10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

Fusion-Driven Approaches for Multimodal Sentiment Analysis Using Machine Learning and Deep Learning

Mayank Devani [1]
Research Scholar, Department of Computer Engineering, Monark University, Ahmedabad
Dr. Harsha Padheriya[2]
Associate Professor, Monark University , Ahmedabad
Vijaysinh Jadeja [3]
Assistant Professor , Sal College of Engineering , Ahmedabad
Mikin Dagli [4]
Assistant Professor , Sal College of Engineering, Ahmedabad

Abstract

Sentiment analysis has changed from traditional text-based methods to multimodal frameworks that integrate text, audio, and visual modalities in response to the exponential growth of user-generated material on social media. By combining information from several sources, Multimodal Sentiment Analysis [3, 4, 13] (MSA) has become a potent paradigm for more correctly interpreting human emotions, views, and emotional states. The development of MSA from early classical methods to more current developments in deep learning is examined in this comprehensive review. It provides a thorough examination of a number of fusion methods, such as tensor-based, attention-based, early, late, hybrid, model-level, and quantum-inspired models. Additionally, it emphasizes well-known datasets that have influenced the benchmark assessments in the field, including CMU-MOSEI, MOSI, and MELD.The performance and applicability of several machine learning and deep learning frameworks, such as SVM, HMM, CNN, LSTM, BiLSTM, and Transformer-based models, are examined. Along with outlining important issues including cross-domain generalization, sarcasm detection, and modality synchronization, the paper also identifies intriguing future approaches in multilingual analysis, real-time sentiment systems, and personalized emotion recognition. For scholars and practitioners working in the field of multimodal sentiment analysis, this paper attempts to offer a foundational guidance.

Keywords

Multimodal Sentiment Analysis [3], Affective Computing, Fusion Techniques, Early Fusion, Late Fusion, Machine Learning, Deep Learning, Audio-Visual Sentiment Analysis, Benchmark Datasets, Opinion Mining

I. Introduction

User-generated content has skyrocketed in the current digital era due to the quick development of social media platforms, websites that share videos, and mobile communication tools. People now communicate their feelings, attitudes, and sentiments through a range of platforms, from brief product reviews and political comments to emotive videos and multimedia memes. The subject of sentiment analysis has shifted from unimodal to Multimodal Sentiment Analysis [3,

10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

4, 13] (MSA) as a result of the expansion of multimodal data, which includes text, speech, and visual material.

The main goal of traditional sentiment analysis techniques was to identify the polarity (positive, negative, or neutral) of opinions by analyzing textual data. However, depending only on text misses important affective cues that are present in gestures, prosody, tone of voice, and facial expressions. For instance, saying "I love this" in a sardonic manner or while rolling one's eyes can express the opposite idea. Researchers are increasingly using MSA, which combines several modalities to enhance accuracy and contextual comprehension of feelings, to overcome these constraints [1].

The combining of various data sources, usually including textual, audio, and visual information, is referred to as multimodal. When combined, these modalities' distinctive and complimentary qualities allow for a more thorough comprehension of human emotions. Such data analysis is a challenging process that calls for advanced feature extraction, modality synchronization, and fusion techniques in order to produce insightful results.

Developments in machine learning (ML) and, more recently, deep learning (DL) have contributed to the expansion of this discipline. Modern techniques use deep architectures like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Transformer-based models for automatic feature learning and end-to-end training, whereas early MSA methods used classical ML models like Support Vector Machines [1] (SVM) and Hidden Markov Models (HMM). These developments have made it possible for researchers to successfully integrate diverse data sources, which has resulted in notable gains in classification robustness and accuracy across domains.

In addition to algorithmic advancements, benchmark multimodal datasets like CMU-MOSEI, CMU-MOSI [1], and MELD have been essential for performance benchmarking and field advancement. Standardized evaluation of MSA models is made possible by these datasets, which include synchronized multimodal information (text, audio, and video) labeled for sentiment or emotion.

Fusion—the process of merging modality-specific data to produce a single sentiment prediction—is one of the main issues in MSA. Early fusion (feature-level fusion), late fusion (decision-level fusion), hybrid fusion, and more sophisticated approaches like tensor-based, attention-based, quantum-inspired, and hierarchical fusion are the general categories into which a number of fusion architectures have been proposed over time. Depending on the task and data characteristics, each technique has unique benefits, drawbacks, and appropriateness [2].

Opinion mining, emotion recognition, human-computer interaction, healthcare (e.g., mental health monitoring), education (e.g., detecting student involvement), and digital marketing are just a few of the real-world applications that multimodal sentiment analysis is relevant for. Because MSA is interdisciplinary, it is a rich study area that crosses computer vision, speech processing, natural language processing (NLP), and cognitive science.

This paper presents a systematic review of the developments in multimodal sentiment analysis. It organizes the literature across three major dimensions:

10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

1. Fusion Techniques – from early to quantum-based methods.
2. Datasets – including their structure, modalities, and application domains.
3. Machine Learning Approaches – both classical and deep learning-based models.

Through this structured overview, we aim to provide a consolidated understanding of the current state of MSA, highlight recent trends and innovations, discuss technical and practical challenges, and suggest future research directions that could push the boundaries of affective computing.

II. Background Concepts

A. Sentiment Analysis and Opinion Mining

The computational study of people's thoughts, feelings, and attitudes as they are expressed in text, audio, or visual media is known as sentiment analysis, or opinion mining. It is essential to comprehending how the general public views goods, services, political organizations, and societal issues. In the past, sentiment analysis has mostly depended on textual content, classifying sentiments as neutral, negative, or positive using methods from machine learning, data mining, and natural language processing (NLP).

But spoken or written words are not the only ways that people can express themselves. Tone, body language, gestures, and facial expressions are frequently used to convey affective states. Neglecting these indicators results in partial or erroneous sentiment prediction, particularly in intricate situations involving comedy, sarcasm, or emotional ambiguity.

B. Emergence of Multimodal Sentiment Analysis (MSA)

Detecting and interpreting human emotions and opinions by combining various data types, or "modalities," mainly text, audio, and video, is the focus of the quickly expanding field of multimodal sentiment analysis [3, 4] (MSA). By examining how people communicate their feelings not only via words but also through body language, tone of voice, and facial expressions, MSA provides a broader, more complex understanding of sentiment than traditional sentiment analysis, which usually only uses textual data. In real-world situations where consumers frequently engage using many channels at once, such product evaluations, customer service encounters, and social media, this holistic approach is quite beneficial.

The text modality in MSA uses natural language processing (NLP) techniques, such as rule-based approaches, machine learning algorithms, and deep learning models like LSTM [3] (Long Short-Term Memory networks) and BERT (Bidirectional Encoder Representations from Transformers), to analyze spoken or written language. In order to decipher the speaker's emotional state, the auditory modality relies on extracting paralinguistic elements including pitch, tone, energy, and rhythm. In this regard, methods such as neural networks, prosodic analysis, and MFCC (Mel-frequency cepstral coefficients) are frequently employed. Utilizing visual clues including gaze, gestures, and facial expressions, the video modality frequently uses computer vision and convolutional neural networks (CNNs) to recognize and categorize emotional signals.

10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

The intricacy of synchronizing data from several modalities, managing missing or noisy information, and creating efficient fusion algorithms to integrate insights from each source are some of the difficulties that MSA faces despite its potential. Common strategies include early fusion, late fusion, and hybrid fusion; each has pros and cons. Furthermore, attaining generalizability across several languages and contexts continues to be a major research challenge [15].

Applications for multimodal sentiment analysis are numerous and include everything from monitoring mental health and analyzing political debate to improving recommendation systems and human-computer interaction. The future of MSA depends on creating more resilient, interpretable, and context-aware systems that can comprehend sentiment like humans do—through numerous, interconnected channels of communication—as deep learning and AI technologies advance [16].

The stages of sentiment analysis utilizing multimodal fusion, which combines diverse audio-visual data from heterogeneous sources, are depicted in Figure 1. First, a variety of data types—unstructured, semi-structured, and structured—are gathered from online resources. After that, this data goes through a pre-processing stage where it is cleansed and pertinent sections are chosen, frequently with the use of dimensionality reduction strategies designed to address the particular issue at hand. During the feature extraction phase, temporal visual features are taken from video frames, while audio features are taken from the spoken information in videos.To extract the text modality from the video, spoken transcripts are also produced. A multimodal feature vector is then created by combining these features that were derived from various modalities. The categorized sentiment is then supplied to the appropriate application once this vector has been fed into classification algorithms to determine the sentiment class [14].

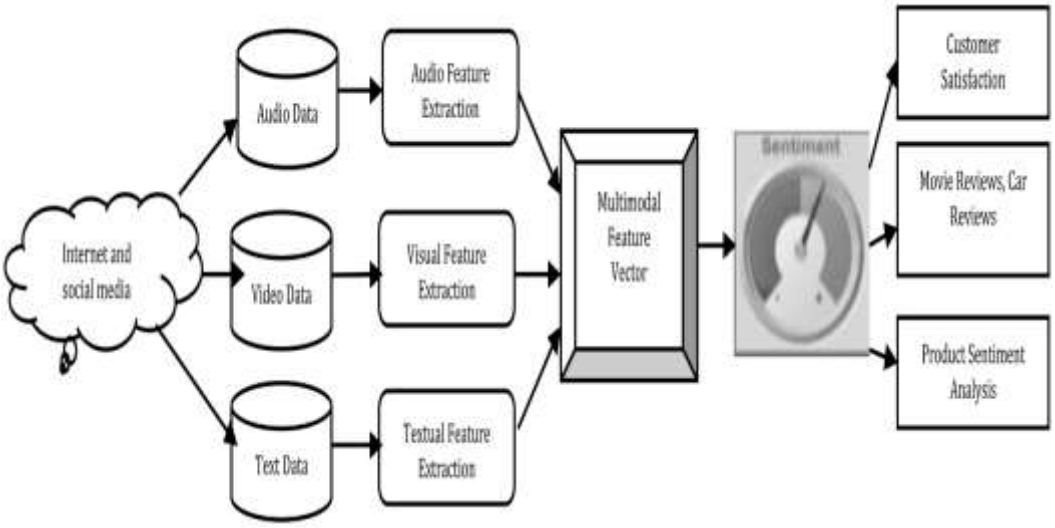

Figure 1.MSA process model.

III. Common Modalities Used in Multimodal Sentiment Analysis

Text, audio, and visual data are the three main modalities that are usually integrated in multimodal sentiment analysis [3, 4, 13]. Every modality adds distinct data that improves the

10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

overallcomprehensionofsentiment.Transcripts of spoken language or written critiques are frequently the source of text modality. Features like polar words (such "great," "terrible"), n-grams, and word embeddings that convey semantic meaning are some of the ways it expresses sentiment. Phoneme sequences and syntactic or semantic patterns also provide additional clues regarding the speaker's intent and emotional tone in spoken text.

Sentiment interpretation can be greatly enhanced by the acoustic qualities of speech, which are captured by audio modality. Pitch and tone are particularly crucial; for example, a high pitch might convey joy or enthusiasm, whereas a low pitch can convey sombreness or melancholy. Subtleties like hesitancy, sarcasm, or emphasis on emotionally laden words can be revealed through pauses, intonation, and changes in energy or intensity.

Information gleaned from video frame analysis is the main emphasis of the visual modality. It entails identifying facial expressions that are clear markers of emotion, like raised eyebrows, frowns, and smiles. The emotional context is further enhanced by body language, such as posture, eye gazing, and gestures. Brief, uncontrollable facial expressions known as micro-expressions are particularly useful for revealing repressed or buried emotions that would not be apparent from language or sound alone.

Together, these modalities provide a comprehensive framework for capturing human sentiment in a more accurate and context-aware manner.

IV. Importance of Modality Fusion

In multimodal sentiment analysis, modality fusion is essential for improving the precision and resilience of sentiment categorization systems. Modality fusion allows for a more thorough comprehension of sentiment conveyed in diverse contexts by combining data from numerous modalities, including text, image, audio, and video. This method improves the overall performance of sentiment analysis models by utilizing the advantages of each modality to offset its shortcomings.

Important Points:

- Improved Contextual Understanding: When text, visual, and aural signals are combined, a richer context for sentiment interpretation is created, allowing for the capture of subtleties that could be overlooked when examining each modality separately.
- Increased Accuracy: By utilizing complementary information sources, fusion approaches such as hybrid fusion (a combination of both), early fusion (integrating features from several modalities at the input level), and late fusion (combining outputs from individual models) help to increase accuracy.
- Robustness to Ambiguity and Noise: By cross-verifying data across modalities, modality fusion produces more accurate sentiment predictions while reducing noise and managing ambiguous sentiments.
- Applications across Domains: Modality fusion expands the applicability of sentiment analysis to a variety of domains, making it flexible enough to accommodate different data characteristics and user requirements. Examples of these domains include social media sentiment analysis, customer feedback interpretation, and healthcare sentiment monitoring.

10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

- Challenges and Future Directions: Despite its benefits, challenges such as feature alignment, scalability, and computational complexity remain. Future research directions include exploring advanced fusion architectures, domain adaptation techniques, and integrating multimodal deep learning approaches for more sophisticated sentiment analysis systems.

V. Fusion Techniques in Multimodal Sentiment Analysis

In Multimodal Sentiment Analysis [3, 4, 13] (MSA), fusion approaches are essential because they combine input from several modalities, including text, audio, and visual data, to generate a single sentiment prediction. Combining the complementing aspects of each modality while managing their inherent disparities and maintaining their individual strengths is the main objective of fusion. Three primary types of fusion strategies have been established over the years by researchers: hybrid (or intermediate) fusion, late fusion, and early fusion. Furthermore, more complex methods like attention-based fusion, tensor fusion, and graph-based fusion have been developed recently; each has unique benefits and applications.

Fusion techniques are crucial to Multimodal Sentiment Analysis [3, 4, 13] (MSA) as they integrate input from several modalities, such as text, audio, and visual data, to produce a single sentiment prediction. The primary goal of fusion is to combine the complementary features of each modality while preserving their unique strengths and managing their inherent differences. Over the years, researchers have identified three main categories of fusion strategies: early fusion, late fusion, and hybrid (or intermediate) fusion. Additionally, more sophisticated techniques have lately been created, such as graph-based fusion, tensor fusion, and attention-based fusion; each has special advantages and uses.

To reach a final choice, late fusion—also referred to as decision-level or score-level fusion—involves processing each modality separately through different models or pipelines and then combining their distinct outputs (such as anticipated sentiment scores or probabilities). Because each input is handled independently, this method is less susceptible to synchronization problems and missing modalities. Simple rule-based voting, averaging, weighted voting, and more complex ensemble techniques like stacking or boosting are examples of common approaches. Late fusion may miss intricate inter-modal interactions that could increase prediction accuracy, despite being simpler to handle in real-world situations [5].

The goal of hybrid fusion, also known as intermediate fusion, is to integrate the advantages of early and late fusion methods. At various points in the model process, such as at intermediate feature layers or with learnt joint representations prior to final decision-making, modalities may be partially fused in hybrid fusion. For instance, utilizing fully connected layers or attention processes, neural architectures may collect modality-specific data, convert them into a shared latent space, and then classify sentiment. Because hybrid fusion makes use of both low-level and high-level interactions among modalities, it frequently performs better than early or late fusion alone. However, it can be difficult and computationally demanding to build such designs.

Because attention-based fusion may selectively weigh key properties across modalities, it has become increasingly popular in recent years. Attention methods are employed to determine which aspects of the input from each modality should be emphasized or suppressed, drawing inspiration from the success of Transformer [7] systems in natural language processing. For

10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

example, if the user's voice and facial expression convey more emotion than the spoken words during a video review, the model may focus more on these aspects. Cross-modal attention makes it possible for modalities to interact dynamically, which improves sentiment classification performance and context understanding. Using this method, models such as Multimodal Transformer and BERT variations with cross-attention modules have demonstrated state-of-the-art outcomes.

Tensor fusion is another sophisticated technique that models all potential interactions by generating high-dimensional outer products between modality features. For instance, Tensor Fusion [9] Networks (TFNs) may capture trimodal, bimodal, and unimodal interactions in a single framework. The resulting feature space can be very extensive and computationally costly, even if this method captures rich inter-modal connections. Tensor fusion complexity can be decreased without noticeably sacrificing performance by employing strategies like compact bilinear pooling and low-rank tensor approximation.

A new technique called graph-based fusion uses graph neural networks (GNNs) to handle multimodal data. Each modality or data segment can be represented as a node in this method, and edges show the semantic or temporal links between them. Graph-based fusion works especially well with hierarchical sentiment structures or sequential data. It enables adaptability to irregular or asynchronous input across modalities and permits flexible modeling of relationships.

Furthermore, Bayesian fusion techniques use probabilistic frameworks that explicitly model uncertainty to integrate modalities. Due to Bayesian approaches' ability to infer missing information and modify prediction confidence appropriately, this is especially helpful when working with noisy or incomplete modalities. Similarly, in domain-specific applications where expert knowledge is available to define modality importance or fusion logic, rule-based and heuristic fusion techniques are still used [8].

Fusion Type	**Fusion Level**	**Key Methods**	**Strength**	**Limitation**
Early Fusion	Feature-level	Concatenation, TFN	Captures inter-modal dependencies early	Sensitivetomisalignment, high dimension
Late Fusion	Decision-level	Voting, Averaging	Robust to missing/noisy modalities	Ignores inter-modal relations
Hybrid Fusion	Mixed	TFN + Voting	Balanced performance	Higher complexity
Model-Level Fusion	Architecture-level	Shared Encoders, Multimodal BERT	Flexible and trainable	Depends on architecture design

10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

Fusion Type	Fusion Level	Key Methods	Strength	Limitation
Tensor Fusion	Feature-level tensor	TFN, MTFN, MRRF	Expressive, high accuracy	Computationally expensive
Attention-based Fusion	Contextual alignment	MMHA, CATF-LSTM	Learns focus dynamically	Needs good attention design
Hierarchical Fusion	Multi-stage	CHFusion, HFNN	Local + global modality interaction	Complex to implement
Quantum Fusion	Decision & feature	QMF, QMN	Captures inter-utterance effects	Theoretical novelty, less widely adopted
Word-Level Fusion	Temporal-word level	MFN, RMFN, RAVEN	Fine-grained alignment	Requires fine temporal annotation

Table 1: Comparison of Fusion Techniques

VI. Popular Datasets for Multimodal Sentiment Analysis

High-quality datasets that include synchronized and annotated data across various modalities, such as text, audio, and video, are crucial for Multimodal Sentiment Analysis [3, 4, 13] (MSA). The advancement of MSA research has been greatly aided by the availability of standardized benchmark datasets, which enable consistent comparison and evaluation of different models and methodologies. In terms of sample size, type and granularity of annotations, number of modalities, language coverage, and intended applications—from sentiment polarity classification to emotion recognition and affective computing—these datasets vary greatly and are typically sourced from a variety of domains, including online video reviews, vlogs, television shows, interviews, and social media content.These datasets allow researchers to train and evaluate sophisticated models that can comprehend human sentiment holistically by providing rich, aligned multimodal information and ground truth labels. The most popular MSA datasets are compared in the section that follows, with an emphasis on their salient features and applicability to different research goals.

Sr. No.	Dataset Name	Year	Modalities	Source	Size/ No. of Videos	Language	Application Domain
1	**CMU-MOSI**	2016	Text, Audio, Visual	YouTube	93 videos, 2199 utterances	English	Opinion review, Vlogs

10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

Sr. No.	Dataset Name	Year	Modalities	Source	Size/ No. of Videos	Language	Application Domain
2	**CMU-MOSEI**	2018	Text, Audio, Visual	YouTube	3228 videos, 23,500 utterances	English	General (General-purpose sentiment analysis including reviews and debates)
3	**MELD**	2019	Text, Audio, Visual	TV Series ("Friends")	1400 dialogues, 13,000 utterances	English	Emotion & Sentiment in Conversations
4	**Memotion**	2020	Text, Visual	Reddit, Facebook	~10k memes	English	Meme Sentiment Analysis
5	**CH-SIMS**	2020	Text, Audio, Visual	YouTube (Chinese)	2281 video segments	Chinese	General
6	**CMU-MOSEAS**	2021	Text, Audio, Visual	YouTube	4000 videos	Spanish, Portuguese, German, French	Multilingual Sentiment/Emotion
7	**MuSe-CaR**	2021	Text, Audio, Visual	YouTube (Vehicle reviews)	291 videos	English	Emotion & Entity Sentiment
8	**FACTIFY**	2022	Text, Visual	Twitter	50,000 tweets	English	Multimodal Fact Checking
9	**B-T4SA**	2021	Text, Visual	Twitter	470k tweets	Multilingual	Image-Text Sentiment
10	**MEMOTION 2.0**	2022	Text, Visual	Reddit, Facebook	10,000 images	English	Meme Emotion and Sentiment Classification

Table 2: Summary of Widely Used MSA Datasets

VII. Machine Learning and Deep Learning Techniques for Multimodal Sentiment Analysis

- Strong classification methods that can comprehend and learn from the fused multimodal representations must be employed in conjunction with fusion to optimize performance in multimodal sentiment analysis (MSA). Over time, MSA research has shifted from employing straightforward, conventional machine learning algorithms to

10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

more intricate deep learning architectures that are capable of accurately modeling contextual linkages, temporal dynamics, and cross-modal interactions.

- **A. Traditional Machine Learning Techniques**
- When fusion techniques were still relatively simple, traditional machine learning models were among the first to be used for MSA problems. These methods are usually used after the early or late fusion stages and frequently rely on manual feature extraction.
- The Support Vector Machine [1] (SVM), which is well-known for its efficiency in managing high-dimensional feature spaces, is one often employed model. Particularly in early MSA systems, it has been extensively utilized in text-based and audio-visual sentiment classification. However, SVMs cannot capture temporal or sequential dependencies and have scalability issues when working with big multimodal datasets.

  The Hidden Markov Model [1] (HMM), another method, is mostly used to model temporal sequences, particularly in audio or video streams like speech or gestures. Although HMMs work well with time-series data, they are less adaptable when it comes to modeling intricate inter-modal connections since they are predicated on strong independence assumptions across states.

- Random forests and decision trees have also been used, frequently in ensemble settings. These models have comparatively short training times and are simple to comprehend. However, when working with multimodal features that are high-dimensional, continuous, or intricately interconnected, their performance frequently deteriorates. As a result, even though conventional approaches have given MSA a strong basis, their drawbacks have prompted the use of more flexible deep learning models.

- **B. Deep Learning Techniques**
- Because deep learning can automatically learn features from raw multimodal data and model complex dependencies, it has significantly advanced MSA. These techniques may process text, audio, and visual inputs immediately and end-to-end, doing away with the requirement for manually created features.In visual sentiment analysis, convolutional neural networks (CNNs) are widely employed, especially for the extraction of spatial data from motions, facial expressions, and video frames [3]. When paired with word embeddings such as Word2Vec or GloVe, CNNs can also be used to classify text or even applied to audio spectrograms.

- One type of recurrent neural network (RNN) that is often used to capture temporal dependencies in sequential data, including speech, text, or audio streams, is the Long Short-Term Memory [3] network (LSTM). They work especially well for sentiment analysis in dialogue-based datasets because comprehension of the present sentiment depends on the context of earlier utterances.

- Bidirectional LSTM [3] (BiLSTM) networks, which process input sequences from both forward and backward directions, are used to further improve contextual awareness. BiLSTMs are particularly helpful for multimodal alignment and conversation-level sentiment analysis because of their dual processing, which enables them to collect richer contextual inputs.

10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

- More recently, sophisticated architectures such as the Recurrent Memory Fusion Network (RMFN) and Memory Fusion Network [12] (MFN) have been presented. Learning modality-specific memories and matching them with gated attention processes are the main goals of MFN. CMU-MOSEI and CMU-MOSI are two fine-grained sentiment datasets on which it excels. By adding recurrence to the fusion process, RMFN expands on this and allows the model to learn inter-modal linkages and long-term dependencies over time.

- The advent of Multimodal Transformer [7] Models, which modify the transformer design from NLP to the multimodal domain, is among the most revolutionary advancements in recent years. Self-attention mechanisms are used by these models, which include MulT, MMTN, and Multimodal-BERT, to process and link information both within and across modalities at the same time. Their ability to concurrently encode text, audio, and visual data sequences enables them to effectively capture intra- and inter-modal interactions. Transformer models presently set the standard for many MSA tasks because of their scalability and parallel processing capabilities.

- In order to capture the interactions between distinct modalities or features organized as graphs, researchers have also started investigating Graph Neural Networks (GNNs). Dependencies between various entities or modalities, as well as within time-series data, can be represented using these models. In addition to GNNs, gated architectures have become popular because they provide selective control over which modality-specific data enters the final decision-making process. This is particularly useful in situations when some modalities are unreliable or noisy.

Technique	Modality Handled	Strengths	Limitations
Support Vector Machine (SVM)	Text, Audio, Visual	Good generalization in high-dimensional space	Not scalable to large data; no temporal modeling
Hidden Markov Model (HMM)	Audio, Visual	Effective for sequential modeling	Strong independence assumptions
Decision Trees / Random Forests	All (after feature extraction)	Fast training, interpretable	Low accuracy with complex multimodal data
Convolutional Neural Network (CNN)	Visual, Audio (Spectrogram), Text	Good for spatial feature extraction	Needs large labeled datasets
Long Short-Term Memory (LSTM)	Text, Audio	Captures temporal dependencies	Needs careful tuning; vanishing gradients
Bidirectional LSTM (BiLSTM)	Text, Audio	Understands full context in sequences	Higher complexity than LSTM

10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

Memory Fusion Network (MFN)	Text, Audio, Visual	Captures word-level alignment using memory	Complex architecture; needs large data
Recurrent Memory Fusion Network (RMFN)	Text, Audio, Visual	Models long-term contextual dependencies	Training can be computationally expensive
Multimodal Transformer Models	Text, Audio, Visual	Captures intra- and inter-modal dependencies	Requires extensive data and training resources
Graph Neural Networks (GNNs) / Gated Models	Text, Audio, Visual	Models structured relationships across features	Interpretability and scalability challenges

Table 3: Comparison of ML and DL Techniques in MSA

Conclusion

Sentiment analysis in multiple modes [3, 4, 13] (MSA), which uses information from text, audio, and visual modalities to provide a more comprehensive understanding of human emotions, has grown in importance in both academic and industry-focused research. The need for more precise and contextually aware sentiment analysis is greater than ever as user-generated material on websites like YouTube, Instagram, and consumer review systems keeps growing. The main elements of MSA approaches—fusion techniques, benchmark datasets, and machine learning techniques—were covered in this paper's systematic review. It is clear that the way modalities are merged is a key factor in determining model performance, ranging from simple early and late fusion strategies to sophisticated tensor-based and attention-driven fusion.Comparably, the development of sentiment analysis systems into increasingly potent and versatile frameworks is demonstrated by the shift from traditional machine learning to deep learning and transformer-based models. Real-world implementation is still constrained by issues including explainability, missing data, generalization across domains, and modality synchronization, despite tremendous advancements. Ongoing research on robust fusion techniques, zero-shot learning, and cross-modal attention, however, points to a future in which MSA can be used in a variety of fields, including intelligent human-computer interaction, healthcare, education, and customer service. In conclusion, the MSA field is headed in a positive direction. Better datasets, more powerful fusion models, and sophisticated learning frameworks are now available to researchers.Now, the emphasis must be on developing effective, flexible, and interpretable systems that can function dependably in a variety of languages and domains. MSA will have a significant impact on the development of affective computing and emotional intelligence in machines in the future with sustained innovation and teamwork.

References

10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

[1] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P., "MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos," *arXiv preprint* arXiv:1606.06259, 2016.

[2] Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P., "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion," *Proceedings of ACL*, 2018, pp. 2236–2246.

[3] Poria, S., Cambria, E., Bajpai, R., & Hussain, A., "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[4] Hazarika, D., Zimmermann, R., & Poria, S., "MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis [3, 4, 13]," ACM Multimedia (MM), 2020.

[5] Majumder, N., Poria, S., Hazarika, D., Gelbukh, A., Cambria, E., & Mihalcea, R., "Multimodal Sentiment Analysis [3, 4, 13] Using Hierarchical Fusion with Context Modeling," Knowledge-Based Systems, vol. 161, pp. 124–133, 2018.

[6] Li, Y., Mao, Y., & Liu, Y., "Multimodal Emotion Recognition Using Deep Canonical Correlation Analysis," *arXiv preprint* arXiv:1510.06066, 2015.

[7] Tsai, Y. H. H., Bai, S., Yamada, M., Morency, L. P., &Salakhutdinov, R., "Multimodal Transformer for Unaligned Multimodal Language Sequences," *Proceedings of ACL*, 2019, pp. 6558–6569.

[8] Liang, P. P., Zadeh, A., Shen, Y. C., & Morency, L. P., "Multimodal Local-Global Ranking Fusion for Emotion Recognition," *ICASSP*, 2018, pp. 4469–4473.

[9] Chen, R., Mao, Q., & Xue, Y., "A Tensor Fusion Network for Multimodal Sentiment Analysis [3, 4, 13]," Information Fusion, vol. 52, pp. 320–328, 2019.

[10] Rahman, T., & Mehta, S., "Fusion Strategies in Multimodal Sentiment Analysis [3, 4, 13]: A Survey," International Journal of Computer Applications, vol. 183, no. 46, pp. 1–6, 2021.

[11] Wollmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., & Morency, L. P., "YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.

[12] Hazarika, D., Zimmermann, R., & Poria, S., "Fusion of Audio, Visual and Text Modalities Using Memory Networks for Multimodal Emotion Recognition," *Proceedings of ICME*, 2018.

[13] Liang, P. P., Zadeh, A., & Morency, L. P., "Foundations and Recent Trends in Multimodal Machine Learning: Principles, Challenges, and Open Issues," *ACM Computing Surveys*, vol. 54, no. 8, pp. 1–38, 2022.

[14] Bhoi, A., & Kumar, A., "Recent Advancements in Multimodal Fusion Techniques for Sentiment Analysis: A Review," *Procedia Computer Science*, vol. 199, pp. 634–641, 2022.

[15] Barros, P., Weber, C., Wermter, S., &Tzirakis, P., "Multimodal Emotion Recognition for Human-Robot Interaction: A Survey," *IEEE Transactions on Affective Computing*, Early Access, 2023.

10.48047/jocaaa.2023.31.04.42
ISSN : 1521-1398

[16] Han, W., et al., "Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis [3, 4, 13]," arXiv preprint arXiv:2107.13669, 2021.

[17] Yu, J., et al., "Self-supervised Multi-task Learning for Multimodal Sentiment Analysis [3, 4, 13]," IEEE Transactions on Affective Computing, Early Access, 2021.

[18] Sun, Y., et al., "Efficient Multimodal Transformer [7] with Dual-Level Feature Restoration," arXiv preprint arXiv:2208.07589, 2022.

[19] Zhang, Y., et al., "Dynamic Invariant Representation-specific Fusion Network for Multimodal Sentiment Analysis [3, 4, 13]," arXiv preprint arXiv:2212.10874, 2022.

[20] Hazarika, D., et al., "Multimodal Transformer [7] Fusion with Explicit Modality Alignment for Sentiment Analysis," IEEE Transactions on Multimedia, vol. 24, pp. 2943–2955, 2022.