real-time or resource-constrained environments. Fifth, the applicability of these systems to other languages, cultures and areas of application is restricted. Primarily English based standards do not reflect the variety of communicative styles, accents, and culturally defined displays of affect. Lastly, the problem of interpretability also remains a research frontier. Though the attention mechanisms provide some level of transparency, scholars still struggle to understand the details of cross-modal interactions that occur in deep architectures . Addressing these hurdles will require the creation of more heterogeneous datasets, the introduction of fairness-conscious modeling paradigms, the creation of more effective architectures, and the improvement of methods that explain cross-modal effects.

13. Future Directions

Future research in multimodal sentiment analysis is expected to developed along several important direction.

First, there is a growing need to assemble large scale, culturally heterogeneous, and multilingual multimodal data. Most of the available benchmarks are mostly English-based, hence limiting cross-cultural generalization. The development of datasets that accurately reflect multiple communicative patterns and demographic subtleties is essential to the development of

models that are both fair and strong [31][32]

Second, future models should aim at refining cross-modal alignment mechanisms. State-of-the-art transformer architectures enhanced with dynamic attention schemes and modality-adaptive weighting schemes are set to capture the asynchronous interaction between textual, auditory and visual cues, in the context of short-form video modalities, more effectively [1][55]

Third, efficiency-conscious designs by the architects of the future systems should significantly reduce the computational overhead. Multimodal transformers that are lightweight, along with knowledge-distillation methods and parameter-efficient fine-tuning techniques, promise to make real-time deployment on mobile and edge devices possible.[34]

Fourth, the research into explainable multimodal artificial intelligence can be of significant benefit to the field. Attention visualization, SHAP-based attribution, and modality-level importance analytics techniques have the potential to significantly improve the transparency and make more people trust automated sentiment systems.[54]

Fifth, researchers ought to question the resilience of multimodal models when in the presence of noisy or intentionally tampered content, which is a common occurrence on short-video platforms where the presence of background music, filters, and post-processing effects can obscure true emotional expressions [1][55]

Lastly, the combination of multimodal sentiment analysis with large-language models (LLMs) and multimodal foundation models is an opportunity to develop contextual reasoning and the overall understanding of emotions in a variety of areas. By following these guidelines, we are bound to develop multimodal sentiment analysis systems, which will not only be more dependable and comprehensible but also socially accountable.

14. Conclusion

This survey presented a comprehensive review of multimodal sentiment analysis in short-form video content, with a focus on datasets, deep learning architectures, fusion strategies, evaluation metrics, and emerging challenges. The analysis highlights the growing dominance of transformer-based hybrid fusion models due to their ability to capture complex cross-modal interactions. However, performance variations across studies are strongly influenced by dataset characteristics, task formulation, class imbalance, and evaluation protocols rather than architectural design alone. The review also emphasizes critical challenges in the field, including temporal misalignment between modalities, dataset bias, computational complexity, interpretability limitations, and the unique constraints of short-form video platforms. These challenges indicate that multimodal sentiment analysis remains an evolving research area rather than a fully matured solution. Furthermore, ethical considerations and fairness-aware modeling are becoming increasingly important as multimodal systems are deployed in real-world social media environments. Addressing bias, ensuring transparency, and improving cross-cultural generalization are essential for responsible AI development. Future advancements are expected to emerge from improved cross-modal alignment mechanisms, efficient transformer architectures, explainable multimodal frameworks, and integration with large-scale foundation models.

Overall, multimodal sentiment analysis represents a promising yet complex domain that requires balanced progress in accuracy, efficiency, interpretability, and ethical responsibility to achieve truly emotionally intelligent AI systems.

References:

[1] H. Shi, "A short video sentiment analysis model based on multimodal feature fusion," Syst. Soft Comput., vol. 6, no. September, 2024, doi: 10.1016/j.sasc.2024.200148.

[2] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines," IEEE Intell. Syst., vol. 33, no. 6, pp. 17–25, 2018, doi: 10.1109/MIS.2018.2882362.

[3] B. Paneru, B. Thapa, and B. Paneru, "Sentiment analysis of movie reviews: A flask application using CNN with RoBERTa embeddings," Syst. Soft Comput., vol. 7, no. January, p. 200192, 2025, doi: 10.1016/j.sasc.2025.200192.

[4] T. Wu et al., "Video sentiment analysis with bimodal information-augmented multi-head attention," Knowledge-Based Syst., vol. 235, pp. 1–12, 2022, doi: 10.1016/j.knosys.2021.107676.

[5] M. Bin Habib, M. F. Bin Hafiz, N. A. Khan, and S. Hossain, "Multimodal Sentiment Analysis using Deep Learning Fusion Techniques and Transformers," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 6, pp. 856–863, 2024, doi: 10.14569/IJACSA.2024.0150686.

[6] M. He, "A Review of Data Fusion and Deep Learning Models for Multimodal Sentiment Analysis," no. Ecai 2024, pp. 5–12, 2025, doi: 10.5220/0013189700004568.

[7] Y. Jin, K. Cheng, X. Wang, and L. Cai, "Frontiers in Business, Economics and Management A Review of Text Sentiment Analysis Methods and Applications," vol. 10, no. 1, 2023.

[8] N. Mandal and Y. Li, "Rethinking Multimodal Sentiment Analysis: A High-Accuracy, Simplified Fusion Architecture," Unkn. J., pp. 1–8, 2025.

[9] "Multimodal Deep Learning."

[10] J. Fu, Y. Fu, H. Xue, and Z. Xu, "TMFN : a text-based multimodal fusion network with multi-scale feature extraction and unsupervised contrastive learning for multimodal sentiment analysis," Complex Intell. Syst., vol. 11, no. 2, pp. 1–16, 2025, doi: 10.1007/s40747-024-01724-5.

[11] W. Xu, J. Chen, Z. Ding, and J. Wang, "Text Sentiment Analysis and Classification Based on Bidirectional Gated Recurrent Units (GRUs) Model," pp. 3–8.

[12] S. W. Lestari, S. Kahar, and T. Dwi, "Deep learning techniques for speech emotion recognition : A review," vol. 3, no. 2, pp. 78–91, 2023.

[13] S. Alizadeh, "Convolutional Neural Networks for Facial Expression Recognition."

[14] Y. Cai, X. Li, Y. Zhang, J. Li, F. Zhu, and L. Rao, "Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning," pp. 1–22, 2025.

[15] S. Li, "Multimodal Alignment and Fusion : A Survey," no. c, pp. 1–20.

[16] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process., no. September, pp. 2539–2544, 2015, doi: 10.18653/v1/d15-1303.

[17] J. Yang, Y. Yu, D. Niu, W. Guo, and Y. Xu, "ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis," Proc. Annu. Meet. Assoc. Comput. Linguist., vol. 1, pp. 7617–7630, 2023, doi: 10.18653/v1/2023.acl-long.421.

[18] G. Tuerhong, F. Fu, and M. Wushouer, "Adaptive multimodal transformer based on exchanging for multimodal sentiment analysis," Sci. Rep., vol. 15, no. 1, pp. 1–13, 2025, doi: 10.1038/s41598-025-11848-4.

[19] J. B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for

emotion recognition and sentiment analysis," Proc. Annu. Meet. Assoc. Comput. Linguist., pp. 1–7, 2020, doi: 10.18653/v1/2020.challengehml-1.1.

[20] H. Lee, S. Suniljit, and Y. S. Ong, "Dynamic Multimodal Sentiment Analysis: Leveraging Cross-Modal Attention for Enabled Classification," 2025, [Online]. Available: http://arxiv.org/abs/2501.08085

[21] A. Kumar and J. Vepa, "Gated Mechanism for Attention Based Multi Modal Sentiment Analysis," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2020–May, pp. 4477–4481, 2020, doi: 10.1109/ICASSP40776.2020.9053012.

[22] J. Qin, F. Liu, and L. Zong, "Hierarchical Adaptive Expert for Multimodal Sentiment Analysis," 2025, [Online]. Available: http://arxiv.org/abs/2503.22715

[23] J. Xie, Y. Wang, T. Meng, J. Tai, Y. Zheng, and Y. I. Varatnitski, "Multimodal Emotion Recognition Method Based on Domain Generalization and Graph Neural Networks," Electron., vol. 14, no. 5, pp. 1–20, 2025, doi: 10.3390/electronics14050885.

[24] Z. Qin, Q. Luo, Z. Zang, and H. Fu, "Multimodal GRU with directed pairwise cross-modal attention for sentiment analysis," Sci. Rep., vol. 15, no. 1, pp. 1–14, 2025, doi: 10.1038/s41598-025-93023-3.

[25] A. Zadeh et al., "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap., vol. 1, pp. 2236–2246, 2018, doi: 10.18653/v1/p18-1208.

[26] Y. Cai, X. Li, Y. Zhang, J. Li, F. Zhu, and L. Rao, "Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning," Sci. Rep., vol. 15, no. 1, pp. 1–21, 2025, doi: 10.1038/s41598-025-85859-6.

[27] J. Gajjar and K. Ranaware, "Multimodal Sentiment Analysis on CMU-MOSEI Dataset using Transformer-based Models," 2025, [Online]. Available: http://arxiv.org/abs/2505.06110

[28] R. Chen, W. Zhou, Y. Li, and H. Zhou, "Video-Based Cross-Modal Auxiliary Network for Multimodal Sentiment Analysis," IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 12, pp. 8703–8716, 2022, doi: 10.1109/TCSVT.2022.3197420.

[29] W. Liu, S. Cao, and S. Zhang, "Multimodal consistency-specificity fusion based on information bottleneck for sentiment analysis," J. King Saud Univ. - Comput. Inf. Sci., vol. 36, no. 2, p. 101943, 2024, doi: 10.1016/j.jksuci.2024.101943.

[30] U. Singh, K. Abhishek, and H. Azad, "A Survey of Cutting-edge Multimodal Sentiment Analysis," ACM Comput. Surv., vol. 56, Mar. 2024, doi: 10.1145/3652149.

[31] Y. Wu, Q. Mi, and T. Gao, "A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions," Biomimetics, vol. 10, no. 7, 2025, doi: 10.3390/biomimetics10070418.

[32] S. Lai, X. Hu, H. Xu, Z. Ren, and Z. Liu, "Multimodal sentiment analysis: A survey," Displays, vol. 80, 2023, doi: 10.1016/j.displa.2023.102563.

[33] Z. Tang, "Review of Multimodal Sentiment Analysis Techniques," Appl. Comput. Eng., vol. 120, no. 1, pp. 88–97, 2024, doi: 10.54254/2755-2721/2025.18747.

[34] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "Recognition : Speech , Text , and Face," MDPI, J. Basel, Switzerland., pp. 1–33, 2023.

[35] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party

dataset for emotion recognition in conversations," ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf., pp. 527–536, 2020, doi: 10.18653/v1/p19-1050.

[36] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The OMG-Emotion Behavior Dataset," Proc. Int. Jt. Conf. Neural Networks, vol. 2018–July, 2018, doi: 10.1109/IJCNN.2018.8489099.

[37] W. Yu et al., "CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotations of modality," Proc. Annu. Meet. Assoc. Comput. Linguist., pp. 3718–3727, 2020, doi: 10.18653/v1/2020.acl-main.343.

[38] R. Mihalcea, L. Morency, and C. Science, "Utterance-Level Multimodal Sentiment Analysis P13-1096.pdf," Acl, pp. 973–982, 2013, [Online]. Available: https://www.aclweb.org/anthology/P13-1096.pdf

[39] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Lang. Resour. Eval., vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.

[40] T. D.-C. (TDC), "No Title TikTok-10M Dataset." Accessed: Oct. 22, 2025. [Online]. Available: https://huggingface.com/dataset/The-data-company/TikTOK-10M

[41] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A Large-Scale Audio-Visual Dataset," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2020–May, pp. 721–725, 2020, doi: 10.1109/ICASSP40776.2020.9053174.

[42] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-Visual Event Localization in Unconstrained Videos," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11206 LNCS, pp. 252–268, 2018, doi: 10.1007/978-3-030-01216-8_16.

[43] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis," MM 2020 - Proc. 28th ACM Int. Conf. Multimed., pp. 1122–1131, 2020, doi: 10.1145/3394171.3413678.

[44] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," Proc. Annu. Meet. Assoc. Comput. Linguist., pp. 2359–2369, 2020, doi: 10.18653/v1/2020.acl-main.214.

[45] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," Sci. Rep., vol. 14, no. 1, pp. 1–14, 2024, doi: 10.1038/s41598-024-56706-x.

[46] M. Torcoli, T. Kastner, and J. Herre, "Objective Measures of Perceptual Audio Quality Reviewed: An Evaluation of Their Application Domain Dependence," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 29, pp. 1530–1541, 2021, doi: 10.1109/TASLP.2021.3069302.

[47] R. Gareus and C. Goddard, "Audio signal visualisation and measurement," Proc. - 40th Int. Comput. Music Conf. ICMC 2014 11th Sound Music Comput. Conf. SMC 2014 - Music Technol. Meets Philos. From Digit. Echos to Virtual Ethos, pp. 1346–1352, 2014.

[48] V. Pandit and B. Schuller, "The Many-to-Many Mapping Between the Concordance Correlation Coefficient and the Mean Square Error," pp. 1–32, 2019, [Online]. Available: http://arxiv.org/abs/1902.05180

[49] M. Dhanalakshmi, K. Bhanu Priya, K. Madhuri, M. Amulya, J. Bhavani Durga, and M. Jyothika, "Signal-to-Noise Ratio (SNR): A Cornerstone Metric for Quality and Reliability in Diverse Applications," Int. J. Res. Publ. Rev., vol. 4, no. 11, pp. 356–359,

2023, [Online]. Available: www.ijrpr.com

[50] Y. Al-Najjar, "Comparative Analysis of Image Quality Assessment Metrics: MSE, PSNR, SSIM and FSIM," Int. J. Sci. Res., vol. 13, pp. 110–114, Mar. 2024, doi: 10.21275/SR24302013533.

[51] J.-F. Féraud, "S–Sim," Dictionaire Crit. la Lang. française, vol. 13, no. 4, pp. 506–575, 2017, doi: 10.1515/9783110914252-043.

[52] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," IEEE Trans. Image Process., vol. 23, no. 2, pp. 684–695, 2014, doi: 10.1109/TIP.2013.2293423.

[53] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks Chuan," arXiv, 1996.

[54] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," Image Vis. Comput., 2017, doi: 10.1016/j.imavis.2017.08.003.

[55] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods : A survey Multimodal Sentiment Analysis Based on Fusion Methods : A Survey," no. March, 2023, doi: 10.1016/j.inffus.2023.02.028.