International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Multimodal Sentiment Analysis on Social Media Big Data: A Survey of Deep Fusion Architectures

Sonali Nimbalkar[1], Om Dandagvhal[2], Om Ganeshpure[3], Vishal Jadhav[4], Omkar Murkute[5]

[1]Associate Professor, Department of Computer Engineering
[2,3,4,5]Student, Department of Computer Engineering
Sandip Institute of Technology and Research Centre, Nashik, Maharashtra, India

---

Abstract - Social media posts combine text, images, audio and video, and the sentiment they express often depends on how these channels interact. This paper surveys deep learning methods for multimodal sentiment analysis (MSA) on social media data, following the progression from lexicon-based classifiers through convolutional and recurrent networks, pre-trained transformers, vision–language and contrastive models to multimodal large language models and state-space architectures. It consolidates the surveyed work in comparative tables, together with the equations that define the core components. Three observations emerge. First, on the main benchmarks the quality of the pre-trained text encoder and of the cross-modal alignment affects accuracy more than the design of the fusion operator. Second, text carries most of the sentiment signal, and visual and acoustic features add small gains. Third, results from different studies are often not comparable because of differences in feature extractors, alignment settings and evaluation conventions. The paper closes with open problems in sarcasm detection, missing modalities, multilingual content and efficiency, and with a practical deployment checklist.

Key Words: multimodal sentiment analysis, social media, multimodal fusion, cross-modal attention, contrastive learning, large language models.

1.INTRODUCTION

1.1 Motivation and Scope

A social media post often combines a caption, an image, a short video with speech, or a meme with overlaid text. The affect expressed in such a post is distributed across these channels, and the channels do not always agree. A post that reads "Best day ever" and is accompanied by photographs of a car accident is sarcastic, and a classifier that reads only the text assigns it the wrong polarity. Multimodal sentiment analysis (MSA) addresses this problem by predicting sentiment or emotion from two or more modalities, usually language, vision and acoustics [1], [2].

Deploying MSA on social media adds constraints that curated research datasets do not have: the data are large and arrive continuously, posts are short, noisy and often code-mixed, modalities may be missing, and sarcasm is common. Most published methods, however, are developed and evaluated on English video corpora recorded under controlled conditions [3], [4], [5]. Here "big data" is used in a practical sense: volume and velocity limit the computational cost a model can afford, and variety and veracity determine how well it copes with missing or unreliable modalities. The question addressed is how well current deep fusion methods handle these conditions, and what separates benchmark performance from deployable systems. This is a narrative survey of representative work from 2011 to 2024 that introduced a widely used benchmark, a fusion or alignment technique, or an evaluation study that changed how results are interpreted; it does not follow a systematic search protocol.

1.2 Contributions and Related Surveys

The paper makes four contributions. First, it gives a taxonomy of MSA tasks (Sec. 2) and fusion strategies (Sec. 4). Second, it presents the core architectures with the equations that define them (Sec. 4). Third, it compares reported results across datasets and explains why they cannot always be compared directly (Sec. 5). Fourth, it discusses open problems and gives a practical deployment checklist (Secs. 6 and 7).

Earlier surveys cover multimodal sentiment analysis in general [1] or multimodal machine learning more broadly [2]. The present survey treats social media conditions (missing modalities, code-mixing, sarcasm and scale) as primary evaluation criteria, relates each method to the equation that defines it so that design differences can be traced to the model and not to implementation details, and pays particular attention to evaluation practice, notably the findings of MERBench [6] and the has-0/non-0 accuracy convention. More recent surveys also exist; a full comparison with them is outside the scope of this paper.

2. TASK FORMULATION AND TAXONOMY

Sentiment analysis assigns a polarity label (positive, negative or neutral) or a continuous intensity score to a piece of content. Early systems used bag-of-words features with lexicon scoring or shallow classifiers. These were replaced by convolutional [7] and recurrent [8] encoders and then by pre-trained transformers such as Bidirectional Encoder Representations from Transformers (BERT) [9] and RoBERTa [10], which now supply contextual word representations as the standard text front end. On social media, text alone is often insufficient because much of the affective content is non-verbal: a smirk, a sarcastic image or a change in tone can reverse the literal meaning of the words.

MSA therefore takes a text sequence X_t, a visual sequence X_v and an acoustic sequence X_a (or, for image–text posts, a text sequence and an image) and learns a mapping $f(X_t, X_v, X_a) \rightarrow y$, where y is a class label or a real-valued score. Four problems recur: aligning asynchronous streams, handling modalities that are missing at inference time, transferring from acted laboratory data to noisy social posts, and keeping computational cost compatible with platform scale. MSA is a family of tasks (Table 1), with granularity ranging from the utterance through the conversation [11] to the whole post. The task type matters because method effectiveness depends on the data regime: cross-modal attention has mainly been evaluated on acted or monologue corpora, whereas contrastive alignment has mainly been reported on noisy image–text data.

Table -1: MSA task–modality taxonomy

Task	Mod.	Typical datasets	Output
Polarity classification	T, V, A	MOSI, MOSEI	2-, 3- or 7-class
Intensity regression	T, V, A	MOSI, MOSEI	Real value in [−3, 3]
Emotion recognition	T, V, A	IEMOCAP, MELD	Categorical
Sarcasm detection	T, I	MMSD2.0, HFM	Binary
Hateful-meme detection	T, I	Hateful Memes	Binary
Aspect-based sentiment	T, I	Twitter-15/17	Target polarity

T: text, V: video frames, A: audio, I: image.

3. EVOLUTION OF METHODS

The literature can be read as four overlapping transitions. The first is from lexicons to learned representations: handcrafted features and sentiment lexicons, which are brittle under negation, sarcasm and domain shift, were replaced by learned encoders. The second is from unimodal to multimodal modeling, with tri-modal evaluation on MOSI [3] and MOSEI [4] becoming the standard setting. The third is from fusion-centered to alignment-centered design: after the tensor and attention operators of 2017–2019, attention moved to pre-trained encoders, alignment signals and auxiliary supervision. The fourth, still in progress, is from architecture to prompting, in which multimodal large language models (LLMs) treat fusion as a matter of instructions and not of network design [12], [13], [14].

3.1 Classical and Context-Based Methods

Lexicon-based and classical classifiers are lightweight and interpretable but unimodal and brittle under negation, sarcasm and domain shift. The first multimodal study on opinion videos combined lexical, prosodic and visual cues and showed that the combination outperformed any single channel, although on a very small corpus [15]. Poria et al. [11] later modeled the sequence of utterances in a video with a long short-term memory (LSTM) network, so that each utterance was classified in the context of its neighbors, and reported clear gains on MOSI and IEMOCAP. Context modeling is simple and consistently helpful, but it needs utterance boundaries, which are not generally available for social posts.

3.2 Tensor, Low-Rank and Attention-Based Fusion

Zadeh et al. [16] proposed the Tensor Fusion Network (TFN), which models unimodal, bimodal and trimodal interactions explicitly through an outer product of the modality embeddings. Liu et al. [17] approximated the resulting weight tensor with low-rank factors (LMF), reducing the cost considerably while retaining accuracy on MOSI, IEMOCAP and POM. Both operate on pooled utterance-level vectors and therefore ignore temporal structure; Tsai et al. [18] addressed this with the Multimodal Transformer (MulT), which applies cross-modal attention directly to unaligned sequences. Re-evaluations under unified settings suggest that the differences among these operators are small once the text encoder and feature extractors are held fixed [6], so they can be viewed as points on one expressiveness–efficiency curve and not as competing paradigms.

3.3 Image–Text Models for Social Media

For image–text posts, early dedicated models include MultiSentiNet [19], which uses scene and object semantics from the image to attend over the words, and a co-memory network [20] in which the two modalities influence each other iteratively; later models add caption views [21] and graph networks [22]. CLMLF [23] aligns token-level text and image representations with a contrastive objective and reports 75.33% accuracy on MVSA-Single; ITIN [24] uses gating to down-weight unreliable visual features, and the multi-view calibration network [25] targets modality heterogeneity. Absolute scores on MVSA are considerably lower than on MOSI, which probably reflects the difficulty of the data (noisy labels, weak image–text correspondence and ironic content) and not weaker models [26], [27]. Models that perform well on scripted video corpora should therefore not be assumed to transfer to these datasets.

3.4 Pre-trained, Contrastive and Unified Models

Rahman et al. [28] injected visual and acoustic information into intermediate BERT layers through a Multimodal Adaptation Gate (MAG-BERT); the gain over text-only BERT was less than one accuracy point on MOSI. Later work improved the representations without changing the fusion operator: Yu et al. [29] generated unimodal pseudo-labels with a self-supervised multi-task scheme, and Han et al. [30] maximized mutual information between modalities hierarchically. ConFEDE [31] decomposes each modality into a shared and a modality-specific component and trains the decomposition contrastively, and UniMSE [32] trains one model for sentiment analysis and emotion recognition in conversations, using datasets such as MELD [33]. Together these studies indicate that custom fusion operators are being replaced by supervision objectives and shared backbones.

3.5 Sarcasm and Hateful Content

Sarcasm and hateful-meme detection are stress tests for MSA because the modalities can contradict each other, so correct classification requires cross-modal reasoning and not only feature combination. Cai et al. [34] introduced a hierarchical fusion model for sarcasm in tweets. Qin et al. [35] showed that the widely used sarcasm dataset contained spurious cues and annotation errors and released the corrected MMSD2.0 benchmark; models score lower on the corrected data, which suggests that part of the earlier progress came from these artifacts. The Hateful Memes challenge [36] was built with benign confounders so that unimodal models are insufficient; its authors report that the best baselines reach about 65% accuracy against about 85% for human annotators. Memotion [37] adds sentiment, humor, sarcasm, offense and motivation labels for memes, although these annotations are subjective. The common lesson is that cross-modal inconsistency must be modeled explicitly and that evaluation should penalize the exploitation of spurious cues.

4. CORE ARCHITECTURES AND FUSION STRATEGIES

This section defines the main components. Encoders (convolutional neural networks (CNNs), LSTM networks, Transformers, BERT, the Vision Transformer (ViT) and Contrastive Language–Image Pre-training (CLIP)) turn raw signals into vectors; fusion operators (TFN, LMF, cross-modal attention, MAG) combine those vectors; supervision objectives (contrastive, mutual-information, pseudo-label) shape the representations; and recent backbones (Mamba, multimodal LLMs) process all modalities in one sequence model. Scalars are written x, vectors x, matrices W, and σ is the logistic sigmoid.

4.1 Encoders and Contrastive Alignment

A one-dimensional text CNN [7] slides k-gram filters over the embedded tokens and max-pools over time, and stacked convolution and pooling layers form the basis of visual encoders. The LSTM [8] counters vanishing gradients in recurrent networks with a gated cell state and serves in MSA as a per-modality sub-encoder and for temporal context modeling [11]. Self-attention replaces recurrence by letting every token attend to every other token in parallel [38]:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d_k})\,V \quad (1)$$

BERT [9] pre-trains a bidirectional encoder with masked language modeling, RoBERTa [10] removes next-sentence prediction and uses dynamic masking, and ViT [39] applies the same architecture to image patches. CLIP [40] trains paired text and image encoders with a contrastive objective: for a batch of N image–text pairs it maximizes the cosine similarity of the N matched pairs and minimizes it for the N^2-N mismatched pairs, which places matched text and images close together in a shared space. In general form, the contrastive loss for an anchor a with positive p is

$$L = -\log\left[\exp(\text{sim}(a, p)/\tau) / \sum_k \exp(\text{sim}(a, k)/\tau)\right] \quad (2)$$

where sim is cosine similarity, the sum runs over the positive and the negatives, and τ is a temperature. These encoders underlie most MSA models published since 2019.

4.2 Cross-Modal Fusion Operators

Cross-modal attention draws the queries from a target modality and the keys and values from a source modality, which allows fusion without prior alignment [18] (Fig. 1a). For source v and target t:

$$CM_{v\to t}(X_t, X_v) = \text{softmax}(X_tW_Q(X_vW_K)^T/\sqrt{d})\,X_vW_V \quad (3)$$

MulT stacks such blocks for all six directed pairs of modalities. TFN [16] forms the outer product of the unimodal embeddings, each extended with a constant 1 so that unimodal and bimodal terms are retained (Fig. 1b):

$$z^m = [z_t; 1] \otimes [z_v; 1] \otimes [z_a; 1] \quad (4)$$

LMF [17] approximates the weight tensor W of the subsequent linear layer by a sum of r rank-one factors,

$$W \approx \sum_{k=1}^{r} w_k^{(t)} \otimes w_k^{(v)} \otimes w_k^{(a)} \quad (5)$$

which reduces the parameter count from the product of the modality dimensions to r times their sum. MAG-BERT [28] injects non-verbal information into intermediate BERT layers (Fig. 1c). In simplified form,

$$H = Z + \beta(W_vx_v + W_ax_a + b)$$
$$\beta = \sigma(W_\beta[Z; x_v; x_a]) \quad (6)$$

where Z is the hidden state of a transformer layer and the gate β limits the influence of the non-verbal path when it carries little information. The small gain of MAG-BERT over text-only BERT is the clearest single indication that, on top of a strong text encoder, non-verbal channels add only a small increment. Contrastive objectives such as Eq. (2) provide further supervision in CLMLF [23] and ConFEDE [31].

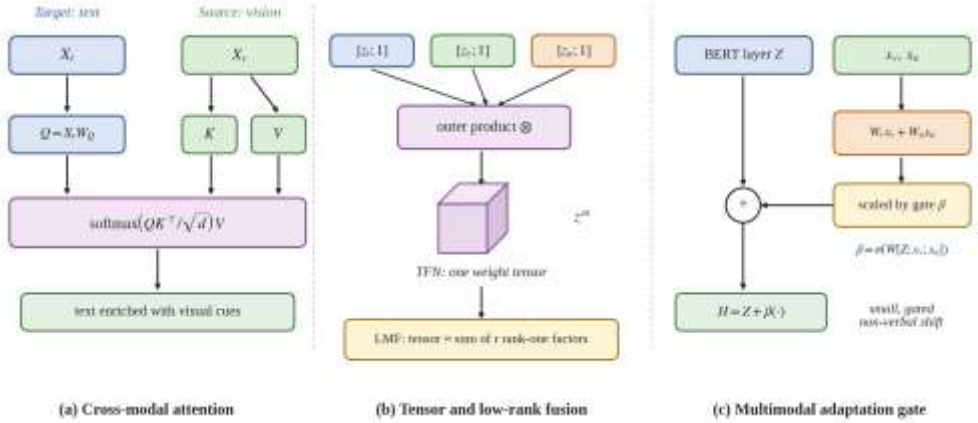


Fig -1: Cross-modal fusion operators: (a) cross-modal attention, (b) tensor fusion and its low-rank approximation, and (c) the multimodal adaptation gate of MAG-BERT.

4.3 Fusion Strategies

Fusion can also be grouped by the point at which the modalities are combined (Fig. 2). Early (feature-level) fusion combines modality features before a single classifier; it can capture rich interactions, as in TFN and LMF, but it is costly and sensitive to missing modalities. Late (decision-level) fusion trains one predictor per modality and combines the scores; it keeps the modalities independent but cannot represent interactions between them. Intermediate (representation-level) fusion, for example cross-modal attention [18] and MAG-BERT [28], combines latent sequences and captures fine-grained interactions between asynchronous streams at a cost that is quadratic in sequence length. Hybrid schemes such as ConFEDE [31] and UniMSE [32] combine a fusion operator with additional contrastive or shared supervision, at the price of a more complex training objective.

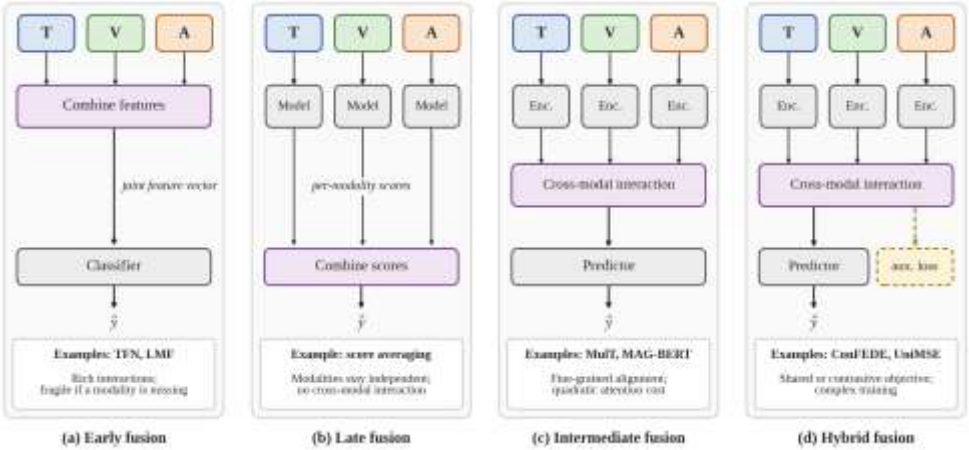


Fig -2: Fusion strategies by the point at which the modalities are combined: (a) early, (b) late, (c) intermediate and (d) hybrid fusion. T: text, V: visual, A: audio, ŷ: prediction.

4.4 State-Space Models and Multimodal Large Language Models

Mamba [41] makes the parameters of a linear state-space recurrence depend on the input (selective state spaces):

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t, \quad y_t = Ch_t \tag{7}$$

This gives time and memory that are linear in sequence length, against quadratic for self-attention, which is attractive for long social streams and video-level temporal modeling. Multimodal LLMs project visual or audio features into the token space of a language model through a learned adaptor and generate labels or reasoning text autoregressively. Zhang et al. [12] benchmarked large language models on 26 datasets in zero-shot and few-shot settings, Lian et al. [13] evaluated GPT-4V on 21 emotion and sentiment datasets, and Cheng et al. [14] instruction-tuned a model on the MERR dataset to produce natural-language reasoning together with labels.

5. DATASETS, REPORTED RESULTS AND EVALUATION

5.1 Corpora

The benchmark landscape does not match deployment conditions. The most widely used corpora, MOSI, MOSEI and IEMOCAP, are English and were recorded in controlled or scripted settings, whereas deployment involves short, multilingual, user-generated posts. The social media corpora, MVSA [26], Twitter-15/17 [42], HFM [34], MMSD2.0 [35] and Hateful Memes [36], are smaller and noisier. CH-SIMS [43] is a notable exception: it provides separate unimodal and multimodal labels, which supports studies of modality disentanglement, but it is limited to Chinese. MERR [14] adds emotion-reasoning annotations suited to instruction-tuned models. No single benchmark therefore supports a claim of general social media competence, and results should be reported together with the provenance of the data (Table 2).

Table -2: Principal corpora referenced in the literature

Dataset	Source and setting	Language	Mod.	Task	Remarks
CMU-MOSI [3]	YouTube opinion monologues	English	T, V, A	Sentiment intensity	Small; reference benchmark
CMU-MOSEI [4]	YouTube monologues	English	T, V, A	Sentiment and emotion	Largest tri-modal corpus; still monologue video
IEMOCAP [5]	Acted dyadic sessions, laboratory	English	T, V, A	Categorical emotion	Acted; not social media
MELD [33]	Television dialogues	English	T, V, A	Emotion and sentiment in conversation	Multi-party conversations
MVSA [26]	Tweets with images	English	T, I	Polarity	Noisy labels; small
Twitter-15/17 [42]	Tweets with images	English	T, I	Target-oriented sentiment	Requires aspect annotations
HFM [34]	Tweets	English	T, I	Sarcasm (binary)	Early multimodal sarcasm dataset
MMSD2.0 [35]	Tweets	English	T, I	Sarcasm (binary)	Corrected version of the original MMSD
Hateful Memes [36]	Memes re-created for the challenge	English	T, I	Hate (binary)	Benign confounders; constructed, not organic
CH-SIMS [43]	Chinese videos	Chinese	T, V, A	Sentiment with unimodal labels	Supports disentanglement studies; Chinese only

5.2 Reviewed Studies

Table 3 compares the studies reviewed in this paper. The figures are those reported in the original papers or in later comparisons. They were obtained with different feature extractors, alignment settings and evaluation conventions and should not be compared across rows.

Table -3: Structured comparison of the reviewed literature

Ref.	Study and method	Dataset	Mod.	Reported result	Limitation
[16]	Zadeh et al. (2017): Tensor Fusion Network	MOSI	T, V, A	Acc-2 ≈ 80.8	Many parameters; no temporal modeling
[17]	Liu et al. (2018): Low-rank Multimodal Fusion	MOSI, IEMOCAP, POM	T, V, A	Acc-2 ≈ 82.5	Sensitive to rank
[18]	Tsai et al. (2019): Multimodal Transformer	MOSI, MOSEI, IEMOCAP	T, V, A	Better than prior methods on unaligned data	$O(n^2)$ cost
[28]	Rahman et al. (2020): MAG-BERT	MOSI, MOSEI	T, V, A	Acc-2 84.2 / 86.1 on MOSI	Modest gain over BERT
[29]	Yu et al. (2021): Self-supervised unimodal pseudo-labels	MOSI, MOSEI, CH-SIMS	T, V, A	Gains over several baselines	Text-heavy
[31]	Yang et al. (2023): ConFEDE contrastive decomposition	MOSI, MOSEI	T, V, A	Gains over earlier methods	Complex objective
[23]	Li et al. (2022): CLMLF contrastive alignment	MVSA, HFM	T, I	75.33% on MVSA-Single	Noisy labels
[24]	Zhu et al. (2023): Image–text interaction network with gating	Social image–text sets	T, I	Gains over image–text baselines	No audio
[25]	Wei et al. (2023): Multi-view calibration network	Social image–text sets	T, I	Gains over earlier methods	Image–text only
[35]	Qin et al. (2023): MMSD2.0 corrected benchmark	MMSD2.0	T, I	Lower scores than on original MMSD	Binary labels; English
[36]	Kiela et al. (2020): Hateful Memes benchmark	Hateful Memes	T, I	About 65% vs about 85% human accuracy	Constructed data
[27]	Cheema et al. (2021): Comparison of tweet sentiment methods	Social image–text sets	T, I	Fair re-comparison of methods	Limited to tweets
[12]	Zhang et al. (2023): LLMs, zero- and few-shot	26 datasets	T	Competitive on simple tasks	Text only
[13]	Lian et al. (2023): GPT-4V, zero-shot	21 datasets	T, I	Strong on visual emotion tasks	Closed model; hard to reproduce
[14]	Cheng et al. (2024): Emotion-LLaMA	MERR, MER2023	T, V, A	F1 0.9036	Large compute
[6]	Lian et al. (2024): MERBench unified evaluation	Multiple emotion datasets	T, V, A	Feature extractor affects results more than fusion	Emotion recognition focus

5.3 Evaluation Metrics and Reporting Practice

Reported metrics fall into three groups: binary or ternary accuracy (Acc-2, Acc-3) and weighted or macro F1 for classification; mean absolute error (MAE) and Pearson correlation for continuous intensity; and binary F1 and the area under the receiver operating characteristic (ROC) curve (AUC) for sarcasm and toxicity benchmarks. A persistent source of confusion is the has-0 versus non-0 convention for binary accuracy on MOSI: has-0 includes samples with a neutral (zero) score and non-0 excludes them, and the choice can shift reported accuracy by several points. Alignment (aligned versus una-

ligned sequences) and the choice of feature extractor add further variance. MERBench [6] re-evaluates methods under unified settings and reports that the feature extractor has a larger influence on results than the fusion strategy, which should change how earlier leaderboards are read.

6. DISCUSSION

6.1 What the Evidence Supports

Four findings are supported by several independent studies. First, fusion usually outperforms the best single modality when all channels are present, although the margin decreases once protocols are unified [6]. Second, pre-trained encoders matter more than fusion topology, as shown by MAG-BERT [28], by the comparison of Cheema et al. [27] and by MERBench [6]. Third, explicit alignment helps most where image–text correspondence is weak, as on social media [18], [24], [25]. Fourth, auxiliary supervision in the form of pseudo-labels [29], mutual information [30] or contrastive decomposition [31] improves representations without changing the fusion operator.

6.2 What the Evidence Does Not Support

Richer fusion algebra does not reliably improve prediction: TFN and LMF perform comparably [16], [17], and differences between fusion operators are often smaller than the effect of the encoder. Progress on sarcasm and meme benchmarks may partly reflect spurious cues and not genuine reasoning [35], [36], [37]. Benchmark fragmentation further obscures real progress, and Lian et al. [6] argue for standardized pipelines comparable to the General Language Understanding Evaluation (GLUE) benchmark in natural language processing.

6.3 Text Modality and the Alignment–Compute Trade-off

On the benchmarks reviewed here, a well-tuned text encoder with a simple fusion operator matches or exceeds more elaborate multimodal pipelines [28], [6], [27]. Three factors probably explain this: language carries the densest and most explicit sentiment signal, it is already discretized into tokens that suit large-scale pre-training, and it has the fewest alignment problems. Vision and audio help mainly when they contradict or complement the text, as with sarcastic imagery or a change in tone, and these cases are a minority of posts. The modeling budget of a social media MSA system should therefore go mostly to text, with vision and audio used as corrective signals when cross-modal disagreement is detected instead of being fused uniformly.

Because the quadratic cost of cross-modal attention is prohibitive for long social streams, alignment should be applied only where it is needed: adaptive gating [24] and multi-view calibration [25] weight or calibrate the visual channel instead of fusing it uniformly, and state-space models [41] replace quadratic attention with a linear-time selective recurrence. A plausible next generation of efficient MSA systems would combine a strong text backbone, selective cross-modal attention and a state-space temporal encoder.

6.4 From Benchmarks to Deployment

Curated benchmarks consist of clean, mostly English data that were acted or carefully collected, whereas deployment involves noisy, multilingual streams with missing modalities and real-time constraints. The gap is operational as well as a matter of accuracy: a system that needs all three modalities, a 512-token window and a very large language model cannot monitor a platform at scale. The remedy lies in shared feature extractors, unified protocols and datasets with controlled cues; until then, authors should include ablations that separate the fusion operator from the encoder and the feature extractor, and claims of state-of-the-art performance should state the benchmark regime and the assumed deployment constraints.

7. OPEN PROBLEMS AND FUTURE DIRECTIONS

7.1 Open Problems

Nine open problems are identified. (1) Missing modalities: social posts routinely lack a channel, and robust imputation, modality-dropout training and disentangled representations remain under-explored. (2) Domain adaptation: models trained on acted corpora degrade on organic posts so zero-shot transfer and social test sets such as MVSA and MMSD2.0 are needed. (3) Multilingual and code-mixed content: most work is on English, and code-mixed text such as Hinglish lacks benchmarks. (4) Low-resource languages, including Dravidian, African and South-East Asian languages, are under-served in both pre-trained coverage and benchmark data. (5) Sarcasm and irony: detecting genuine cross-modal contradiction and evaluating with controlled cues remain open [35]. (6) Robustness to noise, adversarial perturbation and out-of-distribution input is rarely tested. (7) Explainability and fairness: predictions need saliency or concept attributions and audits across demographic groups. (8) Efficiency and edge deployment: the cost of BERT, ViT and multimodal LLM backbones motivates distillation, quantization, pruning and Mamba-based encoders [41]. (9) Integration of multimodal LLMs: fine-tuning, prompting, chain-of-thought reasoning and control of hallucination for structured sentiment output are at an early stage [12], [13].

The problems are related: missing-modality robustness and domain adaptation both call for training on degraded or shifted inputs, multilingual and low-resource coverage both benefit from cross-lingual contrastive pre-training, and explainability, fairness and robustness together determine whether an MSA system can be trusted in consequential uses such as content moderation or mental-health monitoring. Future work should therefore prioritize cross-cutting enablers (standardized evaluation, efficient encoders and reproducible multimodal backbones) over yet another bespoke fusion operator.

7.2 Future Directions

Promising directions are foundation-model and multimodal LLM-based MSA, retrieval-augmented generation for low-resource settings, self-supervised pre-training on unlabeled social media content, federated and privacy-preserving training, responsible-AI auditing, multi-agent fusion systems, cross-lingual benchmarks, real-time streaming at platform scale, and on-device inference with models of fewer than 100 million parameters. Two appear to offer the largest benefit: Mamba-based temporal encoders for efficiency, and retrieval-augmented multimodal LLMs for low-resource and domain-shifted settings.

7.3 Practical Deployment Checklist

For platform-scale social media monitoring, the discussion reduces to five checks. (i) Modality coverage: do not assume that all channels are present; use modality-dropout training and a fallback for missing modalities. (ii) Encoder first: invest in a strong pre-trained text encoder and a contrastive alignment objective before adopting a more elaborate fusion operator. (iii) Evaluation: report results under the same has-0/non-0 convention and feature extractor as the baselines, ideally under MERBench-style unified settings [6]. (iv) Social validity: validate on genuinely social corpora (MVSA, MMSD2.0, Twitter-15/17) and not only on acted video. (v) Efficiency budget: prefer distilled or Mamba-based encoders when latency or memory is constrained. Treating these checks as acceptance criteria should reduce the gap between benchmark scores and deployed behavior.

8. CONCLUSIONS

This paper reviewed deep learning approaches to multimodal sentiment analysis on social media, from lexicons to multimodal LLMs and state-space models. The main finding is that, on the benchmarks surveyed, the quality of the pre-trained encoders and of the cross-modal alignment affects results more than the design of the fusion operator, and that beyond a strong text encoder the visual and acoustic channels add only a small increment: MAG-BERT [28] improved on text-only BERT by less than one accuracy point. Evaluation is a second concern. Differences in feature extractors, alignment settings and the has-0/non-0 convention make published results hard to compare, and MERBench [6] indicates that differences between fusion methods diminish when these factors are controlled. Standardized evaluation protocols are therefore needed.

Open problems remain in sarcasm and meme detection, multilingual and code-mixed content, missing modalities, explainability and fairness, and efficiency at platform scale. State-space models and multimodal LLMs are promising for long sequences and low-resource settings but remain at an early stage.

REFERENCES

1. Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., Pantic, M.: A survey of multimodal sentiment analysis. Image Vis. Comput. 65 (2017) 3–14
2. Liang, P.P., Zadeh, A., Morency, L.-P.: Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. arXiv preprint arXiv:2209.03430 (2022)
3. Zadeh, A., Zellers, R., Pincus, E., Morency, L.-P.: MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259 (2016)
4. Bagher 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 graph. In: Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL), pp. 2236–2246 (2018)
5. Busso, C., et al.: IEMOCAP: Interactive emotional dyadic motion capture database. Lang. Resour. Eval. 42(4) (2008) 335–359
6. Lian, Z., et al.: MERBench: A unified evaluation benchmark for multimodal emotion recognition. arXiv preprint arXiv:2401.03429 (2024)
7. Kim, Y.: Convolutional neural networks for sentence classification. In: Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), pp. 1746–1751 (2014)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8) (1997) 1735–1780
9. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. NAACL-HLT, pp. 4171–4186 (2019)
10. Liu, Y., et al.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
11. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.-P.: Context-dependent sentiment analysis in user-generated videos. In: Proc. 55th Annu. Meeting ACL, pp. 873–883 (2017)
12. Zhang, W., Deng, Y., Liu, B., Pan, S.J., Bing, L.: Sentiment analysis in the era of large language models: A reality check. arXiv preprint arXiv:2305.15005 (2023)
13. Lian, Z., et al.: GPT-4V with emotion: A zero-shot benchmark for generalized emotion recognition. arXiv preprint arXiv:2312.04293 (2023)
14. Cheng, Z., et al.: Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. In: Proc. NeurIPS (2024)
15. Morency, L.-P., Mihalcea, R., Doshi, P.: Towards multimodal sentiment analysis: Harvesting opinions from the web. In: Proc. 13th Int. Conf. Multimodal Interfaces (ICMI), pp. 169–176 (2011)
16. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.-P.: Tensor fusion network for multimodal sentiment analysis. In: Proc. EMNLP, pp. 1103–1114 (2017)
17. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.-P.: Efficient low-rank multimodal fusion with modality-specific factors. In: Proc. 56th Annu. Meeting ACL, pp. 2247–2256 (2018)
18. Tsai, Y.-H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.-P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proc. 57th Annu. Meeting ACL, pp. 6558–6569 (2019)
19. Xu, N., Mao, W.: MultiSentiNet: A deep semantic network for multimodal sentiment analysis. In: Proc. 26th ACM Int. Conf. Inf. Knowl. Manag. (CIKM) (2017)
20. Xu, N., Mao, W., Chen, G.: A co-memory network for multimodal sentiment analysis. In: Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (2018)
21. Yang, X., Feng, S., Wang, D., Zhang, Y.: Image-text multimodal emotion classification via multi-view attentional network. IEEE Trans. Multimedia 23 (2021)
22. Yang, X., Feng, S., Zhang, Y., Wang, D.: Multimodal sentiment detection based on multi-channel graph neural networks. In: Proc. 59th Annu. Meeting ACL and 11th IJCNLP (2021)
23. Li, Z., Xu, B., Zhu, C., Zhao, T.: CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. In: Findings Assoc. Comput. Linguistics: NAACL (2022)
24. Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., Qian, J.: Multimodal sentiment analysis with image-text interaction network. IEEE Trans. Multimedia 25 (2023)
25. Wei, Y., Yuan, S., Yang, R., Shen, L., Li, Z., Wang, L., Chen, M.: Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In: Proc. 61st Annu. Meeting ACL (2023)
26. Niu, T., Zhu, S., Pang, L., El Saddik, A.: Sentiment analysis on multi-view social data. In: Proc. Int. Conf. MultiMedia Modeling (MMM) (2016)
27. Cheema, G.S., Hakimov, S., Müller-Budack, E., Ewerth, R.: A fair and comprehensive comparison of multimodal tweet sentiment analysis methods. In: Proc. Workshop Multi-Modal Pre-Training Multimedia Understanding (2021)
28. Rahman, W., Hasan, M.K., Lee, S., Zadeh, A., Mao, C., Morency, L.-P., Hoque, E.: Integrating multimodal information in large pretrained transformers. In: Proc. 58th Annu. Meeting ACL, pp. 2359–2369 (2020)
29. Yu, W., Xu, H., Yuan, Z., Wu, J.: Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In: Proc. AAAI Conf. Artif. Intell. (2021)
30. Han, W., Chen, H., Poria, S.: Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: Proc. EMNLP, pp. 9180–9192 (2021)
31. Yang, J., Yu, Y., Niu, D., Guo, W., Xu, Y.: ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis. In: Proc. 61st Annu. Meeting ACL, pp. 7617–7630 (2023)
32. Hu, G., Lin, T.-E., Zhao, Y., Lu, G., Wu, Y., Li, Y.: UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In: Proc. EMNLP (2022)
33. Poria, S., et al.: MELD: A multimodal multi-party dataset for emotion recognition in conversations. In: Proc. 57th Annu. Meeting ACL, pp. 527–536 (2019)
34. Cai, Y., Cai, H., Wan, X.: Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In: Proc. 57th Annu. Meeting ACL (2019)
35. Qin, L., Huang, S., Chen, Q., Cai, C., Zhang, Y., Liang, B., Che, W., Xu, R.: MMSD2.0: Towards a reliable multi-modal sarcasm detection system. In: Findings Assoc. Comput. Linguistics: ACL (2023)
36. Kiela, D., et al.: The Hateful Memes challenge: Detecting hate speech in multimodal memes. In: Proc. NeurIPS (2020)
37. Sharma, C., et al.: SemEval-2020 Task 8: Memotion analysis—the visuo-lingual metaphor! In: Proc. 14th Int. Workshop Semantic Eval. (SemEval) (2020)

38. Vaswani, A., et al.: Attention is all you need. In: Proc. NeurIPS (2017)
39. Dosovitskiy, A., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. In: Proc. ICLR (2021)
40. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Proc. ICML (2021)
41. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
42. Yu, J., Jiang, J.: Adapting BERT for target-oriented multimodal sentiment classification. In: Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI) (2019)
43. Yu, W., et al.: CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: Proc. 58th Annu. Meeting ACL (2020)